Comparison of Voice Activity Detection Algorithms for VoIP

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

[#]*CEDT, Indian Institute of Science, Bangalore, India,* [*]*Department of E&C, PESIT, Bangalore.*

Email: [#] *vprasad@cedt.iisc.ernet.in, hsjam@cedt.iisc.ernet.in,* [*] *abhijeetsangwan@netscape.net*

Abstract

We discuss techniques for Voice Activity Detection (VAD) for Voice over Internet Protocol (VoIP). VAD aids in saving bandwidth requirement of a voice session thereby increasing the bandwidth efficiently. In this paper, we compare the quality of speech, level of compression and computational complexity for three time-domain and three frequency-domain VAD algorithms. Implementation of time-domain algorithms is computationally simple. However, better speech quality is obtained with the frequency-domain algorithms. A comparison of merits and demerits along with the subjective quality of speech after removal 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 current trend is to provide this service on data networks [11]. Data networks work on the best effort delivery and resource sharing through statistical multiplexing. Therefore, the cost of services compared to circuit-switched networks is considerably less. However, these networks do not guarantee faithful voice transmission. Voice over packet or 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. Therefore, providing Toll Grade Voice Quality [5] through VoIP systems remains a challenge. In this paper we concentrate on the problem of reducing the required bandwidth for a voice connection on Internet using Voice Activity Detection (VAD), while maintaining the voice quality.

VAD algorithms find the beginning and end of talk spurts. VAD is used in non real-time systems like Voice Recognition systems, Compression and Speech coding [4][13][6]. VAD is also useful in VoIP, in which stringent detection of beginning and end of talk spurts is not needed.

In VoIP systems the voice data (or payload for packet) is transmitted along with a header on a network. The header size for Real Time Protocol (RTP, [10]) 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. Smaller 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 and its 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, various VAD algorithms are presented with varied complexity and quality of reconstructed speech. Time and frequency domain techniques are discussed. Results obtained, and an exhaustive comparison of various algorithms with quantitative measurements of speech quality is presented and shown that it is an improvement over similar work [1]. There are many previous studies on VAD that dealt with energy-based algorithms such as [9]. In this paper, a procedure for choosing the scaling parameter [9] is also given.

1.1. Speech Characteristics

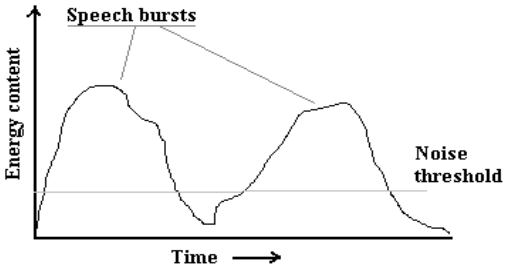

Figure 1. A typical speech signal

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. Silence Periods

The term 'silence segment' does not refer to a period of zero-energy packets, but of incomprehensible sound or background noise. VAD algorithms have to deal with silence periods having small audible content.

1.3. Desirable aspects of VAD algorithms

- A Good Decision Rule: A physical property of speech that can be exploited to give consistent judgment in classifying segments of the signal into silent or voiced segments.
- Adaptability to Changing Background Noise: Adapting to non-stationary background noise improves robustness, especially in wireless telephony where the user is mobile.
- Low Computational Complexity: Internet telephony is a real-time application. Therefore the complexity of VAD al-

Proceedings of the Seventh International Symposium on Computers and Communications (ISCC'02)
1530-1346/02 $17.00 © 2002 **IEEE**

gorithm must be low to suit real-time applications (not more than one packet time).
- Toll quality voice reproduction.
- Saving in bandwidth to be maximized.

2. Parameters for VAD Design

Differentiation of 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 is associated with each frame. For the time-domain algorithms, this parameter is the average energy content and number of Zero Crossings of the frame. For the frequency-domain algorithms, this parameter is the spectrum and variance of the spectrum of a frame. If this parameter exceeds a certain threshold, the signal frame is classified as ACTIVE else it is INACTIVE.

