Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Speaker Recognition and Fast Fourier Transform

Nilu Singh, R. A. Khan
SIST-DIT, Babasaheb Bhimrao
Ambedkar University (Central University),
Lucknow, UP, India

Abstract*— This paper makes available a concise review of to present analysis of speech signal and Fourier transform to representation, for Speaker recognition Technology. Fast Fourier transform used to find the dissimilarity among speakers and speech signals make use of to validate the effectiveness of present methods. Voice have individual characteristic for every human, hence to recognize the human by their voice is a useful technique in recent time and this technique come in to the category of biometric and called Speaker Recognition technology. Spectrum analysis of a 'speech signal' is the method of shaping the 'frequency domain representation' of a time domain signal and this method usually known as Fourier transform. The Discrete Fourier Transform (DFT) is used to find out the frequency content of analog signals. And Fast Fourier Transform (FFT) is a competent process for calculating the DFT.

***Keywords*— Speaker recognition, Fast Fourier Transform (FFT), Frame, Window, Discrete Fourier Transform.**

I. INTRODUCTION

For human being voice is the most natural method to share their thoughts and information to each other. A speech signal contains a number of required information about the person. Speaker recognition is a method of identifying a people. A speech signal holds the information of linguistic communication i.e. a speech signal provide the scientific study of language. Voice/speech signal produced by acoustically exciting cavities of mouth as well as nose and can be used to recognizing the person. The most important examination about Automatic speaker recognitions (ASR), that voice produced naturally. Additional benefits of this technology that, it is not expensive because it does not need any individual equipment. There is required only a head phone to capture a speech signal also the algorithms used for speaker recognition are low cost in addition to memory efficient such as Signal processing and pattern matching algorithms [1]. Now day's security systems require improvements in latest technology at various fields such as communications, banking, networking etc. there are some features of human being which is dissimilar and distinctive of each person such as voice, facial expression, Fingerprints, DNA etc. and these features not possible to reproduction and also univocally authorizes a person [2].

There are several measurements have been proposed and investigated for biometric recognition system such as fingerprint, face, iris, voice etc. amongst the most popular are voice, fingerprint and face [3], [4]. These recognition techniques have their pros and cons in terms of accuracy and user acceptance or deployment. The motive to use voice as biometric which make it convincing are – first since voice produce naturally hence voice called natural signal and second telephone system, which is may be everywhere. Since speech signal are natural hence is not measured unapproachable to provide by users. In many applications for users no need to providing a voice sample for authentication such as for telephone transaction. In case of telephone based applications special signals network/transducers not required at application access point. For non telephonic applications we can use sound cards and microphones at applications access point. FFT is a usual technique to evaluate frequency spectrum of the signal in speech and speaker recognition both. And evaluation of spectrum is an essential operation. For digital signal processing FFT is the fundamental technique relevant for spectrum analysis. FFT is normally used to calculate numerical approximations towards continuous Fourier. For speech signal analysis a transform is used known Fast Fourier Transform, which provides average representation of a signal in the frequency domain. Whereas Short Fourier Transform (SFT) is used to take time frequency changes. FFT, DWT etc. are easy and fast and all used to compute speech spectra [5], [2]. Due to containing a number of parameters in a speech signal make available to recognizing a speaker. It is need of a speech signal that it should be analyzed in a precise way and also the proper representation is required. To make information more visible it is need that use some transform technique because the original representation of a speech signal in the time domain typically provide a few information about the speech signal. Since human voice is time varying due to frequency properties of a speech/voice signal, speech signal enduringly changes by continuous reconfiguration of human vocal tract and resonant chamber [6].

II. SPEAKER RECOGNITION TECHNOLOGY

There are some features required for a good biometric technique such as easy to measure, easy to extract, easy store and compare. And speech signal/voiceprint fulfills all the required features and also does not need very expensive hardware or infrastructure, it just need a microphone for voice recording. Hence the conclusion is that voice is the appropriate biometric technology. For human numerous physical characteristics of speech vary a lot from one to another such as tone or voice intensity, timber, speaking rate, intonation etc. voice is a phenomenon so as to be extremely dependent on speaker. A speech signal contains a lot of properties about the speaker that's why it is an important biometric technique to be used in security systems. Another quality of this technology is, since speech features are easy to measure as compare to other biometric technique [2].

Speaker recognition systems have two main component first is feature extraction and second is feature matching. Feature extraction is the procedure of extracting a small quantity of data from the speech/voice signal that extracted features preserve and is used later for representation of every speaker. Feature matching contains the procedure to classify the strange speaker by comparing the extracted features from their voice input to everyone from a set of known speaker's voice database. After the capturing of speech signal by a microphone a sound signal can be transformed into electrical current Also each voice signal is represented in a waveform. And now continuous oscillations of air pressure turn out to be continuous oscillations of voltage inside an electrical circuit. After that this voltage is changed into a series of numbers using a digitizer. Since digitizer operate like a very speedy digital meter and hence it makes thousands of measurements per second. Now this measurement of speech signal can be stored digitally and this number is called sample of the speech signal and the complete conversion of sound wave is known as sampling. The numbers range depends on the sampling bit-rate such as 16-bit, 8-bit etc [3], [5].

