Voice Activity Detection Algorithm for Speech Recognition Applications

Nitin N Lokhande
Pravara Rural Engineering College, Loni
Ahmednagar-Maharashtra

Navnath S Nehe
JSPM Technical Campus
Narhe, Pune
Pune--Maharashtra

Pratap S Vikhe
Pravara Rural Engineering College, Loni
Ahmednagar-Maharashtra

ABSTRACT
Determining the beginning and the termination of speech in the presence of background noise is a complicated problem. This paper is concerned with labeling sections of speech samples based on whether they are silence, voiced or unvoiced speech. The labeling is done using calculations over the speech samples; zero crossing and short-term energy functions. The short-term energy and zero crossing rate of speech have been extensively used to detect the endpoints of an utterance.

General Terms
Speech Recognition, Voice, Unvoice.

Keywords
Short Term Energy (STE), Short Term Power (STP), Zero Crossing Rate (ZCR), Voice Activity Detection (VAD).

1. INTRODUCTION
For automatic speech recognition, endpoint detection is required to isolate the speech of interest so as to be able to create a speech pattern or template. The process of separating the speech segments of an utterance from the background, i.e., the non speech segments obtained during the recording process, is called endpoint detection. In isolated word recognition systems, accurate detection of the endpoints of a spoken word is important for two reasons, namely: Reliable word recognition is critically dependent on accurate endpoint detection and the computation for processing the speech is less, when the endpoints are accurately located [2].

Classification of speech into voiced or unvoiced sounds provides a useful basis for subsequent processing. A three-way classification into silence/unvoiced/voiced extends the possible range of further processing to tasks such as stop consonant identification and endpoint detection for isolated utterances. The three- state representation is one way to classify the event in speech. The events of interest for the three-state representation are: Silence, when no speech is produced. Unvoiced: Vocal cords are not vibrating, resulting in a periodic or aperiodic speech waveform. Voiced: Vocal cords are tensed and vibrating periodically, resulting in speech waveform that is quasi-periodic Quasi-periodic means that the speech waveform can be seen as periodic over a short-time period (5-100 ms) during which it is stationary[5] as shown in figure.1. It should be clear that the segmentation of waveform into well defined regions of silence, unvoiced, signals is not exact, it is often difficult to distinguish the weak, unvoiced sound (like /f/ or /th/)from silence or weak voiced sound (/v/ or /m/) from unvoiced sounds or even silence. Some of the principal causes of endpoint detection failures are weak fricatives (/f/, /T/, /h/) or voiced fricatives that become unvoiced at the end ("has"), weak plosives at either end (/p/, /t/, /k/), nasals at the end ("gone"), and trailing vowels at the end (“zoo”) [3].

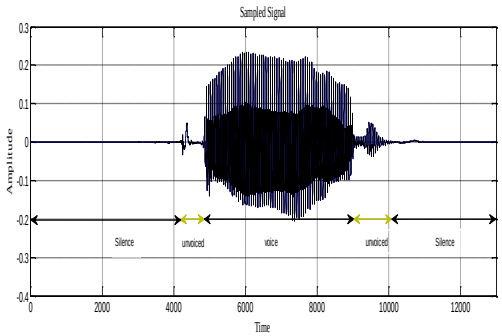

Fig 1: Three state representation of sampled signal.

An environmental condition in which speech is produced, is factor that making reliable speech endpoint detection difficult. The ideal environment for talking is quiet room with no acoustic noise or signal generators other than noise produced by speaker. Such ideal environment is not always practical hence one must consider speech produced in noisy environment [2].

2. VOICE DETECTION USING STE, STP & ZCR
Ordinarily, zero-crossing rate and short-term energy can be combined to form appropriate spatial or temporal features for determining the onset and termination of speech boundaries. These temporal features can be extracted simply from the sample values of speech signal without transforming the signal into the frequency domain [5], [6].

2.1 Short Term Energy
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-times energy measurements in endpoint detection. It is observed that short-term energy is the most effective energy parameter for this task. Voiced speech has most of its energy collected in the lower frequencies, whereas most energy of the unvoiced speech is found in the higher frequencies [8].Different energies used for signal analysis are as per equation 1, 2 and 3. Where, equation 1 represents Logarithmic Short-Term Energy, equation 2 represents the squared short-Term Energy and equation 3 represents Absolute Short-Term Energy.

$$E_{\log} = \sum_{n=1}^{N} \log[s(n)^2] \quad (1)$$

$$E_{sqr} = \sum_{n=1}^{N} [s(n)^2] \quad (2)$$

$$E_{abs} = \sum_{n=1}^{N} \left| s(n)^2 \right| \quad (3)$$

N Length of sampled signal.

