Voice Activity Detection

Dhaval Shah(03307008)
Supervisor:Prof.Preeti Rao

Abstract

Voice activity detection is an ubiquitous problem in speech processing, as it having wide range of application. Various application specific voice activity detection algorithm methods have been devised over the time to detect voice segments on the basis of the speech signal characteristic. Firstly, in this report the requirement of voice activity detection(VAD), and it's applications are discussed. It is followed by desired properties of VAD algorithm, and important parameter for VAD algorithm design. Earlier Time domain and frequency domain VAD algorithms based on energy levels, zero crossing rate, spectral energy, spectral flatness band of speech signal are discussed, Also there comparative performance evaluation is presented. Further, very robust method of VAD based on periodicity, geometrically adaptive energy level which operates reliably in non-stationary noise, are discussed in detail. The fusion of periodicity and geometrically adaptive energy threshold of signal which operates accurate down to -5 dB SNR are described.

1.Introduction

Conversation speech is a sequence of contiguous segments of silence and speech. Where, term silence means pauses between two consecutive words, or intra-word pause. Speech detection consists of the classification of the two clearly distinct signal conditions:periods where the speech signal is present, and periods where the pauses present.

In reality, speech signal is always accompanied by some noise. In most cases background noises of environment where the source speech lies, is the the main component of the noise that adds to the speech signal. Hence, recorded speech signal have presence of background noise during pause period, and during speech period, speech and background noise both are present.

Due to the background noise presence in speech source, it is difficult to differentiate speech period and silence(pause) period. This leads to negative effect on the performance of speech processing application. In addition, it degrades further processing of the speech signal. For instance,in speech recognition, word boundary must be approximately known to trigger recognizer correctly. The main objective of the Voice Activity Detection(VAD) or speech detection is, to detect speech and pause periods correctly from noisy speech. Hence it leads to accurate speech processing for subsequent stages. VAD algorithms take recourse to some form of speech pattern classification to differentiate between voice and silence period.

1.1 Speech degradation at the source of speech

Speech can be degraded at any stage before it reaches the end listener. The different ways in which speech can be degraded broadly categorized below.

1. At the source of speech
2. During the transmission of speech signal
3. Noise at the listener end

VAD is normally applied on speech recorded at the speech source or on real time noisy speech signal from speech source, as a preprocessing technique before transmission of speech, hence we only discuss the speech degradation at the speech source. When the speech source itself is in noisy environment, the background noise may be the noise like such as aircraft cockpit, other moving vehicle or environmental sounds; or it may be speech-like, comprised of competing speakers. Besides this, source of speech is present. Here the noise involved is multiplicative noise, where the noise gets convolved with the speech signal.

1.2 Need for voice activity detection

The background noise generally degrades the quality and intelligibility of speech. Due to intelligibility degradation, content of speech can not be recognized properly or background noise is detected as a speech signal. To avoid the mis-detection of speech signal, there is a need of technique like VAD. To be specific, in the applications where decision control is made on the basis of speech signal, VAD is used as preprocessing technique for speech detection.

1.3 Application of VAD

1. In speech recognition system : In speech recognition, word boundary must be approximately known to trigger the recognizer correctly, boundary detection refers to begin-endpoint detection. Hence, it depends on VAD.
2. Adaptive speech enhancement: Adaptive speech enhancement algorithms typically behave completely different during speech periods than during noise periods. So, correct voice activity detection is crucial for its success.
3. In speech coding and compression: In speech coding and compression, the bit-rate can be lowered drastically during silence period without effect on perceived speech quality. So, VAD helps in efficient coding, also aid in saving bandwidth requirement.
4. In hands-free telephony: It is used to transmit the voice only, when speech is uttered

1.4 Desired properties of VAD algorithms

A properties of VAD algorithm is described in paper by R.Prasad and ahijeet [1].

1. A good decision rule and accuracy : A physical property of speech that can be exploited to give consistent judgment in classifying segments of the signal into silence or speech segments. Also, the rate of mis-detection should be less.
2. Adaptability to changing background Noise : Adapting to non-stationary background noise improves robustness, especially in wireless telephony where the user is mobile.
3. Low computation complexity : It is required for real time processing of speech signal, so speech signal can be transmitted without significant delay.

1.5 Classification of Voice Activity Detection

Voice Activity Detection(VAD) is classified on the basis of characteristics of speech signal is used by particular algorithm. According to characteristics of speech VAD algorithms are based on distance measure, energy levels, pitch, cepstral features, spectrum of noise, periodicity, and zero crossing rate are developed.

Broadly the VAD technique is classified as time-domain VAD, frequency-domain VAD, and fusion Methods VAD on the basis of speech parameter of speech signal is used. Time domain VAD algorithms use the time domain characteristic of speech signal, and time domain parameter like energy, zero crossing rate, periodicity.

Frequency domain VAD algorithms use frequency domain characteristics like variance, energy, and power spectrum of the speech signal. Fusion methods base VAD algorithm uses the two or more than two parameter for speech detection.

Time domain VAD algorithms based on energy, periodicity and the zero-crossing rate of speech signal is discussed in section 2. Frequency domain algorithm based on energy, and variance of speech signal spectrum is describe in section 3. Fusion of method based on geometrically adaptive of energy threshold method, and periodicity measure is discussed in detail in section 4. Performance evaluation of discussed VAD algorithm is described in section 5. Finally, conclusions is described in section 6.

2. Time Domain-VAD Algorithms

Time domain features of speech signal is used to develop, the VAD algorithm. First, the important parameter for VAD design described by R.Prasad and Abhijeet [1] is discussed. After that, following VAD algorithm is discussed in detail.

1. Linear energy-based detector
2. Adaptive linear energy-based detector
3. Weak fricative detector
4. Detector based on periodicity measure

We next discuss the algorithm parameter needs for VAD design.
The speech signal is sliced into contiguous frames. A real-valued non-negative parameter is associated with each frame. If this parameter of frame exceeds a certain threshold, the signal frame is classified as active(having speech content) or else it is inactive(silence or pause period).

