International Journal of Trend in Scientific Research and Development (IJTSRD)
UGC Approved International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 1 | Issue – 5

Voice Activity Detection and Pitch Analysis in Pathological Voices

Mr. G.B. Gour
BLDEA's College of Engineering & Technology,
Dept. of Electronics & Communication Engineering,
Ashram Road, Vijayapur-3, Karnataka, India

Dr. V. Udayashankara
Professor & Head of IT Dept,
Sri Jayachamarajendra College of Engineering,
Mysore-6, Karnataka, India

ABSTRACT

This paper presents the voice activity detection (VAD) using basic parameters. The voice is found to have hoarseness in case of thyroid and laryngeal cancer patients. It is useful for the better analysis of pathological voices in presence of background noise. Automatic voice pathology detection and classification is important in the voice disorder assessment. Here, two pitch detection algorithms are used for the analysis. This enhanced voice enables the early detection and diagnosis of type of pathology related to the patient.

Keywords: *Voice pathology detection and classification, Voice activity detectio, Pitch detection*

1. INTRODUCTION

Speech signals normally contain many areas of silence or noise. Therefore, in speech analysis voice activity detection is important for acquiring ”clean” speech segments. It is useful in applications like, speech-based human-computer interaction, audio-based surveillance systems and in many automatic speech recognition systems. The Voice activity detection is basically a part of speech pre-processing. The most basic use of silence/noise detection is the background noise reduction. The normal tests on differentiating silence/noise frame from the speech one are usually based on the parameters like [1, 3],

- Energy of the signal,
- Zero-crossing rate of the signal,
- Auto correlation coefficients,
- Spectral features, etc.

A basic VAD is an important front end part of pre processing of speech signals. Its working is based on the principle of extracting measured features from the incoming audio signal which is divided into frames of 5-40 ms duration. These extracted features from the speech signal are then compared with a threshold limit estimated from the noise only periods of the input signal and a VAD decision is computed. If the feature of the input frame is more than the estimated threshold value, a VAD decision (VAD = 1) is computed which declares that speech is present. Otherwise, a VAD decision (VAD = 0) is computed which declares the absence of speech in the input frame. The block diagram of a basic VAD is shown in fig 1.

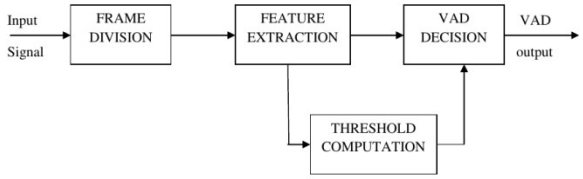

Fig.(1)According to [5], the required characteristics for an ideal voice activity detection are: reliability, robustness, accuracy, adaptation, simplicity. Among

these, robustness against noisy environments has been the most difficult task to achieve. In high SNR conditions, the simplest VAD algorithms can perform satisfactory, while in low SNR environ- ments, all of the VAD algorithms degrade to some extent. At the same time, the VAD algorithm should be of low complexity, which is necessary for real-time systems. Therefore simplicity and robustness against noise are two important characteristics of practicable voice activity detection. In view of voice activity detection, many algorithms have been proposed. The main difference between most of the proposed methods is the features used. The short-term energy and zero-crossing rate have been widely used because of their simplicity. However, they easily degrade by environmental noise.

In order solve this problem, many types of robust acoustic features, like autocorrelation function based features [6,7], spectrum based features [5], the power in the bandlimited region [1,6,7], Mel-frequency cepstral coefficients (MFFC) [8], delta line spectral frequencies [9], have been proposed for VAD. Use of such multiple features however has improved efficiency in different environments but, with increase in its complexity. Some works, propose noise estimation and adaptation for improving VAD robustness [10], but these methods are computationally expensive.

Many pitch detection algorithms (PDAs) have been developed in the past: autocorrelation method [11], HPS [12], RAPT [13], AMDF method [14], CPD [15], SIFT [16], DFE [17] with high accuracy for voiced pitch estimation. But, the PDAs performance degrades as the signal conditions deteriorate [18]. Pitch detection algorithms are grouped into the following basic categories: time-domain based tracking, frequency domain based tracking or joint time-frequency domain based tracking.

