Speech Based Gender Identification Using Empirical Mode Decomposition (EMD)

Mahpara Hyder Chowdhury

Thesis Supervisor

Dr. Hanif Bin Azhar

DEPARTMENT OF COMPUTER SCIENCES & ENGINNERING

SCHOOL OF COMPUTER SCIENCES & ENGINNERING

BRAC UNIVERSITY

MOHAKHALI, DHAKA 1212

BANGLADESH

Speech Based Gender Identification Using Empirical Mode Decomposition (EMD)

A Thesis submitted in partial fulfillment of the requirement for the degree requirement for the degree of Bachelor of Science in Computer Science and Engineering of BRAC University

By

Mahpara Hyder Chowdhury (10101036)

Supervisor

Dr. Hanif Bin Azhar

Co-supervisor

Md. Zahangir Alom

April 2014

DECLARATION

This is to certify that the research work entitled "Speech Based Gender Identification Using Empirical Mode Decomposition (EMD)" is submitted by Mahpara Hyder Chowdhury (10101036) to the Department of Computer Sciences & Engineering in partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering. The content of this thesis have not been submitted elsewhere for the award of any degree or any other publication. I hereby declare that this thesis is my original work based on the results I found. The materials of work found by other researchers and the sources are properly acknowledge and mentioned by reference. I carried out my work under the supervision of Hanif Bin Azhar.

Dated: April 30, 2014

Signature of the Supervisor

Signature of the Author

Dr. Hanif Bin Azhar

Assistant Professor

Department of Computer Sciences & Engineering

BRAC University

Mahpara Hyder Chowdhury(10101036)

BRAC UNIVERSITY

FINAL READING APPROVAL

Thesis Title: Speech Based Gender Identification Using Empirical Mode Decomposition (EMD)

Date of Submission: April 30, 2014

The final form of the thesis report is read and approved by Dr. Hanif Bin Azhar. Its format, citations, and bibliographic style are consistent and acceptable. Its illustrative materials including figures, tables, and charts are in place. The final manuscript is satisfactory and is ready for submission to the Department of Computer Sciences & Engineering, School of Engineering and Computer Science, BRAC University.

Supervisor

_____

Dr. Hanif Bin Azhar

Assistant Professor

Department of Computer Sciences & Engineering

BRAC University

ACKNOWLEDGEMENT

I take this opportunity to express my deep gratitude and profound regards to my guide and the thesis supervisor Dr. Hnif Bin Azhar, Assistant Professor, Department of Computer Sciences and Engineering, for his exemplary guidance, idea, monitoring and constant encouragement throughout the course of this work. Without his supervision and constant help this dissertation would not have been possible. He was abundantly helpful and offered invaluable assistance, support and guidance to complete my work on time.

I also take this chance to convey a deep sense of gratitude to my co-supervisor Md. Zahangir Alom, Lecturer-III, Department of Computer Sciences and Engineering, for helping me to complete my thesis work. His cordial support, valuable information and guidance, helped me in completing this task through various stages.

I am grateful for their cooperation during the period of my thesis work. Without helps of the particular that mentioned above, I would face many difficulties while completing the work.

Abstract

Traditionally for feature extraction, decomposition techniques such as Fourier decomposition are used to capture signals. But those methods have some margins like – it only works for linear and stationary data. On the other hand, in real world, we found data that are in non-linear and non-stationary. EMD or Empirical Mode Decomposition technique is a new approach introduced by Huang et al (1998) that can take any complicated signal and decomposed it to IMF. It extracts the amplitude and frequency information of a signal at a particular time. It is robust for non-linear and non-stationary signal processing. In this paper, I am using EMD as a new approach for gender identification based on speech signal. Gender identification based on the voice of a speaker consists of detecting if a speech signal is given by a male or female. Detecting the gender of a speaker has several applications.

CONTENT

Chapter 1: Introduction

In real world signals are nonlinear and non-stationary, for example signals from our stomach, heart, hand and mostly our voice. In case of signal analysis, previously, the conventional method of FFT is used. Even though FFT is valid under tremendously general circumstances, there are certain critical limitations of the Fourier spectral analysis, system must be linear and the data must be strictly periodic or stationary or else the resulting spectrum will never make more than slight physical sense.

On the contrary, Empirical Mode Decomposition is a time frequency analysis method that is capable of extracting amplitude and frequency information of a signal at a given time incident. From Fourier analysis, one can only detect the frequency contents and their respective amplitude, no information is available as to when the frequencies occurred. This is due to very nature of Fourier analysis in that the information is averaged over time. Unlike Fourier analysis where time resolution is completely lost, EMD method reserves time information [10].

Gender identification based on the voice of a speaker consists of detecting if a speech signal is given by a male or female. Detecting the gender of a speaker has several applications. In the context of speaker recognition, gender detection can improve the performance by limiting the search space to speakers from the same gender.

However, the proposed method which is EMD has been applied on speech based gender identification. This paper describes a novel approach of gender detection exclusively by using EMD.

1.1 Objective:

Speech signal is typical nonlinear and non-stationary data. Speech processing is a diverse field with many applications. In this thesis, enlightened by the successful applications, I try to test the EMD algorithm in speech identification. Here, I must emphasize, that, in this thesis the purpose of the gender based speech identification system is just for testing, analysis and comparison of the new method, not for the performance of the speech identification system itself. In this thesis, I select data from the database recorded internally at Carnegie Mellon University circa in 1991 [1] and all the development work are performed in MATLAB environment. In speech identification, speech can be constrained to be known phrase (text-independent) or totally unconstrained (text-dependent). Text-independent means that the identification procedure should work for any text in either training or testing. In this thesis, attention is focused to the text-independent speech identification system is developed to test the performance of the new method.

1.2 Contribution:

Empirical mode decomposition is a new approach hence not widely used. I chose to work with this method and have analyzed the results in different ways and trained them and the output was compared with standard features.

In the paper entitled "**Gender identification using audio classifier"** by Hadi Hard, Liming Chen, 2002, they used MFSC or Mel Frequency Spectral Coefficients, which is a standard feature. They selected the speech duration of 160 seconds to 2200 seconds. The rate of error decreases from 22% to 11% while time duration increases from 160 to 2200 [9].

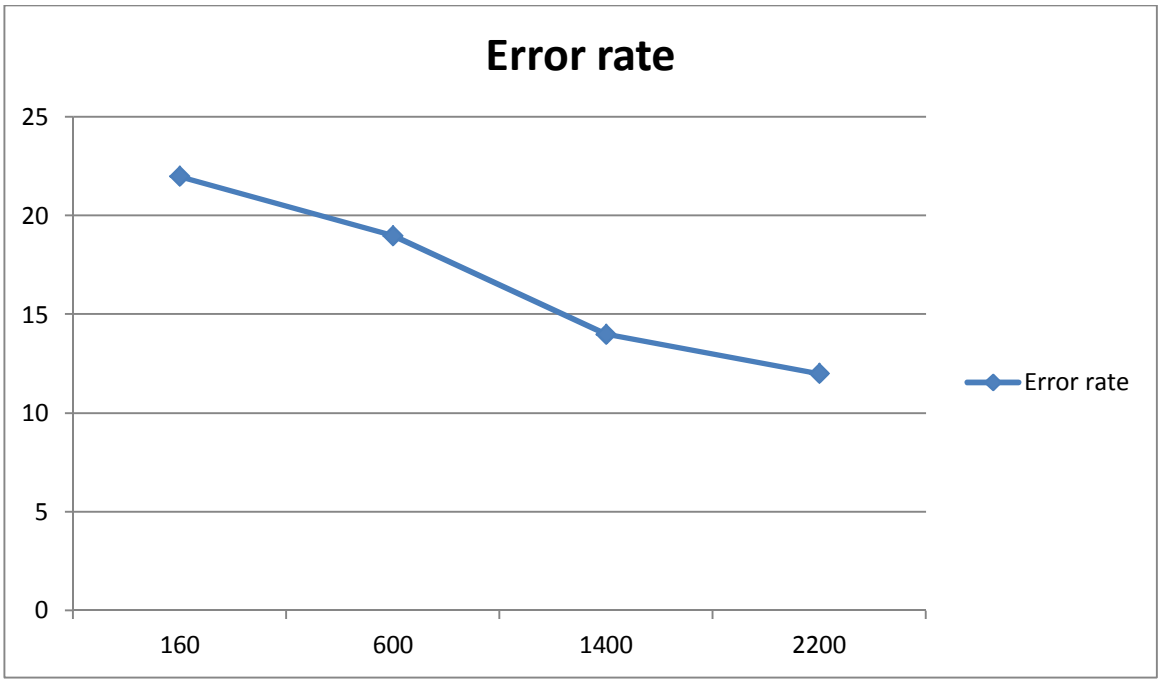


Figure 1.1: The error rate as a function of the duration of the training data [9].

