EXTRACTION OF VOICED AND UNVOICED SEGMENTS IN AN AUDIO USING VAD

A PROJECT REPORT

Submitted by

T.ARIVUSUDAR (810015104011)

D.KOKILA (810015104040)

in partial fulfillment for the award of the degree

Of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

UNIVERSITY COLLEGE OF ENGINEERING – BIT CAMPUS,

TIRUCHIRAPPALLI

ANNA UNIVERSITY::CHENNAI 600 025

APRIL 2019

UNIVERSITY COLLEGE OF ENGINEERING,

BIT CAMPUS,

TIRUCHIRAPPALLI-620 024

BONAFIDE CERTIFICATE

Certified that this project report **"Extraction of Voiced and unvoiced Segments in an audio Using VAD"** is the bonafide work of **" Ms.T.ARIVUSUDAR (810015104011)** and **Ms.D.KOKILA (810015104040) "** who carried out the project work under my supervision.

SIGNATURE	**SIGNATURE**
Dr. D.Venkatesan	**Mrs.G.Revathi**
HEAD OF THE DEPARTMENT	SUPERVISOR
Assistant Professor	Teaching Fellow
Computer Science & Engineering	Computer Science & Engineering
University College of Engineering,	University College of Engineering,
Anna University-BIT Campus,	Anna University-BIT Campus,
Tiruchirappalli-620 024	Tiruchirappalli-620 024

Submitted for the project Viva voce examination held on ……

Internal Examiner **External Examiner**

DECLARATION

We hereby declare the work entitled **"EXTRACTION OF VOICED AND UNVOICED SEGMENTS IN AN AUDIO USING VAD"** is submitted in partial fulfillment of the requirement for the award of the degree in B.E., Computer Science and Engineering, University College of Engineering(BIT Campus), Tiruchirappalli, is a record of our own work carried out by us during the academic year 2018-2019 under the supervision and guidance of Mrs.G.Revathi, Teaching Fellow, Department of Computer Science and Engineering, University College of Engineering(BIT Campus), Tiruchirappalli. The extent and source of information are derived from the existing literature and have been indicated through the dissertation at the appropriate places. The matter embodied in this work is original and has not been submitted for the award of any degree, either in this or any other University.

SIGNATURE OF THE CANDIDATES

1. T.ARIVUSUDAR (810015104011)

2. D.KOKILA (810015104040)

I certify that the declaration made above by the candidate is true.

SIGNATURE OF THE GUIDE

Mrs.G.REVATHI,

Teaching Fellow,

Department of CSE,

University College of Engineering-BIT Campus,

Tiruchirappalli-620 024.

ACKNOWLEDGEMENT

I would like to convey my heartfelt thanks to our honorable Dean **Dr. T. SENTHILKUMAR,** Associate Professor for having provided me with all required facilities to complete my project without hurdles.

I would like to express my sincere thanks and deep sense of gratitude to guide **Mr. D. VENKATESAN,** Assistant Professor and Head, Department of Computer Science and Engineering, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of this project work.

I would like to thank my project guide **Mrs.G.Revathi,** Teaching Fellow, Department of Computer Science and Engineering, for his valuable guidance throughout the phase of the project. It is our responsibility to thank our project coordinator **Mr.C.SANKARRAM,** Assistant Professor, and **Mr.P.KARTHIKEYAN**, Assistant Professor, Department of Computer science and Engineering for his constant inspiration that he has all through the project period.

I would like to thank **Mr. C. SURESH KUMAR,** Teaching Fellow, Department of Computer Science and Engineering, for his encouragement for this work.

I extend my thanks to all other teaching and non-teaching staffs for their encouragement and support.

I thank my beloved parents and friends, for their full support in my career development of this project.

TABLE OF CONTENTS

ABSTRACT

A voice activity detection (VAD) system used in emerging speech technologies such as speech recognition, audio forensics, and wireless communication. An unsupervised VAD is used to classify the audio segments into voiced and unvoiced segments. To make the proposed method efficient and computationally fast, it is implemented by using long term features that are computed by using the Katz algorithm of fractal dimension estimation. The Noisy Speech Corpus (NOIZEUS) database is used to evaluate the performance of the proposed method. The Language of the database is in English, which contain different types of clean and noisy audios of various environments. The evaluation of this method is to label the voiced and unvoiced segments in both clean and noisy audio.

LIST OF FIGURES

LIST OF ABBREVIATIONS

1	VAD	Voice Activity Detection
2	DTW	Dynamic Time Warping
3	HMM	Hidden Markov Model
4	ANN	Artificial Neural Network
5	ASR	Automatic Speech Recognition
6	STT	Speech to Text
7	MOLRT	Multiple Observation Likelihood Ratio Test
8	SNR	Signal to Noise Ratio
9	MMSE	Minimum Mean Square Error
10	FFT	Fast Fourier Transform
11	DFT	Discrete Fourier Transform
12	GMM	Gaussian Mixture Model
13	NOIZEUS	Noisy Speech Corpus

CHAPTER 1

INTRODUCTION

1.1 Speech Processing

Speech processing is the study of speech signals and the processing methods of signals. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signals. Aspects of speech processing includes the acquisition, manipulation, storage, transfer and output of speech signals. The input is called speech recognition and the output is called speech synthesis. Early attempts at speech processing and recognition were primarily focused on understanding a handful of simple phonetic elements such as vowels.

Technologies used in speech processing:

Dynamic time warping:

Dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences, which may vary in speed. In general, DTW is a method that calculates an optimal match between two given sequences (e.g. time series) with certain restriction and rules. The optimal match is denoted by the match that satisfies all the restrictions and the rules and that has the minimal cost, where the cost is computed as the sum of absolute differences, for each matched pair of indices, between their values.

Hidden Markov models:

A hidden Markov model can be represented as the simplest dynamic Bayesian network. The goal of the algorithm is to estimate a hidden variable x(t) given a list of observations y(t). By applying the Markov property, the conditional probability distribution of the hidden variable x(t) at time t, given the values of the hidden variable x at all times, depends only on the value of the hidden

variable x(t − 1). Similarly, the value of the observed variable y(t) only depends on the value of the hidden variable x(t).

Artificial neural networks:

An artificial neural network (ANN) is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it. In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs.

1.2 Speech Recognition

Speech recognition is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields.

Some speech recognition systems require "training". Where an individual speaker reads text or isolated vocabulary into the system. The system analyses the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called "speaker independent" systems. Systems that use training are called "speaker dependent".

Speech recognition applications include voice user interfaces such as voice dialling, call routing, demotic appliance control, search, simple data entry,

preparation of structured documents, determining speaker characteristics, speech-to-text processing, and aircraft.

The term voice recognition or speaker identification refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process.

From the technology perspective, speech recognition has a long history with several waves of major innovations. Most recently, the field has benefited from advances in deep learning and big data. The advances are evidenced not only by the surge of academic papers published in the field, but more importantly by the worldwide industry adoption of a variety of deep learning methods in designing and deploying speech recognition systems.

1.3 Voice Activity Detection

Voice activity detection (VAD), also known as speech activity detection or speech detection, is a technique used in speech processing in which the presence or absence of human speech is detected. The main uses of VAD are in speech coding and speech recognition. It can facilitate speech processing, and can also be used to deactivate some processes during non-speech section of an audio session: it can avoid unnecessary coding/transmission of silence packets in Voice over Internet Protocol applications, saving on computation and on network bandwidth.

VAD is an important enabling technology for a variety of speech-based applications. Therefore, various VAD algorithms have been developed that provide varying features and compromises between latency, sensitivity, accuracy and computational cost. Some VAD algorithms also provide further analysis, for

example whether the speech is voiced, unvoiced or sustained. Voice activity detection is usually language independent.

Algorithm:

The typical design of a VAD algorithm is as follows:

1. There may first be a noise reduction stage, e.g. via spectral subtraction.
2. Then some features or quantities are calculated from a section of the input signal.
3. A classification rule is applied to classify the section as speech or non-speech – often this classification rule finds when a value exceeds a threshold.

There may be some feedback in this sequence, in which the VAD decision is used to improve the noise estimate in the noise reduction stage, or to adaptively vary the threshold. These feedback operations improve the VAD performance in non-stationary noise.

A representative set of recently published VAD methods formulates the decision rule on a frame by frame basis using instantaneous measures of the divergence distance between speech and noise. The different measures which are used in VAD methods include spectral slope, correlation coefficients, log likelihood ratio, Cepstral, weighted Cepstral, and modified distance measures.

Independently from the choice of VAD algorithm, we must compromise between having voice detected as noise or noise detected as voice. A VAD operating in a mobile phone must be able to detect speech in the presence of a range of very diverse types of acoustic background noise. In these difficult detection conditions it is often preferable that a VAD should fail-safe, indicating speech detected when the decision is in doubt, to lower the chance of losing speech segments. The biggest difficulty in the detection of speech in this

environment is the very low signal-to-noise ratios (SNRs) that are encountered. It may be impossible to distinguish between speech and noise using simple level detection techniques when parts of the speech utterance are buried below the noise.

Applications:

- VAD is an integral part of different speech communication systems such as audio conferencing, echo cancellation, speech recognition, speech encoding, speaker recognition and hands-free telephony.
- In the field of multimedia applications, VAD allows simultaneous voice and data applications.
- Similarly, in Universal Mobile Telecommunications Systems (UMTS), it controls and reduces the average bit rate and enhances overall coding quality of speech.
- In cellular radio systems based on Discontinuous Transmission (DTX) mode, VAD is essential for enhancing system capacity by reducing co-channel interference and power consumption in portable digital devices.
- In speech processing applications, voice activity detection plays an important role since non-speech frames are often discarded.