1. Choice of frame duration : Selection of the frame size of speech signal depends on the particular application for which VAD is used. If VAD algorithm is complex, and required more time for computation then, frame size should be at least that much, it can perform it computation during the frame period. The maximum size of the frame should be relevant to minimum word length is assumed approximately 150 ms, and minimum gap length is 100 ms ,as per described by S.V.Gerven and Fei [6].
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 i^{th} sample of speech. If the length of the frame were k samples, then j^{th} frame can be presented in time domain and frequency 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 = energy of the j^{th} frame and
   f_j is the j^{th} frame that is under consideration.

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 [1].
   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 parameter 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

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

   where, E_r = initial threshold estimate,
   v = number of frames in prerecorded sample.
   Method 2: In this method, it is assumed that the initial some duration of signal does not contain any speech. After that, mean energy is calculated by methods described above. A fixed threshold would be remain unchanged to varying acoustic environments of the speaker.

2.1 Linear Energy-Based Detector

Energy of a frame is reasonable parameter on the basis of which frames may be classified as containing speech or pause. The energy of the frame which contains speech is higher than that of frames contain pause period. The classification rule is,

$$\textbf{If} \quad (E_j > kE_r) \quad \textbf{where k} > 1 \quad \textbf{Frame is active} \tag{4}$$
$$\textbf{else} \quad \textbf{Frame is inactive}$$

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 adaption of E_r, and therefore, the threshold.

2.1.1 computation of E_r

computation of E_r is required since background disturbance is non-stationary an adaptive threshold is more appropriate. The rule to update the threshold value is,

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

here, E_{rnew} is the updated value of the threshold,
E_{rold} is the previous energy threshold,
$E_{silence}$ is the energy of the most recent noise frame.
The reference, E_r is updated as convex combination of the old threshold and the current noise update p is chosen considering the impulse response of Eq.(6) as first order filter $(0 < p < 1)$.
The Z-Transform of Equ.(6) is,

$$E_r(Z) = (1-p)\, Z^{-1} E_r(Z) + pE_{noise}(Z) \tag{6}$$

The Transfer function may be determined using,

$$H(Z) = \frac{E_r(Z)}{E_noise(Z)} = \frac{p}{1-(1-p)\, Z^{-1}} \tag{7}$$

For p = 0.2, the fall time is 95% corresponds to 150 ms is observed by R.prasad and Abhijeet [1]. Usually, pasuses between two syllabi are about 100 ms and these pauses should not be considered as silence. The fall-time selected greater than this value, so that pauses do not affect updating of E_r.

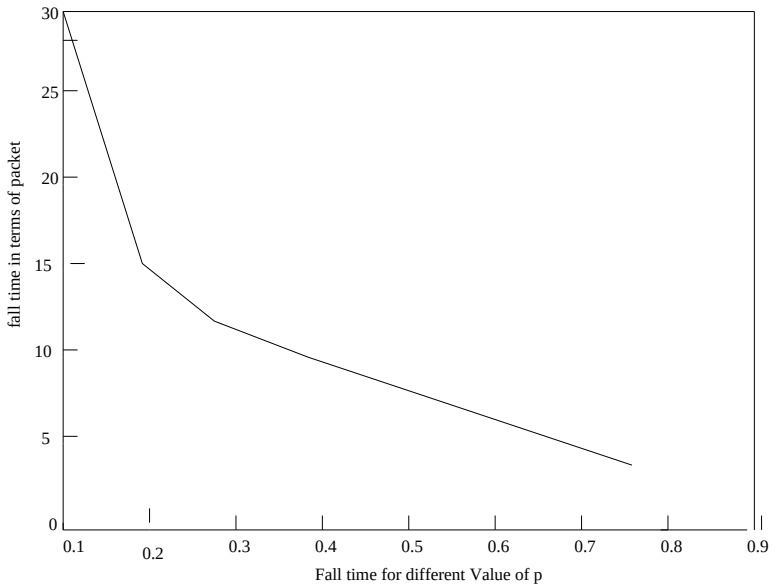

Figure 1: Fall-time for different value of p [1]

2.2 Adaptive Linear Energy-Based Detector

This VAD algorithm calculates value of E_r based on second order statistics of pause period frames. A buffer (linear queue) of the most recent 'm' silence 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 given by

$$\sigma = VAR[E_{silence}] \tag{8}$$

A change in 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 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

$$\sigma_{new} > \sigma_{old} \tag{9}$$

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

$\frac{\sigma_{new}}{\sigma_{old}} \geq 1.25$	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 value of p, is chosen according to ratio of new variance to old variance. AS the value of p is varied the adaption is more profound.

$$E_{rnew} = (1-p)\, E_{rold} + pE_{silence} \tag{10}$$

here, E_{rnew} is the updated value of the threshold,
E_{rold} is the previous energy threshold,

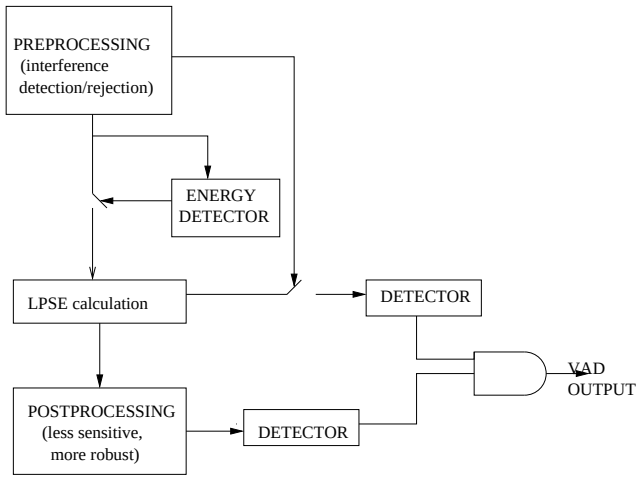

Figure 2: LSPE-based VAD [2]

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

Although, equ.(10) to calculate reference energyE_r is same to previous method, the value of p depend variance of energies of inactive frames. Hence E_r can track faster to quickly changing background noise.

