VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SIGNALS AND SYSTEMS

PROJECT REPORT

FALL SEMESTER 2020

SPEECH RECOGNITION USING CORRELATION

TEAM MEMBERS:

LAVANYA GUNE -19BEC0367

SALONI KOSHE -19BEC0579

RAHUL GOWLAPALLI -19BEC0547

SHREYAS KHANDEKAR -19BEC0541

RITWIK ROSHAN -19BEC0531

MOHAN PRAKASH -19BEC0411

ABSTRACT

Speech is the method of communication among humans, whereas the communication between humans and computers is based on text user interface and graphic user interface. Recognizing the speech is the first step in every security project where our voice is used as the password and is also used for automation. Our project will demonstrate a model that enhances technological advancement where humans and computers interact via voice user interface. In developing the model, cross correlation was implemented in MATLAB to compare two or more voice signals and detect the most accurate one of the all. Cross correlation is implemented here to find the similarity between our recorded signal files and the testing signal. Mel frequency cepstral coefficients (MFCCs) are used as the feature of the recorded speech. For performing the above task an algorithm was designed to use the correlation technique in order to compare the frequency spectra of two voices. Here, user needs to enter the audio file name in the input section then using 'audioread' function it compares the input audio file along with audio file using the algorithm. Thus, we were able to develop a model where machines can differentiate between speech's and perform the respective actions assigned to them.

INTRODUCTION

Speech is the most prominent means of communication amongst humans. Human-to-human interaction is based on speech, emotion and gestures, thereby making it a lot easier to understand one another. On the other hand, the communication between humans and computers is based on either Text User Interface (TUI) or Graphic User Interface (GUI). It is a lot easier for us humans to recognize a person's voice than computers. Speech recognition is a game changer as we can developing machines which can understand and uniquely identify the user's voice which will make Human-Computer interaction more intriguing. Speech recognition is one of advanced technique which facilitates interaction between humans and computer. The research for recognizing speech had begun since late 1950s but due to its complexity and limited computing capabilities at that time, the progress was really slow for a few decades. In ideal conditions set up by laboratories, research show that Automatic Speech Recognition systems (ASR) have achieved high levels of recognition accuracy. Even though this tends to degrade in real world environment, it is still a huge progress and soon high accuracies in real world without ideal conditions will be possible. But still, in today's era, speech technologies have already started to play an important role. This technology is commercially and easily available for a different use. These technologies have been well developed since such that high accuracies which are enough to make machines respond correctly has been developed and it provides valuable services. Nowadays, the computing capabilities have well developed that the speech patterns can be used for security purposes. Speech can be used for the identification of person because every person has different speech characteristic. Thus, if the specific information in speech waves which corresponds to a specific person, speech recognition technology can easily identify the speaker.

CORRELATION

Correlation can be used to describe the mutual relationship that exists between two or more things. This definition holds good even in the case of signals. That is, correlation between signals indicates the measure up to which the given signal bears resemblance another signal.

In other words, if we want to know how much resemblance exists between the signals 1 and 2, then we need to find out the correlation of Signal 2 with respect to Signal 1 or vice versa. For any speech recognition system feature extraction and pattern matching are two very significant terms, here for feature extraction MFCC's is used while cross correlation is adopted for pattern matching.

TYPES OF CORRELATION-

There are two types of corelation

1. Auto-corelation
2. Cross-corelation

- **AUTO-CORELATION**-

  In this type of correlation, the given signal is correlated with itself, usually with the time-shifted version of itself.

  Mathematically, autocorrelation of continuous time signal $x(t)$ is given by

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)\, x^{\star}(t-\tau)\, dt$$

where $\star$ denotes the complex conjugate.

Similarly, autocorrelation of the discrete time signal *x*[*n*] can be expressed as

$$R_{xx}[m] = \sum_{n=-\infty}^{\infty} x[n]\, x^{\star}[n-m]$$

- **CROSS-CORRELATION-**

Cross correlation can be defined as a measure of similarity between two waveforms as a function of a time-lag applied to one of them. It is also referred to as sliding inner-product or sliding dot product. It is extensively used for searching a long-signal for a shorter, known feature. It also has applications in pattern recognition, electron tomographic averaging, single particle analysis, neurophysiology, and cryptanalysis. Mathematically, cross-correlation of continuous time signals x (t) and y (t) is given by

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)\, y^{\star}(t-\tau)\, dt$$

cross-correlation for discrete time signals x [n] and y [n] is given as

$$R_{xy}[m] = \sum_{n=-\infty}^{\infty} x[n]\, y^{\star}[n-m]$$

Hence, by considering these principles we implemented our algorithm of pattern recognition to recognize the spoken words efficiently but in a simple manner.

WHY DID WE USE CROSS CORRELATION?

When a signal sent is sent from a transmitter, the signal arrives at the receiver after being delayed by an unknown interval of time. If we need to find this delay, which is a result of being transmitted over the communication channel. This target could be achieved by cross-correlating the signal sent with the signal received.