CHAPTER 2

LITERATURE SURVEY

2.1 A Robust Voice Activity Detection Method Based On Speech Enhancement

A robust multiple observation likelihood ratio test (MOLRT) based voice activity detection (VAD) method is proposed. At the beginning of this paper, we introduce the Wiener filter to the observed signal in time domain which can help mitigating some noise. The reason why we use the Wiener filter for VAD is that the performance of the VAD is always better in high signal to noise ratio (SNR) range than in low SNR range. Then, some ideas are proposed to improve the performance of the MOLRT based VAD method. As we all know, a conventional MOLRT based method including three modules named likelihood ratio (LR) estimation, threshold setting and hangover technique. To improve the estimation accuracy of LR, we adopt the unbiased minimum mean-square error (MMSE) algorithm for noise power spectrum estimation in every frame, which is very effective for LR estimation in MOLRT-based VAD method. That is because the LR is a function of a prior and a posterior SNR and unbiased MMSE algorithm is very useful for noise estimation. In addition, to make our VAD method more robust, a dynamic threshold setting technique is proposed in our method, which is related to the minimum noise power spectrum. That is because minimum noise power spectrum can help us updating the value of threshold to a suitable level according to the de noised signal. Last but most important, a novel hangover algorithm is introduced in this paper comparing to the conventional HMM based hangover algorithm. In the novel hangover algorithm, the current frame is determined by the statistical result of the following speech/non-speech detections based on the likelihood ratio test. And the evaluation results reveal that proposed method significantly outperform the baseline result of LRT as regards VAD accuracy in both noise variations and low SNR conditions.

Voice activity detection (VAD), which is a scheme to detect the presence of speech in the observed signals automatically, is considered as a crucial aspect for a wide range of speech processing algorithms and their applications, including speech coding, speech enhancement and automatic speech recognition (ASR). Generally, heuristics rule-based VAD and model-based VAD are considered as the most important methods. In the last decade, plenty of heuristics rule-based VAD algorithms have been developed for different kinds of noises. The difference among them is the features used, such as linear predictive coding parameters, energy, formant shape, higher order statistics and Cepstral features. However, heuristics rule-based VAD method is quite difficult to cope with all kinds of noises observed in the real world. Recently, the statistical model based VAD is considered as a more attractive approach for noisy speech, and the likelihood ratio test (LRT)-based VAD algorithm is one of the useful methods. LRT-based VAD method was first proposed by Shon in 1999. In this method, the author assumed the probability density functions of both speech and noise were Gaussian, and a Hidden Markov Model (HMM) was used for hang-over scheme. Later, Gorriz incorporated contextual information in a multiple observation LRT (MOLRT) to overcome the non-stationary noise. Tan only selected the DFT bins that containing harmonic spectral peaks for likelihood ratio (LR) in speech segments, as opposed to non-speech segments. Although many improved LRT-based methods have been proposed, the performance still drop rapidly when the signal-to-noise (SNR) becomes low. In this paper, we further optimize the MOLRT-based method by four different parts, which can obtain more reliable speech/non-speech decision. First, the Wiener filter is used for the noisy speech in the time domain, which can mitigate the noise effectively. Second, unbiased minimum mean square error (MMSE) algorithm is adopt to track the noise power spectrum of each frame. Third, minimum noise power spectrum based threshold is proposed, which makes our method more robustness.

2.2 Voice activity detection using harmonic frequency components in likelihood ratio test.

A new statistical model-based likelihood ratio test (LRT) VAD to obtain reliable speech / non-speech decisions. In the proposed method, the likelihood ratio (LR) is calculated differently for voiced frames, as opposed to unvoiced frames: only DFT bins containing harmonic spectral peaks are selected for LR computation. To evaluate the new VAD's effectiveness in improving the noise robustness of ASR, its decisions are applied to pre-processing techniques such as non-linear spectral subtraction, minimum mean square error short-time spectral amplitude estimator, and frame dropping. From the ASR experiments conducted on the Aurora2 database, the proposed harmonic frequency-based LRTs give better results than conventional LRT-based VADs and the standard G.729B and ETSI AMR VADs.

The performance of an automatic speech recognition (ASR) system degrades with mismatched training and test data. To improve the noise-robustness of an ASR system trained using clean data, techniques such as speech enhancement, noise-robust feature extraction feature enhancement, and model-based noise adaptation can be applied. Most of these techniques require a reliable voice activity detector (VAD) to identify non-speech segments for noise estimation. ASR performance can also improve with a good VAD alone by dropping non-speech segments. Spectral harmonicity has been utilized in noise-robust applications operating in the frequency domain because harmonic peaks are usually preserved in noisy speech. Nonlinear spectral subtraction (NSS) in defined a smaller subtraction factor at harmonic peaks which resulted in higher ASR accuracy at low SNRs. In, regeneration of harmonic structures improved the quality of denoised speech. A harmonic model is used in a generalized LRT for robust voiced/unvoiced detection in, and a periodic-to aperiodic component ratio used for speech/non-speech detection in showed promising results with aperiodic noise interferences. One popular VAD is the statistical model-based LRT VAD

first proposed. Variants of this noise-robust VAD to increase weak speech onset and offset detection have been proposed. For example, a smoothed LR, while used multiple observations of the short-time DFT feature vector to replace the hangover scheme. For these VADs, the high LRs of strong speech frames aid the detection of weak neighbouring speech frames. However, under low signal-to-noise ratio (SNR) conditions, the LRs of the stronger speech frames are not high enough to boost the detection of weaker speech frames. This paper presents a new way for calculating LR to tackle the above issue in LRT-based VADs and is organized as follows. Section 2 reviews the technical background of LRT-based VADs developed, while Section 3 describes the proposed method built on these VADs. The proposed VAD is then compared with the referenced VADs and standardized VADs: ITU's G.729B and ETSI AM. Reference speech and non-speech segments for the Aurora2 database's Test Set A are obtained through manual labelling the clean version every 10 ms. The receiver operating characteristic (ROC) curves are used to evaluate the accuracy of the proposed VAD. In Section 5, non-speech frames are used for noise estimation in speech enhancement algorithms. Hence, the probability of detection, Pd is defined as the percentage of correctly detected reference non-speech frames, while the probability of false alarm, Pf is the percentage of reference speech frames wrongly identified as non-speech. Fig. 1 shows the ROC curves of LRT, MOLRT and the proposed improved versions of these LRT-based VADs, abbreviated as Hmfreq-LRT and Hmfreq MOLRT respectively for subway and babble noise-corrupted data. Mm= 8 is used in MOLRT VADs because it is reported to yield the best performance. The receiver operating points for the standardized VADs are also plotted. Reference speech and non-speech segments for the Aurora2 database's Test Set A are obtained through manual labelling the clean version every 10 ms. The receiver operating characteristic (ROC) curves are used to evaluate the accuracy of the proposed VAD.

2.3 Statistical Model-Based Voice Activity Detection Using Support Vector Machine

From an investigation of a statistical model-based voice activity detection (VAD), it is discovered that a simple heuristic way like a geometric mean has been adopted for a decision rule based on the likelihood ratio (LR) test. For a successful VAD operation, the authors first review the behaviour mechanism of support vector machine (SVM) and then propose a novel technique, which employs the decision function of SVM using the LRs, while the conventional techniques perform VAD comparing the geometric mean of the LRs with a given threshold value. The proposed SVM-based VAD is compared to the conventional statistical model-based scheme, and shows better performances in various noise environments.

Performance of the proposed VAD approach was evaluated on the NTT database that consists of a number of speech Figure 3 Scatter plot for L2 and L3 under the car noise (SNR ¼ 5 dB) Let us explain the NTT database for the training part. All the training data used for building support vectors were recorded in a quiet environment. Utterances of the NTT database spoken by from four male speakers and four female speakers were used to construct 226 s long speech data. It can be pointed that there is consensus on the use of the NTT database since the NTT corpus is faithfully designed for evaluating the performance of the speech coder. Each utterance (8 s) consisted of two different meaningful sentences with silence periods was concatenated. For training, we made a reference decisions on the clean speech materials by labelling manually at every 10ms frame. The proportions of voiced, unvoiced and silence frames of the training materials are 45.6, 13.7 and 40.7%, respectively. After making reference decisions, we added vehicular, babble, street and white noises to whole signal and SNR values of each noise are 5, 10 and 20dB. In our experiments, the input signal was sampled at 8 kHz and the analysis window size was 10ms with 3ms overlap. Each frame of the windowed signal was transformed to its corresponding spectrum through a 128-

point DFT after zero-padding. For the RBF kernel, the kernel width s was set to 1.0. For the test, we made reference decisions on a different clean speech material 349 s long by labelling manually at every 10ms frame by concatenating different speech materials of the NTT database. The percentage of hand marked active speech frames was 56.7%, which consisted of 44.0% voiced sound frames and 12.7% unvoiced sound frames. To simulate noisy conditions, vehicular, babble, street and white noises are added to the clean speech data by 5 dB SNR. To evaluate the performance of the proposed VAD compared with the previous statistical model-based VAD with the geometric mean, we investigated the receiver operating characteristics (ROC's), which shows the trade-off characteristic between the speech detection and false-alarm probabilities (Pd and Pf). We define Pd as the ratio of correct speech decisions to the hand-marked speech frames and Pf as that of false speech decisions to the hand-marked non-speech frames. For fair comparison, we do not consider any hangover scheme, as this can be added after the design of the decision rule. Also, performances of other SVM-based approaches by Enqing et al. and Ramirez et al. with the ITU-T G.729 Annex B (G.729B) are plotted for the purpose of gaining relative performance evaluation. The previous methods are achieved by adopting similar parameter settings such that the frame size is 10ms and the kernel width s is 1.0. In all testing conditions, the proposed SVM-based VAD using the linear kernel yielded higher performance over all than the Shon's method based on the geometric mean. It is evident from the result that SVM is a very desirable way to establish the decision rule for the statistical model-based VAD. Especially, we observe that the RBF kernel in most of the tested conditions significantly improved the performance of the proposed SVM-based VAD while the polynomial based VAD did not give us a consistent performance improvement. This phenomenon is attributable to the fact that the nonlinear drawback of the input data (LRs) is successfully resolved by the RBF kernel, which makes the input data linearly separable.

2.4 Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns

The problem of an adequate quantitative interpretation of epileptic EEG recordings is of great importance in the understanding, recognition, and treatment of epilepsy. In recent years, much effort has been made to develop computerized methods which can characterize different interracial, ictal, and postictal stages. The main issue of whether there exist a preictal phenomenon is unresolved. In the present work we address this issue making use of the m s t basic representation of data complexity, namely, the algorithmic information content. In general this measure, also known as Kolmogorov complexity, represents the compressibility of the data strings. It can also be used to describe properties (linear and nonlinear) of the underlying dynamical system. We analyse Kolmogorov complexity and related characteristics of intracranial EEG recording, containing preictal, ictal, and postictal segments.