2.3 Weak Fricative Detector

Previous two methods in this section of Voice Activity Detection(VAD) are based on the enegy of the frames. 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, thus speech quality is degraded. Weak Fricative Detector(WFD) is designed to overcome this problem [1]. The number of zero crossings of voiced signal lies in a fixed range.to be specific, for voice speech signal zero crossing rate is 500 to 1500 samples/sec. while, the number of zero crossing of the noise is random and unpredictable. This property allows to formulate the decision rule that is independent of energy and therefore, is able to detect low energy phonemes. Zero crossings for each frame are computed by the following decision rule

$$\textbf{If} \quad (N_{zcs}(f_j) \in R) \quad \textbf{Frame is active} \qquad (11)$$
$$\textbf{else} \qquad \textbf{Frame is inactive}$$

where, N_{zcs} is the number of zero crosses detected in a frame.
R is the the set of values of the number of zero crossing for active frame.

The zero crossing detector can detects as active frame that can not be declared by energy-based VAD.

2.4 Least-Square Periodicity Estimator(LSPE) Based Detector

This method is described by Tucker in his paper on voice activiey detection by periodicity measure [2]. The structure of LSPE-based VAD shown in figure 2.In addition to the main LSPE section, there are preprocessing and post-processing sections and energy detector. The LSPE calculation uses non-overlapping 25ms frames on a 200-1000 Hz bandwidth signal. A narrow bandwidth is used to minimize the probability of an in-band interference signal, while still allowing effective periodicity detection. The signal is two-times oversampled at 4 kHz to give extra resolution for the periodicity detection.

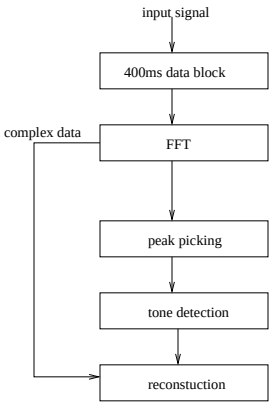

Figure 3: Tone and data elimination system [2]

The energy detector is included to prevent the detection of very low-level signals in the presence of larger signals.If AGC has not been applied to the input signal, the detection threshold needs to adapt to the input signal level. The purpose is to pass signals that might be speech, while rejecting any signal that is definitely not speech.

The preprocessor is needed to detect and, if possible, remove periodic interference. When it detects interference, the preprocessor suppresses the LSPE detector, because there may be residual periodicity even after tone removal. The post-processor is required to provide a less sensitive but more robust detector in the presence of interference.

2.4.1 Preprocessing

A.Requirement

The preprocessor must be able to detect, and if possible to remove, all the expected types of interference.Environment may have broader bandwidth periodic interference(in-car noise, for instance). The preprocessor described here is designed to detect and remove constant-frequency data and tones [2].

B.Data and tone detection elimination

Constant-frequency interference is identified by consistent peaks in the spectrum over a time period. But, speech can also have a constant pitch, so a harmonic speech is needed to differentiate between speech and constant-frequency interference.

Figure 3 shows the detection elimination system. The input signal is formed into 400 ms blocks. Each block is divided into 25 frames, 32 ms long and overlapping by 16 ms, and each frame is windowed with a Hanning window. The input signal is full bandwidth (0-2 kHz) at this point. Using 400 ms blocks allow 25 FFT frames and 16 LSPE frames to be processed in the same interval.The FFT produces two outputs:the complex transform data, which is needed to reconstruct the input data once tone frequencies have been zeroed; and the log magnitude of the complex transform data, required for peak picking.

The peak picker makes a simple binary decision as to where the peaks in each FFT frame, independent of adjacent frames. Each peak p_i consists of five adjacent frequency points in a frame with magnitude that satisfy

$$p_{i-2} < p_{i-1} < p_i$$

$$p_{i+2} < p_{i+1} < p_i \tag{12}$$

The tone detector looks over three consecutive 25-frame (400 ms) blocks, and labels all frequency points that are peaks in more than five frames of each block. Demanding peaks to be present for at least 20% of each blocks keeps spurious detection of noise to an acceptable level. Any labelled frequency with a nearby peak in 73 of the 75 frame is assumed to contain a steady,continuous tone [2].

Remaining labeled frequencies could be either interference or speech pitch bars. To find pitch bars, each frame in three blocks is searched for at least two harmonics, spaced less than 400 Hz and starting peak near the labeled frequency. If harmonics with similar spacing are found in consecutive frames at least twice over the three blocks, the labeled frequency is assumed to be a speech pitch bar, otherwise it is assumed to be interference.

Interference-free data is reconstructed by zeroing all FFT points within two points of interference frequencies. In addition, the data are band-limited to between 200 Hz and 1000 Hz by zeroing all FFT points outside that range. After the inverse FFT, adjacent half-frames are added together to compensate for the windowing that took place before the FFT.

2.4.2 LSPE processing

A.Periodicity calculation

The LPSE calculation produces a periodicity value for each 25 ms frame of the input signal s(i).

$$s\left(i\right) = s_o\left(i\right) + n\left(i\right)$$

for i=1, 2.....N(N = 100 for 25 ms frame)
where
$s_o\left(i\right)$ = periodic component of input signal $s\left(i\right)$
$n\left(i\right)$ = non-periodic component of $s\left(i\right)$
$s_o\left(i\right) = s_o\left(i + kp_o\right)$
P_o = period of $s_o\left(i\right)$

We now let $\hat{P}_o$ is our estimate for P_o and $\hat{s}\left(i, \hat{P}_o\right)$ be the corresponding estimate of the periodic component, which is referred as $\hat{s}_o\left(i\right)$. $\hat{s}_o\left(i\right)$ is obtained form the input signal

$$\hat{s}_o\left(i\right) = \sum_{h=0}^{k_o} \frac{s\left(i + h\hat{p}_o\right)}{K_o} \tag{13}$$

$$1 \leq i \leq \hat{P}_o \qquad P_{min} \leq \hat{P}_o \leq P_{max}$$

where P_{min} and P_{max} are the minimum and maximum number of samples in a pitch period, and $K_o = [\frac{(N-i)}{P_o}]+1$ is the number of periods of $\hat{s}_o\left(i\right)$ in the analysis frame.