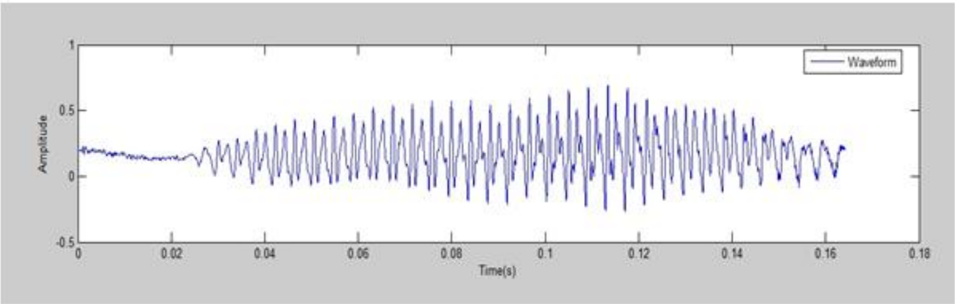

Fig.1. Example of Speech signal

For speaker recognition system there are two phases first is training phase or enrolment and second is testing phase. In case of training period a speaker need to provide an utterance or sample of speech using this utterance system can build a statistical model for that speaker. In case of testing phase the input utterance is matched with models from database and after that make a decision that speaker is recognized or not. An utterance provide by speaker at the time of training is different that the utterance provide at the time of testing because voice is time varying i.e. it change with time. Voice also affected by health condition speaking style, speaking rate, recording environment, channel mismatch etc [3][4].

The principle of ASR system is to be efficiently make available, individual properties of every speaker which is accurate and distinguishable to each other [2].Speaker recognition have two classification First is Speaker identification and second is speaker verification. As discussed in [6] Speaker identification is the process of determining the identity of an unknown speaker by comparing the voice of that unknown speaker from the voice database of speakers it is also entitle one-to-many (1:n) comparison i.e. in this case the purpose is to decide which one of a group of known voices most excellent matches with the input voice samples. While speaker verification is a method of finding whether a speaker is who claims to be and is entitle one-to –one (1:1) comparison [c] i.e. result only be yes or no. Speaker recognition again classified as text dependent and text- independent based on the speech used by the system. Text dependent systems are those that use the same text or word spoken by the speaker in both phases i.e. training and testing phase i.e. speech is constrained. While in case of text- independent system, use different text or word for training and testing phase i.e. speech is natural. The example of text dependent systems such as in case of access control verification application a applicant always make use of the similar personalized cryptogram. But text dependent systems are not reliable because if the applicant personalized cryptogram is somehow recorded then playing it uses to achieve access that system. While text- independent systems are more reliable and capable of being changed. Example of text-independent systems is voice mail retrieval [7].

A. Frame blocking

For doing well spectral analysis it is the requirement that the 'selection of frame length'. Assortment of frame length is an essential parametric quantity. The size of window should be sufficient for frequency resolution. Normally frame length of a speech signal is 10-30 milliseconds are used for recognition process. Frames are overlapped by each other and around 30% to 50% of the frame size is overlapped by neighboring frames [19].

Here continuous speech signal is blocked into frames of N samples and adjacent frames being separated by M, where M < N. first frame of speech signal consists of the first N samples, the second frame started from M samples after the previous frame and it overlaps by N – M samples and so on. This method continuous proceeds until the whole speech signal is blocked within a frame. Normally window size of a speech signal is 30 ms [4].

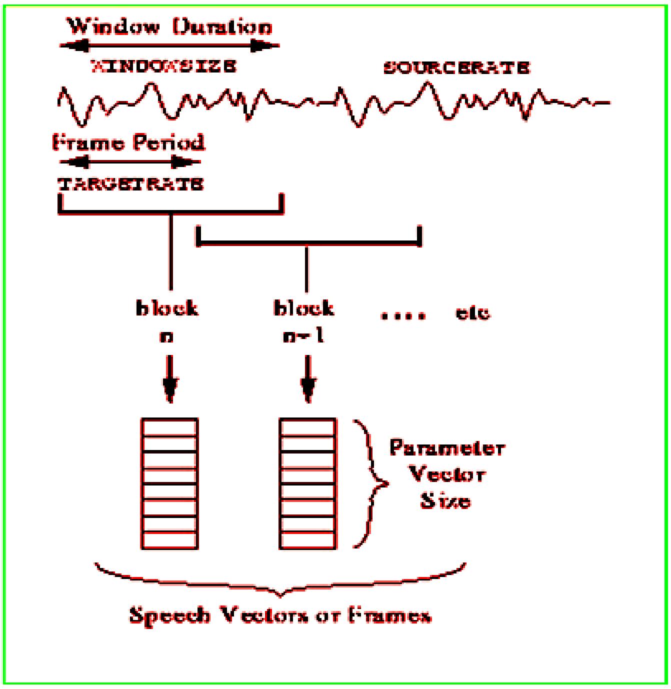

Fig 2. Process of Speech Encoding

B. Windowing

Windows of a speech signal is an investigation of the component parts of FFT i.e. windows are frequency weighting functions which is apply to in time domain data to reduce the spectral outflow associated with finite duration time signals. Window of a speech signal are smooth out functions that crest in the middle frequencies and decrease to zero at the edges, thus to reduce the discontinuity give result of finite duration [8].

