IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 11, Issue 3, Ver. I (May-Jun .2016), PP 07-12
www.iosrjournals.org

Performance Comparison of Various Voice Activity Detection Techniques

V.Adlin Vini [1],Mrs.D.Sharmila [2]

[1] Pg Student , Applied Electronics , C.S.I Institute Of Technology
[2]Assistant Professor, Department of Electronics and Communication,
C.S.I Institute Of Technology

Abstract: *A Signal Processing approach is proposed to highlight the features of speech in degraded signal. The approach is based on Single Frequency Filtering (SFF), where the amplitude envelope of the signal is obtained at each frequency with high temporal and spectral resolution. So this method exploit the fact that speech has high SNR region at different frequency and at different time .Here three methods are discussed and the spectral and temporal resolution is compared .The Mel frequency cepstral coefficients(MFCC's) is computed from Fourier Transform. MFCC features are formed by extracting filter bank energies. The second method is Discrete Wavelet Transform .This method captured the speech and noise ,if the samples are collected (>1 sec).Both the MFCC and Discrete Wavelet Transform are highly complex due to different process. The third method is Artificial Neural Networks (ANNs).It is based on machine learning technique that is it required training data. And the proposed method is SFF based Voice Activity Detection does not use training data to derive the characteristics of speech or noise.*

Keywords: *ANNs, MFCC, SFF, SVM, VAD.*

.

I. Introduction

The voice activity detection is a technique used in speech processing in which presence or absence of human speech is detected. This is also called speech detection. The main objective of voice activity detection is to determine region of speech in acoustic signal even when the signal is corrupted by degradation. The main uses of VAD are in speech coding and speech recognition. It is also used to deactivate some processes during non speech session of an audio session. It can avoid unnecessary coding and transmission of silent packet in voice over Internet Protocol application so that it saves network bandwidth.

VAD is used in speech based application. Therefore various VAD algorithms have been developed that provide varying features and compromises various latency, accuracy, sensitivity and computational cost. The performance of VAD is commonly evaluated on the basis of following parameters,

FEC(front end clipping):clipping introduced in passing from speech to noise activity.

- MSC (mid speech clipping): clipping due to speech is misclassified as noise.
- OVER: noise interpreted as speech due to the VAD flag remaining active in passing speech activity to noise.
- NDS (Noise Detected as Speech) noise interpreted as speech within a silence period.

Although the method described above provide useful objective information concerning the performance of VAD. For example the effect of clipping the speech signal is hidden by the presence of background noise. so clipping with the objective test is not audible. It is therefore important to carry out subjective test on VAD.

II. Voice Activity Detection Techniques

2.1 Mel-Frequency Cepstral Coefficient

MFCC is the representation of short term power spectrum of a sound. It is one of the effective choices of speech features derived from magnitude spectrum.

MFCCs are commonly derived by taking Fourier transform of a signal and map the powers of the spectrum obtained above onto the Mel-scale, using triangular overlapping window. The frequency bands are equally spaced in Mel-scale The logs of the powers is calculated at each Mel frequency and find the discrete cosine transform of log powers for the resulting spectrum.

2.1.1 Drawback Of MFCC Method

Since MFCC technique has different process, it is highly complex. These techniques include various features like voicing and spectral characteristics. But it takes more time to compute the spectral and temporal resolution

2.2 Discrete Wavelet Transform (DWT) And Teager Energy Operator (TEO)

In this technique the speech signal is decompose into four sub band using DWT. This subands helps to develop a robust features parameter called speech activity envelope and then TEO is applied to the DWT coefficient of sub band so that speech signal is extracted even in low signal to noise ratio.

DWT and TEO is mainly used for determining voice activity in presence of noise, especially varying background noise. And also it is based on frame by frame basis. In addition, Teager energy operator is used to eliminate noise component from each sub band. Teager is a powerful non linear operator and has been used in various speech processing application.

2.2.1 Drawback of DWT and TEO method

Discrete wavelet and Teager operator can dynamically work in varying background noise but it cannot reduce effect of noise. This method is higher than that of MFCC only for three types of noise that is car noise, factory noise and white noise. Due to its high temporal variance, most of the VAD algorithms detect machine gun chunks as speech .