The objective of the least-squares method is to find the pitch period $\hat{P}_o$ that minimizes the mean square error $\sum_i^N [s\left(i\right) - \hat{s}_o\left(i\right)]^2$ over each analysis frame. So, the normalized periodicity measure is given by:

$$R_1\left(\hat{P}_o\right) = \frac{I_o\left(\hat{P}_o\right) - I_1\left(\hat{P}_o\right)}{\sum_{i=1}^{p} s^2\left(i\right) - I_1\left(\hat{P}_o\right)} \tag{14}$$

where

$$I_1\left(\hat{P}_o\right) = \sum_{i=1}^{p} \sum_{h=0}^{k_o} \frac{s\left(i + h\hat{P}_o\right)^2}{K_o} \tag{15}$$

and

$$I_0\left(\hat{P}_o\right) = \sum_{i=1}^{N} \hat{s}_o^2(i) = \sum_{i=1}^{P_o} \frac{[\sum_{k=0}^{K_o} s\left(i + h\hat{P}_o\right)]^2}{K_o} \tag{16}$$

For each 25 ms frame, $R_1\left(\hat{P}_o\right)$ is computed for values of $\hat{P}_o$ between $P_{min} = 11$ and $P_{max} = 56$, and the maximum value of $R_1\left(\hat{P}_o\right)$ obtained is the periodicity of the frame [2].

B.LSPE detector

The LPSE calculation produces a periodicity value for each 25 ms frame. As periodicity of around 0.5 are often obtained from white noise, the detector subtracts 0.5 from the periodicity and sets negative values to zero. It then sums periodicities from six consecutive frames to smooth out random periodicity peaks.

2.4.3 Post-processing

A.Requirement

The post-processor must look for particularly speech-like characteristics in the periodicity function$R_1\left(\hat{P}_o\right)$. For this purpose, Speech Signal Quality Estimator(SSQE) is used. The SSQE produces an output based on the harmonic spacing with in each frame and the peak movement between frames.

B.SSQE post-processor

In the beginning, $R_1\left(\hat{P}_o\right)$ is searched for all significant peaks. For speech, the peaks will be harmonically related, with a frequency lying within the normal pitch range of 11-56 samples (71 Hz-363 Hz). For interference, the peak may well be much closer together, and for noise they may not be harmonically related. So, sum s_v of all peaks which are harmonically related at pitch frequencies is calculated. The second is the ratio r_{vm} of sum s_v to sum of all other peaks. From this two parameters, probability of voicing p_v is calculated.

$$p_v = ws_v \tag{17}$$

where the weighting factor w is given by

$$\begin{aligned} w &= 0 \quad \text{if } r_{vn} \leq 1 \\ w &= 0.66\,(r_{vn} - 1) \quad \text{if } 1 < r_{vm} < 2.5 \end{aligned}$$

$$w = 1 \text{ if } r_{vm} \geq 2.5 \tag{18}$$

The peaks found by the peak picker are tracked using a dynamic programming algorithm over four consecutive frame. Associated with each trajectory is a 'winner' function, which is a measure of the steadiness of the pitch function, w_{max} is taken as the measure of peak movement and is combined with p_v to give the SSQE ouptut.

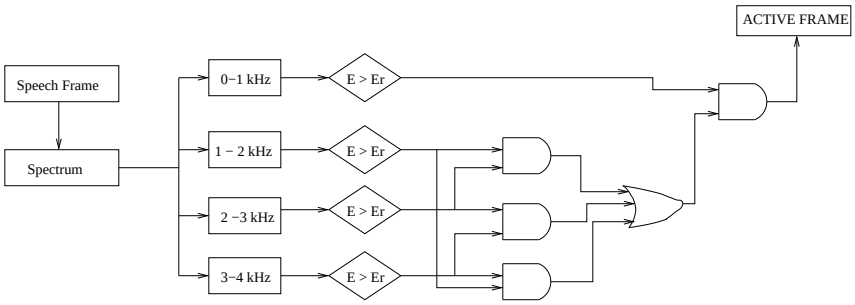

Figure 4: Flowchart for LSED [1]

C.Post-processor detector

In the post-processor detector, the SSQE output is clipped and summed in a similar way to the LSPE detector. Because noise gives very little SSQE output, it is worth summing over as long a period as possible to give maximum detection capability, but at the same time this would add too much delay. Detection threshold is kept such a way that false triggering due to noise may occur very rarely.

3 Frequency Domain-VAD Algorithms

The following algorithms take into consideration the frequency-domain characteristics of speech signals.

1. Linear Sub-Band Energy Detector-LSED
2. Spectral Flatness Detector-SFD
3. Detector based on Power Envelope Dynamics of Speech Signal

3.1 Linear Sub-Band Energy Detector-LSED

This algorithm takes its decisions based on energy comparison of the signal frame with a reference energy threshold in the frequency domain [1]. Energy of the frame in frequency domain is calculated,

$$(f_j) = DCT(f_j) \tag{19}$$

The spectrum of the frame after that, divided into four bands of width 1 kHz, so, four bands are 0-1 kHz,1-2 kHz, 2-3 kHz, 3-4 kHz. and then the energy for each band is calculated as,

$$E_n[f] > F^2\,(f_n) \qquad \text{for } n_{th}\text{band} \tag{20}$$

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

$$E_n[f] > kE_{nth}[f] \qquad \text{for } n_{th}\text{band} \tag{21}$$

The thresholds are computed recursively, but for each band separately by equation,

$$E_nthnew = (1-p)\,E_{nthold} + pE_{nthnew} \tag{22}$$

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

3.1.1 Fraction of energy in lowest frequency band

Most of energy in voice signal tends to be in the lowest frequency band, 0-1 kHz. Selective threshold comparison in the lowest band alone provides good decisions [1].

3.1.2 Decision Rule for Speech

A frame is detected to be ACTIVE or having speech signal if the lowest band is active and any two out of the remaining three bands are active.

3.2 Spectral Flatness Detector

In the beginning, Spectral Flatness Detector calculate the initial threshold for variance of spectrum on initial duration assuming it does not contain any speech signal.Initial threshold of variance is calculated by,

$$\sigma = VAR[F(f_j)] \tag{23}$$

where j is the the variance of the j_{th} frame.

