Survey of the methods for voice activity detector (VAD).

There are numerous VAD methods, but speech/non-speech detection is an unsolved problem in speech processing, becous most of the VAD algorithms fail when the level of background noise increases. During the last time, numerous researchers have developed different strategies for detecting speech on a noisy signal.

(Noise-Robust Voice Activity Detector Based on Hidden Semi-MarkovModels Xianglong Liu, Yuan Liang, Yihua Lou, He Li, Baosong Shan,2010;

Voice Activity Detection. Fundamentals and Speech Recognition System Robustness J. Ramírez, J. M. Górriz and J. C. Segura, 2007).

The main idea of VAD methods is rather simple.

1. The original analogue signal, which to be used by the system is converted from analogue to discrete – s(n), 1 <= n <= N.

When beginning the calculation and estimation of the signal it is useful to do some assumptions.

First we needed to divide the signal into blocks(frames). The length of each block is needed to be 20ms according to the stationary properties of the signal . When using the Fs at 16 kHz, it will give us a block length of 320 samples(L=320). Consider the first 10 blocks to be background noise, then mean and variance could be calculated and used as a reference to the rest of the blocks to detect where a threshold is reached.

Ordinarily, short-term energy and zero-crossing rate can be combined to form appropriate features for determining the onset and termination of speech boundaries. These temporal features can be extracted simply from the sample values of speech signal.

Short-term energy is the principal and most natural feature, that has been used. This is especially to distinguish between voiced sounds and unvoiced sounds or silence, compared the performances of the following three short-term energy measurements in endpoint detection. It is observed that short-term energy is the most effective energy parameter for this task. Equation represents the squared short-Term Energy:

$$E(m)=\sum_{n=(m-1)*L}^{m*L} s(n)^2, 1 \le m < N/L. \quad (1)$$

Next is also the short-term power esimate:

$$P(m)=1/L*E(m) \quad (2)$$

And the short-term zero-crossing rate:

$$Z(m)=\frac{1}{L}\sum_{n=m*L-1}^{m*l}\frac{|sng(s(n))-sng(s(n-1))|}{2} \quad (3)$$

.

Where , sng(x) = -1, if x<0, and , sng(x) = 1, if x>0 .

The short-term zero-crossing rate gives a measure of how many times the signal, s(n), changes sing.

The squared short-term energy is most suitable. For simplicity the frame (block of fixed number of sample) processing used in speech recognition. Instead of calculating this parameter with respect to sample, calculations are done on frame basis.The short-term energy estimate will increase when speech is present in signal.

The number of zero-crossings is also a useful temporal feature in speech analysis. It refers to the number of times speech samples change sign in a given frame. The rate at which zero-crossings occur is a simple measure of the frequency content of a narrowband signal. Zero-crossing rate is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero. It tends to be lager during unvoiced regions.

These measurements need some triggers for making decision about, where the utterance begins ,and, where to end .To create a trigger one need some information about the background noise. This is done by using following procedure ;

1)Assuming that first ten blocks are background noise. With this assumption the mean and variance for the measures are calculated. To make a more comfortable approach following function is used:

$$W(m)=P(m)*(1-Z(m))*S \quad .$$

Where S is scale factor for avoiding small values, in typical application it is 1000 or more.

*2)*The trigger can be described as:

$$t_W=\mu_W+\alpha*\delta_W$$

The μ_W is mean and the δ_W is variance for $W(m)$, calculated for first ten blocks. The term α is constant, that have to be fine tuned according to the characteristics of signal.

3) The voice activation detection function is found as :

$$VAD(m)=\begin{cases}1, W(m)\geq t_W\\ 0, W(m)\leq t_W\end{cases} .$$

VAD(n) is found as VAD(m) in the block of measure.

2. An Algorithm for Determining the Endpoints of Isolated Utterances

An important problem in speech processing is to detect the presence of speech in a background of noise. This problem is often referred to as the endpoint location problem. By accurately detecting the beginning and end of an utterance, the amount of processing of speech data can be kept to a minimum. The algorithm proposed for locating the endpoints of an utterance is based on two measures of the signal, zero crossing rate and energy.

The algorithm is inherently capable of performing correctly in any reasonable acoustic environment in which the signal-to-noise ratio is on the order of 30 dB or better. The algorithm has been tested over a variety of recording conditions and for a large number of speakers and has been found to perform well across all tested conditions.

