CT 478-Speech Technology

Project Report

Project Name:

Relative phase information for detecting human speech and spoofed speech

Guidance: Prof. Hemant Patil

Submitted by : Dhavalkumar Prajapati (201501188)

Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar,Gujarat,India

May 2018

TABLE OF CONTENTS

Acknowledgement

The satisfaction and euphoria on the successful completion of any task would be incomplete without mentioning the people who made it possible whose constant guidance and encouragement crowned out effort with success.

We would like to express our heartfelt gratitude to our esteemed supervisor, **Prof. Hemant Patil** for his technical guidance, valuable suggestions, and encouragement throughout the experimental and theoretical study and in this course project. It has been our honor to work under his guidance, whose expertise and discernment were keys in the completion of this project.

We are grateful to the **DAIICT,** for giving us the opportunity to execute this project, which is an integral part of the curriculum in B.Tech programme **CT478** Speech Technology course at the DAIICT, Gandhinagar.

Many thanks to **Mr. Srinivas bhai** and our friends who are directly or indirectly helped us in our project work for their generous contribution towards enriching the quality of the work.

This acknowledgment would not be complete without expressing our sincere gratitude to **Speech Lab** for the help.

Terminology

RPS : Relative Phase Shift
GMM : Gaussian Mixture Model
MGDCC : Modified Group Delay Cepstral Coefficient
MFCC : Mel frequency cepstral coefficient
GD : Group Delay
DET : Detection Error Trade off

ABSTRACT

Recently, speaker verification technology has been used in many applications, such as a telephone banking, customer services,Authentication of gates, so security is very important and needs to prevent from spoofing attack.Any professional can make synthetic speech which is exactly same as natural speech, Anyone can use replay speech as spoof speech,which is more easy then other attacks. There are many different kinds of methods for detection of human speech and spoofed speech. However, in this project I am going to use relative phase information extracted from a Fourier spectrum is used to detect human and spoofed speech. Because original/natural phase information is almost entirely lost in spoofed speech,so in this project i will analysis the Relative phase shift of signal and extract RPS features and the relative phase information is also combined with the Mel-Frequency Cepstral Coefficient (MFCC) and Modified Group Delay Coefficient(MGDC).To accomplish this project i am going to use “ASV spoof 2015 : Automatic Speaker Verification spoofing and countermeasures challenge” database.

1. Introduction

In this study, we focus on whether speech is natural or spoofed.To detect spoofed speech from human speech, many features (e.g. magnitude spectrum,group delay, pitch,accent,modulation features) have been considered, pitch information, spectrum information was proposed to detect synthetic speech and cosine-normalized phase and MGDC phase spectrum features were used in voice converted speech detection.Modulation features were used in synthetic speech.In this project RPS features used to detect human speech or spoofed speech because the original phase information is entirely lost in spoofed speech.

2. Motivation

Recently many technology using speaker verification base security so, spoof detection is very important in these system,Motivation of this project is make system very effectiveness to words the spoofing detection and give better result with less error equality rate.

3. Application

RPS base feature extraction can be used in various kind of system as described below :

1. These system can be used in baking system.
2. Mobile,laptop and other device can be unlocked using this system
3. Telephones system can be verified using these system.
4. Speaker Verification

4.System Architecture

In this project I implemented three features are RPS,Group Delay and Modified Group Delay Cepstral Coefficient and for classifier I used GMM and for evaluation I took dataset of **“ASV spoof 2017”** for DET curve I used NIST’s toolkit.

Here is the basic diagram for spoofing detection system :

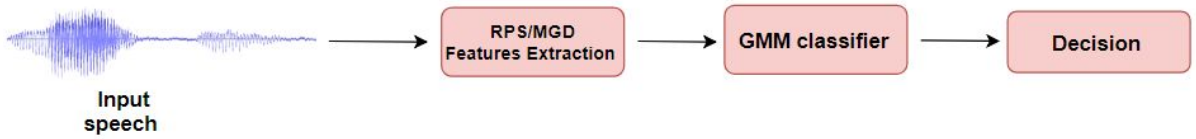


Fig.4.1 Basic diagram of spoofing detection system

We give input as a speech which also known as test speech then extract various features like Mel Frequency Cepstral Coefficient, Relative Phase Shift,CFCC,MGDCC etc.In this project I have used Relative Phase Shift,Group Delay and Modified Group Delay Cepstral Coefficient,After extracting the features it goes to the GMM and In GMM already two model human and spoof already implemented and find the score.At the End we decide whether speech is natural or spoofed.