In his late work Andrei Kolmogorov introduced a new algorithmic approach to the quantitative definition of information that used the concept of recursive functions. He analysed its relation with respect to two other common approaches: combinatorial and probabilistic. Alternative to the probabilistic notion of the information content I(x, y) conveyed by one random object x with respect to another y, he linked a complexity of an object to that of an algorithm which can generate it from another one. Naturally, all these measures are directly related to data compressibility and coding characteristics. It was mentioned in that although the proposed approach yields in principle a correct definition of the quantity of “hereditary information”, it would be difficult to obtain a reliable estimate of it. Later, however, an easily calculable measure for finite sequences was suggested which closely relates to their Kolmogorov complexity. An attempt to make use of such measures to distinguish different EEG signal patterns was made recently in. Despite the numerous positive results and investigations of EEG time series based on purely nonlinear dynamical characteristics (e.g. fractal

dimension of the strange attractor), the limitation of these studies is that a certain brain functional state may often last only for a brief time period, whereas, in order to obtain reliable estimates for the mentioned characteristics, long-term time series are needed. Also, the algorithms involved are extremely sensitive to the noise. Therefore, better alternative EEG analysis methods need to be discovered. The purpose of this work is to explore the relationship between various informational measures, including Kolmogorov complexity, information content, and fractal dimension of epileptic EEG series. The special point of interest is to determine whether the combined analysis of these characteristics can provide a reliable tool for distinguishing preictal stages, and thus may be useful for prediction of epileptic seizures. The EEG data processed were obtained from an epileptic patient with depth electrodes surgically implanted into the epileptic focus. A 32-channel EEG with a sampling rate of 200 Hz was recorded using Stellate's Monitor System. The visual based segmentation of these recordings into interictal, preictal, ictal, and postictal stages was performed by an expert Epileptologist to provide feasible transitional changes. Yet, to designate at what point preictal EEG changes indicated that an epileptic seizure would inevitably follow, the preictal stage was further divided into preictal 1 (seizure non inevitable) and preictal 2 (seizure inevitable) segments.

This study represents an attempt to incorporate the basic informational characteristics for analysis of EEG signals. In particular, it shows the feasibility of combined consideration of those measures for recognition of epileptic EEG patterns. However, the data are preliminary as only one subject was recorded. Obviously, the considered original-to-binary conversion methods can be optimized for a given application by the parameters involved. In addition, the essence of differences of calculated measures when different conversion methods are applied needs to be explored. This, combined with interrelationships of these measures may provide guidelines for distinguishing different preictal EEG patterns.

2.5 Fractal aspects of speech signals: dimension and interpolation

The nonlinear dynamics of air flow during speech production may often result into some small or large degree of turbulence. In this paper we quantify the geometry of Speech turbulence, as reflected in the fragmentation of the time signal, by using fractal models. We describe an efficient algorithm for estimating the short-time fractal dimension of speech signals and use it for speech segmentation and sound classification. We also develop a method for fractal speech interpolation, which can be used to synthesize controlled amounts of turbulence in speech or to increase its sampling rate by preserving not its bandwidth (as classically done) but rather its fractal dimension. There are several mechanisms for the creation of vortices: 1) velocity gradients in boundary layers, 2) separation of flow, which can easily happen at cavity inlets due to adverse pressure gradients (see [GI for experimental evidence for separated flow during speech production), and 3) curved geometry of tract boundaries, where due to the dominant inertia forces the flow follows the curvature and develops rotational components.

Let the continuous function S (t) , 0 I: t 5 T , represent a short time speech signal, let the set XC R2 represent its graph, and let DH be the Hausdorff dimension of X . The signal S is called fractal if its graph is a fractal set [2], i.e., if DH > 1. Next we discuss two other dimensions closely related to DH. Minkowski-Bouligand dimension DMB: Dilate X with a disk of radius E and thus create a Minkowski cover because: 1) It coincides with DH in many cases of practical interest; 2) It is much easier to compute than DH. 3) It will be applied to sampled signals where most approaches can yield only approximate results. 4) It can be more robustly estimated than DE, which suffers from uncertainties due to the grid translation or its spacing E relative to the signal's amplitude. (DE = DMB in the continuous-time case, but they correspond to two different algorithms (with different performances) for sampled signals.), D will not change

if we replace the disks in the Minkowski cover of X with other compact convex symmetric shapes BC R2.

These dilations and erosions create an area-strip as a layer either covering or being peeled off from the graph of the speech signal at various scales. The reason for the higher than usual sampling rate is to preserve the fragmentation of the sampled signal as close as possible to that of the continuous-time speech signal. Thus, for normal conversational speech, we have found that its short time (e.g., over - 10-30msec frames) fractal dimension D (evaluated at a scale < 0.1msec) can roughly distinguish these three broad classes of speech sounds by quantifying the amount of their waveform's fragmentation. However, for loud speech (where the air velocity and Re increase, and hence the onset of turbulence is easier) or for breathy voice (especially for female speakers) the dimension of several speech sounds, e.g. vowels may significantly increase. In general, the D estimates may be affected by several factors including a) the time scale, b) the specific discrete algorithm (usually most algorithms for sampled signals underestimate the true D since some signal's fragmentation has been lost during sampling), and c) the speaking state. Therefore, we often don't assign any particular importance to the absolute D estimates but only to their average ranges and relative differences.

Band limited discrete signal interpolation has been done traditionally by up sampling and passing the up sampled signal through a low-pass linear filter to smooth the abrupt transitions during gaps, while preserving the bandwidth. Given the importance of fractal dimension for speech, we develop here an alternative approach to interpolate speech by synthesizing and up sampling a fractal function that interpolates the given low-rate speech and can have any desired fractal dimension. Before we discuss fractal speech interpolation, we summarize basic ideas from the theory of fractal interpolation functions. The waveform of a word and its short-time fractal dimension, average zero-crossing rate, and energy as functions of time.

2.6 Optimal Detection of Change points With a Linear Computational Cost

We consider the problem of detecting multiple change points in large data sets. Our focus is on applications where the number of change points will increase as we collect more data: for example in genetics as we analyse larger regions of the genome, or enhance as we observe time-series over longer periods. We consider the common approach of detecting change points through minimising a cost function over possible numbers and locations of change points. This includes several established procedures for detecting changing points, such as penalised likelihood and minimum description length method for the minimum of such cost functions and hence the optimal number and location of change points that has a computational cost which, under mild conditions, is linear in the number of observations. This compares favourably with existing methods for the same problem whose computational cost can be quadratic or even cubic. In simulation studies we show that our new method can be orders of magnitude faster than these alternative exact methods. We also compare with the Binary Segmentation algorithm for identifying change points, showing that the exactness of our approach can lead to substantial improvements in the accuracy of the inferred segmentation of the data.

There is therefore a growing need to be able to search for such changes efficiently. It is this search problem which we consider in this paper. In particular we focus on applications where we expect the number of change points to increase as we collect more data. This is a natural assumption in many cases, for example as we analyse longer regions of the genome or as we record financial time-series over longer time-periods. By comparison it does not necessarily apply to situations where we are obtaining data over a time-period at a higher frequency. We now investigate the theoretical computational cost of the PELT method. We focus on the most important class of change point models and penalties and provide sufficient conditions for the method to have a computational cost that is linear in the number of data points. The case we focus on is the set of models

where the segment parameters are independent across segments and the cost function for a segment is minus the maximum log-likelihood value for the data in that segment. More formally, our result relates to the expected computational cost of the method and how this depends on the number of data points we analyse. To this end we dene an underlying stochastic model for the data generating process. Specially we such a process over positive-integer time points and then consider analysing the data points generated by this process. Our result assumes that the parameters. If two change points are identified in this window then one is counted as correct and one false. The number of false change points is then the total number of change points identified minus the number correctly identified. The results are depicted in Figure 3 for a selection of data lengths, n, for the case m = n=100. As n increases the difference between the PELT and BS algorithms becomes clearer with PELT correctly identifying more change points than BS. Qualitatively similar results are obtained if we change how close an inferred change point has to be to a true change point to be classified as correct. Figures for square root increasing and numbers of change points are given in the supplementary material. As the number of change points decreases a higher proportion of true change points are detected with fewer false change points. The supplementary material also contains an exploration of the same properties for changes in both mean and variance. The results are broadly similar to those described above. We now demonstrate increased accuracy of the PELT algorithm compared with BS on an oceanographic data set. In this paper we have presented the PELT method; an alternative exact multiple change point method that is both computationally efficient and versatile in its application. It has been shown that under certain conditions, most importantly that the number of change points is increasing linearly with n, the computational efficiency of PELT is O (n). The simulation study and real data examples demonstrate that the assumptions and conditions are not restrictive and a wide class of cost functions can be implemented.

2.7 Using penalized contrasts for the change-point problem

A methodology for model selection based on a penalized contrast is developed. This methodology is applied to the change-point problem, for estimating the number of change points and their location. We aim to complete previous asymptotic results by constructing algorithms that can be used in diverse practical situations. First, we propose an adaptive choice of the penalty function for automatically estimating the dimension of the model, i.e., the number of change points. In a Bayesian framework, we define the posterior distribution of the change-point sequence as a function of the penalized contrast. MCMC procedures are available for sampling this posterior distribution. The parameters of this distribution are estimated with a stochastic version of EM algorithm (SAEM). An application to EEG analysis and some Monte-Carlo experiments illustrate these algorithms.

Detection of abrupt changes in the characteristics of some physical system is one of the important practical problems arising in signal processing (speech processing, geophysics, EEG, EMG and ECG analysis, etc., see for several examples of application). In a probabilistic framework, we consider a sequence of random variables Y1; . . . ; Yn, that take values in Rp. The changes can affect the marginal distribution of the Yi's (the mean, the variance, or some quantiles for example), or the joint distribution of the sequence (the spectral distribution for example). Among the previously proposed methods for detecting multiple changes, we mention sequential methods and local methods. We shall adopt here a global approach, where all the change points are simultaneously detected by minimizing a penalized contrast measures the fit of s with y. Its role is to locate the change points as accurately as possible. The penalty term only depends on the dimension of the model s and increases. Thus, it is used for determining the number of change points. The penalization parameter b adjusts the trade-off between the minimization of (obtained with a high dimension of s), and the minimization of(obtained with a small dimension of s).Asymptotic results

concerning penalized least squares estimates have been obtained in theoretical general contexts in, extending the previous results of Yao. We shall show that this kind of contrast can also be useful in practice. The main problem is the choice of a good penalty function and a good coefficient b. In the Gaussian case, Yao suggests the Schwarz criterion. A complete discussion of the most popular criteria (AIC, Mallow's Cp, BIC), and many other references can be found in. In a more general context, we can use a contrast other than the least-squares criterion, since the variables are not necessarily Gaussian and independent. Nevertheless, we propose an adaptive procedure for automatically choosing the penalty parameter b in Section 2.In a Bayesian framework, we construct a conditional distribution. Obviously, the mode of this distribution is the minimum penalized contrast estimate previously defined. A MCMC (Markov Chain Monte Carlo) procedure provides a way to sample and examine this posterior distribution, instead of only computing its mode. Furthermore, the artificial introduction of a ''temperature'' parameter allows us to concentrate this posterior distribution around the models s of highest probability.

