International Journal of Soft Computing 2 (2): 257-261, 2007


# Performance Evaluation and Comparison of Voice Activity Detection Algorithms

[1]T. Ravichandran and [2]K. Durai Samy
[1]Velalar College of Engineering and Technology, Thindal,
Erode, Tamilnadu, India
[2]Dean, K.S.R. C.T, Tiruchengode, Erode, Tamilnadu, India

**Abstract:** One of the most difficult problems in speech analysis is reliable discrimination among Silence, Unvoiced and Voiced speech. Although several methods have been proposed for making this three level decision, this study provides a method to detect weak fricatives by Voice Activity Detection (VAD). VAD can be viewed as a decision problem in which detector decides between speech and silence. The signal is sliced into contiguous frames. Energy of a frame indicates the possible presence of voice data as it has high energy. Based on the assumption that zero crossing rates for speech and noise are generally different, low energy speech can be refined. High correlation found in speech samples can be exploited to detect weak fricatives. Various test samples are used by varying accent, loudness of male voice and female voice. A comparison of relative merits and demerits of three time domain VAD algorithms along with the subjective quality of speech after pruning of silence periods is presented .VAD is used in a variety of speech communication systems such as speech coding, speech recognition, hands-free telephony, audio conferencing and echo cancellation.

**Key words:** Voice activity detection, zero crossing rate, weak fricatives

## INTRODUCTION

Conversational speech is a sequence of contiguous segments of speech and silence. In a two way telephone conversation, one party is active for only about 35% of the time. This can be exploited effectively for the reduction of the average bit rate and co-channel interference in digital cellular systems (Khalid and Peter, 1997; Srinvasan and Gersho, 1993). The success of any scheme exploiting this property is critically dependent upon the algorithm used for Voice Activity Detection (VAD), i.e., the process of discrimination of speech from silence or other background noise. 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 period helps to reduce internet traffic (Ravichandren and Duriaswamy, 2005). VAD is used in a variety of speech communication systems such as speech coding, speech recognition, hands-free telephony, audio conferencing and echo cancellation (Jongseo *et al*., 1999).

In this study, various VAD algorithms are presented weth vared complexity and speech quality.

**Speech characteristics:** Speech signals are composed of sequence of sounds. Sounds can be classified into three distinct classes according to their mode of excitation.

- Voiced sounds are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing a quasi-periodic pulse of air which vibrates the vocal tract.
- Fricative or Unvoiced sounds a regenerated by forming a constriction at some point in the vocal tract and forcing air through the constriction at a high enough velocity to produce turbulence.
- Plosive sounds result from making a complete closure and abruptly releasing it

**Vad algorithms:** The basic principle of VAD device is that it extracts some measured features or quantities from the input signal and then compares these values with thresholds. Voice Activity (VAD = 1) is declared if the measured value exceeds the threshold. Otherwise no VAD = 0 is declared for no speech activity. In general, a VAD algorithm outputs a binary decision in a frame by frame basis where a frame of a input signal is a short unit of time such as 20-40 ms. Voice is differentiated into speech or silence based on speech characteristics (Prasad *et al*., 2002). The signal is sliced into contiguous

**Corresponding Author:** T. Ravichandran, Research Scholar, Velalar College of Engineering and Technology, Thindal, Erode, Tamilnadu, India

frames. A real valued nonnegative parameter is associated with each frame. If this parameter exceeds a certain threshold, the signal is classified as active else inactive.

The following are some of the required features of a good VAD algorithm:

**Good decision rule:** A physical property of speech that can be exploited to giveconsistent and accurate judgment in classifying segments of the signal into silence or otherwise.

**Adaptability to background noise:** Adapting to non stationary background noise improves the robustness, especially in wireless telephony.

**Low computational complexity:** The complexity of VAD algorithm must be low to suit real-time applications.

**Energy Based Detection (EBD):** The amplitude of the speech signal varies appreciably with time. The amplitude of unvoiced segments is generally much lower than that of voiced segments. The energy of a signal represents a convenient representation that reflects the amplitude of the signal. Energy of a frame indicates the possible presence of voice data and is an important for VAD algorithms (Prasad *et al*., 2002).