4.1 Overview of spoofing detection system using GMM

The flow chart of Gaussian Mixture Model is shown in figure 1.As per below diagram we have to give utterance of speech as input the extract features (can be MFCC, RPS or other) and then some selected features are given input in GMM.GMM is trained with natural speech and spoofed speech.so, test input feature will make decision.

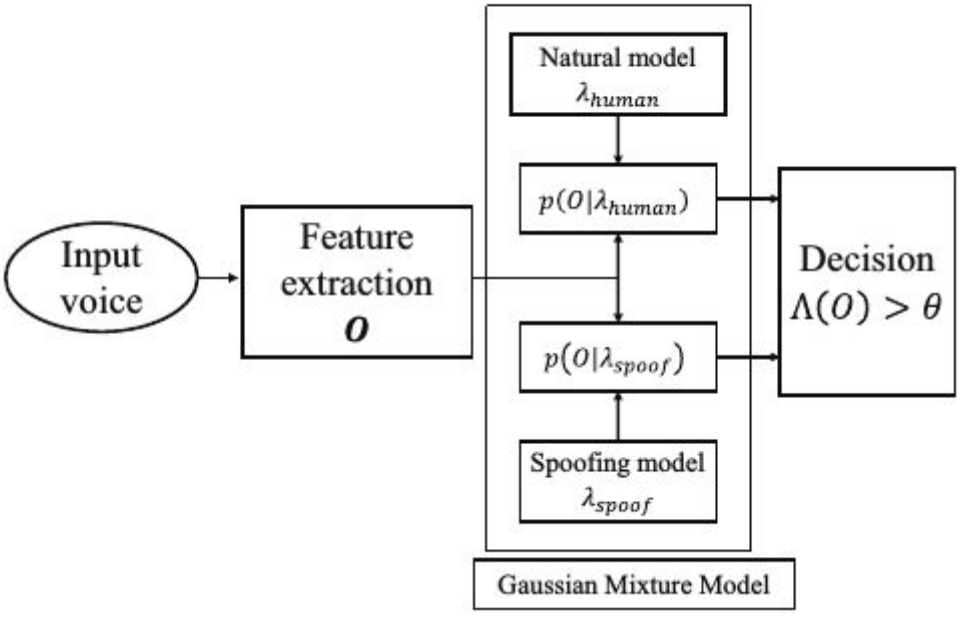


Figure 4.2. Flow chart of spoofing detection system

There are two phase in GMM:

1.Training of GMM

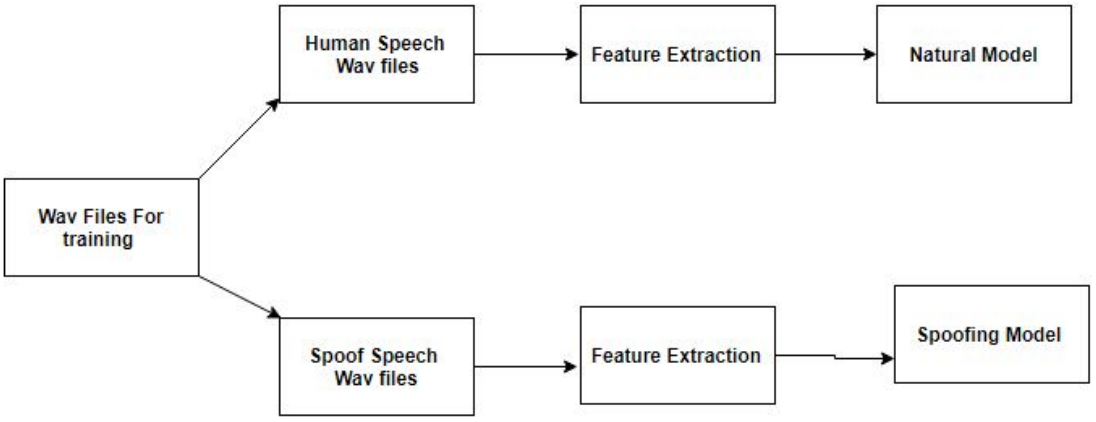


Fig.4.3 Diagram of Training of GMM

for train GMM I have used ASV spoof 2017 dataset in this dataset and protocol as per below

```
T2 T2_1000140.wav human human
T2 T2_1000141.wav human human
T2 T2_1000142.wav human human
T2 T2_1000143.wav human human
T2 T2_1000144.wav human human
T2 T2_1000145.wav human human
T2 T2_1000146.wav human human
T2 T2_1000147.wav human human
T2 T2_1000148.wav human human
T2 T2_1000149.wav human human
T2 T2_1003751.wav S5 spoof
T2 T2_1003752.wav S2 spoof
T2 T2_1003753.wav S1 spoof
T2 T2_1003754.wav S3 spoof
T2 T2_1003755.wav S4 spoof
T2 T2_1003756.wav S5 spoof
T2 T2_1003757.wav S2 spoof
T2 T2_1003758.wav S1 spoof
T2 T2_1003759.wav S3 spoof
```

Fig.4.4 Training Protocol

Last two column are showing whether speech is natural or spoofed. Here s1 to s5 are known attack.As per diagram for all train wav file we will extract features and we created two different model for Human speech we got natural model and for spoof speech we got spoof model.