Let us consider an example-

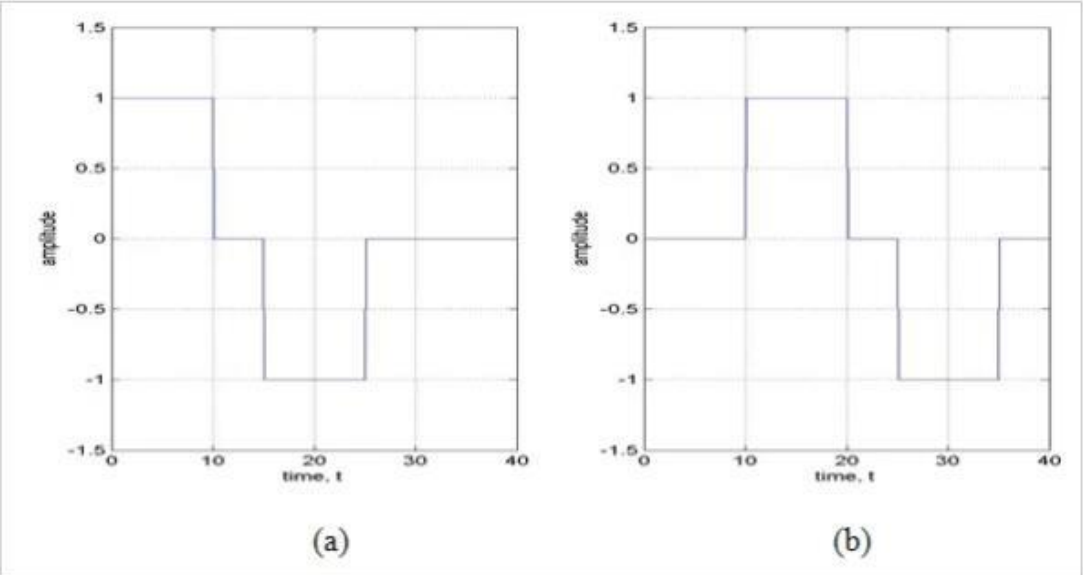
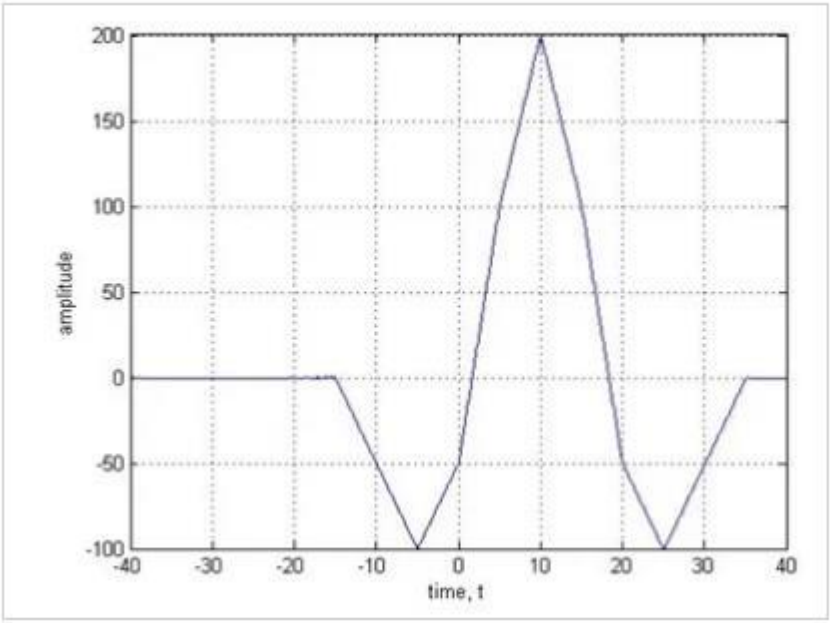


Figure (a) is the trained signal and figure (b) is the received or the test signal.



This is the correlation for the above signals.

The result obtained is shown in the figure below which clearly exhibits a peak at time t = 10. This indicates that the received signal finds best match with the test signal when the test signal is shifted by 10 units along the time-axis.

Another advantage of cross correlation is in case the received signal has not only been shifted but has also been corrupted by noise. Correlation of the signal received remains practically the same, even when the signal received is highly corrupted by noise. Cross correlation algorithm is used to remove the background noise from the human speech and obtained better results.