The task of separating speech from background silence is not a trivial one except in the case of acoustic environments with extremely high signal-to-noise ratio, e.g., an anechoic chamber or a soundproof room in which high-quality recordings are made. For such high signal-to-noise ratio environments, the energy of the lowest-level speech sounds (e.g., weak fricatives, low-level voiced portions, etc.) exceeds the background noise energy and a simple energy measure suffices. However, such ideal recording conditions are not practical for realworld applications of speech-processing systems. Thus, simple energy measures are not sufficient for separating weak fricatives (such as the /f/ in "four") from background silence. In this paper, we propose a fairly simple algorithm for locating the beginning and end of an utterance, which can be used in almost any background environment with a signal-to-noise ratio of at least 30 dB. The algorithm is based on two measures of speech: short-time energy and the zero crossing rate. The algorithm possesses the feature that is somewhat self-adapting to the background acoustic environment in that it obtains all the relevant thresholds on its decision criteria from measurements made directly on the recorded interval.

To eliminate 60-Hz hum, as well as any dc level in the speech, it is assumed that the speech is high-pass filtered above 100 Hz ; similarly, to keep the processing simple, the speech is low-pass filtered at 4 kHz, thereby allowing a 10-kHz sampling frequency.

The problem of locating the endpoints of an utterance in these backgrounds of silence essentially is one of pattern recognition. The way one would attack the problem by eye would be to acclimate the eye (and brain) to the "typical" silence waveform and then try to spot some radical change in the pattern. In many cases this is easy to do. For example, a waveform of the word "eight", in which the silence pattern is easily distinguished from the speech. What one is observing in this case is a radical change in the waveform energy between the silence and the beginning of the speech. What one is observing in this case is a radical change in the waveform energy between the silence and the beginning of the speech. An another example a waveform of the word "six", in which there is a very small difference between the background noise short-time energy and short-time energy of the beginning of the speech. But, In this case, the frequency content of the speech with the background noise is radically different from the frequency content of the noise, as manifested by the sharp increase in the zero crossing (or level crossing) rate of the waveform. For this example, the speech energy at the beginning of the utterance is not radically higher than the silence energy; however, other characteristics of the waveform signal the beginning of the speech.

As a final example, is the waveform for the end of the word "nine." It is quite difficult to say where the final nasal ends and where the silence begins.

Rather than give several more examples of situations in which it is difficult to locate either the beginning or the end of an utterance, lets list below the broad categories of problems encountered. These include :

(i) Weak fricatives (/f, th, h/) at the beginning or end of an utterance.
(ii) Weak plosive bursts (/p, t, k/).
(iii) Final nasals.
(iv) Voiced fricatives at the ends of words which become devoiced.
(v) Trailing off of certain voiced sounds — e.g., the final /i/ becomes unvoiced sometimes in the words "three" (/th-r-i/) or "binary" (/b-ai-n-e-r-i/).

The approach is taken to solve these problems in an automatic endpoint-location algorithm is a pragmatic one. The goal is to isolate enough of the word (utterance) so that a reasonable acoustic analysis of what is obtained is sufficient for accurate recognition of the word. Thus, it is not necessary to locate exactly the point where the word begins or ends, but instead it is important to include all significant acoustic events within the utterance. For a word like "binary," it is of little consequence if the trailing off unvoiced energy is omitted (in fact, it is probably quite helpful for a "phonetic" word-recognition strategy) ; however, for a word like "four" it is important to be able to reliably locate and include the initial weak fricative /f/. For this last example, the word "four," it is not necessary to include the entire initial unvoiced interval; in fact, experience has shown that 30 to 50 ms of unvoiced energy is sufficient for most word-recognition purposes.

With the above considerations in mind, the endpoint location algorithm is based on two simple measurements, energy and zero crossing rate, and uses simple logic in the final decision. Both energy and

zero crossing rate are simple and fast to compute, and, can give fairly accurate indications as to the presence or absence of speech.

Before proceeding to a description of the algorithm, let's firstly define, how the energy and zero crossing rate are measured. The speech "energy," E_mang(m), is defined as the sum of the magnitudes of 10 ms of speech, centered on the measurement interval, i.e.,

With the above considerations in mind, the endpoint location algorithm is based on two simple measurements,
energy and zero crossing rate, and uses simple logic in the final decision

Both energy and zero crossing rate are simple and fast to compute, and, can give fairly accurate indications as to the presence or absence of speech.

Before proceeding to a description of the algorithm, let's firstly define, how the energy and zero crossing rate are measured. The speech "energy," E_magn(m), is defined as the sum of the magnitudes of 10 ms of speech centered on the measurement interval, i.e., (see (1))

$$E_{magn}(m) = \sum_{n=(m-1)*100}^{m*100} |s(n)|,\ 1 \le m < N/100 \quad . \tag{4}$$

Where s(n) is the speech samples and it is assumed that the sampling frequency is 10 kHz. The choice of a 10-ms window for computing the energy and the use of a magnitude function rather than a squared-magnitude function were dictated by the desire to perform the computations in integer arithmetic and, thus, to increase speed of computation. Further, the use of a magnitude de-emphasizes large-amplitude speech variations and produces a smoother energy function. The energy function is computed once every 10 ms, or 100 times per second.