2.Development (Testing)

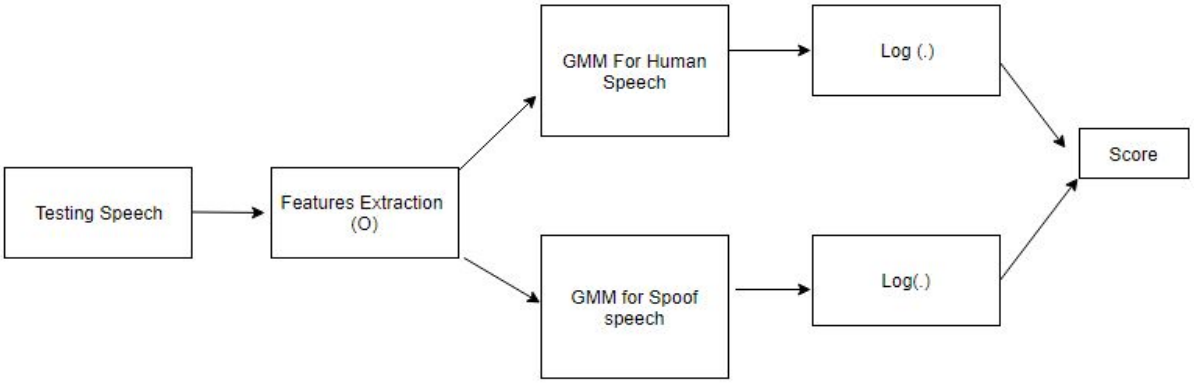


Fig. 4.5 Testing diagram

For Development of GMM, ASV spoof 2017 database snapshot as below :

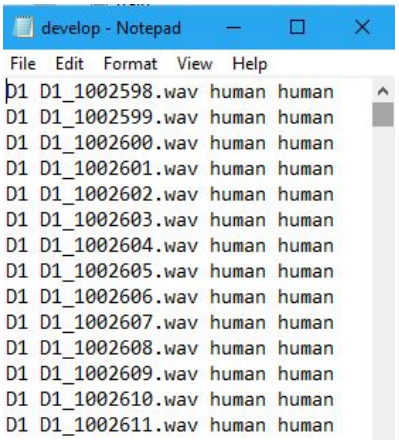


Fig. 4.6 Testing protocol

As above protocol, Test speech features are extracted and base on these features GMM will calculate GMMs for natural speech and and GMMs for spoof speech,After taking log we got log likelihood of GMMs.

The decision(score) about whether speech is natural or spoofed is based on log log likelihood ratio :

$$\Lambda(O) = \log p(O|\lambda_{human}) - \log p(O|\lambda_{spoof}) \text{..... eq(1)}$$

Where O is feature vector of input speech, λ_{human} is GMM for natural and λ_{spoof} GMM for spoofed speech.if GMM take multiple features as input then log likelihood ratio :

$$\wedge comb(O) = \Sigma n\ \alpha n\ \wedge (On) \text{.... eq(2)}$$

Where $\wedge (On)$ is the log likelihood ratio and αn denotes the weighting coefficients corresponding to features.

4.2 Relative Phase Shift

The definition of relative phase shift is The phase of certain base frequency kept constant and the phase of other frequency estimate relative to this base frequency.To understand consider a signal $x(t)$ and it's fourier transform $X(\omega)$ and any fourier transform is written as terms of magnitude and phase as below

$$X(\omega) = |X(\omega)| \, exp(j\theta(\omega)) \ldots \text{eq(1)}$$

The phase changes depending on the clipping position on the input speech even at the same frequency ω. To overcome this problem we kept certain base frequency as constant and calculate relative phase,For example, by setting the base frequency ω to 0. eq(3) is

$$X'(\omega) = |X(\omega)| \, exp(j\theta(\omega)) \, exp(-j\theta(\omega)) \ldots \text{eq(2)}$$

