Speaker Recognition

Deepak Meghwal, Roll No.: 150108009, Branch: EEE;

Aman Kumar, Roll No.: 150108003, Branch: EEE;

Sanjeev Didel, Roll No.: 150108031, Branch: EEE;

Mahaveer Gahlot, Roll No.: 150108020, Branch: EEE;

Abstract

Speaker Recognition systems have become very important these days because this technique makes it possible to use the speaker's voice to verify their identity and control access to many services. We are using pitch and Power Spectral Density as features, which will be used in Pitch analysis, Formant analysis and Waveform comparison. We have used Pitch and Power Spectral Density(PSD) as features. The motivation behind taking Pitch as feature was it is calculated by taking taking average pitches of all the segments. The average pitch reduced the number of trained files to be compared than formant analysis.
But formant analysis produced more accurate results, compare to pitch analysis, so PSD is taken as feature.

1. Introduction

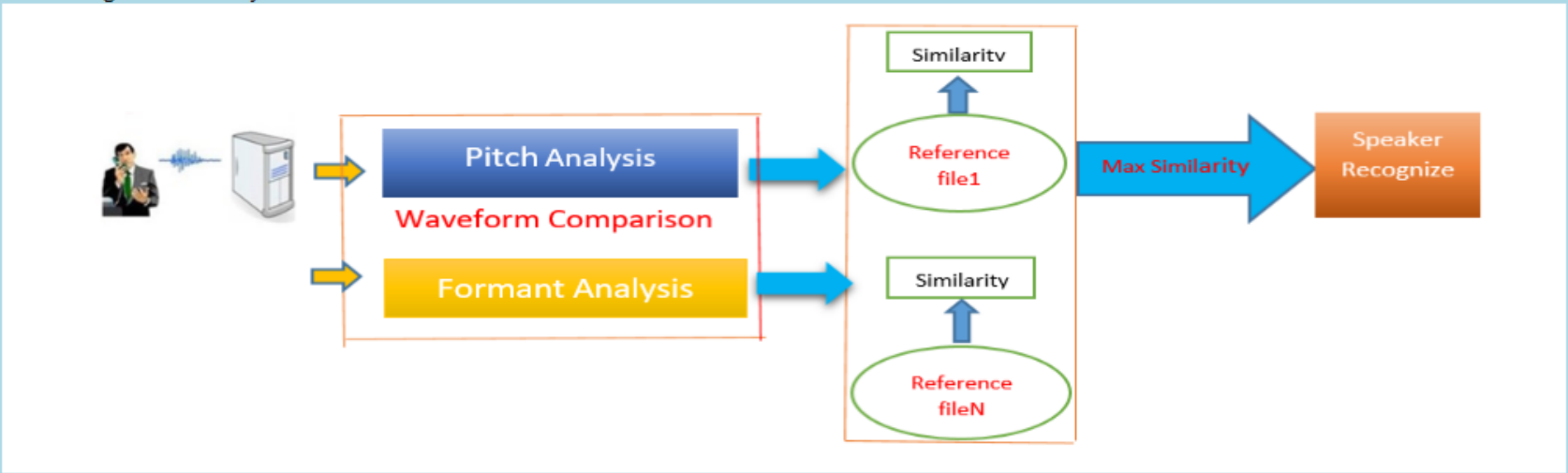
Our project deals with problem of identifying voice of a particular speaker from a given set of sample- voices and arrange them in descending order of resembalance with reference voice, given the speech is same for all the speakers.

1.1 Introduction to Problem

Arranging the all data sample voices in descending order of their matching with the reference voice.It aims to build an algorithm, that can with relative ease, can detect the speaker among the various other speaker.

1.2 Figure

Block diagram of the system->



1.3 Literature Review

1.3 Literature Review

After referring to many voice recognition algorithms we finally decided to do first the pitch analysis of all the given voice samples then on the basis of comparison between the average pitch of the reference voice and all the other voices present in the sample taking few of them for next level analysis that is formant analysis.