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

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

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

The VAD algorithm is trained for a small period by a prerecorded sample that contains only background noise. The initial threshold for various parameters is computed from these samples. The initial energy threshold is obtained by taking the mean of the energies of each sample

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

$E_r$ = initial threshold estimate
U = number of frames in a prerecorded sample

The *classification rule* for speech is as follows
IF

$$E_j \rangle KE_K \quad (K \rangle 1) \quad (4)$$

Frame is Active
ELSE
Frame is Inactive
Here $E_r$ represents the energy of noise frame, while $kE_r$ is the threshold being used in the decision making. Active frames are transmitted while Inactive frames are not transmitted.

Since background noise is non stationary, an adaptive threshold is more appropriate. The rule to update the threshold value is

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

$E_{mew}$ is the updated threshold value
$E_{rold}$ is the previous energy threshold
$E_{moise}$ is the energy in the most recent noise frame

Energy based decisions are not good for low energy phonemes. Weak fricatives are sometimes silenced completely. High energy voiced speech segments are detected in all VAD algorithms even under noisy conditions. However low energy unvoiced speech is commonly missed, reducing speech quality. So to overcome this drawback, the following two methods are proposed.

**Zero Crossing Detector (ZCD):** The notion of zero crossing detector is defined to be the number of times in a sound sample that the amplitude of the sound wave changes sign. For a 10ms sample of clean speech, the zero crossing rates is approximately 5 to 15 for voiced speech and 50 for unvoiced speech. For clean speech, the zero crossing rates should be useful for detecting regions of silence, as the zero crossing rates should be zero. Very few sound samples are recordings of clean speech. This means that often there is some level of background noise that interferes with the speech meaning that silent regions actually have some zero crossing as the signal changes from one side of zero amplitude to the other and back again. This reason allows us to formulate a decision rule that is independent of energy and hence able to detect some low energy phonemes.
IF

$$N_{zcs}(f_j) \in R \quad (6)$$

Frame is Active
ELSE
Frame is Inactive
$N_{zcs}$ is the number of zero crossings detected in $f_j$
R is set to the value of {5-15}.

ZED checks the voice activity of the frames that were declared to be Inactive by Energy Detector. This ZCD detects almost all the low energy speech phonemes.

**Eigen Value Based Detector (EVD):** A drawback of ZCD is that it misclassifies noise frames as Active when zero

crossings of the noise frames satisfy the above equation. This may lead to failure of the algorithm. So a method is required that classifies Eigen Value Detector from noise dependent. This particular problem can be made to overcome by using the Auto correlation function which is exploited by the high correlation found in speech signals.

The unbiased autocorrelation function is given by

$$A[x] = \frac{1}{m} \sum_{m=N-1} y[n] \times y[n-m] \tag{7}$$

A[X] = the autocorrelation vector
y[n] = vector under consideration
N = Frame length

Each frame of the incoming signal is classified into frames of duration 20ms.The energy of each frame is computed as shown in the equation below

$$E_{subframe} = \sum_{index=1}^{4} x^2((subframe-1)\times 4 + index) \tag{8}$$

Where sub framestakes the value from 1 to the total number of sub frames in the sample.
Index denotes each sample in the given vector.

We aim at finding real generalized Eigen values for a definitizable pair of real symmetric matrices A. if there exists a number ë and a non- zero vector x,
such that

$$Ax = \lambda x \tag{9}$$

Then $\lambda$ is said to be an eigen value of A and x the corresponding eigen vector. Note that $\beta x$, where $\beta$ is any real number, is also an eigen vector, since

$$A(\beta x) = \beta Ax$$
$$\beta\lambda x = \lambda(\beta x)$$

this shows that an eigen vector is not unique and may be scaled, if desired Eq. 9.

$$(A - \lambda I)\, x = 0 \tag{10}$$

This equation has a (nontrivial) nonzero solution x if the matrix A - $\lambda$I is singular [i.e., if (A - $\lambda$I) is noninvertible], which is the case if the determinant of (A - $\lambda$I) is zero, that is, if

$$\det\ (A\ - \lambda I) = 0 \tag{11}$$

The eigen values of A are given by the roots of the secular Eq. 1

$$\det(A - \lambda I) \equiv f(\lambda) = 1 + \beta \sum_{i=1}^{M} \frac{q_i^{\,2}}{d_i - \lambda} \tag{12}$$