We propose to adopt the same approach in a more general context. We present here this methodology without giving any more details. The ''tuning'' parameter T is usually called ''temperature''. This parameter controls how the distribution p is concentrated around its mode. It should be chosen small enough to neglect the models s having a low posterior probability and to increase the probability of the most likely models. Here, the MCMC algorithm creates an homogeneous Markov chain since the temperature parameter remains constants. Maximization of the conditional distribution could be achieved using a simulated annealing procedure. In this case, the temperature is not constant but decreases slowly to zero. Simulated annealing is very slow and the dynamic programing algorithm described above should be preferred for computing the mode of this distribution.

CHAPTER 3

SYSTEM ANALYSIS AND DESIGN

3.1 EXISTING SYSTEM

There may first be a noise reduction stage that is pre-processing of audio samples, e.g. via spectral subtraction. Then some features or quantities are calculated from a section of the input signal. Here the long term features are calculated such as fundamental frequency, shimmer, Jitter. A classification rule is applied to classify the section as speech or non-speech – often this classification rule finds when a value exceeds a threshold. In this stage threshold value is calculated to classify the speech presence and speech absence in an audio. If the value above the threshold is classified as speech presence and below is classified as speech absence. Then accuracy is calculated to evaluate the performance of the VAD. Which is minimal accurate for noisy signal. The performance of the proposed method is evaluated by using two databases. One is Texas Instruments Massachusetts Institute of Technology (TIMIT) database another one is King Saud University (KSU). The language of TIMIT is English, while the language of the KSU speech database is Arabic. TIMIT is recorded in only one environment, whereas the KSU speech database is recorded in distinct environments using various recording systems that contain sound cards of different qualities and models. The evaluation of the method suggested that it labels voiced and unvoiced segments reliably in both clean and noisy audio.

Voice activity detection (VAD) is a process that divides speech signals into at least two types of segments, referred to as speech-presence and speech-absence. VAD can be identified as a statistical hypothesis problem, where the purpose is to determine the class of the segment in an audio, i.e., the class of the segment in this case is either speech-presence or speech-absence. The decision of a segment depends on the computed feature vectors, which are a vital part of a VAD system. The feature vector extracts the characteristics of a segment and

serves as the input to a decision rule that assigns a sample vector to one of the given classes. The drawback of such system is the computational cost of the statistical model implemented to label the segments. On the other hand, unsupervised VAD systems automatically detect the acoustic patterns in an audio by using signal properties. Human-annotated data for acoustic model training are no longer needed. The main goal of the current research is to develop an accurate and reliable unsupervised VAD method based on fractal dimension. To estimate the fractal dimension, algorithms such as Katz, Higuchi, Petrosian, Maragos, and the amplitude scale method have been proposed. These algorithms are used in various scientific areas to compute the fractal dimension of time series and waveforms.

This method divides audio samples into shorter frames. Then, fractal dimensions are computed for each frame of an audio to keep the method more efficient. Katz algorithm is employed to calculate the fractal dimension. With the use of the computed fractal dimension, a threshold to detect speech-presence and speech-absence segments is calculated automatically. The threshold varies from one audio to another. Therefore, the method is robust against recording environment and equipment. The method is evaluated by using two speech databases, and it accurately detected speech-presence and speech-absence segments for audio samples from both databases. One of the advantage of this method is VAD is language independent. It is used to deactivate some process during in non-speech section.

3.1.1 Disadvantages

- Existing VAD is system is not efficiently working noisy environment.
- So that we are eliminating the noise in the initial stage of an existing VAD.
- Its computational cost and latency is very high.
- It is very time consuming.
- It has minimal sensitivity and it doesn’t provide better accuracy results.

3.2 PROPOSED SYSTEM

The proposed method for automatic VAD identifies the speech-absence and speech-presence segments in an audio. It performs pre-processing of the audio before extraction of long-term features by using the fractal dimension estimation algorithm. To detect the segments of speech-presence and speech-absence, audios are partitioned into frames of shorter durations. The durations of frames are kept shorter for accurate classification of both types of segments. The drawback of long frame duration is that it may contain speech in some parts, and the remaining parts may contain silence or short pauses. Therefore, each audio is divided into frames of 10 milliseconds. In the next step, the features for the automatic identification of speech-presence and speech-absence segments are computed. The computed features are fractal dimensions of an audio. Fractal dimensions for each divided frame of an audio are calculated by using the Katz algorithm. To calculate the fractal dimension of a frame, the length of the waveform is computed by adding the Euclidean distances between all consecutive points on the waveform. To calculate the fractal dimension, the next step is to calculate the planar extent P, which is the maximum distance between the first and any point on the waveform. Then, length M and planar extent P are normalized by dividing them with the average distance between consecutive points on the waveform. The required fractal dimension T is a ratio between normalized length and planar extent of the waveform. On the basis of the fractal dimensions of each frame, it will be decided whether a certain frame contains speech or silence. The automatic decision for VAD in the proposed method is made by computing a threshold. The fractal dimension above the threshold will determine if it is a speech-presence or a speech-absence segment. To determine the threshold, the computed fractal dimensions are sorted in ascending order, the first most significant change in the sorted fractal dimension is estimated by using the optimal change point detection. The vertical line signifies the position of the

first most significant change in the curve. The threshold is computed by taking the average of sorted fractal dimensions up to the occurrence of the first most significant change. This paper presents an accurate and efficient unsupervised VAD method to classify speech-presence and speech-absence segments in an audio. To the best of our knowledge, an unsupervised VAD method that uses fractal dimension has never been implemented to label segments in audios.

The Noisy Speech Corpus (NOIZEUS) database is used to evaluate the performance of the proposed method. The Language of the database is in English. Which contain different types of clean and noisy audios of various environment. Then the overall performance is evaluated by using linear SVM (support vector machine).

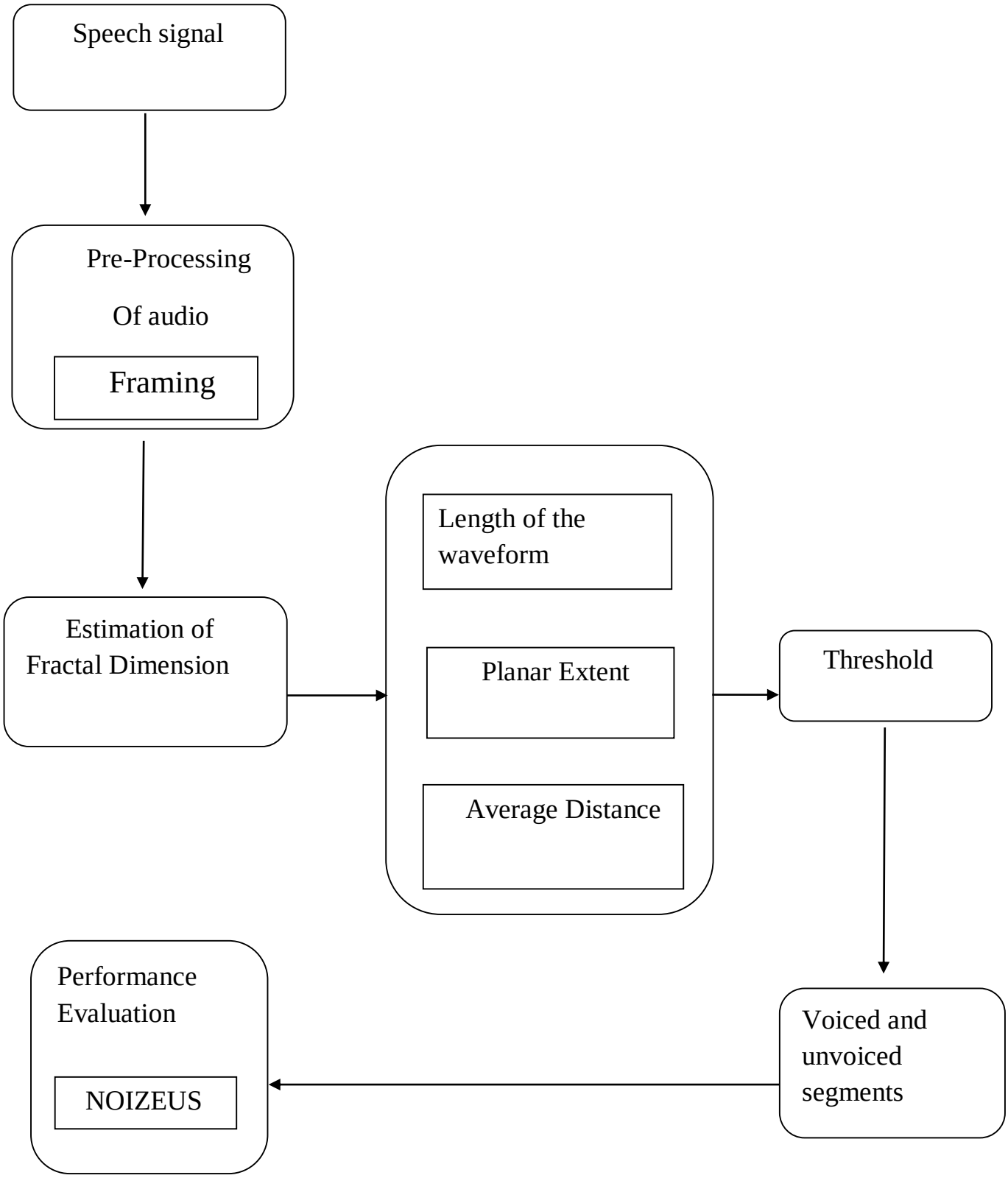
3.2.1 Advantages

- The proposed method is unsupervised and does not need any training data to differentiate between voiced and unvoiced segments.
- The presence of significant noise in an audio does not affect the performance of the proposed VAD method.
- Therefore this method is efficiently working in both clean and noisy signals.
- This method computationally very fast.