Speech signal has non-stationary spectrum with more spectral content in the lower frequencies, while White noise has a flat spectrum. so hight variance implies speech

$$\textbf{If} \quad (\sigma_i > \sigma_{th}) \quad \textbf{Frame is active} \tag{24}$$

$$\textbf{else} \quad \textbf{Frame is inactive}$$

σ_{th} is updated during the silence period, using the equation,

$$\sigma_{thnew} = (1-p)\,\sigma_{thold} + p\sigma_i \tag{25}$$

This VAD method is using statistical approach on the frequency spectrum of the frame, it gives more accuracy than previous frequency domain-VAD technique.

3.3 Detector based on power spectral dynamics of signal

This speech-pause detection technique by using power envelope dynamics of noise is developed by Marzinzik [3].VAD technique calculates the signal's temporal power envelope $E(p)$ by summing up the squares of the spectral components of the input signal in each short time frame p [3].

$$E(p) = \sum_k |X(p,\omega_k)|^2 \tag{26}$$

Where, $X(p,\omega_k)$ denotes the spectral component of the noisy input signal at frequency ω_k at time frame p. In addition, a low-pass band power envelope ans a high-pass band power envelope are calculated by following equation,

$$E_{LP}(p) = \sum_k |X(p,\omega_l)|^2 \tag{27}$$

$$E_{HP}(p) = |X(p,\omega_m)|^2 \tag{28}$$

Where, l runs over 0 to cutoff-frequency of low pass filter, while m runs over remaining spectral frequencies.

For smoothening of the envelope, $E(p), E_{LP}(p)$ and $E_{HP}(p)$ are averaged over a few frames by a recursive low-pass filter of first order with a release time constant τ_E

1. By assuming that, 200 ms initial phase of signal is consist of noise only the minimum and maximum values are set as follows:

$$E_{min}(p) = E(p) \qquad E_{max} = E(p)$$

$$E_{LP,min}(p) = E_{LP}(p) \qquad E_{LP,max} = E_{LP}(p)$$

$$E_{HP,min}(p) = E_{HP}(p) \qquad E_{HP,max}(p) = E_{HP}(p) \tag{29}$$

This guarantees that the minimum envelope values correspond roughly with the noise energy at the beginning [3].

2. The minimum and maximum values are updated for each of the three envelopes in the following manner.

   - If the current envelope value is larger than the maximum value for the corresponding envelope, then the maximum value is set to the current value. Otherwise, the maximum value slowly decays. This is done by a recursive low-pass filter of first order with a release time constant τ_{delay}, which takes as input the current envelope value.
   - If the current envelope value is smaller than the minimum value for the corresponding envelope, then the minimum value is slowly raised. This is done by a recursive low-pass filter of first order with attack time constant τ_{raise}, which takes as input the current envelope value

3. The difference between the maximum and the minimum values are calculated for each envelope

$$\Delta_{HP}(p) = E_{HP,max}(p) - E_{HP,min}(p)$$
$$\Delta_{HP}(p) = E_{HP,max}(p) - E_{HP,min}(p)$$

$$\Delta_{HP}(p) = E_{HP,max}(p) - E_{HP,min}(p) \tag{30}$$

4. Three different criteria are introduced of which only one has to be true for making the decision that target speech is not present in the actual frame: (a) the speech pause decision can be made because of a low signal dynamics in both low-pass and the high-pass band. (Dyn Speech pause); (b) the decision can be based on the low-pass band information (LP Speech Pause); and (c)it can be made upon the high-band information (HP Speech Pause). These decision criteria as follows [3].

   (a) If Δ_{LP} is smaller than some threshold η and also $\Delta_{HP} < \eta$ then it is assumed that only noise is present due to the very small dynamic range of the signal ($\Rightarrow$ Dyn Speech Pause)

   (b) If a) is not true, it is checked whether Δ_{LP} is bigger than η (otherwise the dynamic range in the low-pass band is very small and it should not receive too much attention $\Rightarrow$ no LP Speech pause). Now, if the difference between the current $E_{LP}(p)$ and $E_{LP,min}(p)$ of the low-pass band envelope is smaller than some fraction ν of Δ_{LP} (which means that the actual envelope is near its minimum), a closer look at the high-pass band is necessary to support a speech pause detection.

      i. Δ_{HP} of the high-pass band is smaller than threshold η.

         - In this case no additional information can be obtained from the high-pass band because of its small dynamic range. Now, if at least $E(p)$ (the signal's envelope) lies in the lower half of its dynamic range.[i.e. lower half between $E_{min}(p)$ and $E_{max}(p)$] the current frame can be assumed to be a speech pause because of the closeness of the low-pass band energy to its minimum value($\Rightarrow$ LP Speech Pause) otherwise, however, there is not enough support for a speech pause decision($\Rightarrow$ no LP Speech Pause).

      ii. Δ_{HP} is bigger than two times the threshold η.

- In this case, there is enough dynamic range to pay attention to the high-pass band. Thus, it is demanded that the difference between the current $E_{HP}(p)$ and $E_{HP,min}(p)$ of the high-pass envelope is smaller than two times the fraction ν of Δ_{HP} to support the small envelope value in the low-pass band. Then a noise-only frame is assume($\Rightarrow$ LP Speech Pause). This demand is not as strict as that for the low-pass band, to account for the case that the disturbing noise has a rather high-frequency characteristic. But if this condition is not fulfilled, speech may be presenting the actual frame ($\Rightarrow$ no LP Speech Pause).

iii. Δ_{HP} is smaller than two times the threshold η, but bigger than η.

- In this case, which is not as clear as previous case, it is only demanded that $E_{HP}(p)$ (the high-pass band envelope) lies in the lower half of its dynamic range to support the small envelope value in the low-pass band. Then it is assumed that target speech is absent($\Rightarrow$ Speech Pause). However, if this condition is not fulfilled, speech may be present in the actual frame ($\Rightarrow$ no LP speech pause).

(c) Condition b) accounts for the case that the disturbing noise has a rather high-frequency characteristic, hence the speech pause decision should mainly be made upon the information in the low-pass band.To account also for the case that it has a rather low-frequency characteristic, the same condition as under condition b) have to be checked but now with reverse roles of the low-pass and the high-pass bands to determine whether target speech is absent(HP speech pause).