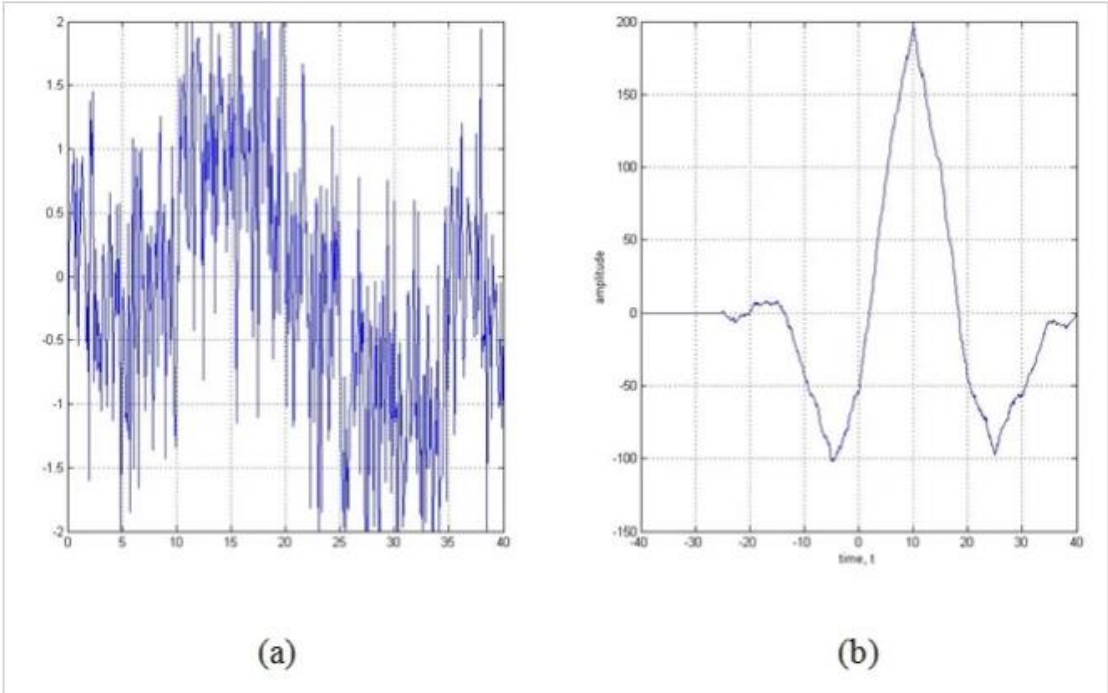


Figure (a) contains lot of noise and figure (b) contains cross correlation between the signal containing noise and the test signal. Here we can find that the even if the test signal contains some noise, the cross correlation in figure (b) is almost similar to the cross-correlation graph obtained for the test signal without noise. Hence by using cross correlation we can overcome the noise factor in the test signals.

Cross-correlation is used to measure the closeness amongst x and moved (slacked) copy of y as a component of the slack.
In the event that x and y have diverse lengths, the capacity annexes zeros toward the finish of the shorter vector so it has a similar length, N, as the other.

WORKING PRINCIPLE

The aim behind the development of this system is to recognize a word issued by the speaker through a microphone. The features we used for comparison are MFCCs (Mel Frequency cepstral coefficients), as it yields good differentiation of speech signal. We have used dynamic programming algorithm in the system to identify the similarity between the stored templates and the test templates for speech identification and specified the optimal distance.

Our entire project has been divided into two parts training phase and recognition phase. We have adopted Template bases Approach for training phase. In this technique, a collection of prototypical speech patterns (training vectors) has been stored in a database for upcoming use in the recognition phase. An unknown spoken utterance is compared with each of these testimonial templates and a category of the most similar pattern is selected. Usually template for each word is fabricated. Hence, Errors due to segmentation of smaller acoustically more small units such as phonemes can be avoided.

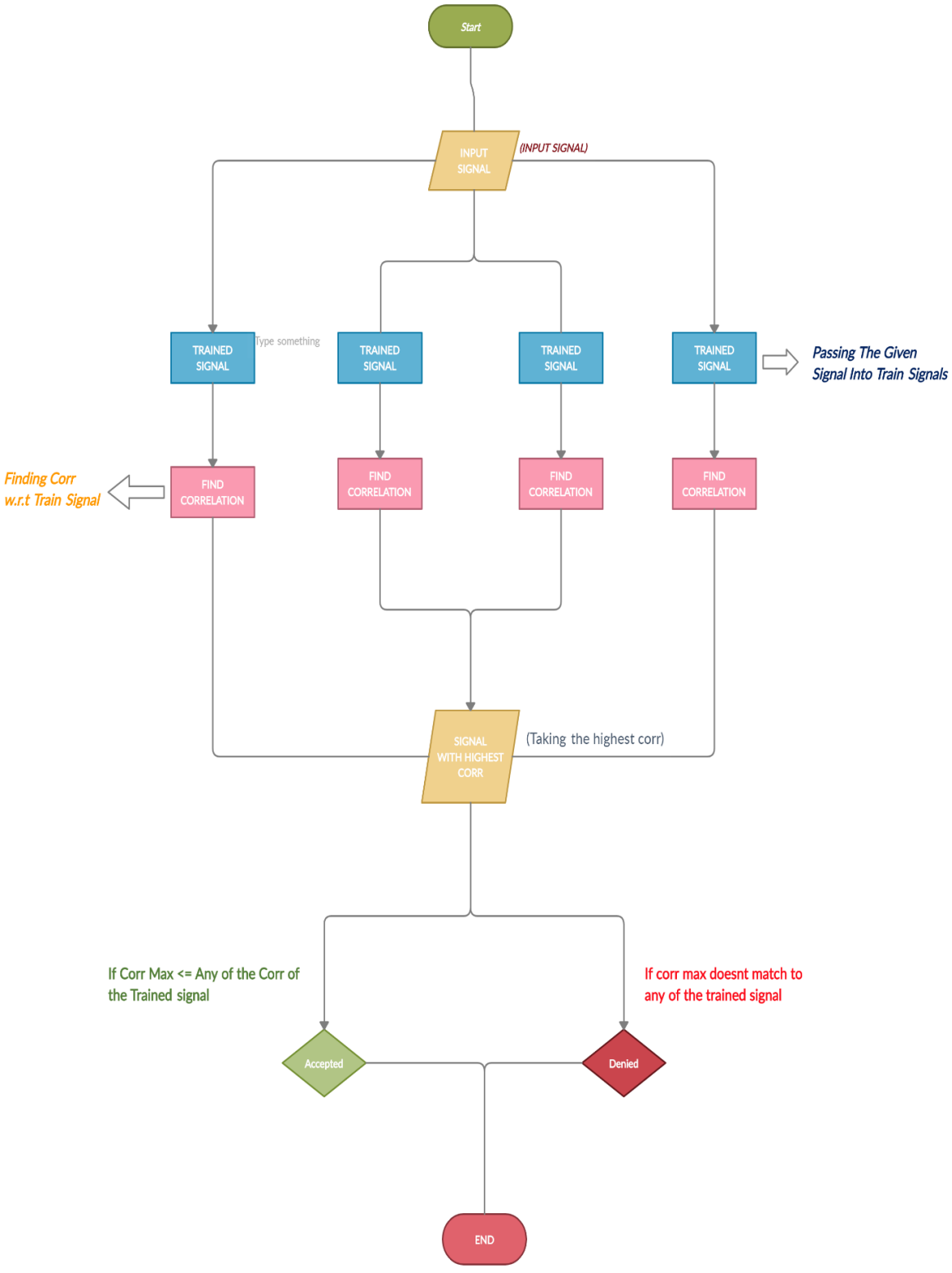
Representation of the short-term power spectrum of a sound, is primarily based on the linear remodel of a log power spectrum. They are extracted from a type of cepstral illustration of the audio clip, this theory uses the .wav format in MATLAB. MFCCs are extensively used as advantage in speech recognition systems which can automatically identify the spoken words from the audio file. MFCCs find their application in audio information retrieval applications such as audio similarity measures, genre classification. MFCC's values are not very powerful in the presence of supplementary noise, so it is easy to normalize their values in speech recognition systems to reduce the influence of noise.