CHAPTER 4

DIAGRAMS AND MODULES

4.1 ARCHITECTURE DESIGN



MODULES:

4.1.1 Pre-processing

The first step of the proposed method is the pre-processing of audio samples. The sampling frequencies of all audio samples are down-sampled to 4 KHz because one of the databases is recorded at 8 KHz. By doing so, all audios of both databases will have a unique sampling frequency and can be used to evaluate the performance of the proposed method.

To detect the segments of speech-presence and speech-absence, audios are partitioned into frames of shorter durations. The durations of frames are kept shorter for accurate classification of both types of segments. The drawback of long frame duration is that it may contain speech in some parts, and the remaining parts may contain silence or short pauses. Therefore, each audio is divided into frames of 10 milliseconds. Consider an audio D, given in Eq. (1), with its samples d_1, d_2, d_3, ……, d_N, where N represents a total number of samples in D

$$D= [\ d_1, d_2, d_3 ,..,d_N]. \quad (1)$$

The i^{th} frame F_i can then be obtained as

$$F_i = [d_{(i-1)l+1},d_{(i-1)l+2},d_{(i-1)l+3}, \ldots..,d_{il-1},d_{il}] \quad (2)$$

Where

$l = s*r$ and $1 <= i <= n (=\frac{N}{l})$

In Eq. (2), s is the sampling frequency of an audio, r stands for the duration of a frame in seconds, l represents the number of samples in a frame, and n provides a total number of frames for an audio D.

After the division of an audio into frames $[F_1, F_2, F_3,\ldots, F_n]$ some samples at the end of an audio may not be a part of any frame because of the insufficient amount of samples. For instance, in the case of a 10 millisecond frame (r = 0.010 seconds) of an audio recorded at 4 KHz (s = 4,000), each frame will contain 40

samples (l = 40), and it is possible that 31 samples at the end of an audio are not part of any frame, when n is not an integer. These samples may contain some important information that is vital for the development of speech-related systems. Zero padding is performed at the end of such audios to generate a complete frame by using the unused samples. In this way, no information in an audio will be lost. The last frame F_n of an audio D can be obtained using Eq. (3):

$$F_n = [\, d_{\left[\frac{N}{l}\right]l+1}, d_{\left[\frac{N}{l}\right]l+2}, d_{\left[\frac{N}{l}\right]l+3}, \ldots\ldots, 0, 0, 0, \ldots\ldots, 0\,] \quad (3)$$

Where [.] is the floor operator. The number of zeros in the frame F_n is determined by using Eq. (4):

$$\text{Numbers of Zeros} = l - \text{mod}\,(N, l) \quad (4)$$

In the next step, the features for the automatic identification of speech-presence and speech-absence segments are computed. The computed features are fractal dimensions of an audio. The steps to calculate the fractal dimension of each frame is described in the following subsection.

4.1.2 Estimation of fractal dimension

Fractal dimensions for each divided frame of an audio are calculated by using the Katz algorithm. To calculate the fractal dimension of a frame, the length of the waveform is computed by adding the Euclidean distances between all consecutive points on the waveform. Each point on the waveform is represented by an ordered pair (X, Y). The length M of the waveform is computed by using Eq. (5):

$$M = \sum_{j=1}^{l-1} \sqrt{\left(X_{j+1} - X_j\right)^2 + \left(Y_{j+1} - Y_j\right)^2} \quad (5)$$

To calculate the fractal dimension, the next step is to calculate the planar extent P, which is the maximum distance between the first and any point on the waveform. The planar extent is given by Eq. 6.

$$P = \max \left(\sqrt{(X_{j+1} - X_1)^2 + (Y_{j+1} - Y_1)^2} \right) \quad (6)$$

Where j = 1,2,3,….,l - 1

Then, length M and planar extent P are normalized by dividing them with the average distance between consecutive points on the waveform. The required fractal dimension T is a ratio between normalized length and planar extent of the waveform. The average distance V is computed by using Eq. (7).

$$V = \text{mean} \left(\sqrt{(X_{j+1} - X_j)^2 + (Y_{j+1} + Y_j)^2} \right) \quad (7)$$

Where j = 1,2,3,….,l - 1

The fractal dimension for all audio samples is computed by following the procedure. On the basis of the fractal dimensions of each frame, it will be decided whether a certain frame contains speech or silence.

4.1.3 Threshold

The automatic decision for VAD in the proposed method is made by computing a threshold. The fractal dimension above the threshold will determine if it is a speech-presence or a speech-absence segment. To determine the threshold, the computed fractal dimensions are sorted in ascending order. The first most significant change in the sorted fractal dimension is estimated by using the process such as Optimal Change Point Detection. The Circle signifies the position of the first most significant change in the curve. The threshold, is computed by taking the average of sorted fractal dimensions up to the occurrence of the first most significant change.

$$\text{Threshold} = \text{mean (sort (fractal (1: Ind)))} \quad (9)$$

Where fractal stands for the fractal dimensions of all segments of an audio, and Ind represents the intersection point of the curve with x-axis. All segments above the threshold will be categorized as speech-presence segments, while those below the threshold will be labelled as speech-absence segments.

4.1.4 Performance evaluation

The effectiveness of the proposed method is evaluated by using a speech database such as NOIZEUS. The method is evaluated by using clean and noisy audios to observe its robustness against noise.

Evaluation by using the NOIZEUS Database

The NOIZEUS English Speech Database is also used to evaluate the proposed VAD method because of its diversity in many aspects. This database conducting the recordings in different environments, airport, babble, car, restaurant, station and street, is one of the significant aspects of the NOIZEUS database. Various recording systems are used to record isolated digits, sentences, paragraphs, and answers to questions.

The audio samples of the NOIZEUS database recorded in the soundproof room are also considered to evaluate the performance of the proposed method. All segments of the audio are labelled correctly without any error. Although, the language of the NOIZEUS database is English, the performance of the proposed method remains high. This clearly indicates that the proposed system works equally well for both clean and noisy environment.

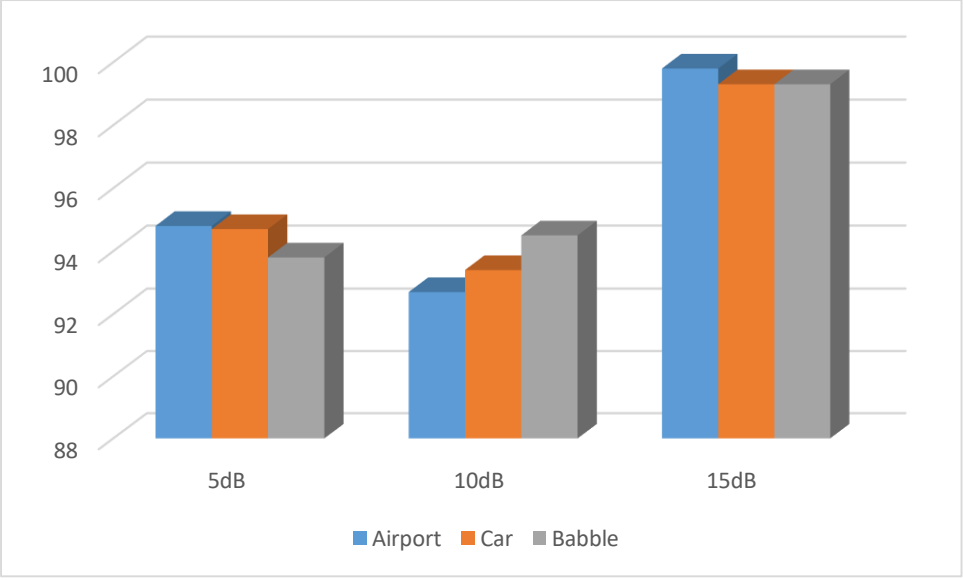
While the performance of the proposed method is good for the clear audio, it may not be practical in most application scenarios. Therefore, investigating the robustness of the proposed method against noise is crucial. Robustness can be observed by adding noise of different SNRs in the audios. Then, artificially generated noisy audio can be used for segment detection. In this study, instead of using artificial noisy audio, speech samples of the NOIZEUS database recorded in a restaurant are used to check the robustness of the proposed method. The audios are recorded in the restaurant in the presence of significant noise. One of the reasons for using the NOIZEUS database is to record in different environments. The labelling of voiced and unvoiced segments of the noisy audio is also carried out with perfection. This outcome suggests that the presence of noise in an audio does not affect the performance of the proposed method. The

proposed system performs well and labels the voiced and unvoiced segment accurately for both clean and noisy audio.