Although a lot number of metods for detecting pitch have been proposed but the autocorrelation pitch detector is still the most reliable and robust method of pitch detection.There are several reasons why autocorrelation method for pitch detection have met with great success .One of them is in autocorrelation method computation is made directly on the waveform.Although a high rate of processing is required and hence the process may be very time consuming than other methods.As autocorrelation method is fairly phase insensitive thus this method is very useful in decting the pitch of voices which have suffered some kind of phase distortion while being transmitted or recorded.Although an autocorrelation pitch detector has some advantages for pitch detection, there several problems associated with it.One problem is to decide which of the several autocorrelation peaks correspond to pitch period.

A wide variety of solutions have been proposed for the problems associated with autocorrelation method.Most methods use a sharp cut off low pass filter of about cut off frequency of 900 Hz.Generally Buttrworth filter.This will in general preserve sufficient number of hormonics for accurate pitch detection.

Now focussing our attention to formant analysis used here as the second level identification.It has been known for many years that formant frequencies are important in determining phonetic content of speech sounds.Several investigations has been done on formant frequencies on the prospect of them being used as a voice recognition tool using various methods for basic analysis such as linear prediction,analysis by synthesis with fourier spectra and peak picking on spectrally smoothed spectra.Here we have used peak picking on spectrally smothed spectra.

However using analysis of formant frequencies for voice recognition can cause problems some times for example examining voices which are not very much different on the basis of formant frequencies can not be differentiated on the basis of formant analysis. To be useful for automatic recognition of speech formant frequencies must be supplemented as general spectral shape information.Whenever the spectrum has peaky nature the phonetic details are better described by formant frequencies than by the more usual high order specrum features.Which have no relationship with the formant frequencies. Sometimes a formant may be so weak as a consequence of weak excitation that it may not cause any peak in the spectrum.All these situations can cause all higher frequency formants to be labelled wrongly.

Futher study refer to below links:- 1.Autocorrelation Analysis for Pitch Detection

2. Formant Analysis

1.4 Proposed Approach

We used Waveform Comparison which includes Pitch analysis and formant analysis. Waveform comparison determined based on the comparison between pitch and formant analysis. Reference file is compared with all others based on average pitch. Top 12 matches then compared by the differences on their formant peak vectors.

1.5 Report Organization

We have divided our project into 3 different sections :
(1)Pitch Analysis
(2)Formant Analysis
(3) Waveform Comparison

(1)Pitch Analysis :->
The algorithm used in it as follows, it takes the average pitches of all the voices present in the given sample and compares with the average pitch of the reference voice which is to be recognized

We used audioread to read the reference file and get the sampling rate and the number of samples.Then choose segments every 30ms of the signal and calculate the total no. of segments(nFrames) and F0=zeros(1:nFrames), then choose segment(i), apply butterworth filters on that segments and then autocorrelation method, after that calculate pitch F0(i), we did the same thing for all segments and take average of all pitches to get average pitch.

a)Butterworth filter :-

The Butterworth filter is a type of signal processing filter designed to have as flat a frequency response as possible in the passband. It is also referred to as a maximally flat magnitude filter.As the Butterworth filter is maximally flat, this means that it is designed so that at zero frequency, the first 2n-1 derivatives for the power function with respect to frequency are zero.
Thus it is possible to derive the formula for the Butterworth filter frequency response:

|Vout|/|Vin|^2 = 1/(1+(f/fc)^2n)

where, f = frequency at which calculation is made,
fo = the cut-off frequency(half power or -3dB frequency) ,
Vin = input voltage,
Vout = output voltage,
n = number of elements in the filter

The equation can be re-written to give its more usual format. Here H(jω) is the transfer function and it is assumed the filter has no gain, i.e. it is not an active filter.

H(jw) = 1/(1+(w/wc)^2n)^1/2

Where: H(jω) = transfer function at angular frequency ω
ω = angular frequency and is equal to 2πf
ωo = cutoff frequency expressed as an angular value and is equal to 2πfo

When wanting to express the loss of the Butterworth filter at any point, the Butterworth formula below can be used. This gives the attenuation in decibels at any point.

