International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 5, May - 2014

Speaker Recognition System using Coefficients and Correlation Approaches in MATLAB

Mr. Sridhar Chandramohan Iyer
Department Of Computer Engineering
St. John College Of Engg. And Tech.
Palghar , Thane

***Abstract*— Speaker Recognition is the methodology through which the deliverer of the speech , diction or audio is detected. The process consists of identifying a particular speaker from a pool of speakers whose voice samples are already been stored inside a data store.**

The main objective of the system is to verify the speaker by identifying the speaker's own voice. To achieve this, the speaker's voice sample is compared with the pre-recorded samples already stored in the database prior to the verification process. The power spectrum density of the speaker's voice is compared with the pre-sampled audio stored in the database and if it matches with the same, then the particular speaker's name is displayed and the same is authenticated. This system can be used to operate multimedia devices or gadgets for personalized usage.

Keywords—Speaker Recognition; Speaker Identification; Speaker Verification;

I. INTRODUCTION

Speaker recognition can be easily confused with voice recognition. Voice recognition is to identify WHAT is being spoken or uttered, whereas speaker recognition is to identify WHO is delivering the speech or audio. Speaker recognition can be broadly classified into identification and verification. It is a process of identifying and then verifying who is speaking on the basis of patterns extracted from his or her own voice. These patterns may be the features such as the pitch, amplitude , frequency variations etc present in the voice.

These features are then stored against the specific speaker in the database . These features are then compared with the current speaker's voice sample to check if any correlation is found or not . If any correlation is found , then the best matching voice sample out of the database is selected and displayed.

This system has a lot of scope for modules involving authentication , which can replace more expensive biometric authentication systems. This system can be further used to provide access control for confidential data where a proper password mechanism is not sufficient enough. Speaker Recognition technologies can be used for daily applications such as a voice controlled remote system , a voice controlled media player ,etc.

The overall recognition process consists of Identification followed by verification. The identification phase consists of a one-to-many comparison to check whether there is any entry available in the database for the corresponding input audio sample.

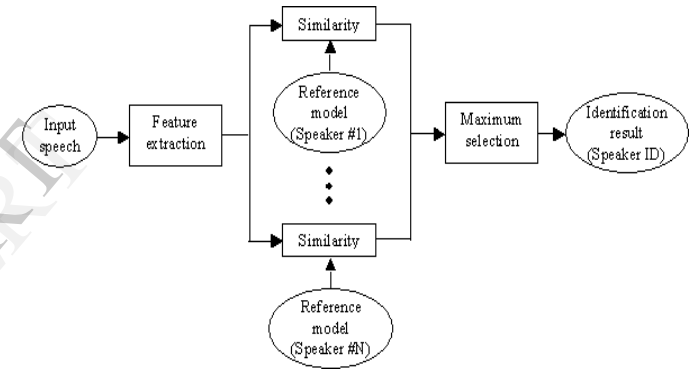

Fig 1: Identification Process [1]

The identification phase is followed by the verification phase in which it is checked whether a successfully identified voice sample is authentic or not i.e whether it belongs to the same speaker or not .

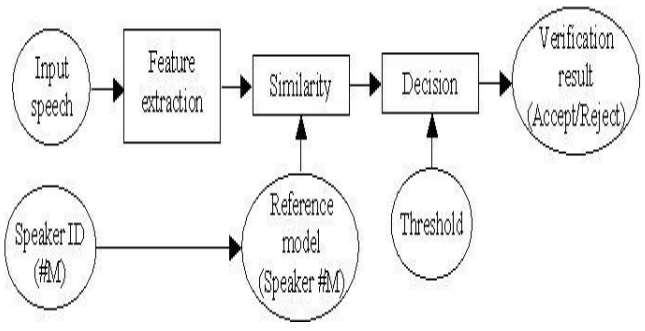

Fig 2 : Verification Process [1]

II. PROPOSED SYSTEM

Speaker recognition systems comprises of two rigorous phases: Enrollment and Test. Enrollment consists of recording a number of feature patterns from the derived voice print or template of the sample. In the test phase , the speaker's voice is matched with the templates or voice models .