2.1. Choice of Frame Duration

ACTIVE Frames that are transmitted are queued up in a packet-buffer at the receiver. This allows them to playing 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 were 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 5-40ms frame sizes.

The specifications for toll quality encoding of speech for all VAD algorithms are [5]:
- 8 kHz sampling frequency
- 256 levels of linear quantization (8 Bit PCM) [12]
- 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 for time domain algorithms and 8ms for frequency domain ($64 = 2^6$), to avoid padding in DCT calculations used in VAD algorithms.

2.2. Energy of a Frame

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

Let **X(i)** be the **ith** sample of speech. If the length of the frame were **k** samples, then the **jth** frame can be represented in time domain and frequency by a sequence as,

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

$$F(f_j) = DCT\{f_j\} \quad (2)$$

We associate energy $\mathbf{E_j}$ with the **jth** frame as

$$E_j = \frac{1}{k} \sum_{i=(j-1)k+1}^{jk} x^2(i) \quad (3)$$

where, $\mathbf{E_j}$ = energy of the **jth** frame and
$\mathbf{f_j}$ is the **jth** frame that is under consideration.

2.3. Initial Value of Threshold

The starting value for the threshold is important for the evolution of the threshold, which tracks the background noise. An arbitrary initial choice of the threshold is prone to a poor performance. Two methods are proposed for finding 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 estimate of energy is obtained by taking the mean of the energies of each sample as in

$$E_r = \frac{1}{\upsilon} \sum_{m=0}^{\upsilon} E_m \quad (4a)$$

where, $\mathbf{E_r}$ = initial threshold estimate,
υ = number of frames in prerecorded sample.

Similarly, the initial threshold for variance of spectrum is obtained using

$$\sigma = VAR\{F(f_j)\} \quad (4b)$$

We have taken a prerecorded sample of 5 seconds, i.e., 500 frames in time domain and 625 frames in frequency domain.

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.4a. We set $\upsilon = 20$.

A fixed threshold would be 'deaf' to varying acoustic environments of the speaker.

3. VAD Algorithms - Time Domain

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 $(\mathbf{E_j > k\,E_r})$ **where** $\mathbf{k > 1}$ (5)
Frame is ACTIVE
ELSE **Frame is INACTIVE**

In this equation, $\mathbf{E_r}$ represents the energy of noise frames, while $\mathbf{kE_r}$ is the **'Threshold'** being used in the decision-making. Having a scaling factor, **k** allows a safe band for the adaptation of $\mathbf{E_r}$, and therefore, the threshold.

3.1. LED: Linear Energy-Based Detector

It is now sufficient to specify the reference noise energy, $\mathbf{E_r}$, for use in Eq (5) to formulate the schemes completely

3.1.1. Computation of $\mathbf{E_r}$. Since background disturbance is non-stationary an adaptive threshold is more appropriate. The rule to update the threshold value can be found in [9] as,

$$E_{rnew} = (1-p)E_{rold} + pE_{silence} \quad (6)$$

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 $\mathbf{E_r}$ is updated as a convex combination of the old threshold and the current noise update. **p** is chosen considering the impulse response of Eq.(6) as a first order filter (**0<p<1**).

The Z-Transform of Eq (6) is,

$$E_r(Z) = (1-p)\,Z^{-1}E_r(Z) + pE_{noise}(Z) \quad (7)$$

The Transfer Function may be determined using,

$$H(Z) = \frac{E_r(Z)}{E_{noise}(Z)} = \frac{p}{1-(1-p)\,Z^{-1}} \quad (8)$$

The impulse response for **H(z)** is given in Fig 2. It is observed that for **p**=0.2, the fall-time (95%) corresponds to 15 delay units, i.e. 150ms. In effect, 15 past INACTIVE frames influence the calculation for $\mathbf{E_{rnew}}$. Usually, pauses between two syllabi are about 100ms and these pauses should not be considered as silence. The fall-time selected is greater than this value, so that these pauses do not affect updating of $\mathbf{E_r}$. For various values of **p** the fall-time is plotted in Fig. 3. **p** in all the algorithms is fixed to 0.2 corresponding to 150ms or 15 packets periods.