But I have used EMD, which is a very new approach and my sample speech's time duration was only 3 seconds. For this my error rate was 21.62%. In comparison to the paper my error rate was much less as I have used totally new approach.

The dataset I took was text independent and the speech consists of alphabet, numbers and few single words. My result arose from the normal and gentle voice of the speech. Anyone can understand by reading my paper that how EMD works on these types of voices.

All the speech samples that I used has time interval of 3 seconds, it showed, very clearly, the performance of EMD on this type of data. None of the paper that I reviewed has this type of less time durational data, even they used standard features, which are highly recognized for speech processing, with data that has more than 10 second time interval.

To the best of my knowledge, I am the first to apply Empirical Mode Decomposition (EMD) approach to identify gender based on their speech signals. I hope that this paper will inspire more research on EMD approaches applied to speech recognition tasks.

1.3 Speech

1.3.1 What is Speech?

Individuals express their opinions, feelings, moods and ideas verbally to one another over a chain of complex actions that change and mold the elementary tone created by voice into specific, decodable sounds. Speech is actually are produced by exactly coordinated muscle actions in the head, neck, chest, and abdomen regions of the body. speech is an immensely information-rich signal exploiting frequency-modulated, amplitude-modulated and time-modulated carriers (e.g. resonance movements, harmonics and noise, pitch intonation, power, duration) to convey information about words, speaker identity, accent, expression, style of speech, emotion and the state of health of the speaker.

1.3.2 Speech Signal Processing:

The formal tools of signal processing emerged in the mid-20^{th} century when electronics gave us the ability to manipulate signals- time-varying measurements – to extract or rearrange various aspects of interest to us i.e. the information in the signal [2].

Major advances in speech signal representations have included perceptually motivated Mel-Frequency Cepstral Coefficients (MFCC) (Krishnamurthy and Childers, 1986), Perceptual Linear Prediction (PLP) coefficients (Hermansky, 1990) as well as Cepstral Mean Subtraction (CMS) (Rosenberg et al., 1994;

Furui,2001), RASTA (Hermansky and Morgan, 1994), and Vocal Tract Normalization (VTLN) (Eide, 1996) [3].

1.3.3 Speech Based Gender identification:

Gender identification based on the voice of a speaker comprises mainly of detecting if a speech signal is given by a male or a female. Automatically detecting the gender of a speaker has many potential applications. Within the context of automatic speech recognition, gender dependent models are more exact than gender independent ones. Therefore gender recognition is required before the application of one gender dependent model.

In the context of speaker recognition, gender detection will improve the performance by limiting the search area to speakers from the same gender. Also, in the context of content multimedia indexing the speaker's gender could be a cue used the annotation. Therefore, automatic gender detection is often a tool in a very content primarily based multimedia indexing system.

1.4 Thesis Layout

The structure of the thesis is as follows: Chapter 2 contains background information of non-linear data, decomposition, why we need to do decomposition, former approach, empirical mode decomposition, EMD algorithm, MFCC, CML. Chapter 3 introduces literature review of some successful applications regarding EMD, speech recognition system, and speech based gender classification. Also what challenges they found and why I am using EMD. In chapter 4 general discussion on proposed approach are described. Chapter 5 tells us about the results and analysis including dataset description, experimental setup1, experimental setup2, experimental setup3 and also discussion about what I observed and possible reasons. Finally in Chapter 5, I concluded with the future works.

Chapter 2: Background Theory

2.1 Non-linear Data:

In nonlinear data structures, data elements are not ordered in a successive manner. A data item in a nonlinear data structure could be attached to several other data elements to reflect a special relationship among them and all the data items cannot be traversed in a single run. Data structures like multidimensional arrays, trees and graphs are some examples of widely used nonlinear data structures. A multidimensional array is simply a collection of one-dimensional arrays. A tree is a data structure that is made up of a set of linked nodes, which can be used to represent a hierarchical relationship among data elements. A graph is a data structure that is made up of a finite set of edges and vertices. Edges represent connections or relationships among vertices that stores data elements [5].

2.2 Decomposition:

The term "decomposition" formally means the breaking down of a compound process or a composite material into separate constituent components. But in many areas related to analysis of different processes, signal analysis, analysis of various sorts of sequences, etc., this term has long been used in a broader meaning very often suggesting not a breakdown into actual initial components but rather a breakdown into certain functions that were not actually present when the initial data was being formed. These functions are sort of artificially formed in the process of data decomposition but despite their "artificial" origin they allow for a deeper analysis of data helping to identify hidden patterns. [4]

2.3 Why we need to do decomposition:

All real processes we've to work with in run-through are complicated, as a rule, comprising of a good variety of components. For instance, weather, once analyzing precipitation charts, we should always bear in mind that they represent interaction between plenty of assorted processes like seasonal changes, universal warming/cooling processes, current changes, dynamics of cyclones and anticyclones, the quantity of carbon dioxide, carbonic acid gas emitted into the atmosphere, solar activity cycles, etc. The list may persist forever.

A chart of this sort is thus quite tough to be analyzed as its components, once interacting with one another, mask and deform the regularities we would prefer to establish. This gives rise to a rightful need to interrupt down the method into consideration into individual components and analyze every of it individually. Analysis of individual components and contemplation of the contribution they create into the method at hand helps us higher perceive the method in progress, as well as, e.g. increase the forecast reliableness [4].

And there is no exception once it involves varied data on mercantilism, together with currency quotes that are formed based on a great number of a variety of various factors. That is why it's quite natural to expect that a direct upfront breakdown into individual components will significantly facilitate their additional analysis.

2.4 Former Approaches:

Data analysis for pure research and practical applications, we face one or more following problems: a) the data are non-stationary; and b) the data represent non-

linear processes. Using those types of data, we have inadequate alternatives to use them in the study.

Historically, Fourier spectral analysis has provided a general method for examining the global energy-frequency distributions [12]. Somewhat due to its competency and because of its easiness, Fourier analysis has subjugated the data analysis works since quickly after its introduction, and has been applied to all types of data. Although the Fourier transform is valid under extremely general conditions (see, for example, Titchmarsh 1948), there are some crucial restrictions of the Fourier spectral analysis: the system must be linear; and the data must be strictly periodic or stationary; otherwise the resulting spectrum will make little physical sense [12]. So the consequence for non-stationary and nonlinear data is misleading energy-frequency distribution.

Another approach is Wavelet Transformation. It is also a widely accepted approach. But the negative side is, it is non-adaptive means same basic wavelet is used for all data [13].

But for real-world purposes, it would be well enough to have a transform that would not just let to deal with non-stationary processes but also use an adaptive transform origin determined by initial data. This type of transform does exist and will be concisely considered below thus addressing the main subject of this article.

2.5 Empirical Mode Decomposition (EMD):

The Empirical Mode Decomposition (EMD) was proposed as the fundamental part of the Hilbert–Huang transform (HHT). The Hilbert Huang transform is carried out, so to speak, in 2 stages. First, using the EMD algorithm, we obtain intrinsic mode functions (IMF).

Then, at the second stage, the instantaneous frequency spectrum of the initial sequence is obtained by applying the Hilbert transform to the results of the above step. The HHT allows obtaining the instantaneous frequency spectrum of nonlinear and non-stationary sequences. These sequences can consequently also be dealt with using the empirical mode decomposition.

However, I am not going to cover the plotting of the instantaneous frequency spectrum using the Hilbert transform. Here, I will focus only on the EMD algorithm.

EMD decomposes any given data into intrinsic mode functions (IMF) that are not set analytically and are instead determined by an analyzed sequence alone. The basic functions are in this case derived adaptively directly from input data.

An IMF resulting from the EMD shall satisfy only the following requirements:

1. The number of IMF extrema (the sum of the maxima and minima) and the number of zero-crossings must either be equal or differ at most by one;
2. At any point of an IMF the mean value of the envelope defined by the local maxima and the envelope defined by the local minima shall be zero [4].

2.6 EMD Algorithm:

The algorithm as proposed by Huang is based on producing smooth envelopes defined by local maxima and minima of a sequence and subsequent subtraction of the mean of these envelopes from the initial sequence. This requires the identification of all local extrema that are further connected by cubic spline lines to produce the upper and the lower envelopes.