By making assumption that, the recorded speech signals for the same word are the same, the spectrums of two recorded speech signals are also identical. When performing the cross-correlation of the two similar spectrums and plotting the cross-correlation, the cross-correlation should be symmetrical according to the definition of the cross-correlation. The level of symmetry in cross-correlation of two signals indicates the matching level of both signals. In addition, the more symmetric is the cross-correlation, the smaller is the value of the mean square error. By contrasting the mean square errors of processing trained words and the target word, the system decides the trained word that is better matched with the test signal based on the minimum mean square error of the cross-correlation differences at different lags.
In MATLAB the cross-correlation function is xcorr to find the correlation between two given functions. The Syntax applied for Correlation in MATLAB is expressed as:

r = xcorr (x, y)

this returns the cross-correlation of two discrete-time sequences, x and y.

ALGORITHM AND WORKFLOW



MATLAB CODE

```
function spr(file)
voice=audioread(file);
x=voice;
x=x';
x=x (1, :);
x=x';
y1=audioread('one.wav');
y1=y1';
y1=y1(1, :);
y1=y1';
z1=xcorr (x, y1);
n1=max(z1);
p1=length(z1);
u1=-((p1-1)/2):1:((p1-1)/2);
u1=u1';
%subplot(3,2,1);
plot(u1,z1);
y2=audioread('two.wav');
y2=y2';
y2=y2(1,:);
y2=y2';
z2=xcorr(x,y2);
n2=max(z2);
p2=length(z2);
u2=-((p2-1)/2):1:((p2-1)/2);
u2=u2';
%subplot(3,2,2);
figure
plot(u2,z2);
```