Merits: This algorithm is simple to implement. It gave an acceptable quality of speech after compression.

Shortcomings

- This algorithm cannot give a good speech quality under varying background noise. This was because, the threshold of Eq. (6) 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.
- Non-plosive phonemes as in the words such as "high" and "flower" were clipped completely. This is because the algorithm was based exclusively on the energy content of the frames.
- Low SNR conditions caused undue clippings, there by deteriorating the performance.

3.1.2. Comment. The calculation of $\mathbf{E_r}$, and in turn the threshold, explained above, is used in all the algorithms that follow. We use the same formulation for calculating **p** throughout this paper for all the algorithms whenever there is a convex sum of the old and new noise energy.

3.2. ALED: Adaptive Linear Energy-Based Detector

The sluggishness of **LED** is a consequence of **p** in Eq. (6) being insensitive to the noise statistics. We compute $\mathbf{E_r}$ based on second order statistics of INACTIVE frames. A buffer (linear queue) of the most recent ***'m'*** silence frames is maintained. The buffer contains the value of $\mathbf{E_{silence}}$ 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 by

$$\sigma = \mathrm{VAR}[E_{silence}] \quad (9)$$

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 σ_{old}. After the addition of the new INACTIVE frame, the variance is σ_{new}. A sudden change in the background noise would mean

$$\sigma_{new} > \sigma_{old} \quad (10)$$

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

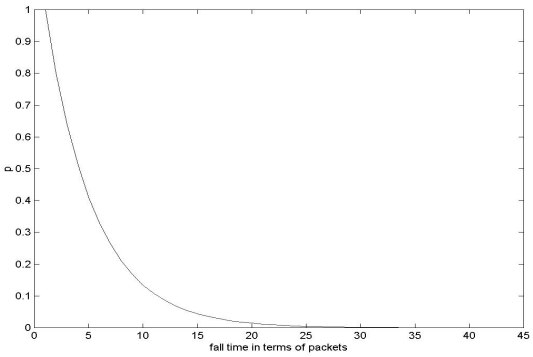

Figure 2. Impulse Response of H(Z) for p = 0.2

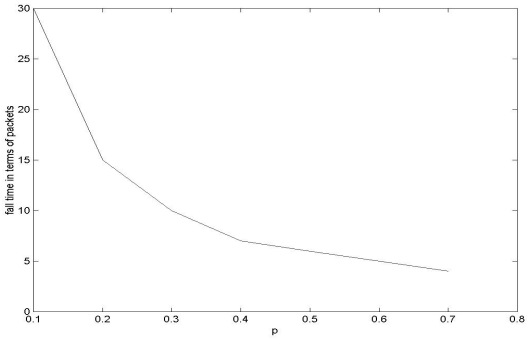

Figure 3. Fall-time for different values of p

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

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

The coefficients of Convex Combination (Eq. (6)) now depend on variance of energies of INACTIVE frames. We are able to make the otherwise sluggish $\mathbf{E_r}$ respond faster to sudden changes in the background noise. The classification rule for the signal frames continues to be Eq (5). Therefore, detection of ACTIVE frames is still energy-based.

Shortcomings: **A.** Inability to detect non-plosive phonemes persisted. **B.** Low SNR conditions caused undue clippings in the compressed signal, as in **LED** Algorithm.

3.3. WFD: Weak Fricatives Detector

LED and **ALED** are exclusively energy-based. Low energy phonemes are sometimes silenced completely. It is observed that high energy voiced speech segments are always detected in all

VAD algorithms under very noisy conditions. However low energy unvoiced speech is commonly missed [9], thus reducing speech quality. This algorithm is designed to overcome this problem. 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 therefore, is able to detect low energy phonemes in quite a number of cases.