The squared short term energy is most suitable, hence used for implementation. 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. According to this approach the [4];

$$E_{s1}(m) = \sum_{n=m-L+1}^{M} S(n)^2 \quad (4)$$

The short term energy estimate will increase when speech is present in signal $s(n)$ .This is the also case with short term power estimate only thing that separates by $\frac{1}{L}$.

$$P_{s1}(m) = \frac{1}{L} \sum_{n=m-L+1}^{M} S(n)^2 \quad (5)$$

2.2 Zero Crossing Rate

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. A zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a 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. The number of zero-crossings, which is also a useful temporal feature in speech analysis, refers to the number of times speech samples change sign in a given frame. This short term zero crossing rates tend to be larger during unvoiced regions. The multiplication by a factor of $\frac{1}{L}$ is intended to take the average value of the zero-crossing measure [6].

$$Z_{s1}(m) = \frac{1}{L} \sum_{n=m-L+1}^{M} \left| \frac{\text{sgn}(s(n)) - \text{sgn}(s(n-1))}{2} \right| \quad (6)$$

Where $\text{sgn}(s(n)) = {}^{+1, s(n) \geq 0}_{-1, s(n) < 0}$

3. METHODOLOGY

The algorithm is used as first step in speech recognition system. The database used for experimentation was ZERO to NINE digits in English language. The data was collected from 12 adult speakers out of which six speakers are male & six are female speakers. Each digit was uttered 10 times by each speaker and recorded with sampling frequency 16 kHz. Recording is done using condenser microphone in closed room & clean environment. Block (frame) length is used 320 which correspond to 20ms for 16 KHz.

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 [4];

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

$$W_{s1}(m) = P_{s1}(m).(1 - Z_{s1}(m)).S_c \quad (7)$$

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

2) The trigger can be described as,

$$t_w = \mu_w + \alpha \delta_w \quad (8)$$

The μ_w mean and δ_w is the variance for $W_{s1}(m)$ calculated for first ten blocks.

3) The α term is constant that have to be fine tuned according to the characteristics of signal. After some testing the following approximation of α give good voice activation detection according various background noises;

$$\alpha = 0.3 \cdot \delta_w^{-0.92} \quad (9)$$

4) The voice activation detection function is found as

$$VAD(m) = \begin{cases} 1, W_{s1} \geq t_w \\ 0, W_{s1} < t_w \end{cases} \quad (10)$$

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

4. EXPERIMENTAL RESULTS

This algorithm is used to find endpoint of different isolated word i.e. digit (zero to nine) in our experiment uttered by different speakers. It gives fine results over manual endpoint detection of speech signal. Figure.2.shows the sampled speech signal, its STE estimate, STP estimate and ZCR estimate with voice activity detection ,in which part labeled as '1'indicates voice region, while '0' indicates unvoiced and silence region. For programming we have used MATLAB Language.

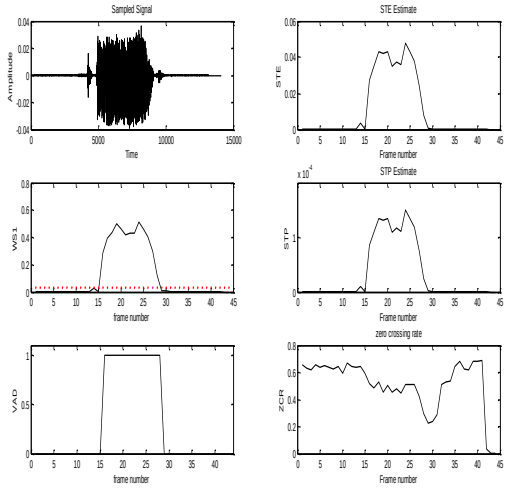

Fig 2: Different measures used to find speech content.

Figure.3 shows endpoint detected speech signal in which no. of samples required to represent the speech activity is subsequently reduced.