The zero (level) crossing rate of the speech, Z(m),(see(3)), is defined as the number of zero (level) crossings per 10-ms interval. Although the zero crossing rate is highly susceptible to 60-Hz hum, dc offset, etc., in most cases it is a reasonably good measure of the presence or absence of unvoiced speech.

For the endpoint-location algorithm, the speech waveform is filtered prior to sampling at 10 kHz by a bandpass filter with a 100-Hz low-frequency cutoff and a 4000-Hz high-frequency cutoff and having 48 dB per octave skirts. It is assumed that during the first 100 ms of the recording interval there is no speech present. Thus, during this interval, the statistics of the background silence are measured. These measurements include the average and standard deviation of the zero crossing rate and the average energy. If any of these measurements are excessive, the algorithm halts and warns the user. Otherwise, a zero crossing threshold, IZCT, for unvoiced speech is chosen as the minimum of a fixed threshold, IF (25 crossings per 10 ms), and the sum of the mean zero crossing rate during silence, mean_IZC, plus twice the standard deviation of the zero crossing rate during silence, i.e.,

$$IZCT = Min(IF, mean_{IZC} + 2\sigma_{IZC}) \tag{5}$$

The energy function for the entire interval, E_magn(m), is then computed. The peak energy, IMX, and the silence energy, IMN, are used to set two thresholds, ITL and ITU, according to the rule Equation(6) shows I1 to be a level which is 3%(percent)of the peak energy (adjusted for the silence energy), whereas (7) shows I2 to be a level set at four times the silence energy. The lower threshold, ITL, is the minimum of these two conservative energy thresholds, and the upper threshold, ITU, is five times the lower threshold.

$$I1 = 0.03 * (IMX - IMN) + IMN \tag{6}$$

$$I2 = 4 * IMN \tag{7}$$

$$ITL = min(I1, I2) \tag{8}$$

$$ITU = 5 * ITL \tag{9}$$

The algorithm for a first guess at the endpoint locations is following. The algorithm begins by searching from the beginning of the interval until the lower threshold is exceeded. This point is preliminarily labeled the beginning of the utterance unless the energy falls below ITL before it rises above ITU. Should this occur, a new beginning point is obtained by finding the first point at which the energy exceeds ITL, and then exceeds ITU before falling below ITL; eventually such a beginning point must exist. A similar algorithm is used to define a preliminary estimate of the endpoint of the utterance. We call these beginning and ending points N1 and N2, respectively.

Until now, we have only used energy measurements to find the endpoint locations ; and these endpoint locations are conservative in that fairly tight thresholds are used to obtain these estimates. Thus, at this point, it is fairly safe to assume that, although part of the utterance may be outside the (N 1,N 2 interval, the actual endpoints are not within this interval. In relation to this, the algorithm proceeds to examine the interval from N1 to N1- 25, i.e., a 250-ms interval preceding the initial beginning point, and counts the number of intervals where the zero crossing rate exceeds the threshold IZCT. If the number of times the threshold was exceeded was three or more, the starting point is set back to the first point (in time) at which the threshold was exceeded. Otherwise, the beginning point is kept at N1. The rationale behind this strategy is that for all cases of interest, exceeding a tight threshold on zero crossing rate is a strong reliable indication of un-voiced energy. Of course, it is still possible that a weak fricative will not pass this test, and will be missed. However, in these cases there is no simple, reliable method of distinguishing such a weak fricative from background silence.
A similar search procedure is used on the endpoint of the utterance to determine if there is unvoiced energy in the interval from N2 to N 2 + 25. The endpoint is readjusted based on the zero crossing test results in this interval.

3. Voice Activity Detection In Noisy Environments.
   Statistical algorithms.

An important problem in many areas of speech processing is the determination of presence of speech periods in a given signal. This task can be identified as a statistical hypothesis problem and its purpose is the determination to which category or class a given signal belongs. The decision is made based on an observation vector, frequently called feature vector, which serves as the input to a decision rule that assigns a sample vector to one of the given classes. The classification task is often not as trivial as it appears since the increasing level of background noise degrades the classifier effectiveness, thus leading to numerous detection errors. The selection of an adequate feature vector for signal detection and a robust decision rule is a challenging problem that affects the performance of VADs working under noise conditions. Most algorithms are effective in numerous applications but often cause detection errors mainly due to the loss of discriminating power of the decision rule at low SNR (signal-to-noise ratio) levels . For example, a simple energy level detector can work satisfactorily in high signal-to-noise ratio (SNR) conditions, but would fail significantly when the SNR drops. VAD results more critical in non-

stationary noise environments since it is needed to update the constantly varying noise statistics affecting a misclassification error strongly to the system performance.