Zero Crossings for each frame are computed by the following decision rule:

If $(N_{zcs}(f_j) \in R)$ **(11)**

Frame is 'ACTIVE'

Else **Frame is 'INACTIVE'**

Here,

$\mathbf{N_{zcs}}$ is the number of Zero Crosses detected in a frame.

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

This is incorporated in **ALED**. The Zero Crossing Detector (ZCD) checks the voice activity of the frames that were declared to be INACTIVE by **ALED**. Thus, ZCD recovers almost all the low-energy speech phonemes that were otherwise silenced.

Shortcoming

- A ZCD often makes incorrect decisions as noise frames may have the same number of zero crossings as in speech frames.

4. VAD Algorithms - Frequency Domain

The following algorithms take into consideration the frequency-domain characteristics of speech signals. DCT is used for computation of the spectrum for the following reasons: -

a) Computationally less complex as compared to DFT.

b) Real-valued transform.

4.1. LSED: Linear Sub-Band Energy Detector

This algorithm takes its decisions based on energy comparisons of the signal frame with a reference energy threshold in the frequency domain. The frequency domain counterpart of the frame is obtained by Eq (2).

The spectrum obtained is divided into four bands of width 1kHz, i.e., the bands are 0-1kHz, 1-2 kHz, 2-3kHz, 3-4kHz. The energy for each band is calculated as,

$$E_n[f] = F^2(f_n) \qquad \text{for } n^{th} \text{ band} \qquad (13)$$

And the condition for presence of speech in each band is given by

$$E_n[f] > k\,E_{nth}[f] \qquad \text{for } n^{th} \text{ band} \qquad (14)$$

The thresholds are computed recursively, but for each band separately as a Convex Combination (Eq. 6). For the n^{th} band,

$$E_{nthnew} = (1-p)\,E_{nthnew} + p\,E_{nthnew} \qquad (15)$$

Thus, in each band, the energy threshold is computed based on the previous energy threshold and the latest noise update of the current band.

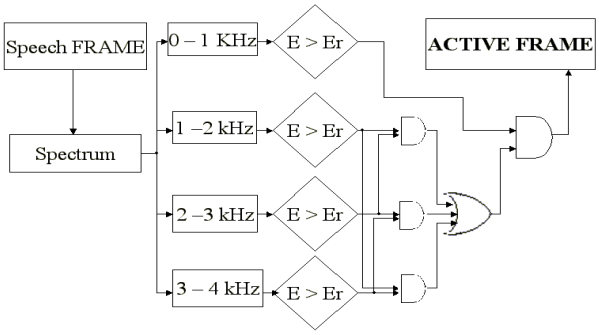

Figure 4. Flowchart for LSED

4.1.1. Fraction of Energy in Lowest Frequency Band. Most of the energy in voice signal tends to be in the lowest frequency band, i.e., 0-1kHz. Selective threshold comparison in the lowest band alone provides good decisions. This condition embedded in the algorithm **WFD** improves the performance of the VAD.

4.1.2. Decision Rule for Speech. A frame is declared to be ACTIVE if the lowest frequency band is ACTIVE and any two out of the remaining three bands are ACTIVE.

Demerits : Performance is not satisfactory when SNR is low.
Low energy phonemes can't be detected.

4.2. SFD: Spectral Flatness Detector

The algorithms proposed so far are inefficient at low SNR. The following algorithm is intended to work even with low SNR. White noise has a flat spectrum while voiced signals have a non-stationary spectrum with more spectral content in the lower frequencies. Thus high variance implies speech content while low variance implies noise alone.

$$\sigma_i = \mathbf{VAR}\{X[f]\} \qquad (16)$$

Variance of each frame is compared against the variance threshold (σ_{th}) to determine its 'ACTIVITY'. An INACTIVE frame is used to update threshold value. The condition for presence of speech in the given frame is

IF $(\sigma_i > \sigma_{th})$ (17)

Frame is ACTIVE

ELSE **Frame is INACTIVE**

σ_{th} is updated during silence using the Convex Combination,

$$\sigma_{thnew} = (1\text{-}p)\,\sigma_{thold} + p\,\sigma_i \qquad (18)$$

This algorithm works well in low SNR conditions because the algorithm uses a statistical approach to the energy distribution in the spectra, unlike energy-based algorithms.