Working

The input signal is digitized with a sampling frequency of 22050 Hz and partitioned in Hann-windowed segments of length 8 ms with 4 ms overlap [3]. These segments were padded with zeroes and a 256-point was FFT performed. The cut-off frequency between low-pass and high-passband was set to 2 kHz, by considering the fact that excluding speech frequencies above 1.9 kHz has roughly similar effect on speech intelligibility as including those below 1.9 kHz value. The time constant τ_E for the envelope smoothing was set to 32 ms. The time constants τ_{decay} and τ_{raise} were both set to 3 s. These constants are determined by experimentally.

4 Fusion method based Voice Activity Detectors

Fusion methods in speech detection is used to improve the accuracy of detector. So,better speech quality of reconstructed speech can be achieved.This methods applies more stringent decision rule which is depends on two or more than two parameters of speech signal. Here, following method is discussed.

1. Comprehensive VAD
2. Geometrically Adaptive Energy Threshold Method
3. Fusion method based on periodicity measure and geometrically adaptive energy threshold of signal

4.1 Comprehensive VAD

This algorithm use spectral energy and the variance of speech spectrum in frequency domain, as well use zero crossing rate of signal in time domain for detection of speech. So,this algorithm is the fusion of time domain and frequency domain method. The decision rule of the algorithm is based on priority of the parameter.

As shown in figure 5, at first stage multi-band energy comparison is done, as described in Linear Sub-Band Energy Detector(LSED). If frame is active than no further verification required. but, in case of low energy speech signal, if frame is detected is inactive, zero-crossing rate of the signal is measured by method described for Weak Fricative Detector(WFD). If frame detected as active frame, than in third stage variance of frame spectrum is compared as method described in Spectral flatness detector. if at last stage frame is detected as

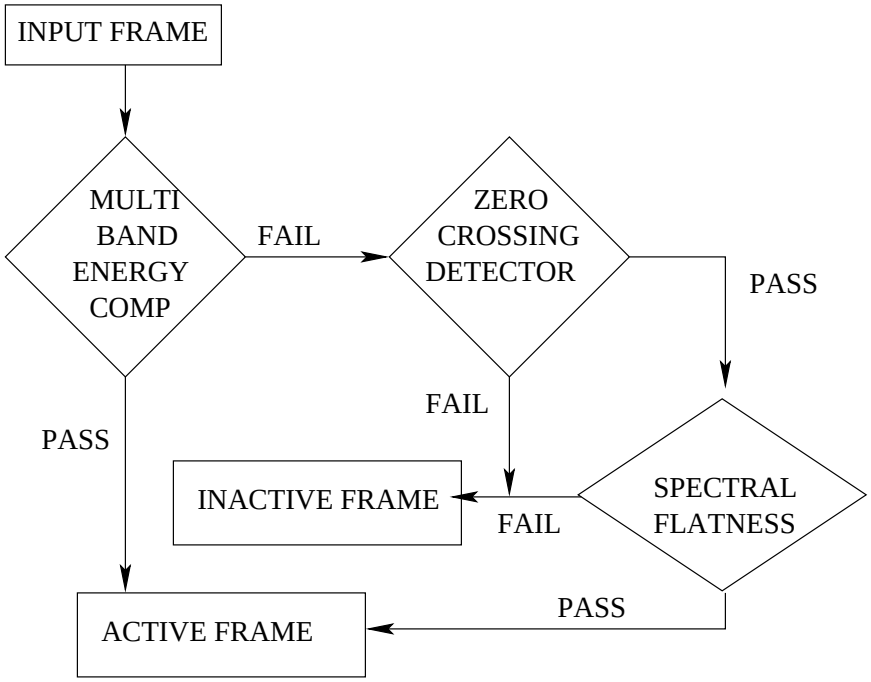

Figure 5: Flowchart for CVAD [1]

active, then decision is active frame. This algorithm can detect low energy speech signal, While it's energy is comparable with background noise. Although CVAD gives better speech detection compared to previous method, it's computation complexity is increased compared to previous method. Which may cause more delay in speech processing application.

4.2 Geometrically Adaptive Energy Threshold Method(GAET)

This method is presented by Ozer and Tanyer for speech detection in non-stationary background noise [4]. In the classical energy threshold method, the threshold value is re-calculated at each voice-inactive segment. When the background noise is non-stationary, the algorithm often can not track the threshold value accurately, especially when speech signal is mostly voice-active and noise level changes considerably before the next noise level re-calibration instant. The geometrically adaptive energy threshold(GAET) method set the threshold level adaptively without the need of voice-inactive segments by using the amplitude probability distributions of the speech signal [4].

4.2.1 Amplitude Probability Distributions(APD)

The Amplitude probability distribution(APD) is useful tool for statistical analysis of noise. For graphical analysis, assume that the input signal to VAD can be written by,

$$s(t) = c(t) + n(t) \tag{31}$$

where, $c(t)$ is the clear speech signal
$n(t)$ is noise signal.

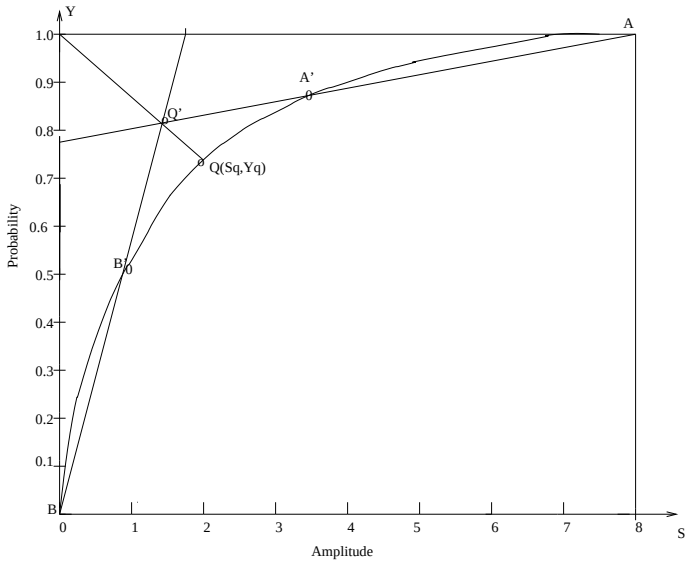

Figure 6: Amplitude probability distribution graph and the geometrical technique to calculate the noise level [4]

The amplitude probability distribution (APD) function $F_s\,(m)$,and the amplitude probability density (apd) function $f_s[m]$ of the discrete time random variable $s[k]$ of speech signal can be defined as

$$F_s[m] = \sum_{i=0}^{m} f_s[i] \tag{32}$$

where $f_s[i]$ is the number of samples of $s[k]$ satisfying

$$i\Delta s \leq |s[k]| \leq (i+1)\,\Delta s \tag{33}$$

is normalized by total number of samples N. If the APD of the signal and noise, $F_s\,(s)$ and $F_n\,(n)$, are different then, $F_s[m]$ and $F_n[m]$ are expected to be different. So, for a noise corrupted signal noise will partially occupy different regions on the APD.