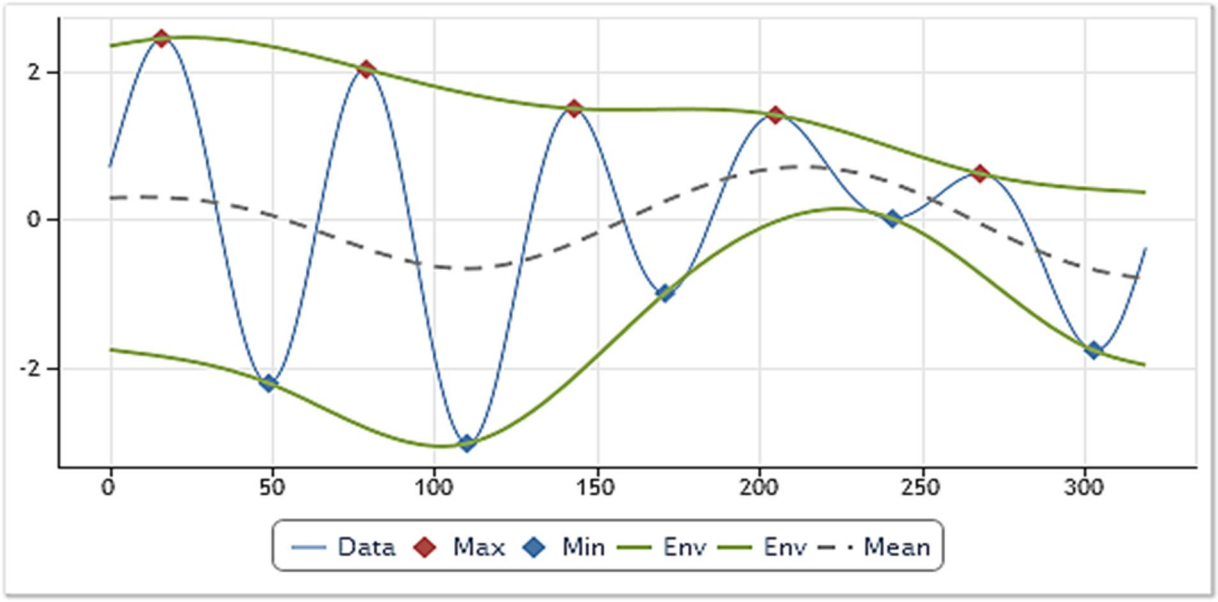


Figure 2.1:-The procedure of plotting the envelopes

The procedure of extracting an IMF is called sifting. The sifting process is as follows:

1. Identify all the local extrema in the test data.
2. Connect all the local maxima by a cubic spline line as the upper envelope.
3. Repeat the procedure for the local minima to produce the lower envelope.

The upper and lower envelopes should cover all the data between them. Their mean is m_1. The difference between the data and m_1 is the first component h_1:

$$X(t) - m_1 = h_1$$

Ideally, h_1 should satisfy the definition of an IMF, for the construction of h_1 described above should have made it symmetric and having all maxima positive and all minima negative. After the first round of sifting, a crest may become a local

maximum. New extrema generated in this way actually reveal the proper modes lost in the initial examination. In the subsequent sifting process, h_1 can only be treated as a proto-IMF. In the next step, it is treated as the data, then

$$h_1 - m_{11} = h_{11}$$

After repeated sifting up to k times, h_1 becomes an IMF, that is

$$h_{1(k-1)} - m_{1k} = h_{1k}$$

Then, it is designated as the first IMF component from the data:

$$c_1 = h_{1k}$$

The stoppage criterion determines the number of sifting steps to produce an IMF. Two different stoppage criteria have been used traditionally:

- 1. The first criterion is proposed by Huang et al. (1998). It similar to the Cauchy convergence test, and we define a sum of the difference, SD, as

$$SD_k = \frac{\sum_{t=0}^{T} |h_{k-1}(t) - h_k(t)|^2}{\sum_{t=0}^{T} h_{k-1}^2(t)}$$

Then the sifting process is stop when SD is smaller than a pre-given value.

- 2. A second criterion is based on the number called the S-number, which is defined as the number of consecutive siftings when the numbers of zero-crossings and extrema are equal or at most differing by one. Specifically, an S-number is pre-selected. The sifting process will stop only if for S consecutive times the numbers of zero-crossings and extrema stay the same, and are equal or at most differ by one [4].

Once a stoppage criterion is selected, the first IMF, c_1, can be obtained. Overall, c_1 should contain the finest scale or the shortest period component of the signal. We can, then, separate c_1 from the rest of the data by

$$X(t) - c_1 = r_1$$

Since the residue, r_1, still contains longer period variations in the data, it is treated as the new data and subjected to the same sifting process as described above.

This procedure can be repeated to all the subsequent r_1's, and the result is

$$r_{n-1} - c_n = r_1$$

The sifting process stops finally when the residuer$_n$, becomes a monotonic function from which no more IMF can be extracted. From the above equations, we can induce that

$$X(t) = \sum_{j=1}^{n} c_{j+r_n}$$

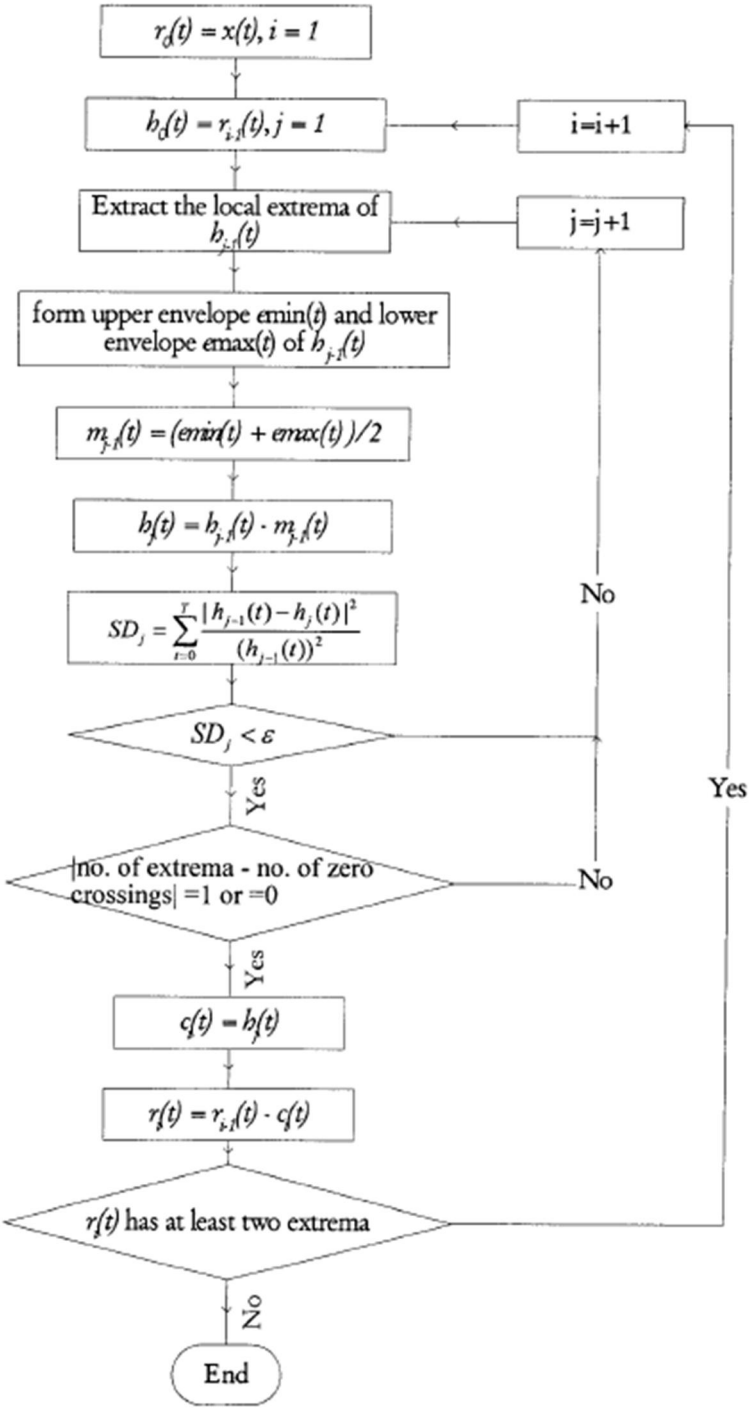


Figure 2.2: Flow chart of the EMD algorithm

Figure 2.3 shows another example featuring decomposition. It can be seen, that the decomposition of this sequence resulted in extraction of four IMFs and the residue