4.3. CVAD: Comprehensive VAD

It was observed that in the previous algorithms, only a few characteristics of speech are exploited. To obtain a better speech quality of reconstructed speech, the ideas discussed earlier are all incorporated into one algorithm. This VAD algorithm is capable of identifying white noise as well as frequency selective noise and maintaining a good quality of speech. The calculations of parameters for the previous algorithms remain the same but the decision rule is changed based on high priority for the Energy comparison. The decision flowchart for this algorithm is shown in

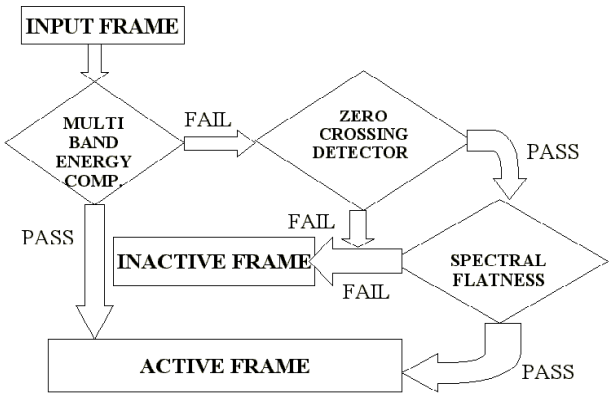

Figure 5. Flowchart for CVAD

Fig. 5. The decision rules are the same as in previous algorithms. Skipping the calculation of ZCD and Spectral flatness once the multi-band energy comparison passes the test can reduce computation.

Although the quality of speech is better compared to all other previous algorithms, its performance is poor for low SNR speech with variable background noise at the cost of higher complexity.

5. Results, Discussions and Comparisons

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 used. 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.
3. 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 randomly for an unbiased decision.
4. Objective Assessment of 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 is taken as the **MISDETECTION** percentage. This gives a quantitative measure of VAD performance.
   Though this number represents in a sense the quality of speech after applying a VAD technique, the quality of speech has to be assessed only by the **MOS *(Mean Opinion Score)***. This number gives an approximate assessment of the performance of an algorithm.