2.3 Statistical Model-Based Voice Activity Using Support Vector Machine (SVM)

SVM based VAD is proposed effective feature vector to improve the performance of VAD employing a SVM, which is known to incorporate an optimized non linear decision.SVM is a discriminative classifier defined by separating hyper plane.SVM based VAD technique is used to divide the speech signal into frames. It has been observed that the SVM based solution is computationally efficient and provides around 90% accuracy for speech signal directly recorded using a microphone and an accuracy of over 85% for noisy speech. By using this technique it is easy to discriminate speech and non speech.

2.3.1 Drawback of SVM Based VAD

Even though it minimizes the decision error still there is a problem in detecting speech and non speech. This method is not able to work in low signal to noise ratio.

2.4 Single Pole Filtering Based Voice Activity Detection

A signal processing approach is proposed for both speech and non speech. This approach is based on Single Pole Filtering. In order to design the Single Pole Filter, the spectrum of the signal is calculated. Consider a audio signal as a input signal which is obtained by taking the difference of desire signal and previous signal.

$$x(n) = s(n) - s(n-1) \qquad (1)$$

The resulting operation in the time domain is given by,

$$x_k(n) = x(n)e^{j\omega_k n} \qquad (2)$$

Since x(n) is multiplied by $e^{j\omega_k n}$,the resulting spectrum of $x_k(n)$ is a shifted spectrum of x(n) that is,

$$X_k(\omega) = X(\omega - \omega_k) \qquad (3)$$

The signal $x_k(n)$ is passed through a single pole filter, whose transfer function is given by,

$$H(z) = \frac{1}{1+z^{-1}} \qquad (4)$$

The single pole filter has a pole on the real axis at a distance of r from the real axis. The location of the root is at z = -r in the z-plane. The output of the filter is given by,

$$y_k(n) = -ry_k(n-1) + x_k(n) \qquad (5)$$

Hence the envelope of the signal $y_k(n)$ is,

$$e_k(n) = \sqrt{y_{kr^2}(n) + y_{ki^2}}(n) \qquad (6)$$

Where $y_{kr^2}(n)$ and $y_{ki^2}(n)$ are the real and imaginary component of $y_k(n)$. The envelope and the corresponding weighted envelope of the signal is calculated at every 20 Hz. The weighted envelope is calculated to reduce the effect of noise. The normalized weight value at each frequency is given by,

$$\omega_k = \frac{1/\mu_k}{\sum_{l=1}^{N}\frac{1}{\mu_l}} \qquad (7)$$

At each time instant, mean $\mu(n)$ and variance $\sigma(n)$ is computed across frequency. The mean $\mu(n)$ is expected to be higher for speech than for noise. Generally $\sigma(n) + \mu(n)$ is higher in speech region and lower in non speech region. Since the spread of noise after compensation is expected to be lower. In order to highlight the contrast between speech and non speech region,

$$\delta(n) = \sqrt[M]{|(\sigma^2(n) - \mu^2(n))|} \qquad (8)$$

Hence multiplying $\sigma(n) + \mu(n)$ with $(\sigma(n) - \mu(n))$ gives $\sigma^2(n) - \mu^2(n)$,which highlight the contrast between speech and non speech where M is chosen as 64.The value of M is not critical. Any value of M in the range of 32 to 256 seems to provide good contrast. The $\delta(n)$ value with M=64 are used for further processing for decision making. The binary decision of speech and non speech at each time instant is denoted as 1 and 0 respectively, is further smoothed using an adaptive window to arrive at the final decision.

Block Diagram

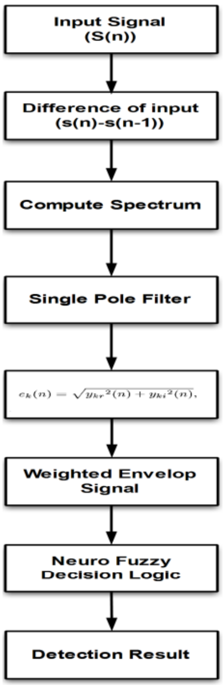
Input Signal (S(n))
Difference of input (s(n)-s(n-1))
Compute Spectrum
Single Pole Filter
$e_k(n) = \sqrt{y_{kr}{}^2(n) + y_{ki}{}^2(n)},$
Weighted Envelop Signal
Neuro Fuzzy Decision Logic
Detection Result

Figure 1: Single Pole Filtering based VAD

The steps involved are,
- Apply input signal.
- Find the difference of the input signal and previous signal.
- Sample the signal.
- Find the Fourier transform of sampled signal, and the output is spectral output.
- The spectral output is used to design the filter.
- The envelope of the filtered output and the corresponding weighted envelope is calculated.
- Determine the Neuro Fuzzy Decision logic.