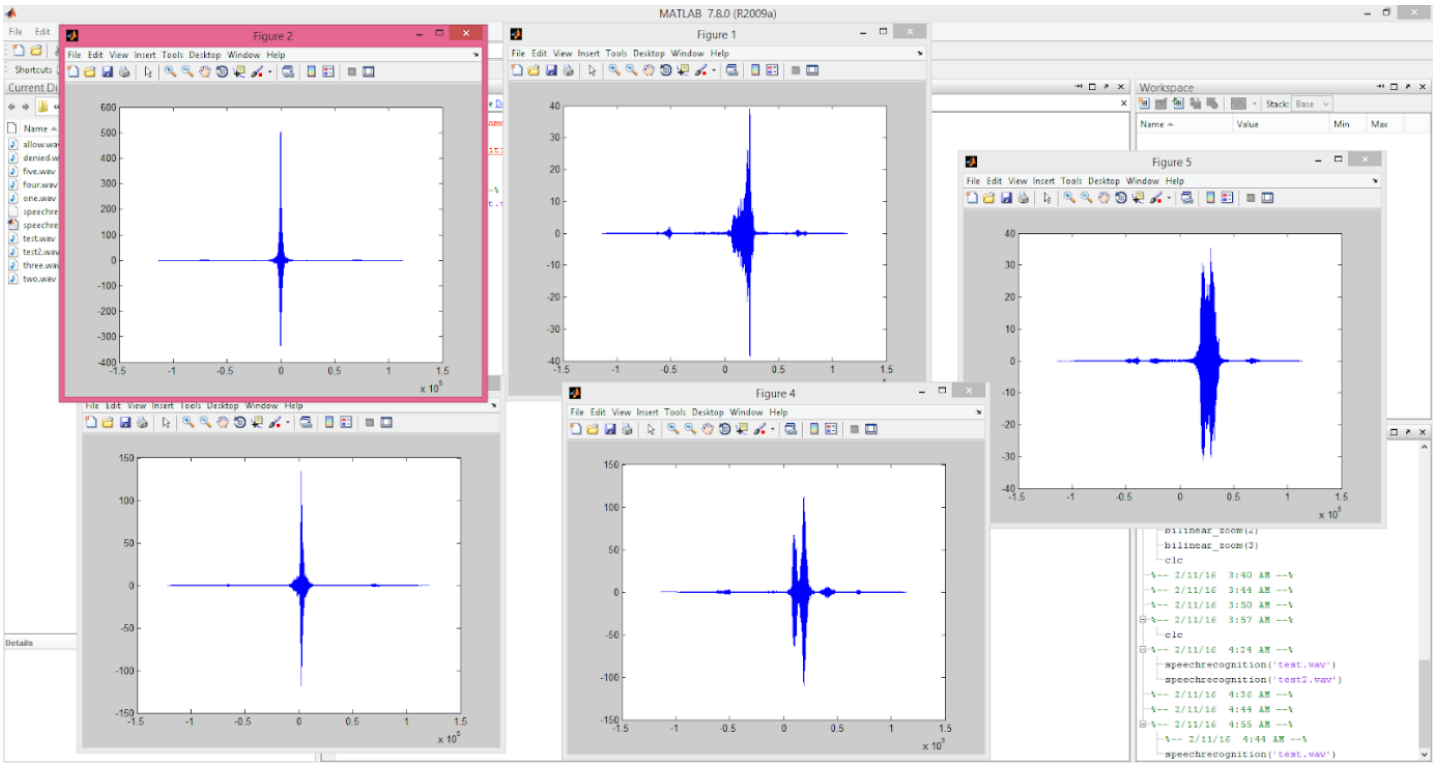
```
y3=audioread('three.wav');
y3=y3';
y3=y3(1,:);
y3=y3';
z3=xcorr(x,y3);
n3=max(z3);
p3=length(z3);
u3=-((p3-1)/2):1:((p3-1)/2);
u3=u3';
%subplot(3,2,3);
figure
plot(u3,z3);
y4=audioread('four.wav');
y4=y4';
y4=y4(1,:);
y4=y4';
z4=xcorr(x,y4);
n4=max(z4);
p4=length(z4);
u4=-((p4-1)/2):1:((p4-1)/2);
u4=u4';
%subplot(3,2,4);
figure
plot(u4,z4);
y5=audioread('five.wav');
y5=y5';
y5=y5(1,:);
y5=y5';
z5=xcorr(x,y5);
n5=max(z5);
p5=length(z5);
u5=-((p5-1)/2):1:((p5-1)/2);
u5=u5';
%subplot(3,2,5);
```

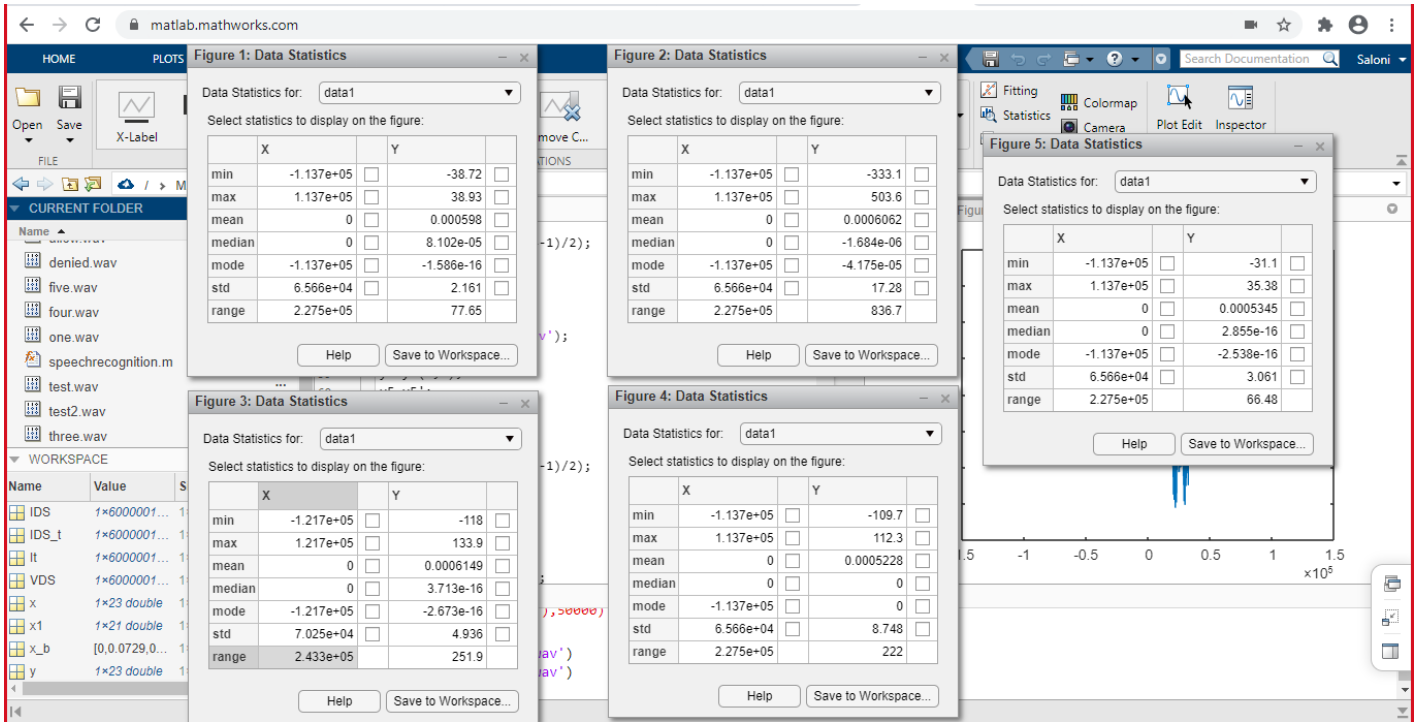
```
figure
plot(u5,z5);
n6=300;
a=[n1 n2 n3 n4 n5 n6];
n=max(a);
h=audioread('allow.wav');
if n<=n1
    soundsc(audioread('one.wav'),50000)
        soundsc(h,50000)
elseif n<=n2
    soundsc(audioread('two.wav'),50000)
        soundsc(h,50000)
elseif n<=n3
    soundsc(audioread('three.wav'),50000)
        soundsc(h,50000)
elseif n<=n4
    soundsc(audioread('four.wav'),50000)
        soundsc(h,50000)
elseif n<n5
    soundsc(audioread('five.wav'),50000)
        soundsc(h,50000)
else
   soundsc(audioread('denied.wav'),50000)

end
```