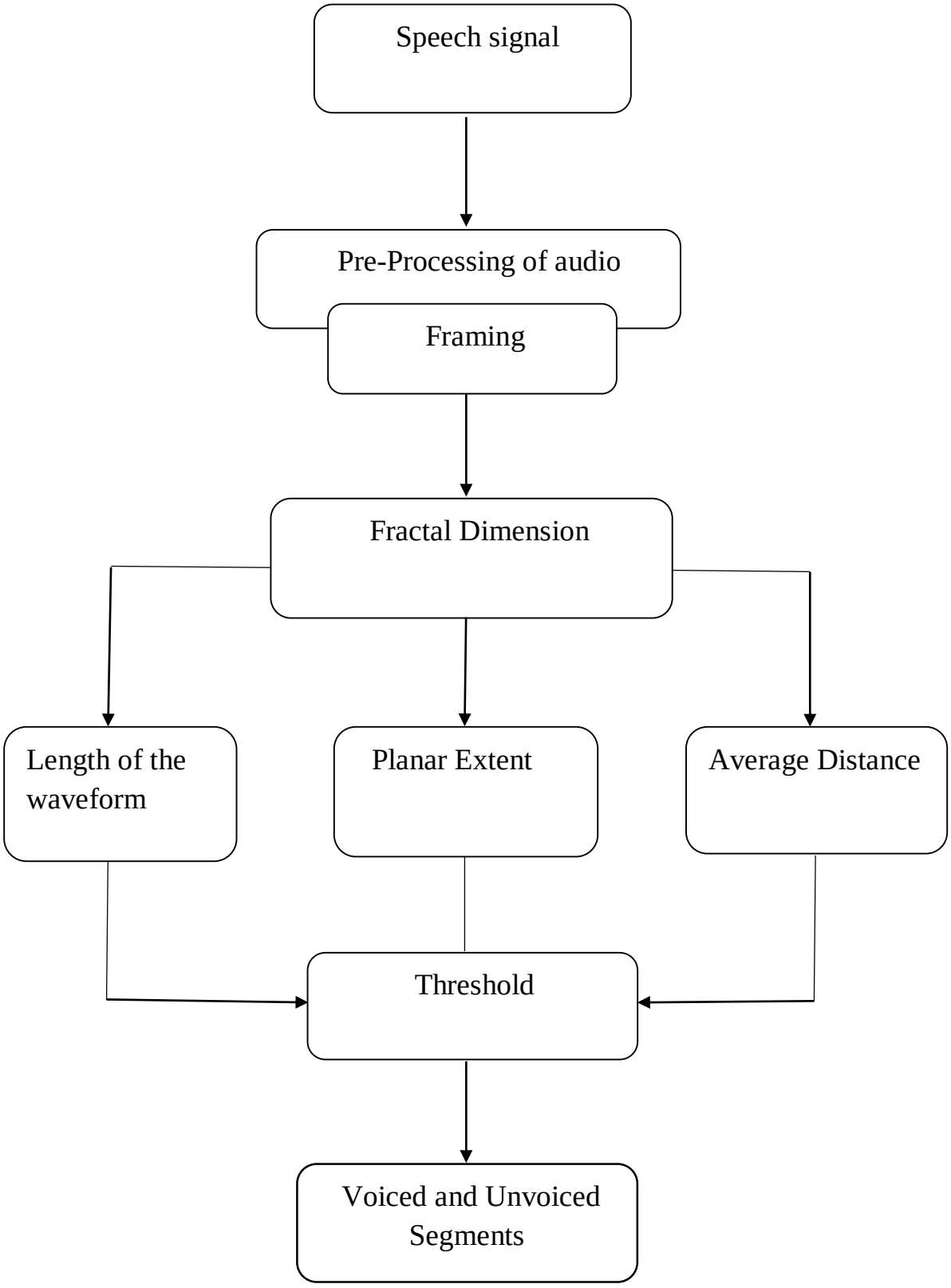
The method is evaluated by using SNR's of 5, 10 and 15 dB. The obtained maximum accuracy values are 94.7%, 94.5% and 99.8% for 5, 10 and 15 dB, respectively. Results show that the performance of the proposed method in the case of noisy audio also good.

Table 1. Accuracy for noisy audios in the case of NOIZEUS database

Noise	SNR		
	5 dB	10 dB	15 dB
Airport	94.8%	92.7%	99.8%
Car	94.7%	93.4%	99.3%
Babble	93.8%	94.5%	99.3%



4.2 DATA FLOW DIAGRAM:



CHAPTER 5

SYSTEM REQUIREMENTS

5.1 HAREWARE REQUIREMENTS

System : Intel Pentium CPU A1018@ 2.10GHz

Hard Disk : 320GB

Monitor : Generic PnP Monitor

RAM : 2GB

Input Device : Standard Keyboard and Mouse

Output Device : Monitor

5.2 SOFTWARE REQUIREMENTS

Operating system : Windows 10

Tool : MATLAB R2014a

Frond End : Mat lab

Backend : NOIZEUS

CHAPTER 6

IMPLEMENTATION AND RESULTS

6.1 SAMPLE CODING

Reading the wave file

```
[x1,fs] = audioread ( E:\ Wavfile);
```

Applying FFT

```
x = fft (x1);
x = abs(x);
```

Plotting the signal

```
Plot ( x );
```

Down Sampling

```
Fs = Fs/2;
```

Frame Duration

```
F_d = 0.01;
```

No of Samples in a frame

```
N_s = round( Fs* F_d);
```

No of Samples in an audio

```
L_s = length(x);
```

No of frames in an audio

```
N_f = floor( L_s / N_s);
```

Framing

```
temp = 0;
for i=1:N_f;
        Frames( i , : ) = x (temp+1: temp+N_s);
         temp = temp+N_s;
end
```

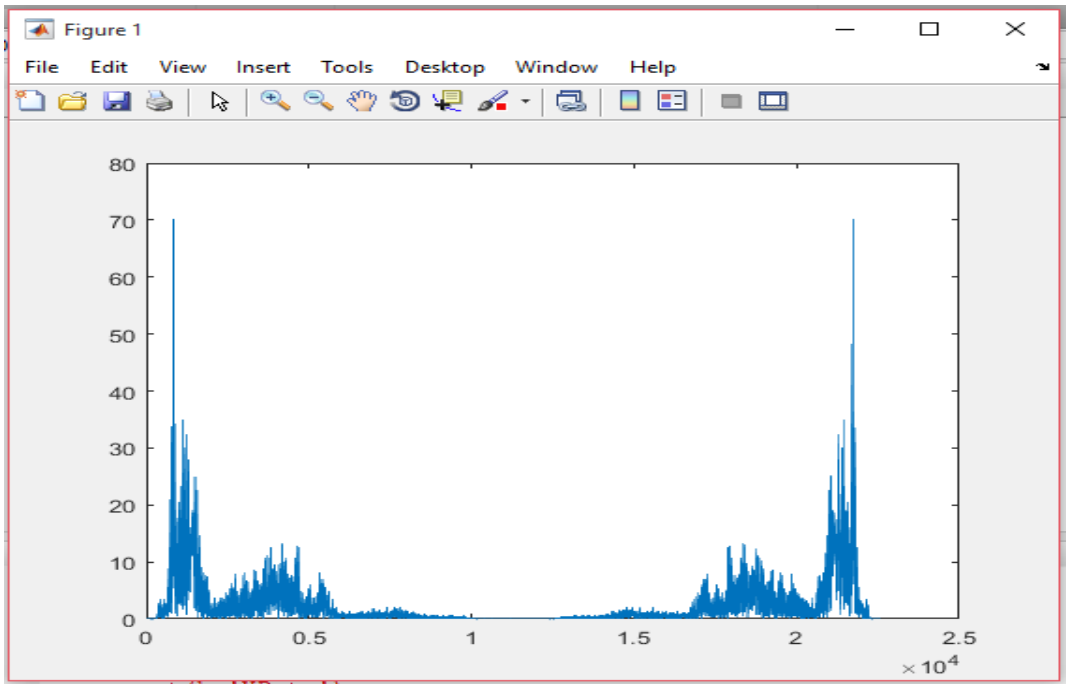
Fractal dimension Estimation

```
        Length of the Waveform
        F= sqrt(((x(j+1)-x(j)^2)+((y(j+1)-y(j))^2));
        M=sum(F);
        Planar Extent
        E= sqrt(((x(j+1)-x(1))^2)+((y(j+1)-y(1))^2));
        P= max(E);
        Average Distance
        V= mean(M);
        Fractal Dimension
        T1=log10(M/V);
        T2=log10(P/V);
        T=T1/T2;
Threshold
th=mean(sort(fractal(1:ind));
```