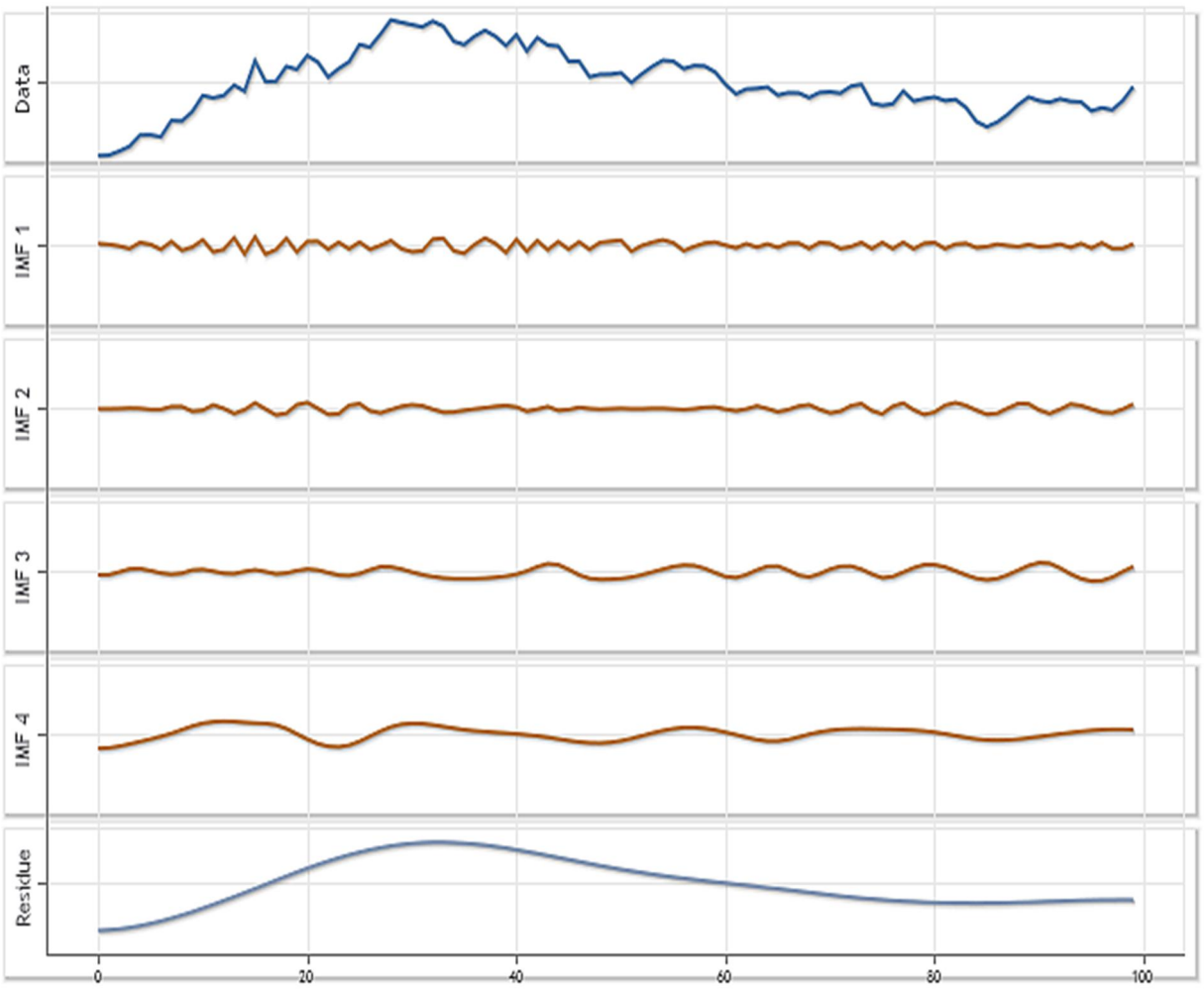


Figure 2.3: Example of EMD algorithm

2.7: MFCC (Mel-frequency cepstral coefficient):

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transformof a log power spectrum on a nonlinear mel scale of frequency.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

MFCCs are commonly derived as follows:

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

MFCCs are commonly used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken into a telephone. They are also common in speaker recognition, which is the task of recognizing people from their voices [4].

2.8: RASTA:

RASTA is a separate technique that applies a band-pass filter to the energy in each frequency sub band in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel e.g. from a telephone line.

2.9: CML(Coupled Mapped Lattice):

"A CML is a dynamical system in which there is some interaction ('coupled') between continuous state elements, which evolve in discrete time ('map') and are distributed on a discrete space ('lattice')" [14].

CMLs provide a novel approach to the study of spatiotemporal chaos, pattern formation, and nonlinear biological information processing.

In addition, because CMLs are adaptive to semi-macroscopic conditions and are numerically efficient, they are useful tools for simulating pattern formations, nonlinear waves and biological information processing.

Here, CML or Coupled Map Lattice is used for pattern formation.

The general formula is:-

$$NL(x,y) = L(x-1,y) + L(x+1,y) + L(x,y-1) + L(x,y+1) - 4 * L(x,y)$$

Where,

L = lattice

NL = new lattice after applying CML

m = row of the lattice

n = column

x and y are the co-ordinates

There are 8 boundary conditions of this formula,

I. If x == 1 & y == 1 then

$$NL(x,y) = L(m,y) + L(x+1,y) + L(x,n) + L(x,y+1) - 4 * L(x,y)$$

II. If x == 1 & y == n then

$$NL(x,y) = L(m,y) + L(x+1,y) + L(x,y-1) + L(x,1) - 4 * L(x,y)$$

III. If x == 1 & y>1 & y< n then

$$NL(x,y) = L(m,y) + L(x+1,y) + L(x,y-1) + L(x,y+1) - 4 * L(x,y)$$

IV. If x == m & y == 1 then

$$NL(x,y) = L(x-1,y) + L(1,y) + L(x,n) + L(x,y+1) - 4 * L(x,y)$$

V. If x == m &y == n then

$$NL(x,y) = L(x-1,y) + L(1,y) + L(x,y-1) + L(x,1) - 4 * L(x,y)$$

VI. If x == m & y> 1 & Y < n then

$$NL(x,y) = L(x-1,y) + L(1,y) + L(x,y-1) + L(x,y+1) - 4 * L(x,y)$$

VII. If y == 1 & x > 1 & x < m then

$$NL(x,y) = L(x-1,y) + L(x+1,y) + L(x,n) + L(x,y+1) - 4 * L(x,y)$$

VIII. If y == n & x > 1 & x < m then

$$NL(x,y) = L(x-1,y) + L(x+1,y) + L(x,y-1) + L(x,1) - 4 * L(x,y)$$

Chapter 3: Literature review

3.1 Applications using EMD for non-linear data

3.1.1: Artifact reduction in electrogastrogram based on empirical mode decomposition method by H.Liang, 2000- this paper is based on EMD method with which the contaminated EGG signals were reduced and hence giving the exact stomach signal, helping to find out the problems associated with stomach with an ease.

3.1.2: Hand Tremor detection via adaptive Empirical Mode Decomposition and Hilbert-Huang Transformation by James Z. Zhang, Robert D. Adams, Kenneth Burbank, 2009- In this research paper, the main work is done with using EMD and Hilbert-Huang Transform (HHT). The EMD is modified by adaptively changing its stopping criteria and therefore more accurately extract a tremor signal's amplitude, frequency and time information. These results are expected to be helpful for real-time tremor detection and suppression.

From Fourier analysis, one can only observe the frequency contents and their respective amplitude, no information is available as to when the frequencies occurred. This is due to very nature of Fourier analysis in that the information is

averaged over time. Unlike Fourier analysis where time resolution is completely lost, EMD method reserves time information of the tremors.

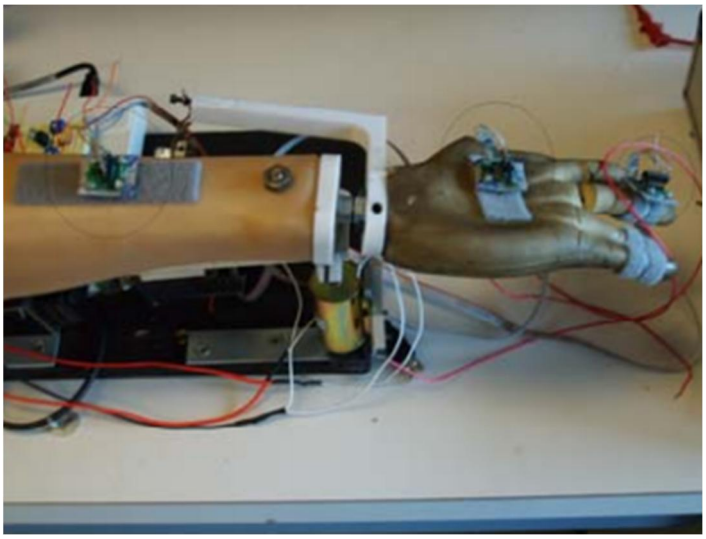


Figure 3.1: Human hand tremor stimulation test bed

3.2: Speech recognition

3.2.1: Speech processing using the empirical mode decomposition and the Hilbert transform by Ke Gong, 2004- this paper explores the full interpretation of the EMD and the Hilbert transform for complicated data. Associated properties of the marginal spectrum and various definitions of the degree of stationarity also are explored and some practical problems are pointed out. From this paper I have learned how to work with EMD. They developed a text independent speaker identification system to test the performance of the new method. This feature inspired me to work with text independent speech signals.