Adb = 10log(1+(w/wc)^2n)

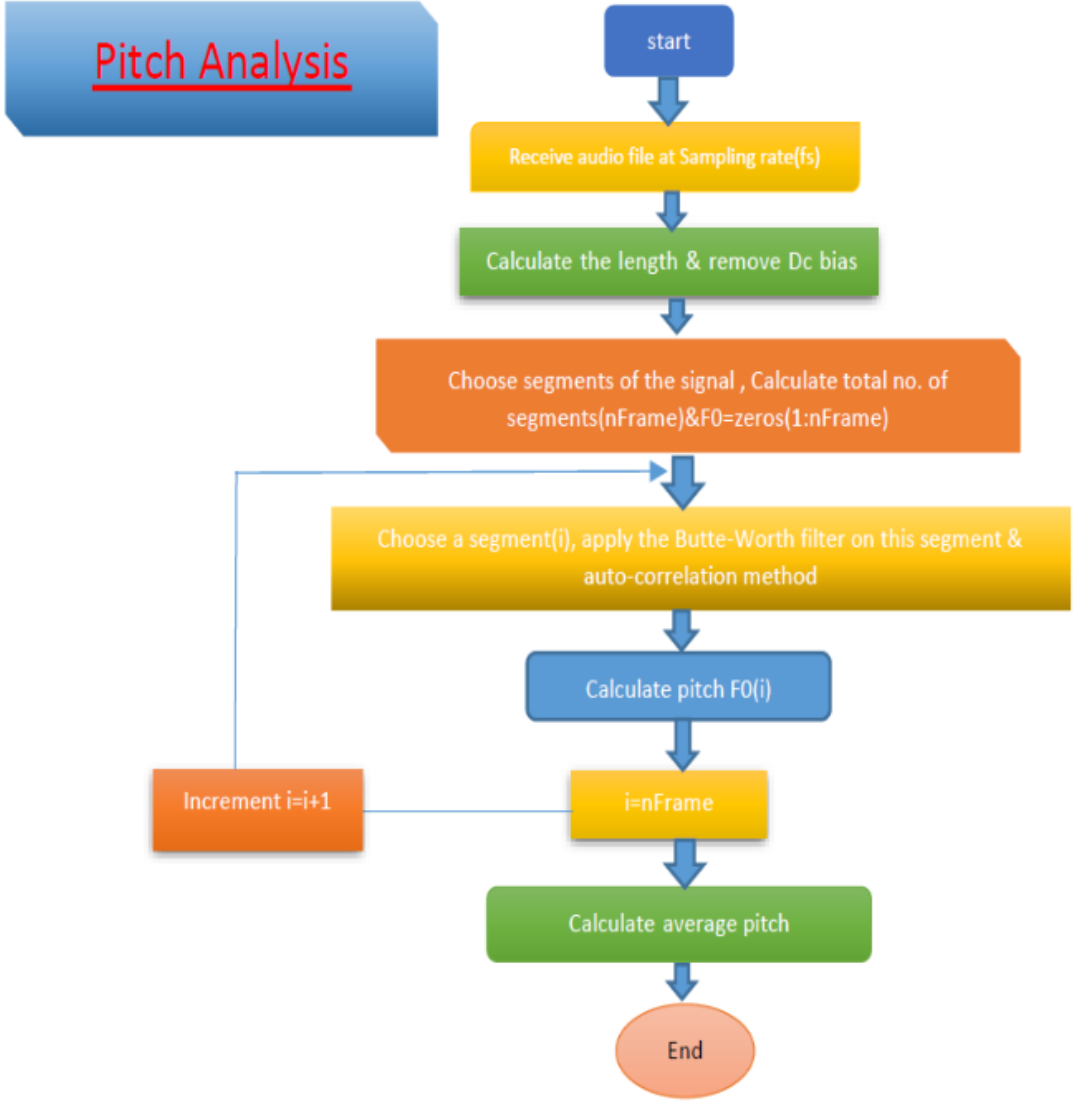
(b)Auto-correlation :-

"Autocorrelation" is used to compare a signal with a time-delayed version of itself. If a signal is periodic, then the signal will be perfectly correlated with a version of itself if the time-delay is an integer number of periods. That fact, along with related experiments, has implicated autocorrelation as a potentially important part of signal processing in human hearing.