So, other frequency fourier transform relative to base frequency ω can be written as below

$$X'(\omega') = |X(\omega)| \, exp(j\theta(\omega')) \, exp(-j(\omega/\omega')\theta(\omega)) \ldots \text{eq(3)}$$

In eq(5) we will get normalized phase $\theta normalized(\omega')$ information can be written as below

$$\theta normalized(\omega') = \theta(\omega') + (\omega/\omega')(-\theta(\omega)) \ldots \text{eq(4)}$$

4.3 Group Delay

Group delay is define as a negative derivative of phase fourier transform with respect to frequency

$$\tau(\omega) = -d(\theta(\omega))/d(\omega)$$

Group delay can be directly derived from signal, consider a signal $x(n)$ and $y(n) = nx(n)$. Take fourier transform of both equation the we will get $X(\omega)$ and $Y(\omega)$ this two quantity are complex.so group delay defined as below equation

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2}$$

Here R is represent the real term and I is represent Imaginary term.Now we modified this equation we won't touch numerator term we change numerator term as per below

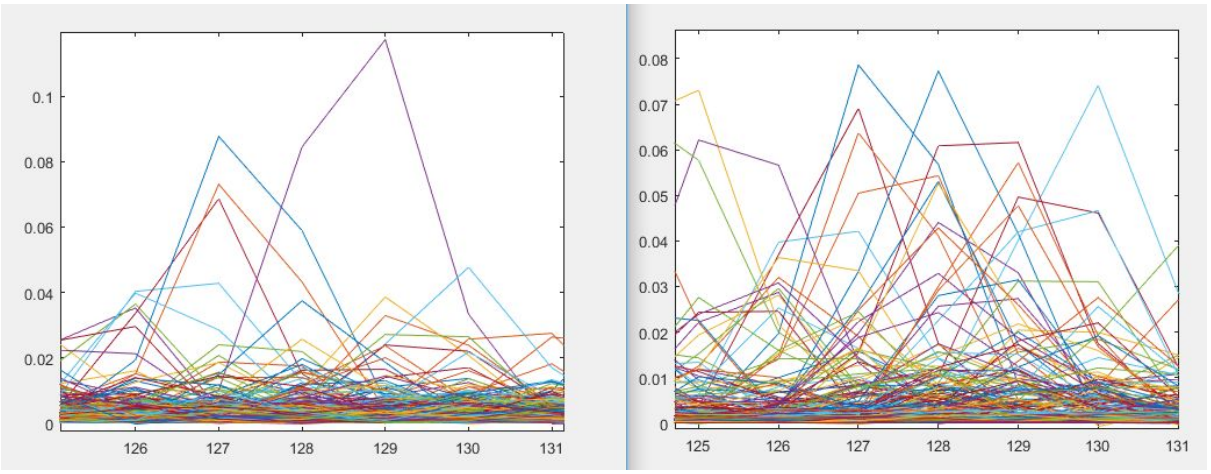
$$\tau(e^{j\omega}) = \left(\frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{S_c(e^{j\omega})^{2\gamma}}\right).$$

$$\tau_c(e^{j\omega}) = \left(\frac{\tau(e^{j\omega})}{|\tau(e^{j\omega})|}\right)(|\tau(e^{j\omega})|)^{\alpha},$$

Here γ and α are control the shape of spectrum and these parameter make curve very smoothly.

5. Novelty of My work

- Found alpha and gamma parameter for better result
- Comparison between group delay and modified group delay



Group Delay MGDCC

6. Algorithm Development

6.1 Algorithm Development of Relative phase shift

6.1.1 Pitch tracking

In RPS we first found fundamental frequency by using below function this algorithm is called a srh pitch tracking algorithm.The Summation of Residual Harmonics (SRH) method is a pitch tracker exploiting a spectral criterion on the harmonicity of the residual excitation signal.

the speech signal s(t) and the residual signal e(t) is obtained by inverse filtering. This whitening process has the advantage of removing the main contributions of both the noise and the vocal tract resonances. For each Hanning-windowed frame, covering several cycles of the resulting residual signal e(t), the amplitude spectrum E(f) is computed. E(f) has a

relatively flat envelope and, for voiced segments of speech, presents peaks at the harmonics of the fundamental frequency F0. From this spectrum, and for each frequency in the range [F0,min, F0,max], the Summation of Residual Harmonics (SRH) is computed as:

$$[srh\_f0, srh\_time] = SRH\_PitchTracking(wave, Fs, F0min, F0max)$$

For voicing speech male min frequency is 85 Hz and female max frequency is 255 Hz. so we will get fundamental frequency in these range.after getting fundamental frequency we can do sin analysis of signal.