3.2.2: In Honglak Lee, Yan Largman, Peter Pham and Andrew's paper **Unsupervised feature learning for audio classification using convolutional deep belief networks**, they worked on speaker identification, speaker gender classification, phone classification, music genre classification, music artists

classifications. They used a “deep learning” approach named Convolutional deep belief networks (CBDNs) feature reorientations. The deep belief network is a generative probabilistic model composed of one visible (observed) layer and many hidden layers. Recently Convolutional deep belief networks have been developed to scale up the algorithm to high dimensional data. This high dimensional data applied to images, Lee et al. (2009) and showed good performance in several visual recognition tasks. For speaker identification, they took 10 random trails and averaged. Then measured the classification accuracy with the widely used standard feature MFCC then evaluate the features using standard supervised classifiers such as SVM, GDA and KNN.

3.2.3: In P. Dhanalakshmi, S. Palanivel, V. Ramalingam’s paper **Classification of audio signals using SVM and RBFNN,** they used they discriminated the six categories of audio namely music, sport, advertisement, cartoon and movie, a number of features such as LPC, LPCC, MFCC are extracted to characterize the audio content. They use SVM to train and compared to RBF network. They took audio signals up to 10s and the performance is-

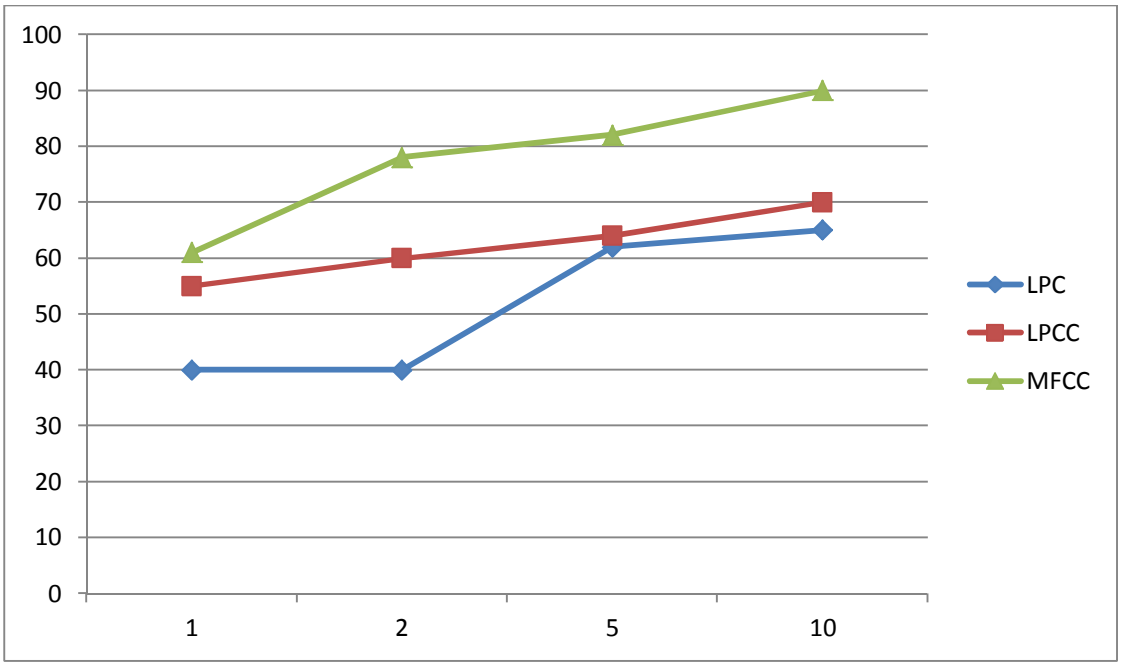


Figure 3.2: performance of SVM audio classification

SVM(%)	RBFNN(%)
92.1	93.7

Table 3.1: Audio classification performance using SVM and RBFNN

3.3: Speech based gender identification

3.3.1: In Honglak Lee, Yan Largman, Peter Pham and Andrew's paper **Unsupervised feature learning for audio classification using convolutional deep belief networks**, they also work on speaker gender classification. They reported the classification accuracy for various quantities of training examples per gender. They reported the test classification accuracy averaged over 20 trials. They

have showed, both the first and second CDBN features outperformed the base line features, especially when the numbers of training examples were small.

3.3.2: Gender identification using audio classifier by Hadi Hard, Liming Chen, 2002- I have studied this paper and it tells about how they use FFT with Hamming Window for feature extraction. Those windowing was non-overlapping. The mean and variance were calculated from those windows. Then merged those windows and made a new vector. Then the data were trained and tested for the result. I have taken most of my suggestion from the paper.

3.4: Challenges found in literature review:

I. In **Unsupervised feature learning for audio classification using convolutional deep belief networks**, they took 168 speakers and 10 sentences per speaker, total 1680 sentences. they faces several challenges-

- For speaker gender identification, they reported the test classification accuracy averaged over 20 trials. I belief it will be a challenge to reduce the trial number.
- For 180 speaker, when they took only one sentence per speaker and test the classification accuracy for speaker identification, they got poor result in CDBN layer two (62.8%).
- In case of music classification, they got 73.1% in CDBN Layer one and 69.2% in CDBN Layer two when their sampled was 3secs.

II. In **Classification of audio signals using SVM and RBFNN,** they optimized the audio signals recorded for 60s at 8000 samples per second. To minimize the time duration for the audio signal was a challenging job for them.

III. In the paper entitled "**Gender identification using general audio classifier"** by Hadi Hard, Liming Chen, 2002, they used MFSC or Mel Frequency Spectral Coefficients, which is a standard feature. They selected the speech duration of 160 seconds to 2200 seconds. The rate of error decreases from 22% to 11% while time duration increases from 160 to 2200. It remains a challenge for them to deduce the time duration of speeches.

IV. In **Hand Tremor detection via adaptive Empirical Mode Decomposition and Hilbert-Huang Transformation,** the tremor detection was one-dimensional. Further investigation and research is needed for simultaneous multi-dimensional detection of tremor.

3.5: Justification over using of EMD method

I chose to work with the new approach of EMD as because,

- Firstly if we want to work with real world signals, generally these are nonlinear and non-stationary data. The conventional approach of FFT strictly works on linear and stationary data. If we work with another approach of Wavelet, it has certain drawbacks as well. It is non-adaptive as the same basic wavelet is used for all data. But for practical purposes, it would be good to have a transform that would not only allow dealing with non-stationary processes but would also use an adaptive transform basis determined by initial data. This type of transform does exist and here comes the necessity of using EMD.
- In Fourier analysis, one can only observe the frequency contents and their respective amplitude, no information is available as to when the frequencies

occurred. This is due to very nature of it in that the information is averaged over time. Unlike Fourier analysis where time resolution is completely lost, EMD method reserves time information. EMD generally is time-frequency analysis method that is capable of extracting amplitude and frequency information of a given time incident.

- Most of the literature that I reviewed showed extensive work with EMD thus providing a positive result in most of the cases.
- It is possible to achieve high performance on different audio/speech signal recognition based work if it is applied in a way it works.

There are other existing approaches available and widely used, for example MFCC, MFSC, RASTA, LPC, LPCC, PLP but I chose to work with EMD as it is a new approach and I was curious to see how it perform on this speech based gender identification work.