4.2.2Modified Amplitude Probability Distribution (MAPD) Function

Modified Amplitude Probability Distribution (MAPD) function is define by $R_s[m]$.$R_s[m]$ is implicitly obtained by the setting the x and y-axis by $y = \frac{k}{N}$ and x = sort(s[k]) respectively while sorting is done in ascending order. The MAPD is corrupted by Gaussian noise. $R_s[m]$ is similar to $F_s[m]$. it is used to describe MAPD in place of $F_s[m]$

4.2.3 Geometrical Technique to calculate the Noise Level

It is observed on $R_s[m]$ plot that the samples s[k] and n[k] are partially separated, and the amplitude of the zero mean Gaussian noise samples n[k] locate closer to the origin whereas, the clear signal c[k] dominate the higher values. It can be notice from the figure 6 that noise level approximately corresponds to bending point, and bending point shifts higher amplitude level as the noise increases.

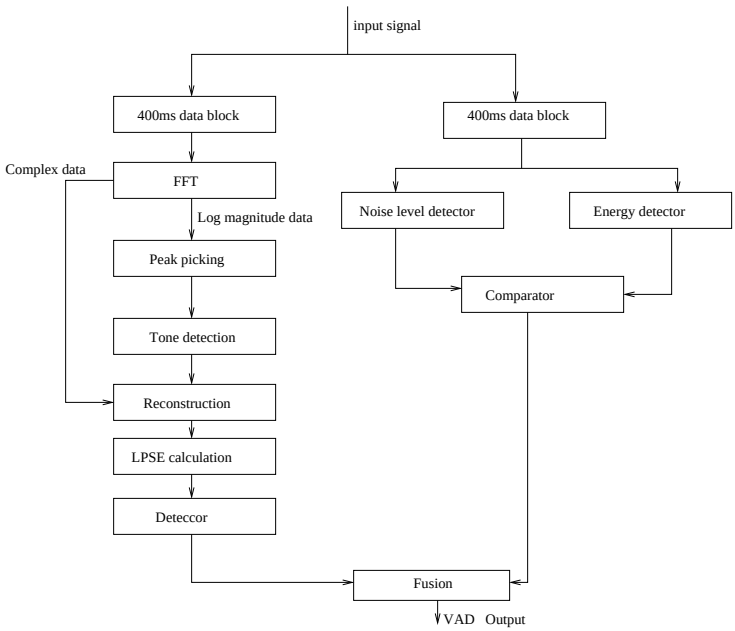

Figure 7: Voice activity detection using the fusion of the LPSE(left side) and the geometrically adaptive energy threshold(right side) method [5].

A Geometrical technique can heuristically be used to find the bending point on the MAPD graph which represents the noise level. As shown in figure, the point Q' can be found by intersecting the two tangent lines passing through the points A-A' and B-B' respectively. Then third line passing through the top left corner and point Q' crosses the MAPD graph at the "optimum point", $S(s_q, y_q)$. The noise level further can be multiplied by a safety coefficient α $(0.8 < \alpha < 1.2)$ which is constant throughout the detection process. The SNR estimation error is calculated for Gaussian, water, public restaurant and the traffic noise, and is observed to be less than 3dB for $-15dB < SNR < 20dB$.

4.3 Fusion method based on Periodicity measure of signal and Geometrically adaptive energy threshold of signal

The Least Square Periodicity Estimation(LSPE) method is accurate down to 0 dB SNR, but it has a problem of false triggering by periodic interference [5]. The Geometrically Adaptive Energy Threshold (GAET) method is robust to non-stationary noise but false triggering occurs when noise has short bursts. The fusion of LSPE and GAET method is observed to yield more accuracy and reliability [5]. The GAET method keeps track of the non-stationary back-ground noise while the LSPE analyzes the periodic content of the incoming signal.

In the GAET branch of the fusion method, the speech signal is processed in 1200 ms analysis blocks, and the noise level is recalculated for each block. Each block is divided into 32 ms long 75 frames, overlapping 16 ms. Voice-active decision for a frame is made if speech signal is above the threshold more than 50 % of the time, and the branch returns 1;otherwise, it returns 0.

In the LSPE branch, the size of the analysis block and frame lengths are the same as in GAET branch. The algorithm returns 1 for a frame if peaks are present at least 20 % of its FFT frame, otherwise it returns 0. Hanning window is used for both baranches.

In the fusion algorithm, the digital output for both branches are monitored. The weighted sum of the two outputs using the weight α and $\beta = (1 - \alpha)$,is compared to 0.5 for the overall voice-active decision. One algorithm can be dominated by the other with proper adjustment of weights.

5.Evaluation of performance

Performance evaluation of different VAD methods is done on the basis of following quantitative measures.

1. Speech detection rate or accuracy
2. Computational complexity
3. Sensitivity

For some VAD techniques may have better accuracy while others have less computation complexity and low processing delay. There is a trade-off between accuracy and computational complexity. So, by evaluating performance for above quantitative measure, one could able to make choice of VAD application suitable to application requirement. For instance, fusion of geometrically adaptive energy threshold and periodicity measure gives highest accuracy for low SNR value among the VAD method discussed here, but it's computational complexity is also higher than other simple method, which occurs into more signal processing delay. Although it's accuracy is higher, it's computational complexity may deter it's use in real time application where processing delay is crucial factor.