6.1.2 Sin Analysis

In sin analysis estimate the parameter like amplitube,phase and frequency.For estimation we will use peak picking algorithm

```
option = sin_analysis();
option.fharmonic  = true; % Use harmonic model
option.use_ls = false; % Use Peak Picking
```

From previous step we got value of F_o .It will use in sin analysis as in below function

$$frames = sin\_analysis(x, fs, fos, option);$$

Frames contain phase so we can found RPS :

$$pk = frames(n).sins(3, :); \ \% \ 1.Amp \ , \ 2.fre, \ 3.The \ instantaneous \ phase$$
$$rps = pk - Ks. * pk(1);$$

Hence, This way we can find rps features and we also found delta and double delta features.

6.2 Algorithm Development of Modified Group Delay cepstral coefficient

In Matlab we can implement function as we see in system architecture for modified group delay cepstrum:

$$modified\_group\_delay\_feature(filePath, alpha, gamma, num\_coeff)$$

Here alpha and gamma are parameter for shaping the curve.By trial and error i got alpha value as 0.2 and gamma value is 0.9 for good result.

6.3 Algorithm Development of Group Delay

It's algorithm is same as modified group delay algorithm but take alpha and gamma value is equal to one.

7.pseudo code

7.1 pseudo code for RPS

```
[RPS,option] = phase_rps(frames, fs, option);
```

Function phase_rps algorithm:

```
for n=1:size(f0s)
 Ks = (0:size(frames(n).sins,2)-1)
 pk = frames(n).sins(3,:); % 1.Amp , 2.fre, 3.The
instantaneous phase
rps = pk - Ks.*pk(1);
dctsignal = dct2(rps);
static = dctsignal(1:60,:);   % First 60 features
del = deltas(static);
deldel = deltas(del);
feature = [static;del;deldel];
```

Metrix Output :

7.2 pseudo code for Modified Group Delay

Modified Group delay cepstrum:

x(n)=speech signal

y(n)=nx(n)

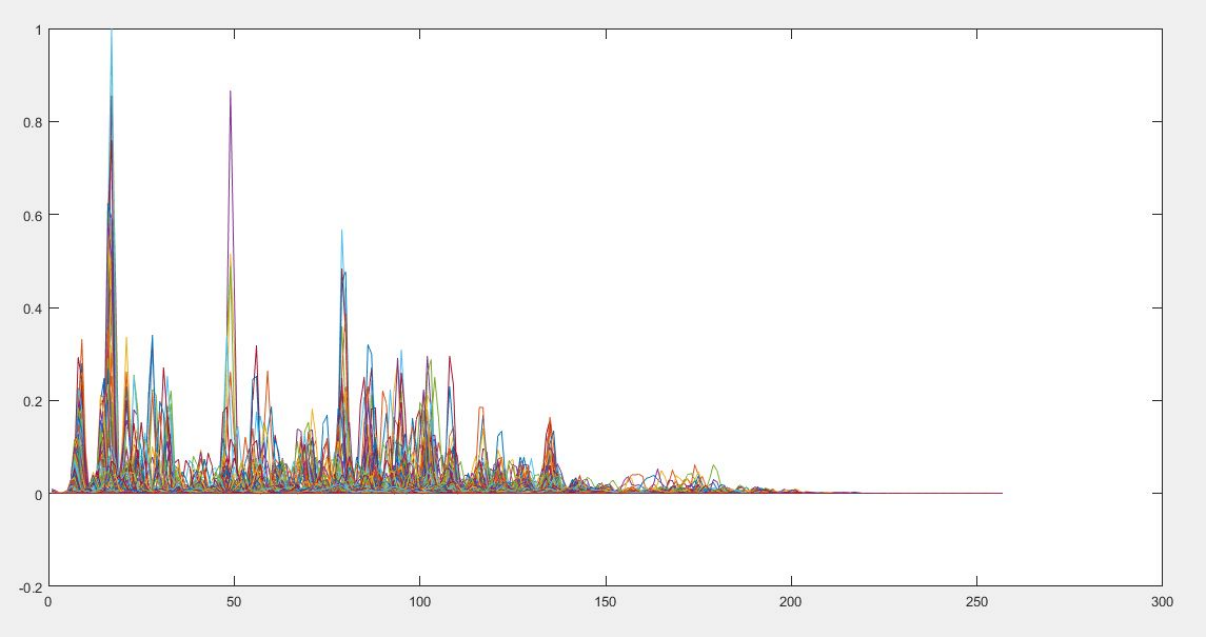
$X(\omega) = FFT(x(n))$

$Y(\omega) = FFT(y(n))$

```
grp_phase1 = (real(x_spec).*real(y_spec) + imag(y_spec) .*
imag(x_spec)) ./(exp(smooth_spec).^ (2*rho));
grp_phase = (grp_phase1 ./ abs(grp_phase1)) .*
(abs(grp_phase1).^ gamma);
```