To obtain n approximation for the maximum eigenvalue, we rewrite (12) as

$$f(\lambda) = 1 + \beta \sum_{i=2}^{M} \frac{q_i^{\,2}}{d_i - \lambda} + \frac{\beta q_1^{\,2}}{d_1 - \lambda} \tag{13}$$

Since $f(\lambda)$ has a pole at $d_1$, the last term dominates the behaviour of $f(\lambda)$
For $\lambda \approx d_1$. The second term, which is the sum of terms associated with the other eigenvalues, is almost constant in the region around $\lambda_{max}$. Thus, we can write

$$f(\lambda_{max}) \approx 1 + \beta \sum_{i=2}^{M} \frac{q_i^{\,2}}{di - dl} + \frac{\beta q_1^{\,2}}{dl - \lambda_{max}} \tag{14}$$

The solution of $f(\lambda_{max}) = 0$ yields

$$\tilde{\lambda}_{max} = \gamma\lambda_{max}^{0} + \alpha + \frac{\beta q_1^{2}}{1 + \beta \sum_{i=2}^{M} \frac{q_i^{2}}{(d_i - d_1)}}$$

Where we have replaced $d_1$ in the last term (14) by its definition given in (12). Note that
$\tilde{\lambda}_{max} < \lambda_{max}$ where $\lambda_{max}$ is the true maximum
eigenvalue of R
The voice signal(x), R is set through is Eigen value ($e_m$) to assign the threshold value of the function threshold is
eigen $(e_m)^T$ = eigen value $e_m(x)$
IF
eigen $(e_m)^T(T>0)$ Then Active voice
ELSE
Inactive voice
END

Eigen Value Based Detector, VAD algorithm are used to all template voice signal evaluated the performance. Eigen value based detector has well classified the noise and silence in the voice signals. The quality of service is to improve bandwidth saving of suppression on silence ratio is highly used to the all templates and rejection of the noise.

**Median Value Based Detector (MVD):** Median value based detector is a generalization of the median filter, where a non-voice signal is assigned to each input sample in the observation of the filter. Given a discrete time sequence {x(n)}, a filter observation

x = [x(n), x(n-1),……x(n- N+1)] and a corresponding set of noise signals are
($w_0$, $w_1$, ……$w_{N-1}$), the output form of a filter is given by
y(n) = Median[w*x]
= Median[x(n)$w_0$x(n)$w_1$.....x(n-N+1)$w_{N-1}$]each sample x(n-i) to the number of the corresponding noise $w_i$ and choose the median value calculated from the new sequence. The sequence after the FIR filter can be written as

$$Q(n) = 1/R \sum_{I=1}^{R} q(n- i + 1)$$

The over sampled filtered signal q(n) is next decimated down by R yielding.
$x_d(n) = q(nR)$

$$= 1/R = R \sum_{I=1}^{R} q(Rn- i + 1)$$

Next, define theR- using vector Q(n) as
Q(n) = [ q(n),q(n-1),……, q(n-R+1)] T

$$= [ Q_1(n),Q_2(n),\ldots,Q_R(n)]$$

$$\text{The singal } x_d(n) = 1/R \sum_{I=1}^{R} Q_i(Rn)$$

Median filtering a signal at the Nyquist rate can be writte
y(n) = M [ $x_d$(n-N),…,$x_d$(n+N)]

$$= 1/R\ M[\sum_{I=1}^{R} Q\ i(R(n-N)),\ldots,\sum_{I=1}^{R} Q\ i(R(n+N))]$$

The median value is used to reference[y(n) = $m_{ref}(x)^T$] threshold of speech signals divides into voice segments and un-voice segments.
Threshold is [y(n) = $m_{ref}(x)^T$]
IF
Median ($m_{ref}(x)^T$ ) Then Active voice
ELSE
Inactive voice
END

The lower median value of the voice signals are eliminated noise and silence in the voice transmission.