The final step is Neuro Fuzzy decision logic which produces a clear decision output. There are three steps,
- Fuzzification
- Rule base engine
- Defuzzificaion

2.4.1 Fuzzification

Fuzzification is the process of changing a real scalar value into a fuzzy value. This is achieved with the different types of fuzzifiers. There are generally three types of fuzzifiers which are used for the fuzzification process are,
1. Singleton Fuzzifier,
2. Gaussian Fuzzifier, and
3. Trapezoidal or Triangular Fuzzifier.

Trapezoidal / Triangular Fuzzifiers

For the simplicity of discussion only the triangular and trapezoidal fuzzifiers are presented here. Fuzzification of a real-valued variable is done with intuition, experience and analysis of the set of rules and conditions associated with the input data variables. There is no fixed set of procedures for the fuzzification.

Fuzzy systems are built to replace the human expert with a machine using the logic a human would use to perform the tasks. Suppose we ask someone how hot it is today. He may tell us that it is hot, moderately hot or cold. He cannot tell us the exact temperature. Unlike classical logic which can only interpret the crisp set such as hot or cold, fuzzy logic has the capability to interpret the natural language. Thus, fuzzy logic can make human-like interpretations and is a very useful tool in artificial intelligence, machine learning and automation. Fuzzy logic operates on the basis of rules which are expressed in the form of If-Then constructs, also known as horn clauses.

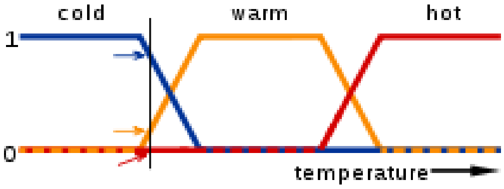

Figure 2: Fuzzy Logic Membership

The concept of linguistic variable was introduced to process the natural language. The linguistic variable is temperature. The linguistic variable can take the verbal values such as hot, moderately hot or cold. The terms temperature is hot and temperature is cold and temperature is moderate are known as fuzzy propositions

2.4.2 Rule Base Engine

Build a set of rules into the knowledge base in the form of IF-THEN-ELSE structures. Fuzzy logic consists of a mapping between an input space and an output space by means of a list of if-then statements called rules. These rules are useful because they refer to variables and the adjectives that describe these variables. The mapping is performed in the fuzzy inference stage, a method that interprets the values in the input vector and, based on some set of rules, assigns values to the output.

2.4.3 Defuzzification

Fuzzy logic is a rule-based system written in the form of horn clauses (i.e., if-then rules). These rules are stored in the knowledge base of the system. The input to the fuzzy system is a scalar value that is fuzzified. The set of rules is applied to the fuzzified input. The output of each rule is fuzzy. These fuzzy outputs need to be converted into a scalar output quantity so that the nature of the action to be performed can be determined by the system. The process of converting the fuzzy output is called defuzzification. Before an output is defuzzified all the fuzzy outputs of the system are aggregated with an union operator. The union is the max of the set of given membership functions and can be expressed as

$$\mu_A = \bigcup_i \left(\mu_i(x)\right) \qquad (9)$$

There are many defuzzification techniques but primarily only three of them are in common use.

Maximum Defuzzification Technique

This method gives the output with the highest membership function. This defuzzification technique is very fast but is only accurate for peaked output. This technique is given by algebraic expression as

$$\mu_{\mathcal{A}}(x^*) \geq \mu_{\mathcal{A}}(x) \text{ for all } \mathbf{x} \in \mathbf{X} \qquad (10)$$

where *x** is the defuzzified value.

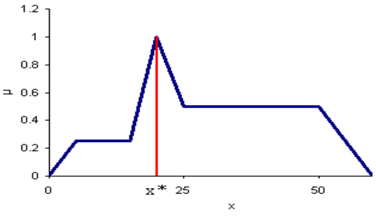

Figure 3 Max-membership Defuzzification Method

Centroid Defuzzification Technique

This method is also known as center of gravity or center of area defuzzification. This technique was developed by Sugeno in 1985. This is the most commonly used technique and is very accurate. The centroid defuzzification technique can be expressed as

$$x^{*} = \frac{\int \mu_{i}(x)\, x\, dx}{\int \mu_{i}(x)\, dx} \qquad (11)$$

where x^{*} is the defuzzified output, $\mu_i(x)$ is the aggregated membership function and x is the output variable. The only disadvantage of this method is that it is computationally difficult for complex membership functions.