A typical voice print and a corresponding database entry is as shown:

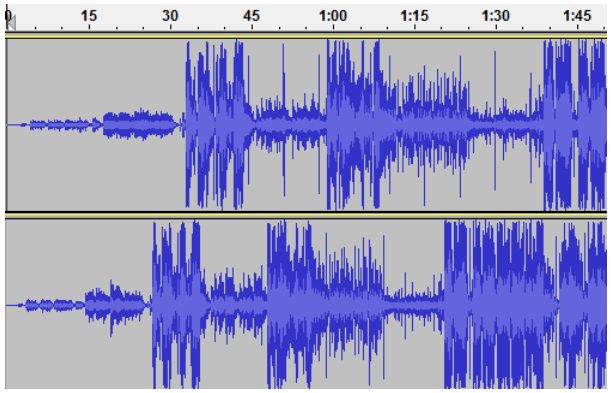

Fig 3 : Digital Voice-Print

Speaker Recognition Systems use three styles of spoken audio sample : text-dependent, text-prompted and text-independent.Text-dependent systems instructs the speaker to speak exactly the same text sample which was used during the enrollment phase or else the user will not be authenticated. A high level of accuracy is achieved in this as the text to be spoken is always fixed. Text-prompted systems prompts to the speaker a piece of text which he or she should speak into the system. This could be any kind of a text. The System selects a subset out of the larger voice sample and prompts the user to re-enact the selected textual part.

Text-Independent Systems do not consider any textual restrictions for the recognition process. It basically considers the variations in the voice samples based on the variations in the pitch , frequency , amplitude , etc and compares it with the pre-recorded sample to carry out the recognition.The System under consideration uses a combination of both text-independent and text-dependent approach for speaker recognition.

III. ALGORITHM

The Speaker recognition algorithm begins with dividing the input voice sample and the pre-recorded database sample into arrays of a fraction equivalent to $(1/1000)^{th}$ of the original sample. These arrays represent their respective voice signals. The algorithm can be broadly divided into the following four phases.

- A function - SoundSignal is used to convert the voice signals into an array.
- A function – SumCoefficient is used to calculate the sum coefficient of the voice array.
- A function – Correlation is used to calculate the correlation coefficient
- Selecting and matching the database entry.

IV. WORKING OF THE ALGORITHM

A . Converting a voice signal into an array (***SoundSignal*** *)*

This function is used to convert the voice sample into its corresponding frequency domain representation. This can be achieved using MATLAB's inbuilt function fft(Fast Fourier Transform).A real time voice signal can be converted into its fft representation using Y=fft(X) where fft(X) will return the Discrete Fourier transform (DFT) of the signal X .

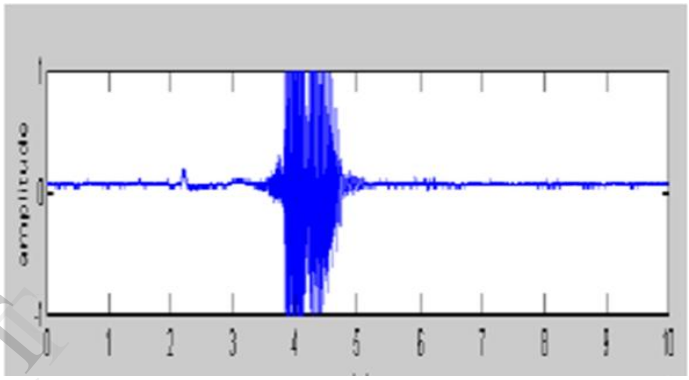

Fig:4 Real Time Voice Signal in time domain

The main requirement of performing this step is to convert the signal into a 100000- point DFT .The next step is to achieve the power spectral density of the signal, which can be obtained by multiplying each DFT point with its conjugate.

As power = Square of amplitude.

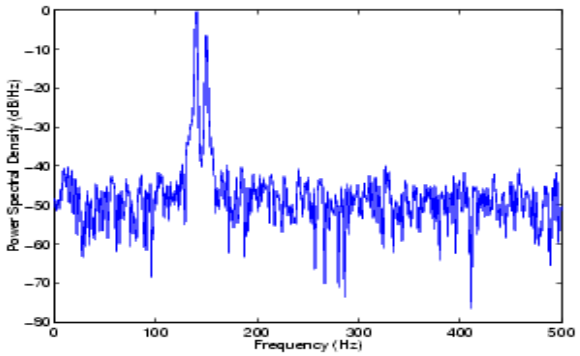

Fig 5:Power Spectrum Density of real Time Voice Signa[2]

The above figure represents the power spectral density graph of the real time voice signal, but this does not provide a conclusive result as the DFT obtained is divided into

100000 sample points but the actual voice sample is divided into voice arrays of size 1000 points. Hence we again need to convert this spectral density into an equivalent form using which we can carry out the further processing of the interim signal. To obtain this , the actual 100000 point spectrum is converted into a 1000 point modified power spectrum(MPS).This is summarized as the 1st point of the modified spectrum represents the sum of the powers of the first 100 points and so on. In general we can say that for the ith point , the MPS will represent the sum of the powers of the (i-1)*100 to (i*100).This 1000 point MPS voice signal is known as the voice array.