In this paper a VAD algorithm is proposed which is easy-to-implement. As spectral features adopted are found to be suited for low SNR signals or pathological speech signals with low back ground noise. This introduction follows by a discussion on short-term features which are used in the proposed method. In Section 2, the proposed VAD algorithm is explained in detail. Section 3 further explains the algorithm in detail with extraction of features. The Section 4 discusses the results of VAD and Pitch detection algorithms (PDA) in detail. Finally, the conclusions and future works are mentioned in Section 5.

2. ALGORITHM DESCRIPTION

The following steps are to be followed:

1) Two feature sequences are extracted from the speech signal.
2) For each sequence two thresholds are estimated.
3) A simple thresholding criterion is applied on the sequences.
4) Speech segments are detected based on the above criterion and finally a simple post-processing stage is applied.

3. FEATURE EXTRACTION AND VOICE DETECTION

3.1. Feature Extraction

For the feature extraction the signal is first divided into non-overlapping short-term-frames of 50 milli seconds length. Then for each frame, the two features, described below, are calculated.

1) **Signal Energy:**
Let x_i (n), n = 1, . . . , N the speech samples of the i^{th} frame, of length N . Then, for each frame i the energy P is calculated according to N^{th} equation:

$$E(i) = \frac{1}{N}\sum_{n=1}^{N}\left(\left(\mathrm{xi}(n)\right)\right)^2$$

This simple feature can be used for detecting silent periods in speech signals, but also for discriminating between speech classes.

2) **Spectral centroid:**
The spectral centroid of a sound is the midpoint of the spectral energy distribution of that sound. The spectral centroid, C_i , of the i^{th} frame is defined as the center of "gravity" of its spectrum, i.e.,

$$\mathrm{Ci}=\frac{\sum_{k=1}^{N}(\mathrm{k+1})\,\mathrm{Xi}(k)}{\sum_{k=1}^{N}\mathrm{Xi}\,(k)} \circ \mathrm{Xi}(k)$$

k = 1 . . . , N , is the Discrete Fourier Transform (DFT) coefficients of the i^{th} short-term frame, where N is the frame length. It is calculated by taking the sum of the frequencies weighted by (i.e. multiplied by) the linear amplitudes, divided by the sums of the linear amplitudes alone. This feature is a measure of the spectral position, with high values corresponding to "brighter" sounds. Experiments have indicated that the sequence of spectral centroid is highly varied for speech segments [1, 2].

The reasons that these particular features were selected

(Apart from their simplicity in implementation) are:

1) For simple cases, (where the level of background noise is not very high) the energy of the voiced segments is larger than the energy of the silent segments.
2) If unvoiced segments simply contain environmental sounds, then the spectral centroid for the voiced segments is again larger, since these noisy sounds tend to have lower frequencies and therefore the spectral centroid values are lower.

3.2. Speech Segment Detection

As long as the two feature sequences are computed, as simple threshold-based algorithm is applied, in order to extract the speech segments, at a first stage two thresholds (one for each sequence) are computed. Towards this end, the following process is carried out, for each feature sequence:

1) Compute the histogram of the feature sequence's values.
2) Apply a smoothing filter on the histogram.
3) Detect the histogram's local maxima.
4) Let M_1 and M_2 be the positions of the first and second local maxima respectively.

The threshold value is computed using the following equation:

$$T=\frac{W.M1+M2}{W+1}$$

W is a user-defined parameter. Normally the value of W chosen is 5. Large values of W lead to threshold values closer to M_1.

The above process is executed for both feature sequences, leading to two thresholds: T_1 and T_2 , based on the energy sequence and the spectral centroid sequence respectively. As long as the two thresholds have been estimated, the two feature sequences are thresholded, and the segments are formed by successive frames for which the respective feature values (for both feature sequences) are larger than the computed thresholds. For speech signal as reliable step size in range of 1 ms and frame length of 25 ms is chosen.

3.3 Post processing

As a post-processing step, the detected speech segments are lengthened by 5 short term windows (i.e., 250 milli seconds), on both sides. Finally, successive segments are merged.

4. RESULTS AND DISCUSSION

4.1 VAD Analysis

The main function is implemented in linux platform using Octave. When this function is called, the algorithm finished detecting the voiced segments. The second argument in the function provides a figure and is plotted that contains: 1) the energy sequence and the respective threshold 2) the spectral centroid sequence and the respective threshold and 3) the speech signal, plotted with different colours for the areas of the detected segments as shown in Figure 2. It shows the sequence of the signal's energy, followed by the spectral centroid sequence is presented. In both cases, the respective thresholds are also shown. The third sub figure presents the whole speech signal. Red color represents the detected voiced segments.