5.1 Measurement of speech detection rate

Speech detection rate or mis-detection rate is quantitative measure to check accuracy of VAD algorithm. Mis-detection refers to the total of number of signal frames which have the speech content, but were classified as silence(Inactive) frame, and number of frames without speech content is classified as voiced(Active) frame [1]. Mis-detection rate refers to ratio of number of mis-detection to the total number of frames in sample signal. Lower the mis-detection rate, better is the accuracy for given algorithm. VAD accuracy is measured by either subjective assessment or objective assessment technique.

5.2 Objective assessment method

In this method clean speech signal, which should include possible variation in speech characteristics is selected as sample speech signal. This sample speech signal pause and speech detection frames calculated manually. After that, sample speech signal is corrupted by background noise from noise database at particular SNR. This background noise can be car noise, jet engine noise, factory noise, babble noise or any other type of environment noise. This noise may be stationary or nonstationary. If sample speech signal is corrupted by non- stationary noise it's SNR may vary in some range with time. On corrupted speech signal VAD technique is applied. After that, number of pause and speech detection segments is compared with manually calculated pause and speech detection segments of clean speech. So, mis-detection rate can be found.

5.3 Subjective assessment method

Speech detection rate can be measured by subjective assessment. It involves qualitative assessment of speech segments by listener after applying VAD algorithm for detection of speech segments from a noisy speech signal. Because, if false detection rate is higher than, it degrades speech quality and intelligibility. Quality degradation can be felt by listener. This experimentation is repeated for number of persons and their subjective opinion is taken. On the basis of their opinion VAD algorithm accuracy is decided. As this procedure is highly subjective,opinion may very person to person.

5.4 Performance of different VAD algorithm

We have broadly classified the VAD techniques as time domain techniques, frequency domain techniques and fusion of both these technique. For performance evaluation, we are restricting ourselves to the technique discussed in earlier section in this report.

Time domain VAD algorithms : Performance evaluation of energy based VAD algorithm revealed that, this algorithm accuracy is higher, only when SNR is above 10 dB. Further, it performance degrades while background noise is non stationary and changing very fast. It can not detect low energy non-plosive phonemes, because of their energy is comparable with background noise. Weak fricative detector based on ZCR performs better than energy base detectors. It can detect most of non-plosive phonemes which are not detected by energy base detector, but it performance degrade when noise has same number of zero crossings as speech frames. Periodicity measure base algorithm is more robust than energy base, and zero crossing base algorithm. It can operates satisfactory up to 0 dB SNR, but it's performance degrades when noise is periodic, which increase it's mis-detection rate.

Frequency domain VAD algorithms : In frequency domain, spectral energy base algorithm can not detect low energy phoneme, and it performance degrades at low SNR. It's performance is good down to 10 db SNR. Spectral flatness detector performs good at low SNR when background noise is white, than spectral energy base algorithm, but when it's noise variance is same range speech frame variance performance degrades. VAD algorithm based on power envelope dynamics of signal performs best among the methods discussed previously. It's performance is good for all type of background noise like babble, drilling machine noise, car engine noise down to -5 dB SNR. but, it's computation complexity is higher than the energy base, and spectral flatness detectors.

Fusion methods based VAD algorithms : Fusion of spectral energy base algorithm and zero crossing based algorithm is CVAD performs good speech detection than time domain and frequency domain algorithm. But, it's performance degrades at low SNR speech with variable background noise. Fusion of geometrically adaptive energy threshold algorithm, and periodicity based algorithm performs accurate down to -5 dB. SNR in non-stationary background noise condition. It is most robust algorithm among the time domain algorithm discussed here. Fusion VAD algorithms computation complexity is higher than time domain and frequency domain VAD algorithms.

6.Conclusions

Following conclusions are drawn after studying the different voice activity detection methods.

1. The algorithms based on energy of speech signal fail to deliver accurate speech detection in non stationary background noise and at low SNR, also non-plosive phonemes misses from the speech signal. It's accuracy is acceptable when SNR is above 10 dB.
2. Zero crossing method based VAD provides some improvement over the energy base VAD, with same computational complexity.
3. The geometrically adaptive energy threshold method is robust to non stationary noise, but false detection occurs when background noise has short bursts. It operates satisfactory down to -10 dB SNR.
4. The method based on periodicity measure of speech signal is accurate down to 0 dB SNR, but mis-detection occurs when background noise is periodic.
5. The fusion of periodicity measure based method and geometrically adaptive energy threshold method gives more accurate and reliable results for non stationary fast varying noise. But, it's computational complexity is higher among the all other VAD algorithms.

6. Power envelope dynamics based methods gives accurate, and consistent output for various types of background noise down to -5 dB SNR.

Although the various robust techniques are devised for voice activity detection, there is still space to have consistent and robust method for voice activity detection to increase the accuracy. The performance of any VAD detector totally depends on the characteristics of background noise, variation in background noise with time, intensity of background noise. VAD technique works well for one type of background noise, may not be able to perform well for another type of background noise. So, it is difficult task to get consistent performance for all the type of background noise by using single parameter based VAD technique.

Acknowledgment

I would like to express my sincere gratitude to Prof. Preeti Rao for her valuable guidance and support given throughout the course of my seminar.

References

[1] R.Venkatesha Prasad, Abhijeet Sangwan, H.S.Jamadagm, M.C.Chiranth and Rahul Sah, "Comparison of voice activity detection algorithms for VOIP, "IEEE Symposium on Computers and Communications, July.2002.

[2] R.Tucker, "Voice activity detection using Periodicity measure, " Proc.Inst.Elec.Eng., vol.139, pp.377-380, Aug.1992.

[3] M.Marzinzik and B.Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics, " IEEE. Trans. Speech And Audio Processing, vol.10, no.2, pp.109-118, Feb.2002.

[4] H.Ozer and S.G.Tanyer, "Voice activity detection in non stationary noise, " IEEE Trans. Speech And Audio Processing, vol. 8, no.4, pp.478-482, July.2000.

[5] H.Ozer and S.G.Tanyer, "A geometric algorithm for voice activity detection in non stationary Gaussian noise, " in *Proc.EUSIPCO'98*, Rhodes, Greece, Sept.1998.

[6] S.V.Gervan and F.Xie, "A comparative study of speech detection methods, " European Conference on Speech, Communication and Technology'97, vol. 3, pp.1095-1098.