RESULTS

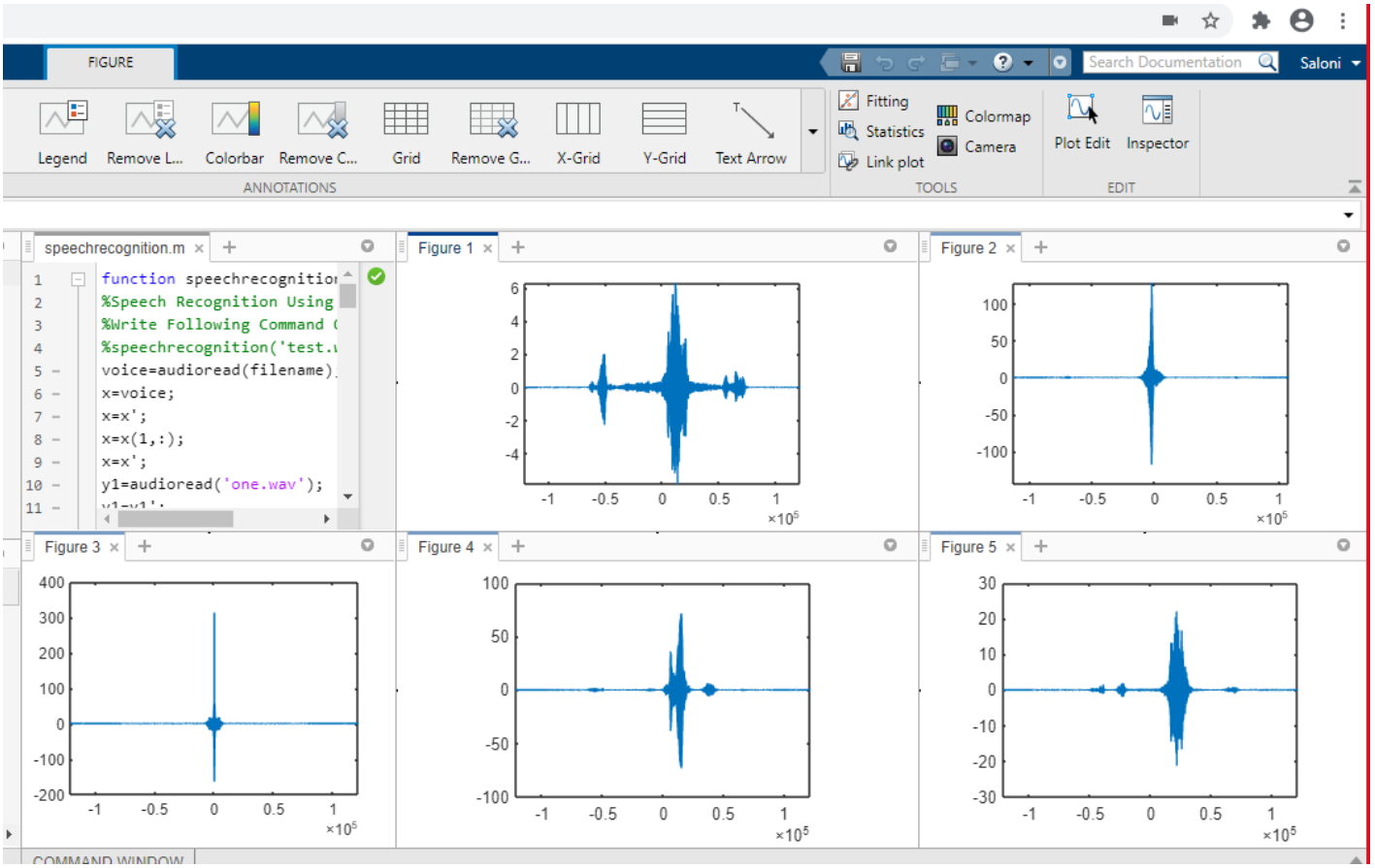
OUTPUT FOR TEST.WAV:

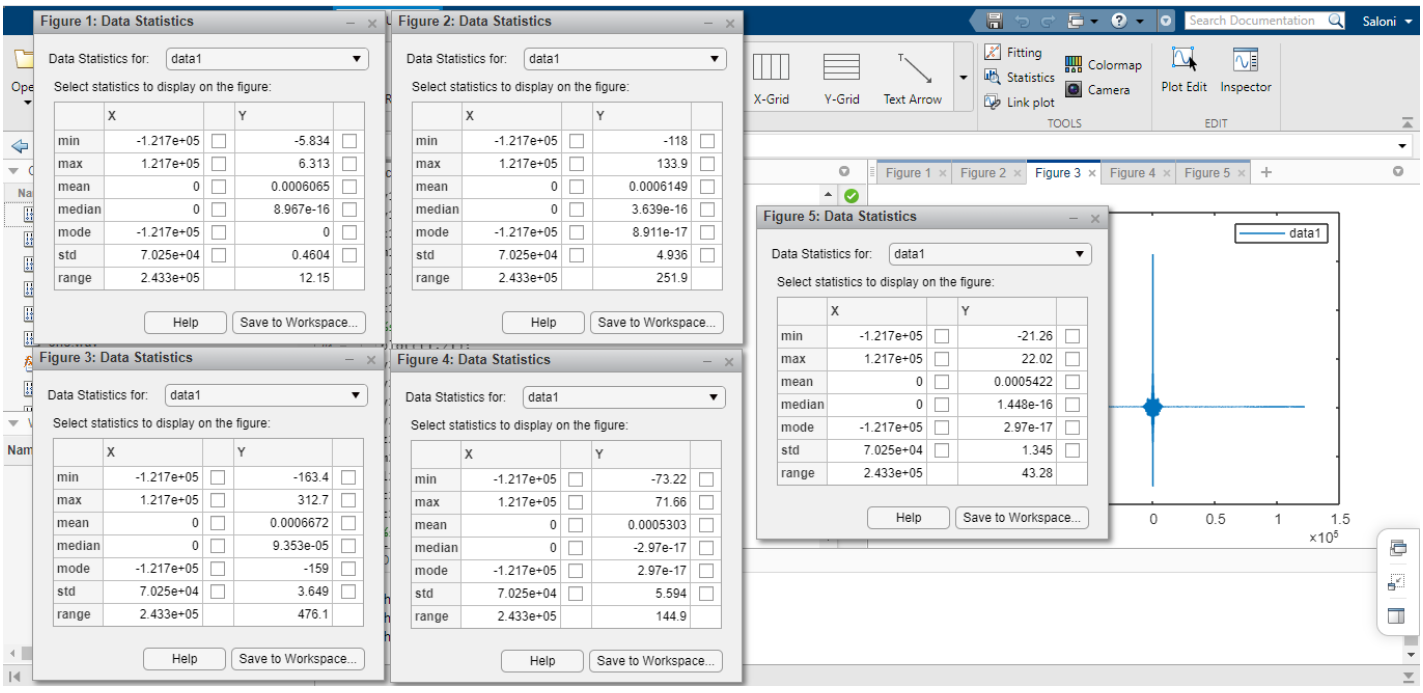




Here we can observe that the maximum value obtained at the y axis (cross correlation) is for the figure 2. Hence this is the signal that matches the best with the test signal.

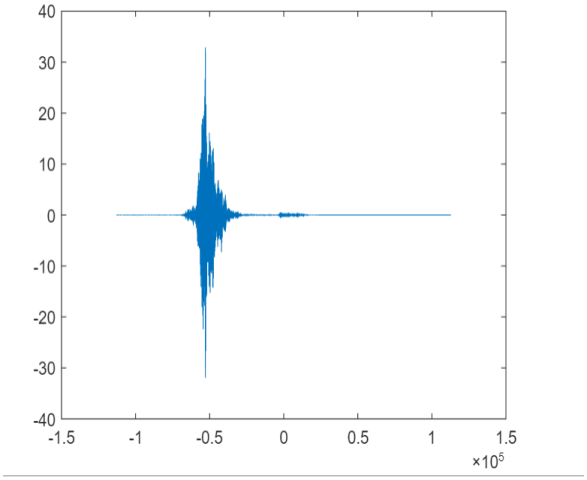
OUTPUT FOR TEST2.WAV:

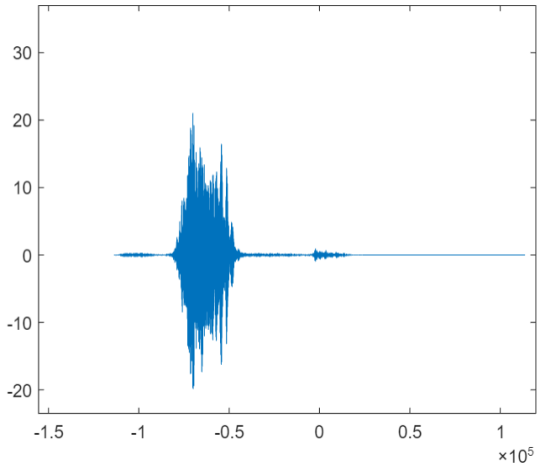


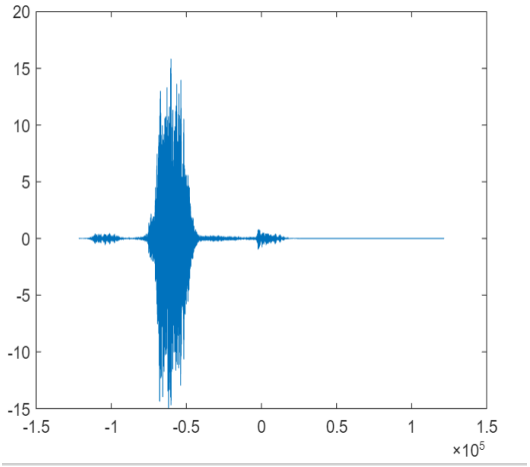


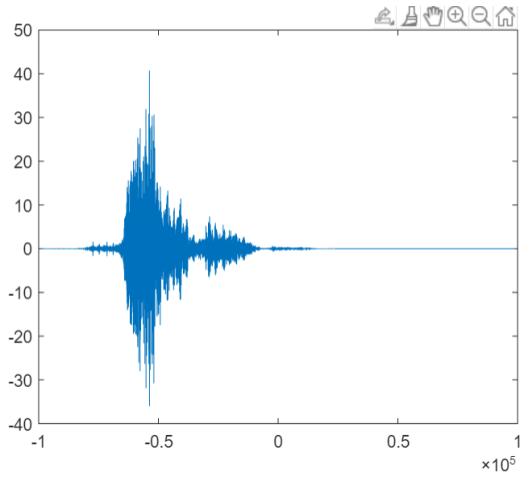
Here we can observe that the maximum value obtained at the y axis (cross correlation) is for the figure 3. Hence this is the signal that matches the best with the test signal.

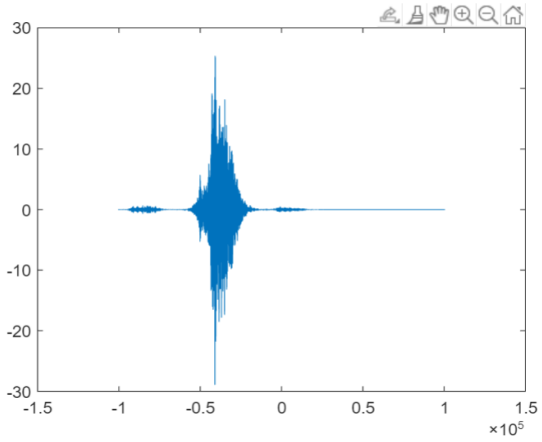
OUTPUT FOR DENIED.WAV:











The graph with less distortion has more correlation, from the above graphs we can conclude that the given voice is denied as there is more distortion in every graph.

TABULATION:

Test	One	Two	Three	Four	Five
test.wav	38.93	**503.6**	133.9	112.3	35.38
test2.wav	6.313	133.9	**312.7**	71.66	22.02

LIMITATIONS OF THIS MODEL

A speech recognition system is easily disrupted by noises and the way a person speaks. When training and test signals are recorded by the same person and that too with the same accent, this speech recognition system created by us would work well for identifying different words, regardless of who the person is. However, when the training and test signals are recorded by different persons, both these cases or systems would perform before the level of our expectations. In order, to improve the performance in terms of pattern recognition, the system must be made noise immune and the trained voices should have a larger database. This means that they should have certain distinctive features of speech which are recorded by various people.

Nevertheless, we must also take care about the time duration while one speaks in the microphone. The maximum delay in the recording of the speech should not be more than 1 to 2 sec. In case of more delay, it may result in incorrect match and the system might not work as expected. The main reason of this discrepancy in result may depend upon the pronunciation. We can say that both the words "kept" and "slept" ends with the letter "t". We can observe that during pronunciation we are emphasizing on "t" in both cases. But the ending pronunciation of the letter 't' is almost similar. This pronunciation matches in both the cases. However, it would show a lesser lag in the other case and lesser error as well, thus making it as the produced output. These types of similarities can also be treated as a type of noise which is misleading and is responsible for not showing the exact match.

APPLICATIONS

Virtual Assistants:

Many MNC's have developed their own virtual assistant's like Alexa, Cortana etc. The user's voice is analysed each and every time and respective commands are carried out. Voice assistants can be used for
Scheduling meeting
Calling someone
Search for files or documents on our computer

Banks:

Banks and many start-ups have invested in speech recognition technology to identify their customers. Some of them have implemented this in their apps so that the user can transfer money by using voice assistants.
Request information regarding your balance, transactions, and spending without opening your cell phone
Making payments
Receive information about user's transaction history

Voice biometry:

Voice biometry is used by companies who want to offer services based on user's voice. For this they create a digital profile where someone's voice is analysed and then is pitch, intensity, tone and dominant frequencies is identified and stored.

REFERENCES

https://courses.physics.illinois.edu/phys406/sp2017/Student_Projects/Spring16/Josh_Kable_Gilad_Margalit_Physics_406_Final_Report_Sp16.pdf

https://research.ijcaonline.org/volume41/number8/pxc3877646.pdf

http://www.asel.udel.edu/icslp/cdrom/vol2/343/a343.pdf

http://www.inf.ed.ac.uk/teaching/courses/asr/2018-19/asr02-signal-handout.pdf

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.9600&rep=rep1&type=pdf

https://jonathan-hui.medium.com/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9

https://www.ijert.org/analysis-of-voice-recognition-algorithms-using-matlab