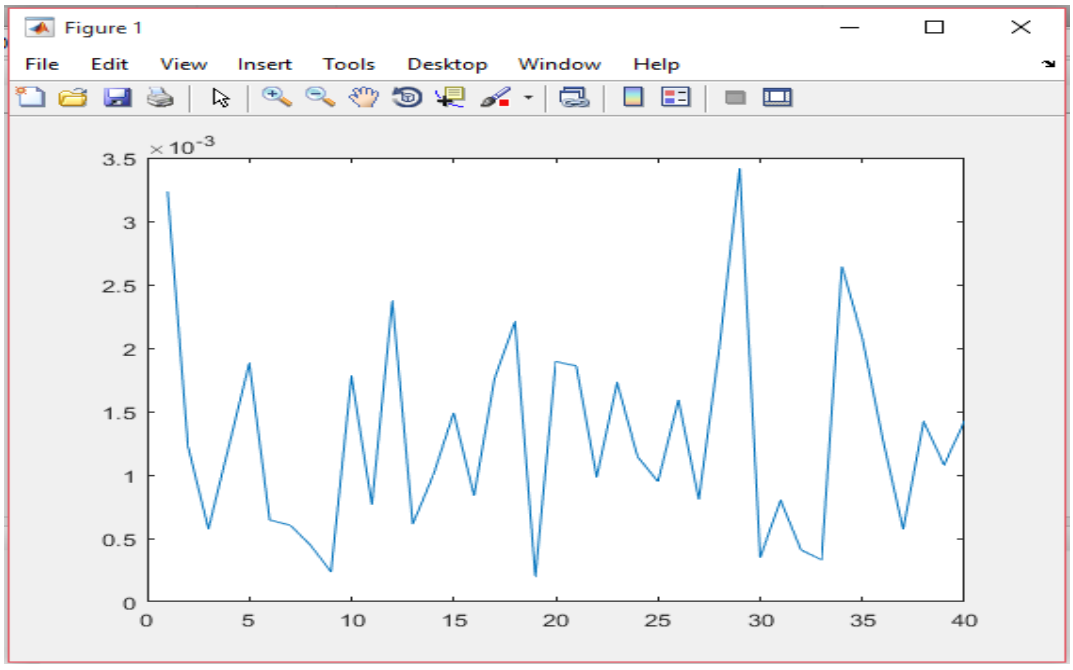
6.2 SCREEN SHOT

6.2.1 Clean signal in NOIZEUS database

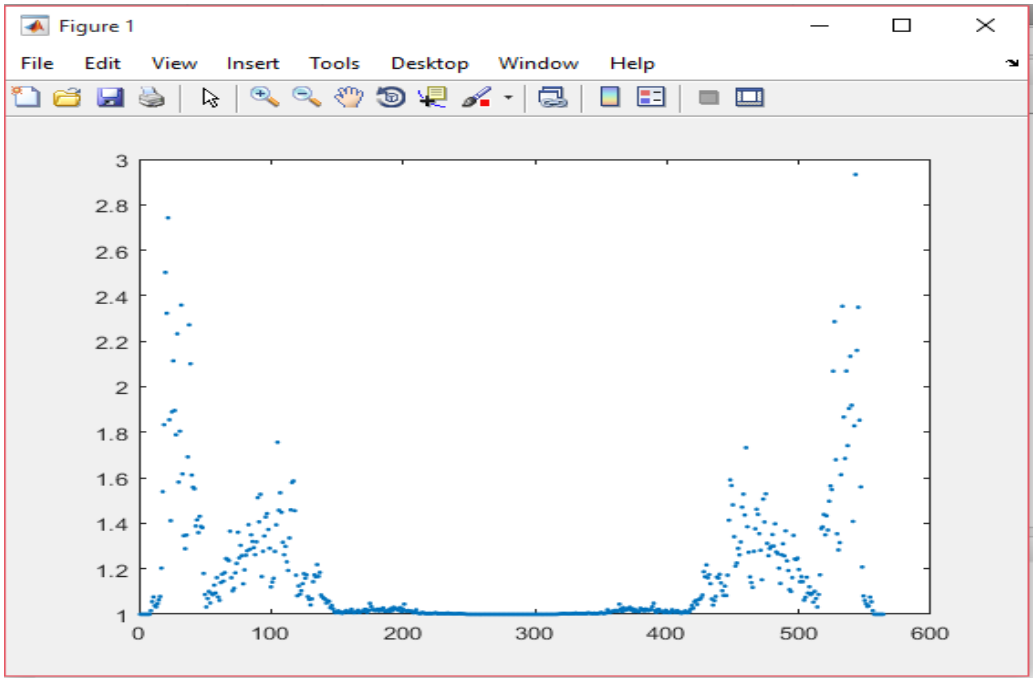
Input signal



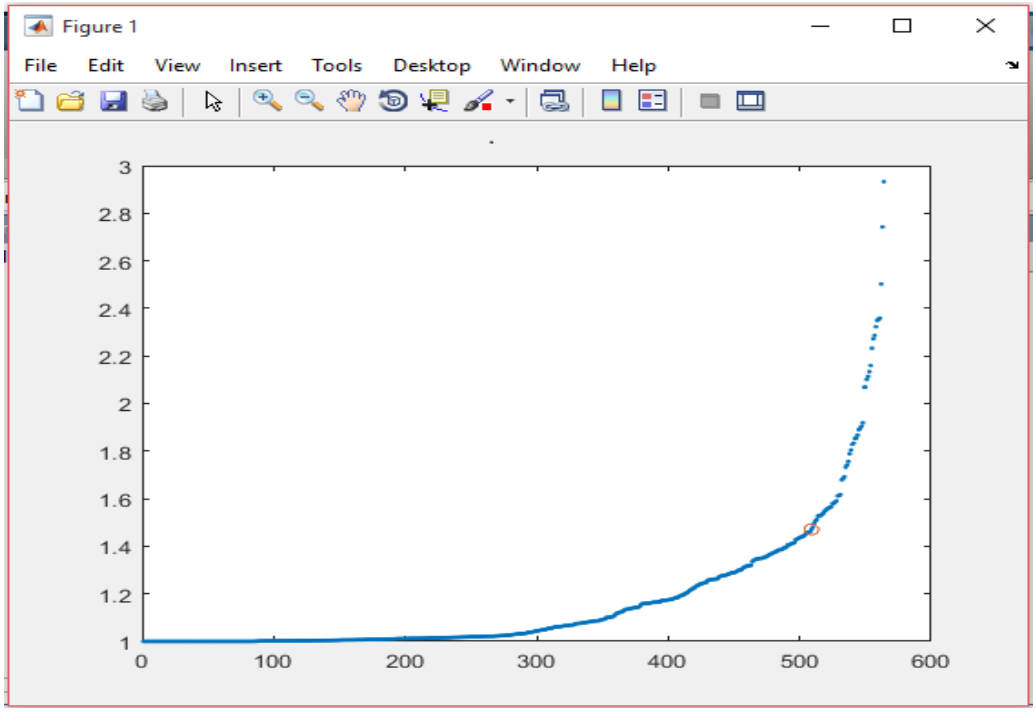
After Framing



Fractal dimension of all frames in an audio

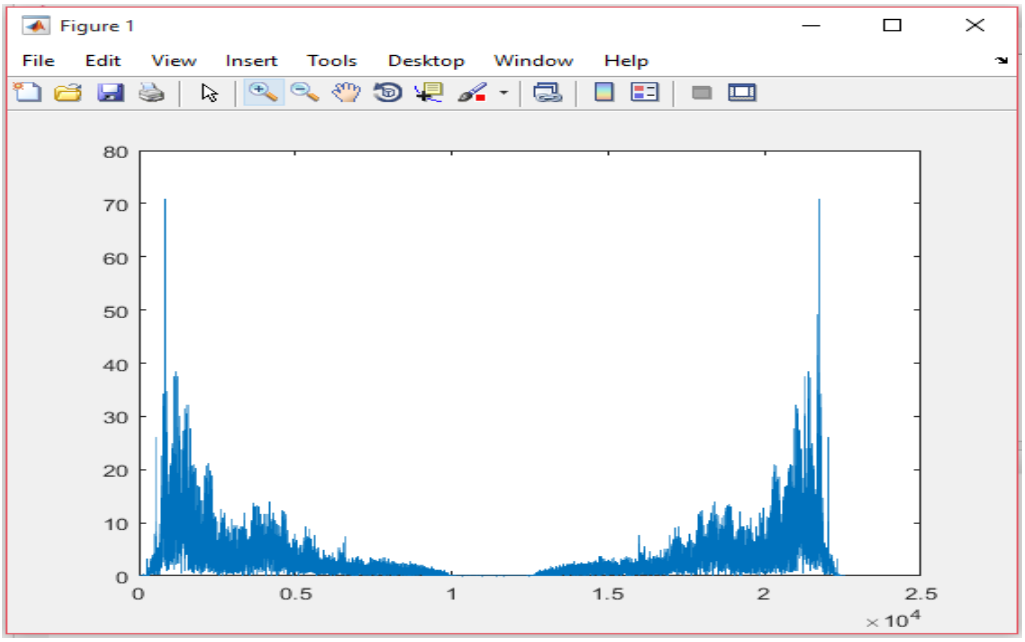


Segmentation of voiced and unvoiced segments (Threshold)

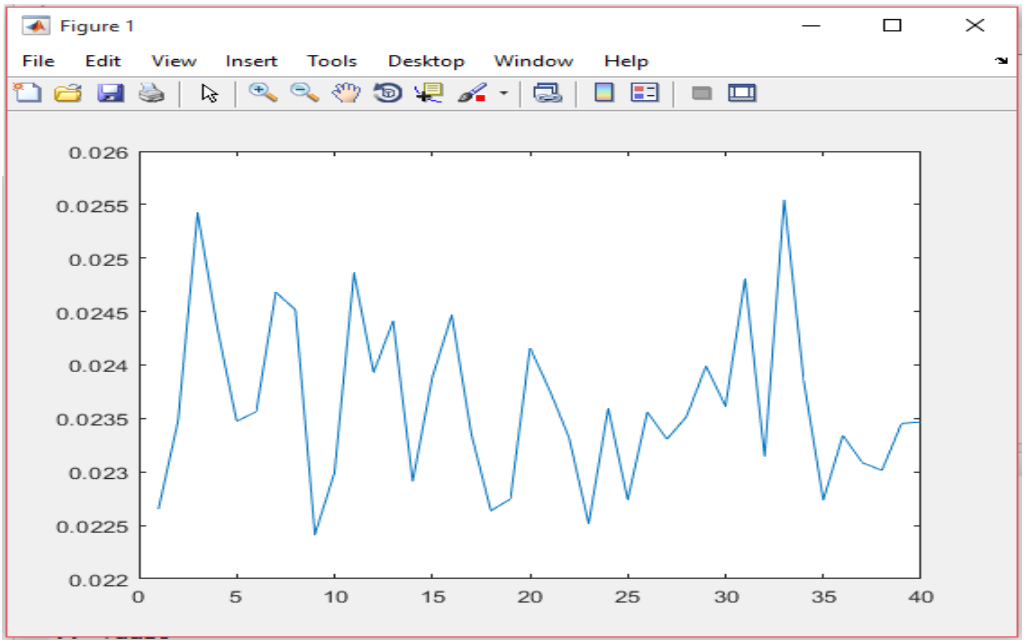


6.2.2 Noisy signal in NOIZEUS database

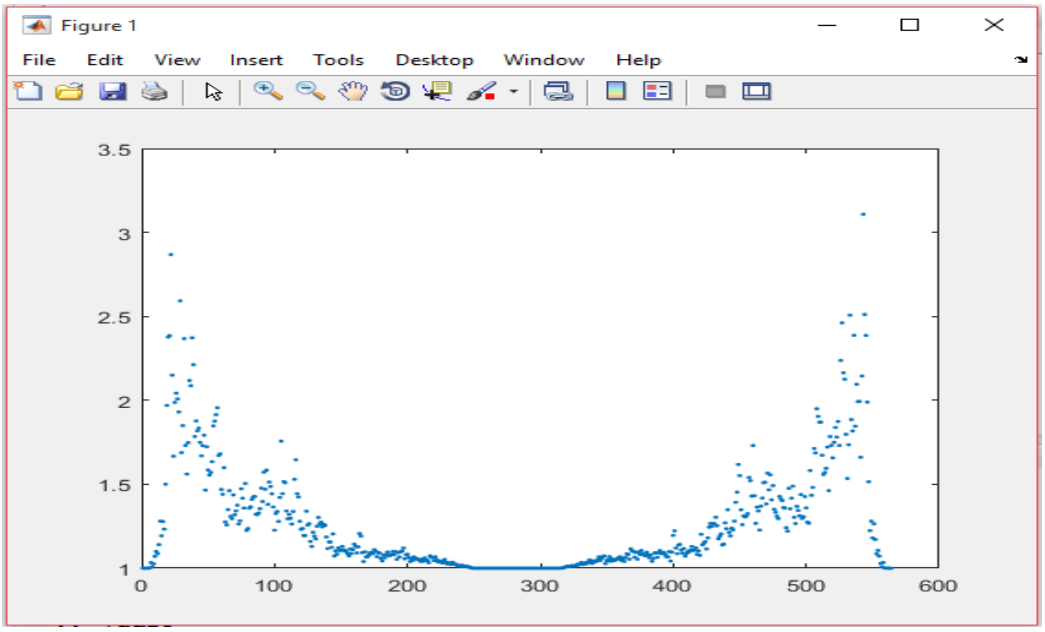
Input (Car_5db)