Chapter 4: Proposed Approach

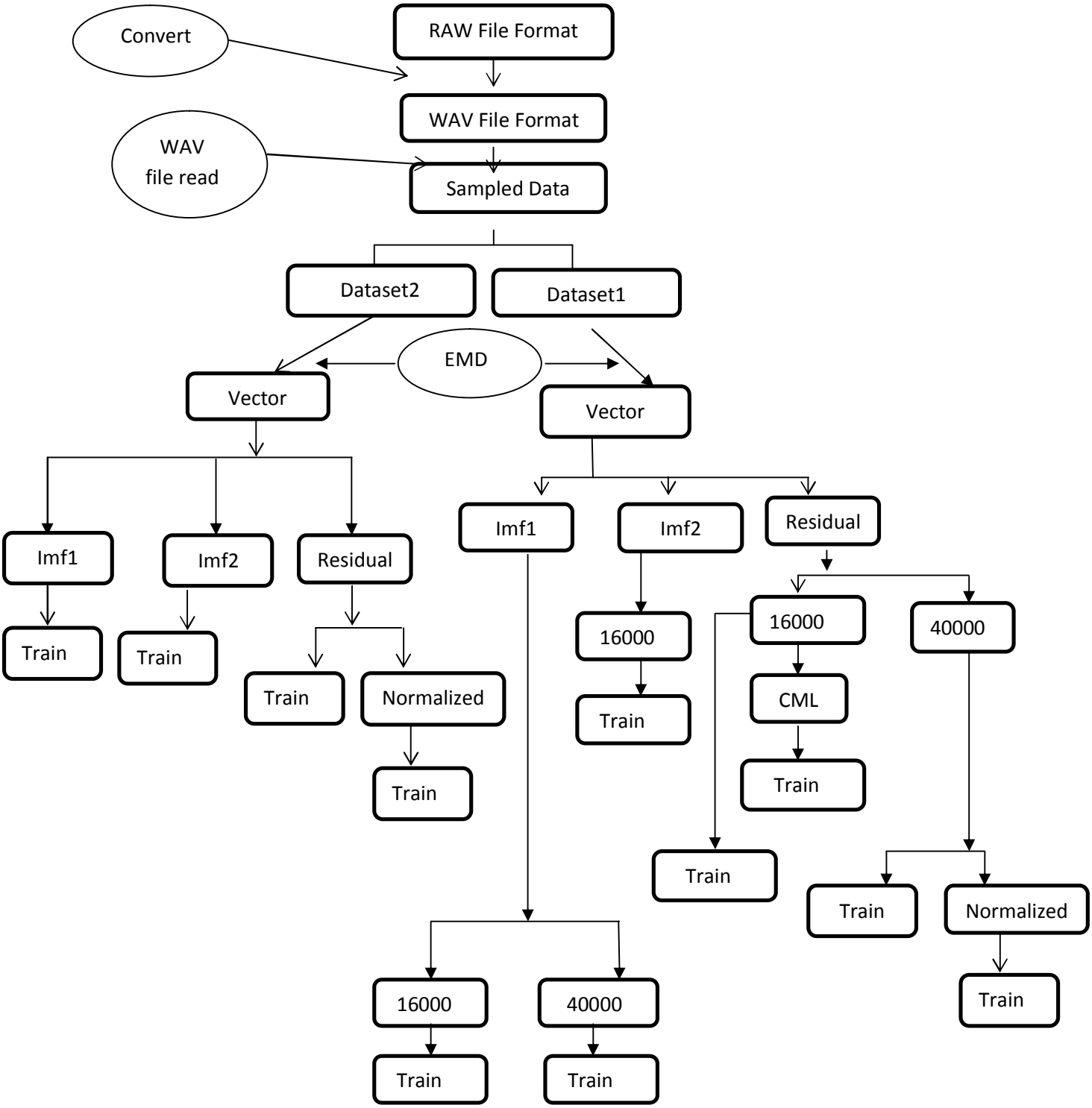
At first I have taken data which was text independent from the data base [1]. The data has the frequency of 16 kHz. I made two dataset. In dataset1 I took 3 speeches per speaker and total speaker were 25, in total 75 speeches for training and testing. Total female and male speaker were 10 and 15 respectively. In dataset2 I took 8 speeches were taken per speaker and total 48 speeches. Total 6 speakers including 3 male and 3 female speakers were there. All of the speeches were in RAW format then I converted into WAV format. And I used it in MATLAB environment. Then I applied EMD in dataset. If you look into the figure 4.1 you can easily see that I have taken the first, second and last row which represents IMF1, IMF2 and Residual respectively from the resulted 2D vector.

From IMF1 I have taken two types samples including 16,000 and 40,000 and trained. For training I have used SVM. Then from IMF2 I chose only 16,000 and trained. No 40,000 data were trained in IMF2 as there were poor results with 16,000 data. Coming into Residual, 16,000 data were trained along with CML was applied on it and trained. Again 40,000 data were taken and trained. On the other hand these 40,000 data were normalized and trained.

I applied MFCC and RASTA on dataset1 by taking the mean of each and normalized and then trained. Finally I compared with the results.

In dataset2 I applied EMD taking 40,000 samples.

Figure 4.1: Overview of the analysis



Chapter 5: Results and Analysis

5.1 Data preparation

For this, I take data from the database recorded internally at Carnegie Mellon University circa in 1991 [1]. This database is described in details in "Acoustical and environmental robustness in automatic speech recognition", by Alex Acero, published by Kluwer Academic Publishers, 1993. The database used internally at CMU has 1018 training and 140 test utterances, whereas the database provided there has 948 training and 130 test utterances.

The data which has been chosen was some audio raw little_endian and big_endian files of 16 KHz. So, those have been converted to .wav format by software and read by MATLAB to get the signals.

Two types of data samples were taken from the database-

- In dataset1- I have selected 10 female and 15 male, so total person were 25. Three speech samples were collected from each of 25 person, in total 25*3= 75 samples.
- In dataset2-I have selected 3female and 3 male, so total person were 6. Eight speech samples were collected from each of 6 person, in total 6*8= 48 samples

5.2 Experimental Setup1

I applied EMD in dataset1 and taken IMF1, IMF2, Residual of 16,000 samples and trained. Also I applied CML on the Residual and then trained. The details of the results are given below

- Taking IMF1: This is the first row of the vector. I trained IMF1 in SVM and the percentage was 67.57%.
- Taking IMF2: This is the second row of the vector. I trained IMF2 in SVM and the percentage was 64.86%.
- Taking Residual: this is the last row of the vector. We know in residual, most of features of the main signal are present. So I took the residual. I trained the residual in SVM and the percentage was 70.27%.
- Using CML (Coupled Map Lattice): A lattice has been created with three residuals of the same speaker each taking 16000samples per case. Next, CML (Coupled Map lattice) has been run using those lattices and make a new lattice of each. CML helps us to determine a unique pattern for each speaker. I got 56.76%

Sample	Residual	IMF1	IMF2	CML
16,000	70.27%	67.57%	64.86%	56.76%

Table 5.1: Results of Experimental Setup1

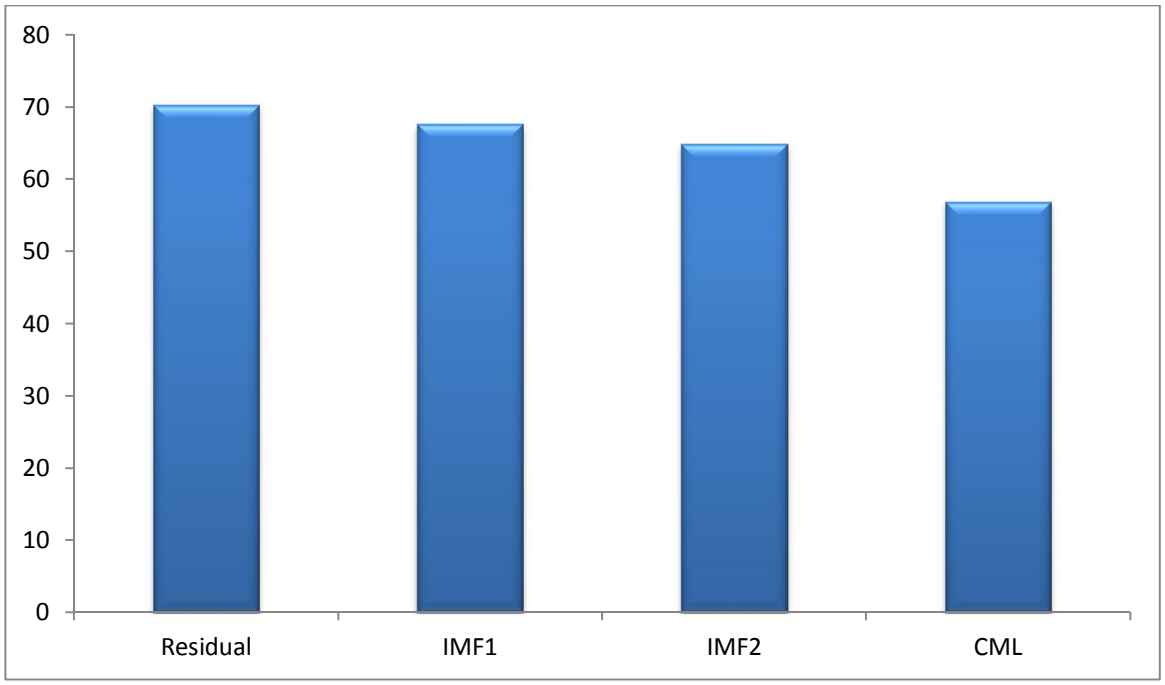


Figure 5.1: Bar chart of Experimental Setup1

5.3 Experimental Setup2

I have applied EMD in dataset1 and taken IMF1 and Residual of 40,000 samples and trained. Also normalized the Residuals and then trained. Here the noticing part is that, I have not taken IMF2 and do not use CML anymore as both of them shows poor results before. The results are given below

- Taking IMF1: I trained IMF1 in SVM and the percentage was 61.16%.
- Taking Residual (without normalization): I trained the residual in SVM and the percentage was **78.38%.**
- Taking Residual (with normalization): I trained the residual in SVM and the percentage was 61.54%.