Mathematically, for a continuous signal, s(t), the autocorrelation, R(τ) is calculated using:

$R(\tau) = 1/(t_{max} - t_{min}) \int s(t)s(t-\tau)dt$

Pitch Analysis



(2)Formant Analysis->

Formant is defined as the spectral peaks of the sound spectrum and in formant analysis PSD(power spectral density) of each voice of the sample and reference voice is calculated, considering only 3-4 peaks, their positions and peak differences is calculated, if these quantities matched with the reference signal, then the reference signal is identified in the sample.We read the reference file, get sampling rate and sampled data.Then applied Yule Walker method for PSD(Power Spectral Density) P, then convert P to db and calculate normalized frequency axis, calculate difference between consequece.

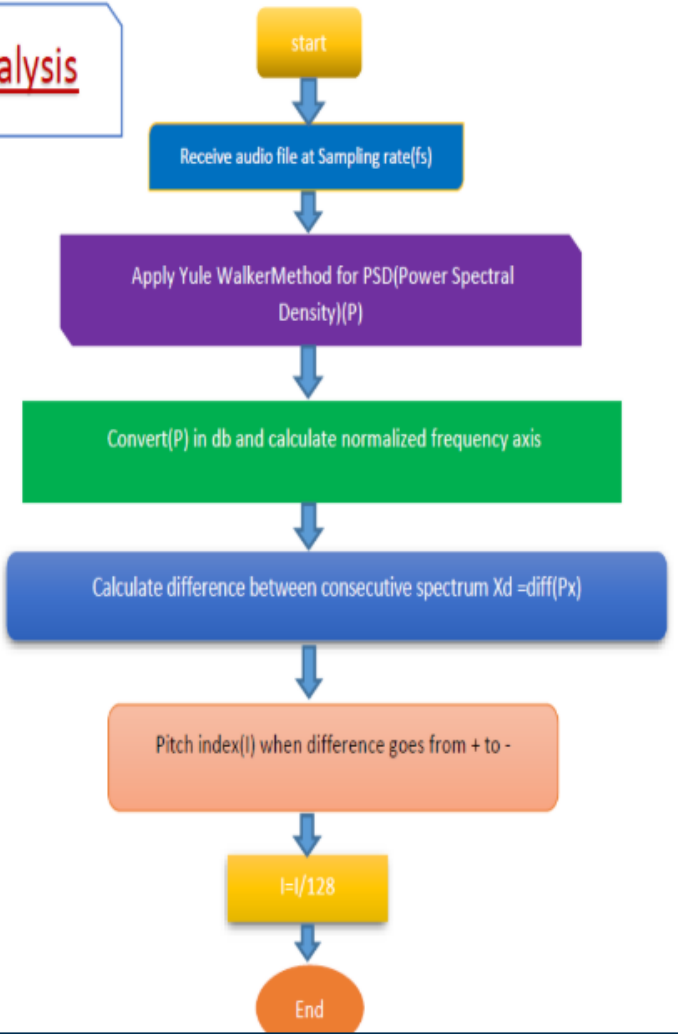
Yule-Walker Method:-The Yule-Walker Method block estimates the power spectral density (PSD) of the input using the Yule-Walker AR method. This method, also called the autocorrelation method, fits an autoregressive (AR) model to the windowed input data. It does so by minimizing the forward prediction error in the least squares sense. This formulation leads to the Yule-Walker equations, which the Levinson-Durbin recursion solves. Block outputs are always nonsingular.

The input must be a column vector. This input represents a frame of consecutive time samples from a single-channel signal. The block outputs a column vector containing the estimate of the power spectral density of the signal at Nfft equally spaced frequency points. The frequency points are in the range [0,Fs), where Fs is the sampling frequency of the signal.When you select Inherit estimation order from input dimensions, the order of the all-pole model is one less that the input frame size. Otherwise, the Estimation order parameter value specifies the order. To guarantee a valid output, the Estimation order parameter must be less than or equal to half the input vector length. The block computes the spectrum from the FFT of the estimated AR model parameters.