Graph of modified Group delay:

$\tau(\omega)$

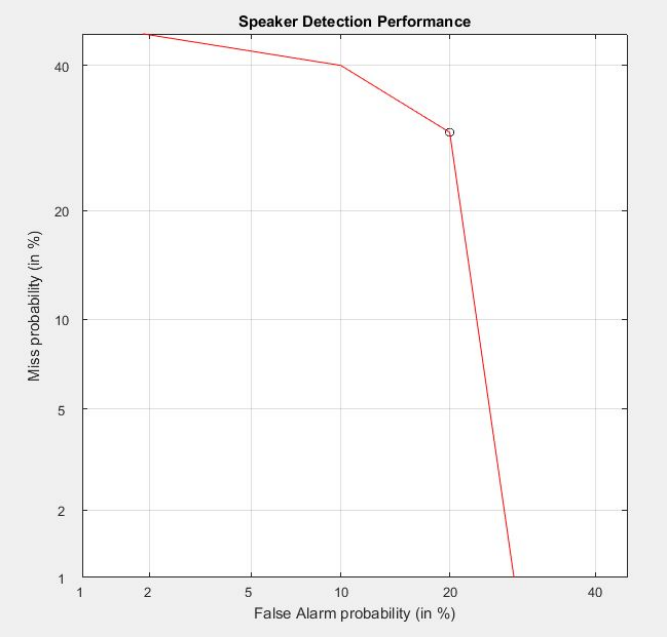


ω

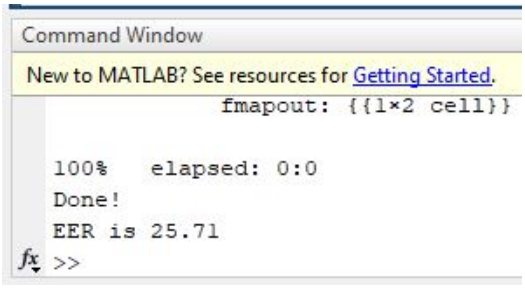
8. Result and Analysis

8.1 Result and Analysis for Relative Phase shift

DET :

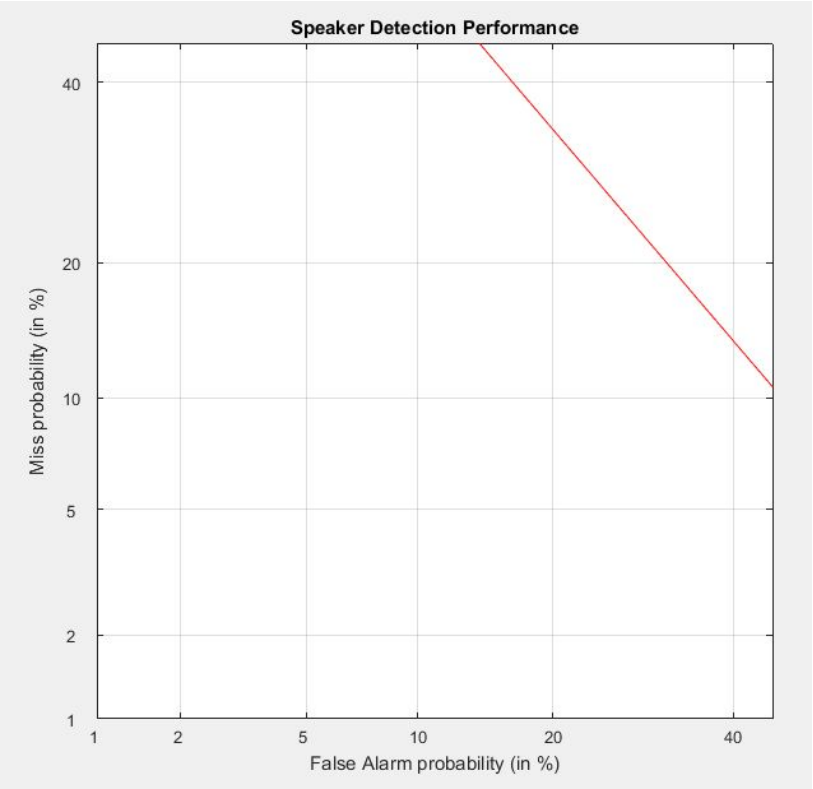


EER:

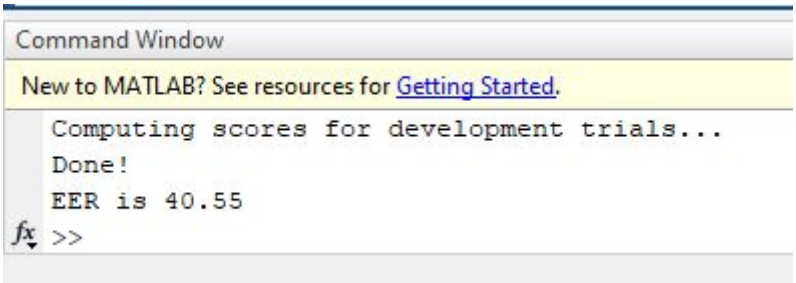


8.2 Result and Analysis for Group delay

DET:

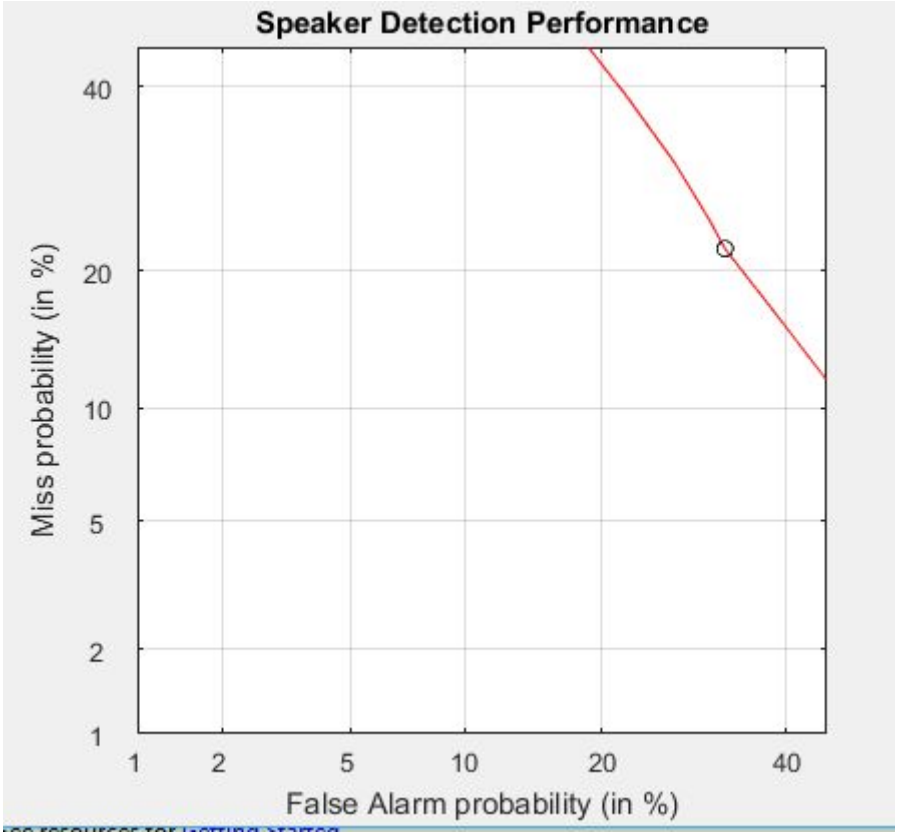


EER:

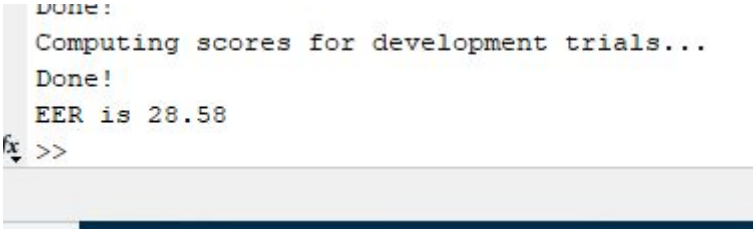


8.3 Result and Analysis for Modified Group Delay Cepstral Coefficient

DET:



EER:



8.4 comparison of result

Feature Extraction Method	EER(Error Equality Rate)
Relative Phase Shift	25.71
Group Delay	40.55
Modified Group Delay Function	28.58

Hence RPS is giving good result compare to modified group delay and group delay

9.Limitations

Limitation of this project is that we can't extract RPS features for unvoiced speech