Sample	IMF1	Residual(without normalization)	Residual(with normalization)
40,000	61.16%	**78.38%**	61.54%

Table 5.2: Results of Experimental Setup2

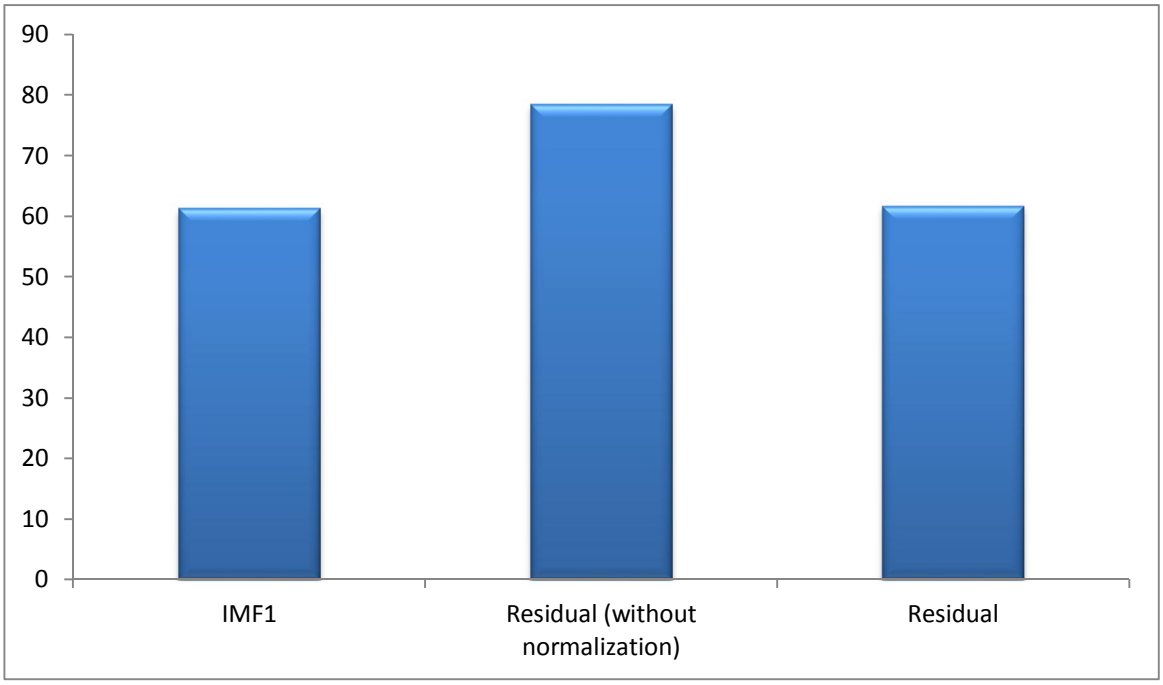


Figure 5.2: Bar chart of Experimental Setup2

5.3.1: Comparison

Here, I have found a better result taking Residuals of 40,000 samples. Now, I have applied standard features MFCC and Rasta to my dataset1 and trained with SVM.

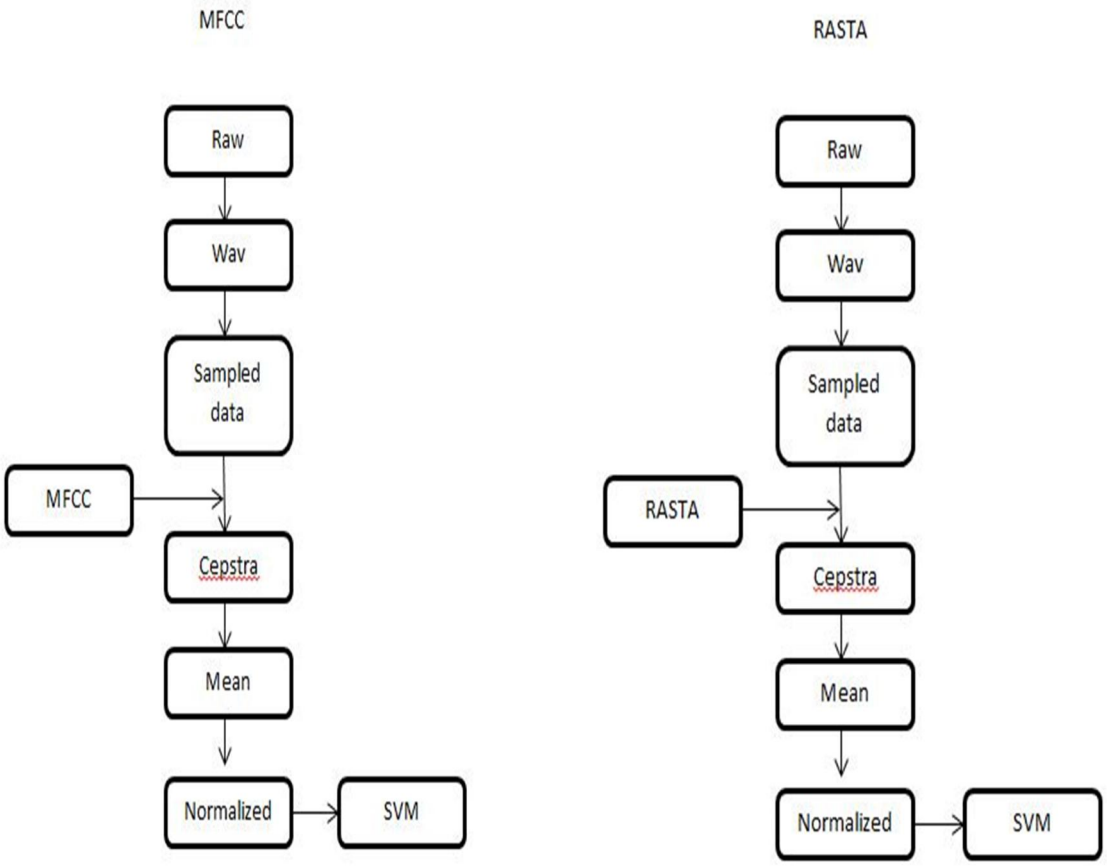


Figure 5.3: flowchart of applying MFCC and RASTA

I used MFCC and RASTA used on the dataset1 and compared the result with my new approach.