Selecting the Inherit FFT length from estimation order parameter specifies that Nfft is one greater than the estimation order. Clearing the Inherit FFT length from estimation order check box allows you to use the FFT length parameter to specify Nfft as a power of 2. The block zero-pads or wraps the input to Nfft before computing the FFT.When you select the Inherit sample time from input check box, the block computes the frequency data from the sample period of the input signal. For the block to produce valid output, the following conditions must hold:

(1)The input to the block is the original signal, with no samples added or deleted (by insertion of zeros, for example).
(2)The sample period of the time-domain signal in the simulation equals the sample period of the original time series.

Formant Analysis



2. Proposed Approach

We have studied three techniques:
(1)Pitch Analysis
(2)Formant Analysis
(3) Waveform Comparison
(1)Pitch Analysis-> The algorithm used in it as follows, it takes the average pitches of all the voices present in the given sample and compares with the average pitch of the reference voice which is to be recognized. we used audioread to read the reference file and get the sampling rate and the number of samples.Then choose segments every 30ms of the signal and calculate the total no. of segments(nFrames) and F0=zeros(1:nFrames), then choose segment(i), apply butterworth filters on that segments and then autocorrelation method, after that calculate pitch F0(i), we did the same thing for all segments and take average of all pitches to get average pitch.
(2)Formant Analysis->Formant is defined as the spectral peaks of the sound spectrum and in formant analysis PSD(power spectral density) of each voice of the sample and reference voice is calculated, considering only 3-4 peaks, their positions and peak differences is calculated, if these quantities matched with the reference signal, then the reference signal is identified in the sample.We read the reference file, get sampling rate and sampled data.Then applied Yule Walker method for PSD(Power Spectral Density) P, then convert P to db and calculate normalized frequency axis, calculate difference between consequtive spectrum Xd= diff(P), then we pick the index(I), when difference goes from + to -(at this index there will be maxima), and by these differences with reference file the speaker is identified in the sample data.
Now we use Waveform comparison which involves above both technique. Waveform comparison determined based on the comparison between pitch and formant analysis. Reference file is compared with all others based on average pitch. Top 12 matches then compared by the differences on their formant peak vectors.

3. Experiments & Results

3.1 Dataset Description

We recorded around 90 voices of all students of our class and used it as data set, refrence voice is taken any of these files. We recorded two different text data sets each of 45 persons. The link of both data-set is given below:

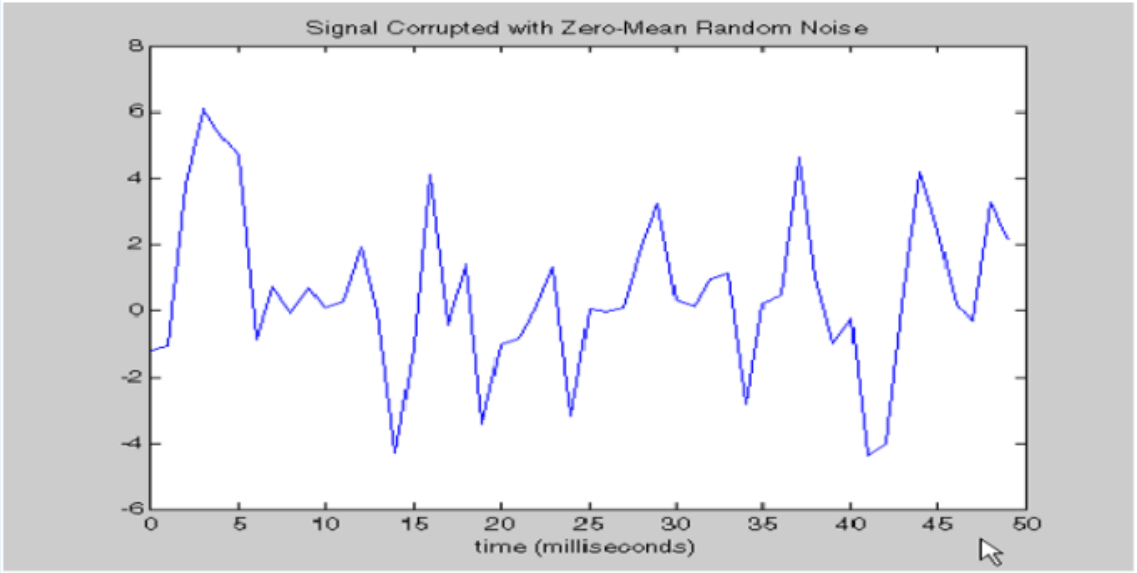
Visit our First Data-set

Visit our Second Data-set

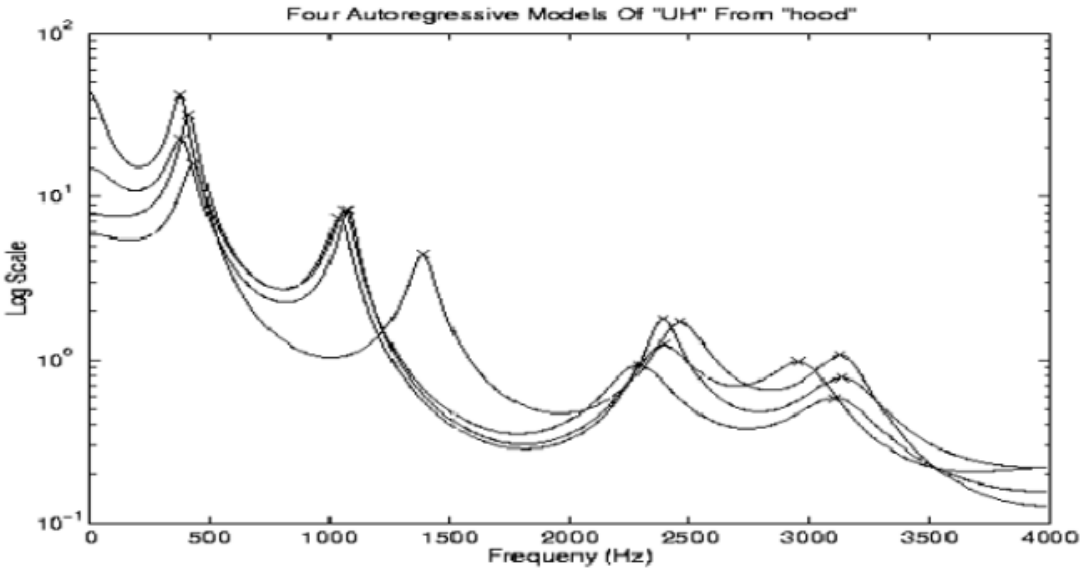
Link of All Codes

Graphs Output->

(1) Pitch Analysis



(2) Formant Analysis



3.2 Discussion

As a result we get the file arranged in descending order of their matching with given file.If we do text dependent speaker identification, then this thing can be done by only pitch analysis, since if we get the same pitch of any sample file with reference file, then we identified the speaker, else reference file is not present in data-set. Some time results where not consistance bacause of the reason discuss in literature review .

4. Conclusions

4.1 Summary

While starting this project of voice recognition we went thorugh some voice recognition algorithms and we found the algorithm of pitch analysis using autocorrelation of sample voices a useful tool do the job and hence we used it as a first level of processing of the sample voices we where do conduct the our project. Now we have some of the voices which had avarage pitch closest to the reference voice , then on these selected voices we perform the comparison of peaks of their spectrum of formant frequency components based on both these test . We arranged the given sample voices on the basis of their resembalance with the reference voice .

4.2 Future Extensions

(1) Speaker Recognition can be used for Security purposes, Control access.
(2) By recognizing particular person's voice,we can also find the involvement of this person in meeting or in debate. Like in a Meeting, we can record all the voice data that the head person said.
(3) In Singer Replacement or actor voice replacement can be done by our technique, as we are rearranging the all data voices in descending order of their matching with the referenc voice.
(4) So after removing noise from song, singer identification, Commercial Detection from Sound Tracks can be done easily.