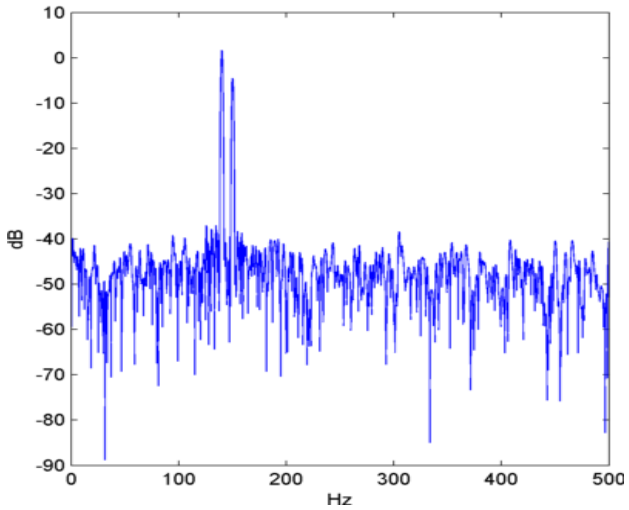

Fig 6: Modified Power Spectrum Density of Real time voice Signals [3]

B. *Calculating the **SumCoefficient***

The sum coefficient can be calculated as :[5]
$S=\sum X(i)*Y(i)$

Where X and Y are the voice arrays to be compared for correlation.

If we say that X and Y are two voice samples belonging to the same person , then it clearly indicates that the value of the sum coefficient will be maximum for that person as compared to others , but we cannot come to any conclusion whether the voice samples does not belong to the same person or not because we do not have any lower bound threshold value below which we can definitely say that it is a false positive.

Hence to achieve more accuracy, we have implemented another approach such as the Correlation Coefficient

C. *Calculating **Correlation Coefficient***

It is another method which is used to find out the similarity between any two signals. It can be expressed as : [4]

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{Sx}\right)\left(\frac{y-\bar{y}}{Sy}\right)$$

Where X and Y are the voice samples to be compared.
Sx and Sy are the standard deviations
n is the number of data which is 1000 in our case.

After testing this program in various conditions such as a closed room , a noisy room , outdoors , we came to the conclusion that the value of r should be around 0.5 or 0.6 which is an optimum value for carrying out the correlation coefficient . i.e. if the correlation coefficient is over 0.5 it is an indication that the two voice samples belong to the same person , else not. Hence if the value of r is set too low , then it will give false positives , on the other hand if it is set too high , it will give false negatives.

D. *Selecting and matching the database entry.*

The steps for selecting , retrieving and matching of the database entry.

1. Record the real time voice sample, say a.
2. Let b be the database voice sample.
3. Apply SoundSignal to a.
4. Apply SoundSignal to b.
5. $S=\sum a.b.$
6. Apply Correlation to (a,b)
7. Select b with the maximum value of S.
8. If r<0.5 → No Match found
9. Else Match Found

V. FACTORS AFFECTING ACCURACY OF THE PROGRAM

Since we have the value of r which has been set to an optimum value of 0.5 , here we will have to compromise with the efficiency of the program . If set too low , then it will match almost all the voice samples , even if such a voice sample is not present in the database , It will match it with any random entry. On the other hand , if set to too high , it will reject even a correct voice sample.

If we consider the external disturbances or any noise signal is added , then it will add up to the unwanted power signals , which will ultimately raise up the power spectral density peaks , as a result unwanted results will creep in. Due to this , one person in the database may be recognized as someone else.

VI. SUCCESS RATE

We have tested this program under various conditions . We tried both the text-dependent as well as text-independent methods and came up with the following results.

For text-independent	70%
For text-dependent	80%
Total Database Entry	15

VII. CONCLUSION

In this paper we have implemented a MATLAB based speaker recognition system which uses statistical approaches such as Coefficient and Correlation to identify and verify the speaker based on his or her own voice . The testing phase comprised of testing the program under various conditions to find out the optimum threshold value for comparison and the success rate of the program under various conditions. It was concluded that text-dependent version was more efficient as compared to the text-independent one.

VIII. REFERENCES

[1]"Speaker Identification and Verification"
http://www.cslu.ogi.edu/HLTsurvey/img151.gif
[2]"Power Spectral Density"
http://radio.feld.cvut.cz/matlab/toolbox/signal/pmtm_hir.gif
[3]"Modified Power Spectral Density"
http://www.mathworks.com/help/signal/ug/periodogram_psd.png.
[4]"Correlation"
http://www.stat.yale.edu/Courses/1997-98/101/correl.html
[5]Prof.Kumbhojkar's,"Applied Mathematics-5",
published byC.Jamnadas&Co.