Description of the problem

The VAD problem considers detecting the presence of speech in a noisy signal. The VAD decision is normally based on a feature vector x. Assuming that the speech signals and the noise are additive, the VAD module has to decide in favour of the two hypotheses:

$$\begin{aligned} H_0 &: X = N \\ H_1 &: X = N + S \end{aligned} \tag{10}$$

Formulation of the decision rule based on a statistical likelihood ratio test (LRT) involving a single observation vector. The method considered a two-hypothesis test where the optimal decision rule that minimizes the error probability is the Bayes classifier. Given an observation vector to be classified, the problem is reduced to selecting the class (H0 or H1) with the largest posterior probability P(Hi|x):

$$\begin{aligned} p(H_1/X) > p(H_0/X) &\quad \rightarrow H_1 \\ p(H_1/X) < p(H_0/X) &\quad \rightarrow H_0 \end{aligned} \tag{11}$$

Using the Bayes rule:

$$\begin{aligned} p(H_1/X) &= \frac{p(X/H_1) * p(H_1)}{p(X)} \\ p(H_0/X) &= \frac{p(X/H_0) * p(H_0)}{p(X)} \end{aligned} \tag{12}$$

From (11) we get:

$$\frac{p(X/H_1) * p(H_1)}{p(X)} > \frac{p(X/H_0) * p(H_0)}{p(X)} \quad \rightarrow \quad H_1 \tag{13}$$

$$p(X/H_1) * p(H_1) > p(X/H_0) * p(H_0) \quad \rightarrow \quad H_1 \tag{14}$$

$$\frac{p(X/H_1)}{p(X/H_0)} > \frac{p(H_0)}{p(H_1)} \quad \rightarrow \quad H_1 \tag{15}$$

Analogy:

$$\frac{p(X/H_1)}{p(X/H_0)} < \frac{p(H_0)}{p(H_1)} \quad \rightarrow \quad H_0 \tag{16}$$

It (see (15),(16))is statistical likelihood ratio(SLR) test, which is used in statistical VAD algorithms.

In order to evaluate this test, the discrete Fourier transform (DFT) coefficients of the clean speech (Sj) and the noise (Nj) are assumed to be asymptotically independent Gaussian random variables. Then the probability density functions conditioned on H0 and H1 are given by:

$$p(X/H_0)=\Pi_{j=0}^{L-1}\frac{1}{\pi\lambda_N(j)}\exp\left[-\frac{|X_j|^2}{\lambda_N(j)}\right]$$
$$p(X/H_1)=\Pi_{j=0}^{L-1}\frac{1}{\pi[\lambda_N(j)+\lambda_S(j)]}\exp\left[-\frac{|X_j|^2}{[\lambda_N(j)+\lambda_S(j)]}\right] \quad (17)$$

Value of L is the number of the coefficients discrete Fourier transform, which are considerable. S, N, and X are L-dimensional discrete Fourier transform (DFT) coefficient vectors of speech, noise, and noisy speech with their j-th elements Sj, Nj, and Xj, respectively. We adopt the Gaussian statistical model that the DFT coefficients of each process are asymptotically independent Gaussian random variables. Where $\lambda_N(j)$ and $\lambda_S(j)$ denote the variances of Nj and Sj, respectively. The likelihood ratio for the k-th frequency band is

$$\Lambda_k=\frac{p(X_k/H_1)}{p(X_k/H_0)}=\frac{1}{1+\xi_k}\exp\left[\frac{\gamma_k\xi_k}{1+\xi_k}\right] \quad (18)$$

where $\xi_k=\lambda_S(k)/\lambda_N(k)$ and $\gamma_k=|X_k|^2/\lambda_N(k)$, and they are called the *a priori* and *a posteriori* signal-to-noise ratios (SNR's), respectively . The decision rule is established from the geometric mean of the likelihood ratios for the individual frequency bands, which is given by

$$\log\Lambda=\frac{1}{L}\sum_{k=0}^{L-1}\log\Lambda_k \quad > \quad \eta \quad \rightarrow \quad H1$$
$$\log\Lambda=\frac{1}{L}\sum_{k=0}^{L-1}\log\Lambda_k \quad < \quad \eta \quad \rightarrow \quad H0 \quad (19)$$

We assume that $\lambda_N(k)$'s are already known through the noise statistic estimation procedure, and have to estimate the unknown parameters, ξ_k 's.
The ML estimator for ξ_k can easily be derived as follows:

$$\xi_k^{(ML)}=\gamma_k-1 \quad (20)$$