## RESULTS AND DISCUSSION

MATLAB was used to test the algorithm developed on various sample signals. Various test samples were used by varying accent, loudness, male voice and female voice .The performance and results are shown below.

**A criteria for assessing VAD performance:** Performance of VAD was analyzed basedupon the following criteria:

- Compression: The ratio of total INACTIVE frames detected to the total number of frames formed.
- Compression Ratio = Inactive Frames / Total Frames
- Subjective Speech Quality:The quality of the samples was related on a scale of 1 to 5 with for poorest and 5 for best. The input signal is assumed to have a good quality of rating 5.The speech samples after compression were played to users.
- Misdetection values are calculated inverse of quality to values assign MOS (Mean Opinion Score).

An effective VAD algorithm should have high compression while maintaining the speech quality.

**Graphical representation of result:** Performances of various VAD algorithms were analyzed and their results were shown in graphical form Fig. 1-3. Three templates were taken for comparison namely monologue, discontinuous monologue and rapidly spoken monologue

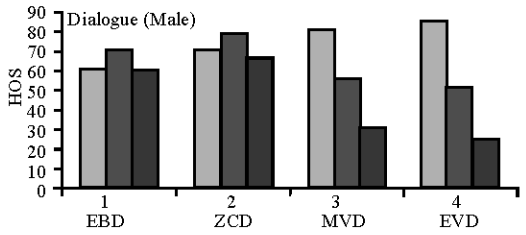


Fig 1: Discontinuous dialogue (Male voice)

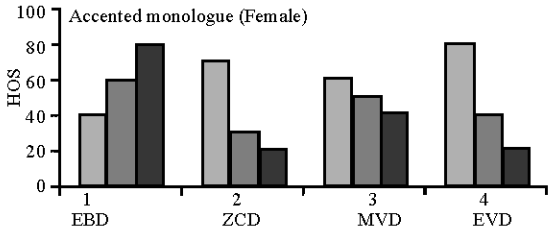


Fig 2: Accented monologue (Female voice)

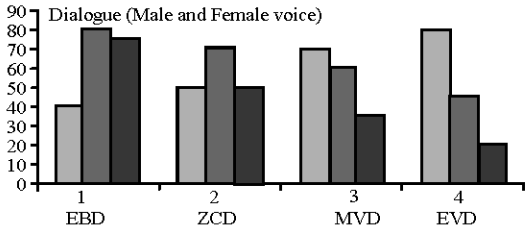


Fig 3: Rapidly spoken dialogue (Male and female voice)

These templates were used with both male voice and female voice. All these were tested for their quality and compression.

**Observations:** The following were some observations that were observed during the implementation

- The algorithm based on energy did not give an acceptable speech quality for the all the templates taken.
- Zero crossing detector was used to recover some low energy phonemes that were rejected by energy based VAD.
- Median Value based Detector was more effective on silence suppression and less quality of speech.
- Autocorrelation Vector method provides better results compared to zero crossing detector and median value based detector in detecting eigen value based detector.

## CONCLUSION

We have presented in this study a comparative study of three VAD algorithms under different cases.With these schemes, good speech detection and noise immunity were observed.

## REFERENCES

Jongseo Sohn, Nam Soo Kim and Wonyong Sung,1999. A statistical model based voice activity detection, IEEE Signal Processing Letters.

Khaled El-Maleh and Peter Kabal, 1997. Comparison of Voice Activity Detection for Wireless Personal Communications Systems, IEEE Canadian Conf. Elec. Computer Eng., pp: 470-473.

Prasad, V. *et al*., 2002. Comparison of Voice Activity Detection Algorithms for VoIP, submitted to the Seventh IEEE Symposium on Computers and Communications, Tacrmina.

Ravichandran, T. and K. Duraiswamy, 2006. Factor Influencing voice Transfer using VOIP-A Performance Enhancement Technique, accepted for the publication in ACCST Res. J., Vol-IV.

Srinivasan, K. and A. Gersho, 1993. Voice activity detection for cellular networks, in Proc. IEEE Speech Coding Workshop, pp: 85- 86.