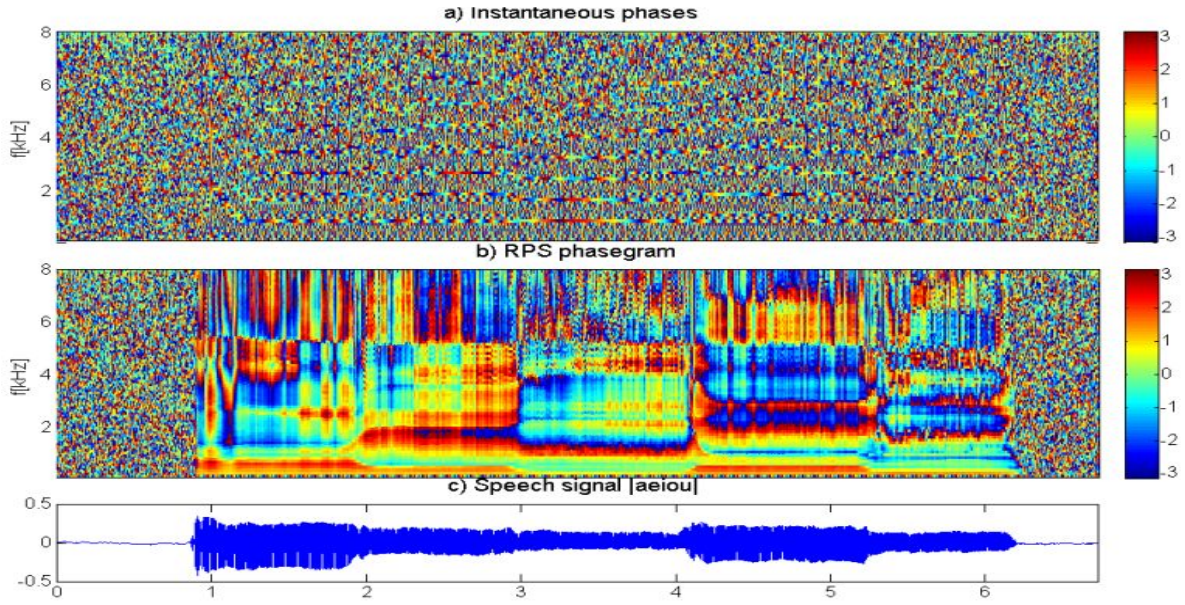


Fig a)instantaneous phases b)RPS phase gram c) speech signal

Above speech signal is utterance of vowels.This speech utterance contain voiced segment and unvoiced segment unvoiced means where zero crossing is very high.By analysing phase gram we can say that speech signal is not showing very good result so RPS features only for voiced segment.

10.Future work

- Large Dataset would give very good results
- Relative phase shift and Modified Group Delay cepstral coefficient combination would give better result
- Combination of Relative Phase Shift and MFCC
- Combination of RPS + MFCC + MGDCC

11.References

- Longbiao Wang , Yohei Yoshida, Yuta Kawakami1and Seiichi Nakagawa, “Relative phase information for detecting human speech and spoofed speech”,ASVspoof Challenge 2015

- Longbiao Wang, Member, IEEE, Seiichi Nakagawa, Zhaofeng Zhang, Yohei Yoshida, and Yuta Kawakami, “Spoofing Speech Detection Using Modified Relative Phase Information”,IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 11, NO. 4, JUNE 2017

- Hema a Murthy and B Yegnanarayana “Group delay functions and its applications in speech technology”,Sadhana Vol. 36, Part 5, October 2011, pp. 745–782. c Indian Academy of Sciences
- Simple representation of signal phase for harmonic speech models by I. Saratxaga, I. Herna´ez, D. Erro, E. Navas and J. Sa´nchez , ELECTRONICS LETTERS 26th March 2009 Vol. 45 No. 7
- Features and Classifiers for Replay Spoofing Attack Detection by Cemal Hanilci
- Speech Analysis/Synthesis Based on a Sinusoidal Representation by Robert J McAULY , THOMAS QUATIERI and IEEE Members,IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-34, NO. 4, AUGUST 1986
- T.Drugman, A.Alwan, "Joint Robust Voicing Detection and Pitch Estimation based on residual harmonics” interspeech 2011
- Synthetic Speech Detection Using Phase Information by Ibon Saratxagaa , Jon Sanchez , Zhizheng Wub , Inma Hernaez
- ASVspoof 2017: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan by Tomi Kinnunen , Nicholas Evans , Junichi Yamagishi , Kong Aik Lee , Md Sahidullah , Massimiliano Todisco , H´ector Delgado
- DETware_v2.1 DET-Curve Plotting software for use with MATLAB https://www.nist.gov/file/65996