The function returns:

1) Cell array "segments": each element of that cell is a vector of speech samples of the corresponding detected voiced segment.

2) The sampling frequency of the speech signal.

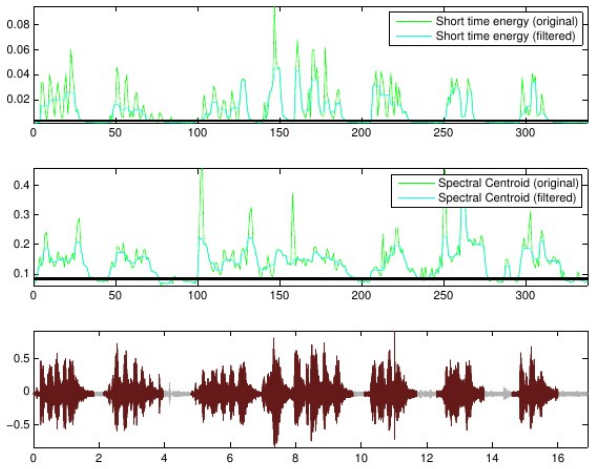

Fig.(2)

4.2 Pitch Detection Algoritms (PDA) Analysis

(i) Modified Auto correlation Function (MACF) Method:

This method is based on detecting the highest value of the auto correlation function in the region of interest. For given discrete signal x(n), the auto correlation function is defined as,

$$R(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} x\,(n).x(n+m), 0 \leq m < Mo$$

The variable m in above equation is called lag or delay, and the pitch is equal to the value of 'm' which results in the maximum R(m). The modified auto correlation pitch detector MACF [16] differs from the common auto correlation method by using center-clipping technique in a pre processing stage.

(ii) Average Magnitude Difference Function (AMDF) Method:

The average magnitude difference function (AMDF) [14] is another type of auto correlation analysis. Instead of correlating the input speech with different delays (where multiplications and summations are formed at each value), a difference signal is formed between the delayed speech and original, and at each delay value the absolute magnitude is taken. For the frame of N samples, the short-term difference function AMDF is defined as,

$$Dx(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} \Big(\big(x(n) - x(n+m)\big)\Big), 0 \leq m < Mo$$

where x(n) are the samples of analyzed speech frame, x(n-m) are the samples time shifted on 'm' samples and N is the frame length. The difference function has a local minimum if the lag 'm' is equal to or very close to the fundamental period. PDA based on average magnitude difference function has relatively low computational cost and simple implementation as it does not involve any multiplications. Along with the pitch estimation the ratio between the maximum and minimum values of AMDF (MAX/MIN) is obtained. This measurement with the frame energy is used to make a voiced/unvoiced decision.

(iii) Cepstrum Pitch Determination (CPD):

This is frequency domain method [11]. The sample is passed through Hamming window to get flat and smooth frequency peaks. Next Cepstrum is calculated this is Fourier analysis of the logarithmic spectrum. The equation is as follows,

$$C = \left(\left(F^{-1}\left(\log\left(\left(F(x(t))\right)\right)\right)^2\right)^2\right)$$

If the peak value exceeds threshold it is classified into voiced sample and location of the peak is the pitch period. If in case the peak value doesn't exceed threshold and zero-crossing count will determine whether sample is voiced or unvoiced.

The accuracy of the different pitch detection algorithms was measured according to the following criteria [18]:

1. Classification Error (CE): it is the percentage of unvoiced frames classified as voiced and voiced frames classified as unvoiced.
2. Gross Error (GE): percentage of voiced frames with an estimated fundamental frequency value that deviates from the reference value more than 20%.

Method	GE%		CE%	
	Male	Female	Male	Female
MACF	0.82	2.5	2.98	8.21
AMDF	3.5	7.08	17.15	25.64
CPD	0.79	3.2	10.9	18.7

Table .1

As shown in Table (1), for MACF only 3.07% of voiced frames were misclassified as unvoiced and unvoiced frames misclassified as voiced. The CPD

algorithm gets the good results in pitch estimation for male or female speech. Results of experiments show that AMDF method is the most inaccurate one. It has the biggest values of gross error and classification error parameters.