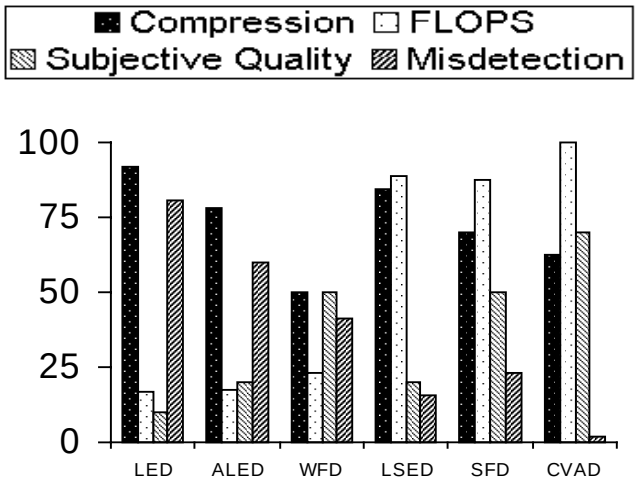

Figure 6. Dialogue

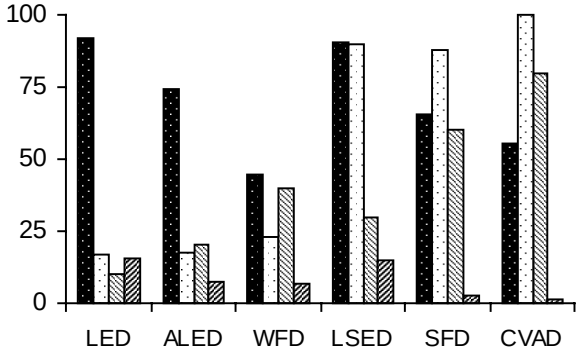

Figure 7. Discontinuous Monologue with low-energy phonemes

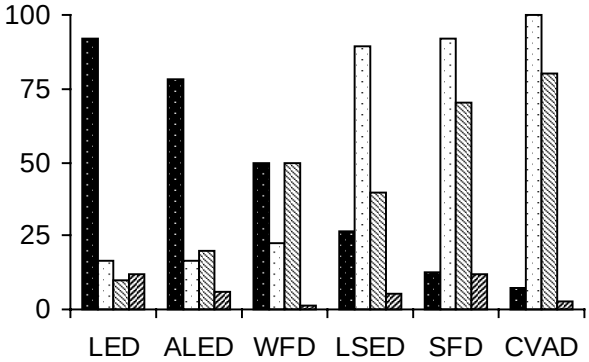

Figure 8. Rapidly spoken accented monologue

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

The figures given below are graphical presentation of the six algorithms with respect to Percentage Compression, number of FLOPS, Subjective Speech Quality and Misdetection for three different speech samples (or templates). We have taken three types of templates for comparison namely, *Dialogue*, *Monologue* and *Rapidly spoken Accented monologue*. All data have been normalized and scaled to 100 with respect to **CVAD** whenever normalization can't be done. For example, parameter FLOPS will be always high for **CVAD**, therefore the normalization is done with respect to **CVAD**. 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).

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

a. The time domain algorithms had the lowest FLOPS. This was expected, as the implementation was straightforward and not as complex as the frequency domain algorithms.
b. The Percentage Compression was low for the speech quality to be high. This is because some algorithms resulted in high compression rates at the cost of front-end clipping and non-detection of low energy phonemes.
c. The algorithms based solely on energy failed to deliver better speech quality with all the test templates. Spectral flatness and zero crossing detection gave better speech quality.
d. 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.
e. SNR affected all the algorithms except the last two. The spectral flatness concept was very effective in speech detection at low SNR.
f. Misdetection follows inversely with subjective speech Quality.

The algorithms are compared with each other for each template and then across the templates. In time domain algorithms, the **LED** has less computational requirement, the quality is poor compared to other algorithms and the percentage of compression is high. **ALED** improves quality but reduces the compression and has increased number FLOPS requirement. The **WFD** has the same trend and has better quality than the first two. In frequency domain solutions, the **CVAD** offers better speech quality compared to **LSED** and **SFD**. But the computational requirement is higher. **SFD** offers a better quality compared to **LSED** at the cost of less compression.

For all the speech templates we observe that compression reduces and quality increases from **LED** to **CVAD**. The time domain solutions are computationally less demanding but the quality of speech suffers, as misdetection is more. Quality of speech is high for **SFD** compared with **LSED** though the FLOPS are most often approximately the same.

6. Conclusions

VoIP has become a reality, though not yet very popular. This is predominantly due to existing systems being not very satisfactory or dependable. One solution lies in efficient VAD scheme used for VoIP systems. The time domain VAD algorithms are found to be computationally less complex but the quality of speech is poor compared to frequency domain algorithms. The frequency domain algorithms have better immunity to low SNR compared to time domain algorithms, however have higher computational complexity. We have proposed six VAD algorithms in time and frequency domain. The results consistently show superiority of the Comprehensive VAD scheme above all other algorithms. With this scheme good speech detection and noise immunity were observed. There is still performance degradation under low SNR conditions. This can be overcome using Cepstral methods [8]. The algorithms presented in this paper are found to be suitable for real-time applications.

7. 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, Pavel Sovka, and Jan Uhlir, "Cepstral Speech/Pause Detectors", *proc. of IEEE Workshop on Nonlinear Signal and Image Processing*, Neos Marmaras, Greece, June 1995, pp 388-391.

[9] 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.

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

[11] Stefan Pracht and Dennis Hardman, Agilent Technologies - "Voice Quality in Converging Telephony and IP Networks", *Ciscoworld Magazine* - White Paper 2001.

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

[13] 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.