Weighted Average Defuzzification Technique

In this method the output is obtained by the weighted average of the each output of the set of rules stored in the knowledge base of the system. The weighted average defuzzification technique can be expressed as

$$x^{*} = \frac{\sum_{i=1}^{n} m^{i} w_{i}}{\sum_{i=1}^{n} m^{i}} \qquad (12)$$

where x^{*} is the defuzzified output, m^{i} is the membership of the output of each rule, and w_i is the weight associated with each rule. This method is computationally faster and easier and gives fairly accurate result.

2.4.4 Advantage of Single Pole Filtering based VAD

The performance of the proposed method is higher than of all the above methods for all types of noises. The technique uses the Neuro Fuzzy decision hence it determines all the intermediate values of the speech and also the detection accuracy is high in this method. Since most of the VAD technique has different process hence it required more time for computation but the single pole filtering based VAD consumes less time.

III. Simulation Result

The following result is obtained by carrying out the simulation using MAT lab.

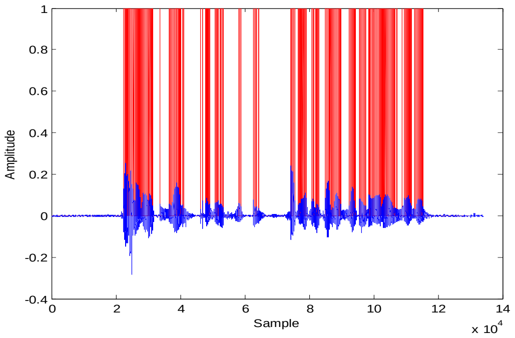

Figure 4 Output waveform of Single Pole Filtering based VAD, here speech signal is detected and the red line indicate voice activity detection

S.NO	**VAD Technique**	**Average score for speech across different noise**	
		Pink Noise (10db)	**Machine Gun (5db)**
1)	MFCC (Mel-Frequency Technique)	96.54	30.17
2)	Discrete Wavelet Transform and Teager Energy	69.76	77.60
3)	Support Vector Machine	87.55	82.20
4)	Single Frequency Filtering using VAD	98.72	98.16

Table 1 Comparison of speech across different noise in VAD technique

Hence from the above table it is observed that the single pole filtering based VAD consumes less time and the detection accuracy is high.

IV. Conclusion

A new VAD method is proposed based on Single Frequency Filtering approach introduced in this paper. The method exploits the fact that speech has high SNR regions at different frequencies and at different times. The mean and variance of speech at each frequency is higher than that for noise, after compensating for spectral characteristics for noise. The spectral characteristics of noise are determined using the floor of the temporal envelope at each frequency, computed by the SFF approach. The clear speech signal is obtained even in the presence of background noise. Future work is to apply Neuro-Fuzzy Decision logic. The input to the fuzzy operator is two or more membership values from fuzzified input variables. Any number of well defined methods can fill in for the AND operation or the OR operation. The product for AND, the maximum for OR and the weighted average as the defuzzification method. Finally, the output of the system is compared to a fixed threshold (η). If the output is greater thanη, the current frame is classified as speech otherwise it is classified as non-speech or silence.

References

[1]. A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, and S. Narayanan, “Multi-band long-term signal variability features for robust voice activity detection,” Aug. 2013, pp. 704–708.

[2]. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, and S. Narayanan, “Multi-band long-term signal variability features for robust voice activity detection,” inProc. Interspeech, Aug. 2013, pp. 718–722.

[3]. N.Dhananjaya and B. Yegnanarayana, “Voiced/nonvoiced detection based on robustness of voiced epochs,”IEEE Signal Process. Lett., vol. 17, no. 3, pp. 273–276, Mar. 2011.

[4]. T. Pham, M. Stark, and E. Rank, “Performance analysis of wavelet subband based voice activity detection in cocktail party environment,” inProc. Int. Conf. Comput. Commun. Technol., Oct. 2010,pp.85–88.

[5]. P.Ghosh,A.Tsiartas,and S.Narayanan,” Robust voice activity detection using long-term signal variability,” IEEE Trans. Acoust., Speech, Language Process., vol. 19, no. 3, pp. 600–613, Mar. 2010.