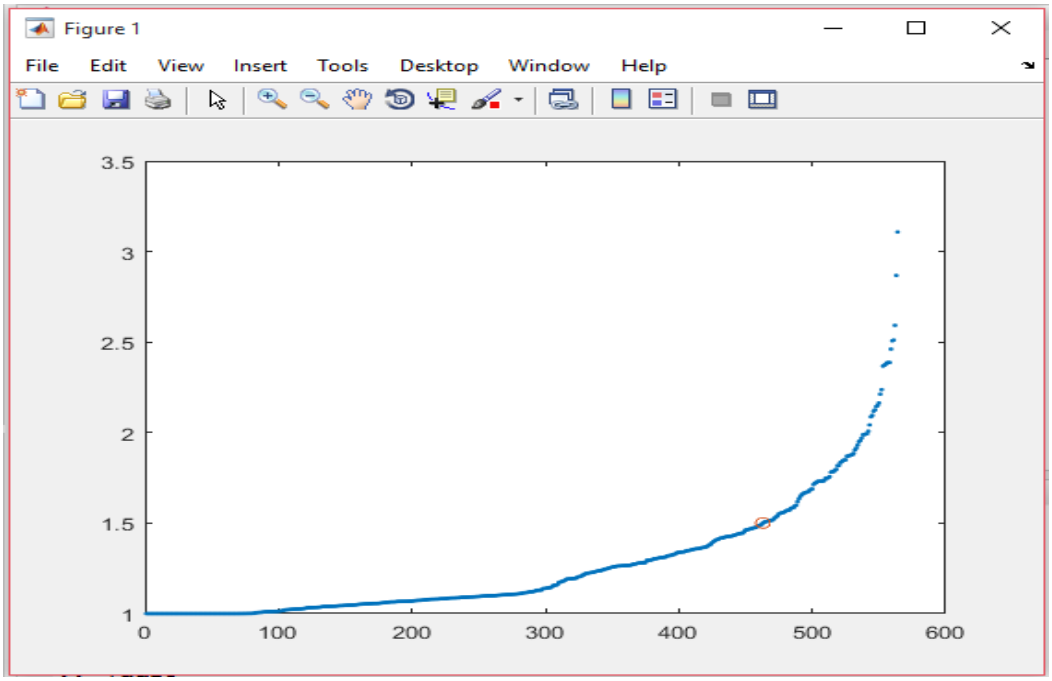
After Framing



Fractal Dimension of all frames in an audio



Segmentation of voiced and unvoiced segments (Threshold)



Confusion matrix

CHAPTER 7

CONCLUSION AND FUTURE ENHANCEMENT

7.1 CONCLUSION

A method to detect the speech-presence and speech-absence segments of an audio is presented. The proposed method is unsupervised and does not need any training data to differentiate between voiced and unvoiced segments, a feature that is a positive aspect of the method. The Noisy Speech Corpus database is used to evaluate the performance of the proposed method. The performance evaluation of the method suggests that it labels the segments accurately even for noisy environment. The presence of significant noise in an audio does not affect the performance of the proposed VAD method. This method is tested for NOIZEUS database.

7.2 FUTURE ENHANCEMENT

In Future, the method can be implemented in various applications related to continuous speech recognition. This method can be used reliably to automatically generate forged audio where the detection of boundary points is very crucial and vital.

APPENDIX

MATLAB R2014a

MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment and proprietary programming language developed by Math Works. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, C#, Java, Fortran and Python.

Although MATLAB is intended primarily for numerical computing, an optional toolbox uses the MuPAD symbolic engine, allowing access to symbolic computing abilities. An additional package, Simulink, adds graphical multi-domain simulation and model-based design for dynamic and embedded systems.

Syntax

The MATLAB application is built around the MATLAB scripting language. Common usage of the MATLAB application involves using the Command Window as an interactive mathematical shell or executing text files containing MATLAB code.

Variables

Variables are defined using the assignment operator. MATLAB is a weakly typed programming language because types are implicitly converted. It is an inferred typed language because variables can be assigned without declaring their type, except if they are to be treated as symbolic objects, and that their type can change. Values can come from constants, from computation involving values of other variables, or from the output of a function.

Matrices

A square identity matrix of size n can be generated using the function eye, and matrices of any size with zeros or ones can be generated with the functions zeros and ones.

Transposing a vector or a matrix is done either by the function transpose or by adding prime after a dot to the matrix. Without the dot MATLAB will perform conjugate transpose.

Structures

MATLAB has structure data types. Since all variables in MATLAB are arrays, a more adequate name is "structure array", where each element of the array has the same field names. In addition, MATLAB supports dynamic field names (field look-ups by name, field manipulations, etc.). Unfortunately, MATLAB JIT does not support MATLAB structures, therefore just a simple bundling of various variables into a structure will come at a cost.

Functions

When creating a MATLAB function, the name of the file should match the name of the first function in the file. Valid function names begin with an alphabetic character, and can contain letters, numbers, or underscores. Functions are often case sensitive.

Function handles

MATLAB supports elements of lambda calculus by introducing function handles, or function references, which are implemented either in .m files or anonymous/nested functions.

Classes and Object Oriented Programming Language

MATLAB supports object-oriented programming including classes, inheritance, virtual dispatch, packages, pass-by-value semantics, and pass-by-reference semantics. However, the syntax and calling conventions are significantly different from other languages. MATLAB has value classes and reference classes, depending on whether the class has handle as a super-class (for reference classes) or not (for value classes).

NOIZEUS

A noisy speech corpus (NOIZEUS) was developed to facilitate comparison of speech enhancement algorithms among research groups. The noisy database contains 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different real-world noises at different SNRs. The noise was taken from the AURORA database and includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise. This corpus is available to researchers free of charge.

Noise is artificially added to the speech signal as follows. The IRS filter is independently applied to the clean and noise signals. The active speech level of the filtered clean speech signal is first determined using the method B of ITU-T. A noise segment of the same length as the speech signal is randomly cut out of the noise recordings, appropriately scaled to reach the desired SNR level and finally added to the filtered clean speech signal.

Noise signals were taken from the AURORA database [4] and included the following recordings from different places:

- Babble (crowd of people)
- Car
- Exhibition hall
- Restaurant

- Street
- Airport
- Train station

The long-term spectra of the above noises are given in. The noise signals were added to the speech signals at SNRs of 0dB, 5dB, 10dB, and 15dB.

REFERENCES

[1] I. Hwang and J. H. Chang, "Voice Activity Detection Based on Statistical Model Employing Deep Neural Network," in 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2014, pp. 582-585.

[2] M. Imran, Z. Ali, S. T. Bakhsh, and S. Akram, "Blind Detection of Copy-Move Forgery in Digital Audio Forensics," IEEE Access, vol. 5, pp. 12843-12855, 2017.

[3] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux," IEEE Signal Processing Letters, vol. 20, pp. 197-200, 2013.

[4] D. Cournapeau and K. Tatsuya, "Using variational bayes free energy for unsupervised voice activity detection," in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 4429-4432.

[5] M. J. Katz, "Fractals and the analysis of waveforms," Computers in Biology and Medicine, vol. 18, pp. 145-156, 1988.

[6] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," Physica D: Nonlinear Phenomena, vol. 31, pp. 277-283, 1988.

[7] A. Petrosian, "Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns," in Computer-Based Medical Systems, 1995., Proceedings of the Eighth IEEE Symposium on, 1995, pp. 212-217.

[8] P. Maragos, "Fractal aspects of speech signals: dimension and interpolation," in Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on, 1991, pp. 417-420 vol.1.

[9] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal Detection of Changepoints With a Linear Computational Cost," Journal of the American Statistical Association, vol. 107, pp. 1590-1598, 2012/12/01 2012.

[10] M. Lavielle, "Using penalized contrasts for the change-point problem," Signal Processing, vol. 85, pp. 1501-1510, 2005/08/01/ 2005.
[11] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, N. Dahlgren, et al., "TIMIT acoustic-phonetic continuous speech corpus," ed: The linguistic data consortium catalog. Philadelphia: The Linguistic Data Consortium. ISBN:1-58563-019-5., 1993.
[12] Y. Liang, X. Liu, Y. Lou, and B. Shan, "An improved noise-robust voice activity detector based on hidden semi-Markov models," Pattern Recognition Letters, vol. 32, pp. 1044-1053, 2011/05/01/ 2011.
[13] T.-J. Park and J.-H. Chang, "Dempster-Shafer theory for enhanced statistical model-based voice activity detection," Computer Speech & Language, 2017.
[14] G. Friedland, O. Vinyals, H. Yan, and C. Muller, "Prosodic and other Long-Term Features for Speaker Diarization," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, pp. 985-993, 2009.
[15] R. J. Moran, R. B. Reilly, P. De Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," Biomedical Engineering, IEEE Transactions on, vol. 53, pp. 468-477, 2006.
[16] Y. Wang and L. Lee, "Supervised Detection and Unsupervised Discovery of Pronunciation Error Patterns for Computer-Assisted Language Learning," Audio, Speech, and Language Processing, IEEE/ACM Transactions on, vol. PP, pp. 1-1, 2015.
[17] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," Computer Speech & Language, vol. 14, pp. 333-353, 2000/10/01/ 2000.
[18] M. Alhussein, Z. Ali, M. Imran, and W. Abdul, "Automatic Gender Detection Based on Characteristics of Vocal Folds for Mobile Healthcare System," Mobile Information Systems, vol. 2016, p. 12, 2016.

[19] B. Xulei, Z. Jie, and C. Ning, "A robust voice activity detection method based on speech enhancement," in IET Intelligent Signal Processing Conference 2013 (ISP 2013), 2013, pp.1-4.

[20] M. Alsulaiman, Z. Ali, G. Muhammed, M. Bencherif, and A. Mahmood, "KSU Speech Database: Text Selection, Recording and Verification," in 2013, pp. 237-242.

[21] M. M. Alsulaiman, G. Muhammad, M. A. Bencherif, A. Mahmood, and Z. Ali, "KSU Rich Arabic Speech Database," Information, vol. 16, pp. 4231 -4253, 2013.

[22] M. Alsulaiman, G. Muhammad, B. Abdelkeder, A. Mahmood, and Z.Ali, "King Saud University Arabic Speech Database," ed: The linguistic data consortium catalog LDC2014S02. Philadelphia: The Linguistic Data Consortium. ISBN:1-58563-669-X.,2014.

[23] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice Activity Detection using harmonic frequency components in likelihood ratio test," in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 4466-4469.

[24] Q. H. Jo, J. H. chang, J. W. Shin, and N.S. Kim, "Statistical model-based voiced activity Detection using Support Vector Machine," IET Signal Processing, vol. 3, pp. 205-210,2009.