EMD	MFCC	RASTA
78.38%	90%	55%

Table 5.3: Comparing with Standard features

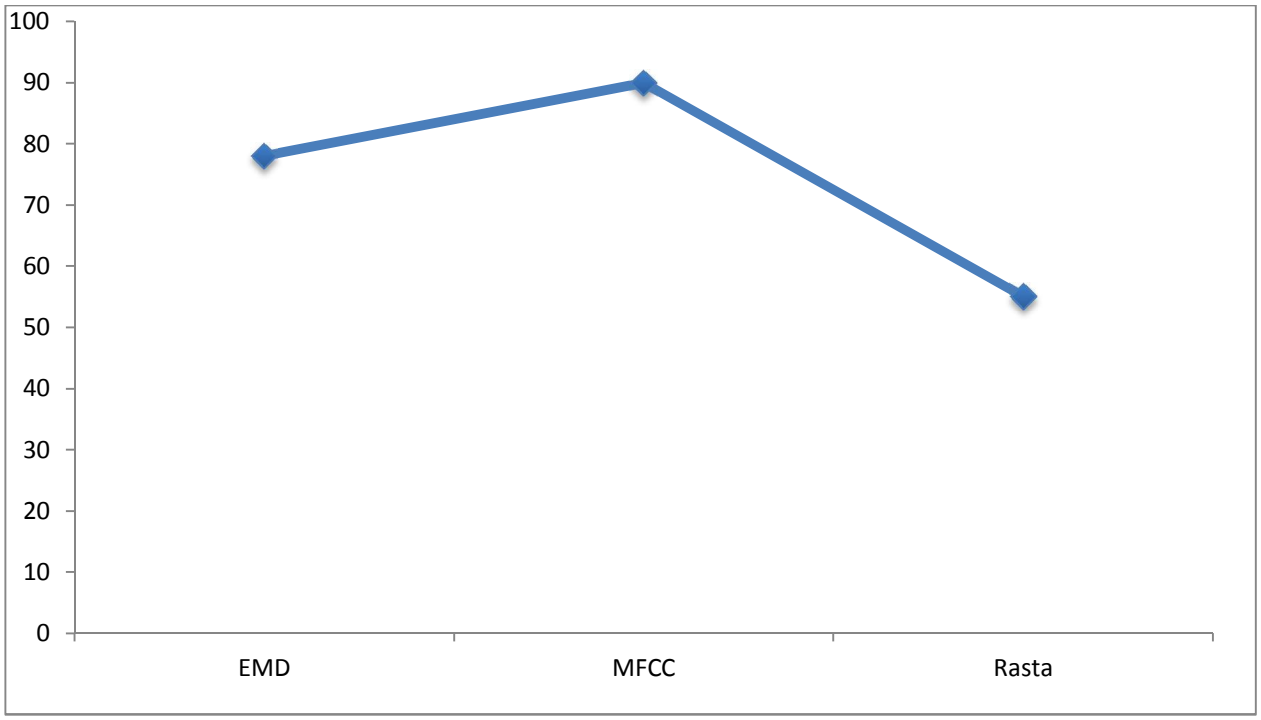


Figure 5.4: New Approach (EMD) VS Standard Approach VS RASTA

5.4 Experimental Setup3

Here I used dataset2, where 8 speeches per speaker were taken. Then I applied EMD and I have taken IMF1, IMF2, Residuals both with and without normalization from the resulted 2D vector.

- Taking IMF1: I trained IMF1 in SVM and the percentage was 50.00%.
- Taking IMF2: I trained IMF2 in SVM and the percentage was 45.83%.
- Taking Residual (without normalization): I trained the residual in SVM and the percentage was 62.50%.
- Taking Residual (with normalization): I trained the residual in SVM and the percentage was 54.17%.

Sample	IMF1	IMF2	Residual (without normalization)	Residual (with normalization
40,000	50.00%	45.83%	62.50%	54.17%

Table 5.4: Results of Experimental Setup3

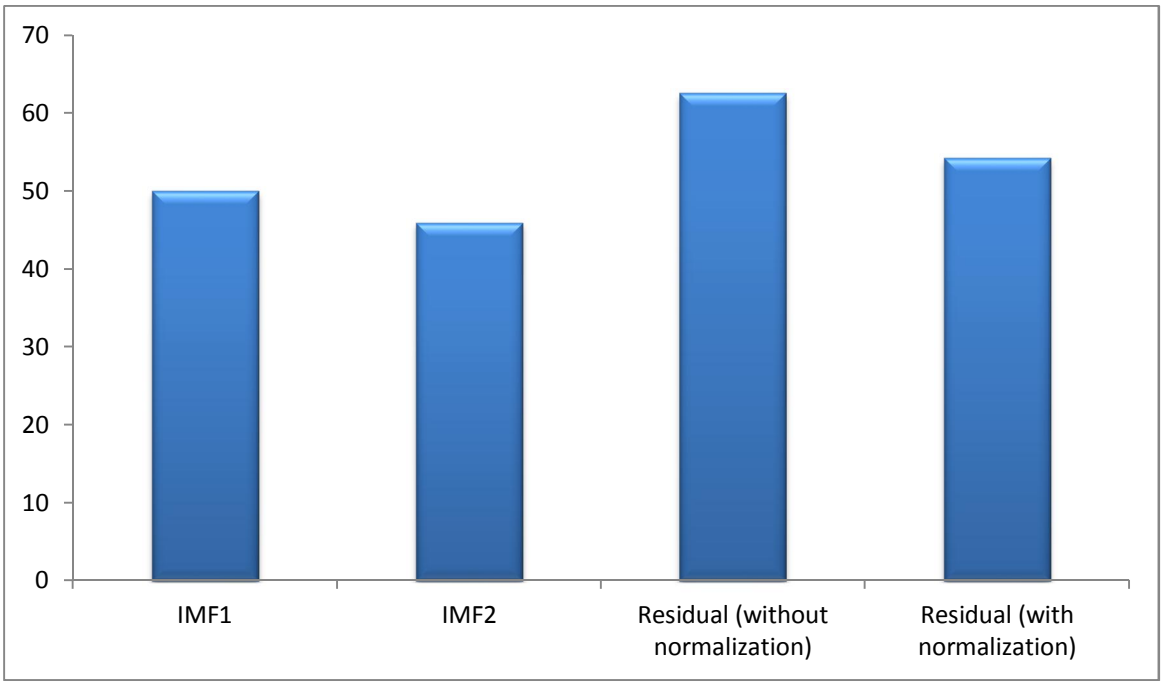


Figure 5.5: Bar chart of Experimental Setup3

5.5 Challenges observed and possible reasons:

In Hadi Hard, Liming Chen's paper named "**Gender identification using general audio classifier"** 2002, they applied MFSC or Mel Frequency Spectral Coefficients, which is a standard feature. They chose the speech length of 160 seconds to 2200 seconds. The rate of error drops from 22% to 11% while time duration rises from 160 to 2200 (Figure 1.1).

But I have used EMD, which is a very new approach and my sample speech's time duration was only 3 seconds. For this my error rate was 21.62%. It was a challenge for me to get the positive results with those data sets. In comparison to the paper my error rate was much less as I have used totally new approach. It proves that if we take speech samples of much time duration, I believe, though it is a time consuming process it would produces a result with less error rate.

When I worked with sample duration of 3 seconds in MATLAB environment it took minimum 15 minutes to maximum 30 minutes each time I applied EMD on single speaker. So it was very time consuming as EMD approach uses huge numbers of calculation and iteration rate is very high.

The sample I have collected, the voices were normal and gentle, did not fully fell under non-linear data category. To work with those samples was challenging as EMD mostly uses non-linear data for calculation.

During the sample collection, it was very hard to get noise free speech samples to continue my work. These are called contaminated speech samples. One cannot get positive results using these types of samples.

Chapter 6: Conclusion and future work

6.1 Conclusion:

In this paper I applied Empirical Mode Decomposition technique to speech data and evaluate on various aspects of gender identification. Furthermore, although my feature did not outperform MFCC but it has outperformed the conventional standard feature of RASTA. I hope that my approach will inspire more research automatically identification of gender by speech signal.

When I worked with sample duration of 3 seconds in MATLAB environment it took large time duration to complete each time I applied EMD on single speaker. So it was very time consuming as EMD approach uses huge numbers of calculation and iteration rate is very high.

In this thesis, enlightened by the successful applications, I try to test the EMD algorithm in speech identification. I have used EMD, which is a very new approach and my sample speech's time duration was only 3 seconds. For this my error rate was 21.62%. In comparison to the paper my error rate was much less as I have used totally new approach.

6.2 Future Work:

Further research and study must be required to improve the work. I will be working with large durational speech sample also with different sets of data.

References:

Internet:

[1] http://www.speech.cs.cmu.edu/databases/an4/index.html

[2] http://www.ee.columbia.edu/~dpwe/pubs/Ellis10-introspeech.pdf

[3] http://research.microsoft.com/pubs/194580/MINDS-report.pdf

[4] http://www.mql5.com/en/articles/439

[5] http://www.differencebetween.com/difference-between-linear-and-vs-nonlinear-data-structures/#ixzz30GI4Lpzm

Publication

[6] Speech processing using the empirical mode decomposition and the Hilbert transform by Ke Gong, 2004.

[7] Unsupervised feature learning for audio classification using convolutional deep belief networks by Honglak Lee, Yan Largman, Peter Pham and Andrew, 2009.

[8] Classification of audio signals using SVM and RBFNN by P. Dhanalakshmi, S. Palanivel, V. Ramalingam, 2008.

[9] Gender identification using audio classifier by Hadi Hard, Liming Chen, 2002.

[10] Hand Tremor detection via adaptive Empirical Mode Decomposition and Hilbert-Huang Transformation by James Z. Zhang, Robert D. Adams, Kenneth Burbank, 2009.

[11] Artifact reduction in electrogastrogram based on empirical mode decomposition method by H.Liang, 2000.

[12] The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis by Norden E. Huang, Zheng Shen, 1998.

[13] Empirical mode decomposition, a useful technique for neuroscience? Review of Huang et al. by Robert Liu, 2002.

[14] Theory and Applications of Coupled Map lattices by Kunihiko Kaneko, 1993.

APPENDIX

If anyone is interested to carry on the research further or require and assistance regarding this thesis the contact information of the researcher and author is given below.

Mahpara Hyder Chowdhury

Email: semonti27@gmail.com

The source codes used in this thesis are open source. Please go through the references to find the citations. The source codes, data files and results of my thesis can be found upon request. Please contact the author if required.