5. CONCLUSIONS

The present work is useful in pathological voice analysis recorded in low back ground noise, where spectral features provides better results. It is implemented in linux platform using Octave, an open source software in comparison to Mat Lab. The work provides a simple VAD implementation, an important front end part of a pre processing in speech signal analysis. The Cepstral methods are found to be useful in speech signal analysis with low back ground noise. Hence, the inclusion of this work as a part of speech signal pre processing and use of Cepstral methods may provide a better way of classification of voice disorders. Besides perceptual evaluation, the objective analysis of voice is a better classical approach. There are many time and frequency domain algorithms for the back ground noise reduction in speech signals.

Each of the described PDA algorithms have their advantages and drawbacks. From the experimental results, the MACF method is more convenient for common usage. This algorithm gives accurate results of pitch estimation and low computational complexity. The CPD shows good pitch estimation accuracy. Fundamental frequency estimation in this algorithm is immune to errors due to effects of vocal tract. But, CPD method is computationally complex; it needs additional parameters for voiced/unvoiced decision. The AMDF method has great advantage in very low computational complexity, it possible to implement it in real-time applications. This work provides provides better emphasis for the working such algorithms.

REFERENCES

[1] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Third Edition. Orlando, FL, USA: Academic Press, Inc., 2008.

[2] Atal B. S., Rabiner L. R.: A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-24, no. 3, June 1976

[3] T. Giannakopoulos, “Study and application of acoustic information for the detection of harmful content, and fusion with visual information,” Ph.D. dissertation, Dpt of Informatics and Telecommunications, University of Athens, Greece, 2009.

[4] K. Li, N. S. Swamy and M. O. Ahmad, “An improved voice activity detection using higher order statistics,” IEEE Trans. Speech Audio Process. , 13, pp. 965-974, 2005.

[5] M. H. Savoji, "A robust algorithm for accurate end pointing of speech," Speech Communication , pp. 45–60, 1989.

[6] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan and R. Sarikaya, “Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system,” Proc. ICASSP , 1, pp. 53-56, 2002.

[7] T. Kristjansson, S. Deligne and P. Olsen, “Voicing fea- tures for robust speech detection,” Proc. Interspeech, pp. 369-372, 2005.

[8] R. E. Yantorno, K. L. Krishnamachari and J. M. Lovekin, “The spectral autocorrelation peak valley ratio (SAPVR) – A usable speech measure employed as a co-channel detection system,” Proc. IEEE Int. Workshop Intell. Signal Process . 2001.

[9] M. Marzinzik and B. Kollmeier, “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics,” IEEE Trans. Speech Audio Process , 10, pp. 109- 118, 2002.

[10] B. Lee and M. Hasegawa-Johnson, "Minimum Mean Squared Error A Posteriori Estimation of High Variance Vehicular Noise," in Proc. Biennial on DSP for In-Vehicle and Mobile Systems , Istanbul, Turkey, June 2007.

[11] A. M. Kondoz , "Digital speech: Coding for low bit rate communication systems", 2nd Edn, John Wiley&Sons, England, 2004.

[12] W. J. Hess, Pitch Determination of Speech Signals. New York: Springer, 1993.

[13] D. Talkin, “A robust algorithm for pitch tracking (RAPT)”.Speech Coding and Synthesis, Elsevier Science, Amsterdam, pp.495-518,1995.

[14] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, “Average magnitude difference function pitch extractor,” vol.ASSP-22, no. 5, pp. 353–362, Oct. 1974.

[15] A. M. Noll, "Cepstrum Pitch Determination", Journal of the Acoustical Society of America, Vol. 41, No. 2, pp. 293-309, 1967

[16] L. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms,"

IEEE Transactions on ASSP, vol. 24, pp. 399-417, 1976.

[17] H. Bořil, P. Pollák, "Direct Time Domain Fundamental Frequency Estimation of Speech in Noisy Conditions". Proc. EUSIPCO2004, Wien, Austria, vol. 1, p. 1003-1006, 2004.

[18] B. Kotnik, H. Höge, and Z. Kacic, "Evaluation of Pitch Detection Algorithms in Adverse Conditions". Proc. 3rd International Conference on Speech Prosody, Dresden, Germany,pp. 149-152, 2006 .