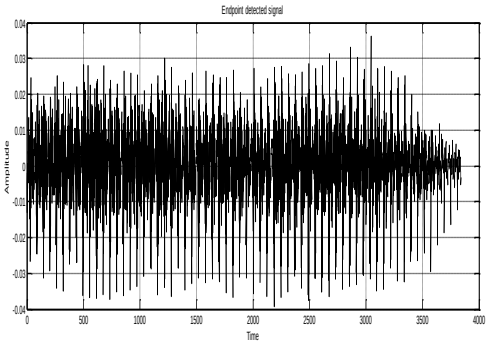

Fig 3: Endpoint detected signal.

Table.1 shows the percentage of sample reduction using manual segmentation & VAD algorithm for each digit. Figure 4 shows comparison between manual segmentation results & algorithm results in terms of percentage of sample reduction.

Table 1. Performance Index

Digit	Percentage of Sample Reduction using Manual Segmentation	Percentage of Sample Reduction using algorithm
Zero	54.75	64.07
One	56.87	70.00
Two	54.37	73.90
Three	63.12	71.87
Four	60.00	75.81
Five	54.68	79.93
Six	50.06	59.90
Seven	43.13	55.75
Eight	66.25	77.87
Nine	48.06	77.93

Total number of sample required to represent specific digit uttered by different speaker is vary according to spectral characteristic of speech.

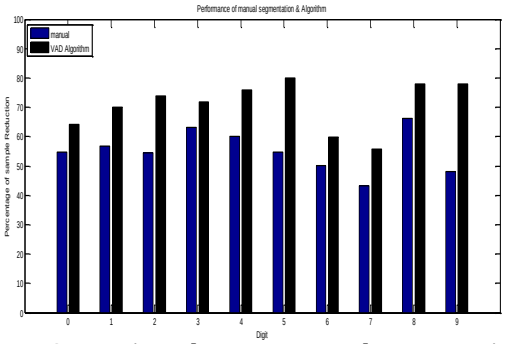

Fig 4: Comparison between manual segmentation & proposed method for one speaker.

5. CONCLUSION

We have presented an approach for separating the voiced & unvoiced part of speech in a simple and efficient way. The algorithm shows good results in classifying the speech as we segmented speech into many frames. It is effective for detection of endpoints of different languages digits & isolated words. This reduces the memory requirement & computational time. It is also observed that the performance of this algorithm is better than manual segmentation.

6. ACKNOWLEDGMENTS

We would like to thank Mikael Nilsson, Marcus Ejnarsson Department of Telecommunications and Speech Processing, Blekinge Institute of Technology for technical note on Speech recognition using HMM.

7. REFERENCES

[1] Lori F. Lamel, Lawrence R. Rabiner, Aaron E. Rosenberg, Jay G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition" IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. Assp-29, No. 4, August 1981.

[2] Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of speech Recognition", Prentice Hall, Englewood Cliffs, N.J., 1993.

[3] John R. Deller, Jr., John H. L. Hansen, John G. Proakis, "Discrete-Time Processing Of Speech Signals", John Wiley & Sons, inc., publication, IEEE Press.

[4] Mikael Nilsson, Marcus Ejnarsson. "Speech Recognition using Hidden Markov Model". Department of Telecommunications and Speech Processing, Blekinge Institute of Technology. 2002

[5] K.R. Aida–Zade, C. Ardil and S.S. Rustamov, Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems", World Academy of Science, Engineering and Technology 19 2006.

[6] L.R. Rabiner, M.R Sambur, "An Algorithm for determining the endpoints of Isolated Utterances", The Bell System Technical Journal, February 1975, pp 298-315.

[7] B. Atal, and L. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," IEEE Trans. On ASSP, vol. ASSP-24, pp. 201-212, 197.

[8] Rabiner, L. R., and Schafer, R. W., Digital Processing of Speech Signals, Englewood Cliffs, New Jersey, Prentice Hall, 512-ISBN-13:9780132136037, 1978.

[9] L. Siegel, "A Procedure for using Pattern Classification Techniques to obtain a Voiced/Unvoiced Classifier", IEEE Trans. on ASSP, vol. ASSP-27, pp. 83- 88, 1979.

[10] Y. Qi, and B.R. Hunt, "Voiced-Unvoiced-Silence Classifications of Speech using Hybrid Features and a Network Classifier," IEEE Trans. Speech Audio Processing, vol. 1 No. 2, pp. 250-255, 1993.