Hamming window is used normally in speech analysis to reduce the unexpected changes and undesirable frequencies which is occurring in the framed speech signal. It is defined as [5]:

$$w(k) = 0.54 - 0.46 \cos \frac{2\pi k}{L-1} \qquad (1)$$

The significant windowed segment is-

$$x(k) = S_n\, w(k) \qquad (2)$$

Where-

L= width of S_n.
k= is an integer, 0 k L-1.
S_n = signal function.
w(k) = window function.

C. Frequency Analysis

In general signals are come from measurement acquired with a selected sampling interval Δt while generally not described by mathematical functions. Hence signals are not continuous in nature but discrete. Discrete signals also produced by using simulation models (Matlab Simulink). Generally the duration of a signal is finite in nature also the majority cases it will not be the equal as the period requisite by the Fourier Theorem. A signal contains different frequency (f) components to be identify therefore selection of sampling frequency (f_s) is randomly. The maximum measurable frequency is the 'frequency equal to half of the sampling frequency' and it is entitled the Nyquist or folding frequency. If sampling frequency is low than the frequency i.e. $f_s <= 2f$, then a lower incorrect frequency will be supposed. And this phenomenon is called the aliasing. To avoid the aliasing most efficient approach is by filtering it with a low pass filter through the cut off frequency less than half of the sampling rate [9].

Voice signal have a very complex waveform because of the superposition of various frequency components in the speech signal. For speech recognition and speaker recognition, to determine a representation with the purpose of extracting information from speech signal is an important problem. A speech signal incurred two types of information i.e. a signal can be represented as time domain as well as frequency domain. In time domain sharp variations in signal amplitude are generally a good number of meaningful features. In frequency domain, dominant frequency channels of speech signal are located in the middle frequency region and in this each speakers may have different responses in all frequency regions. The usually methods which consider fixed frequency channels possibly will lose some needed information at the process of feature extraction. For that reason use the multi-resolution decomposing technique using this, speech signal decompose into different resolution levels. The features of multiple frequency channels and any alteration in the smoothness of the signal after that are detected to completely represent the signals [10].

III. FOURIER TRANSFORM

For speech signals analysis the ordinary transform used called the Fast Fourier transform (FFT). FFT provides the standard representation of a speech signal in the frequency domain. While Short Fourier transform be able to hold time frequency changes. The drawback of FFT that it is not appropriate for the signals whose frequencies are time varying hence in case of FFT is assumes that the signals are stationary in nature [11]. The FFT allows working in frequency domain and therefore using the frequency spectrum of the speech signal as a substitute of waveform. Frequency domain provides more information about the speech signal and hence can be more efficient to distinguish between speakers. For speaker recognition some techniques use the vice signal acquire directly by the sampling phase and some techniques use transformed form of the speech signal [3].

When a speech signal represented in time domain then it gives a little information regarding speech signal properties hence suitable transformation of speech signal is an essential problem. For this purpose generally Fourier or wavelet transform are used [2]. Speech signal/audio processing techniques begin by converting the raw speech into a sequence of acoustic feature vectors carrying features of the signal. And this is known as pre-processing i.e. feature extraction is completed here and is also called front-end processing. Mel Frequency Cepstral Coefficients (MFCC) is the most usable acoustic vectors and MFCC features are derived from the FFT power spectrum. The acoustic features are based on the spectral information which is derived from 'a small time window segment' of a speech signal [12].

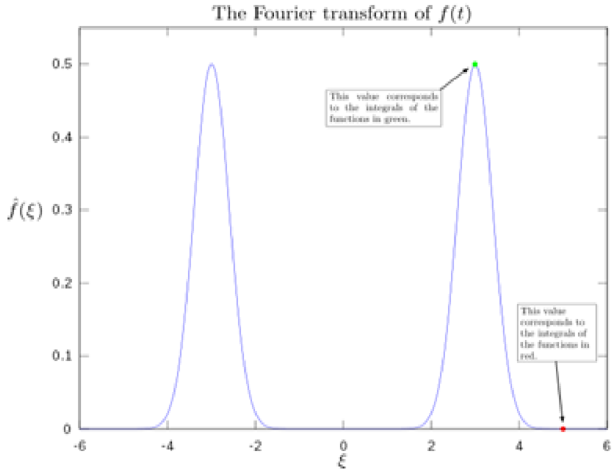

Fig.3. Fourier Transform of oscillating function

Fourier Transform is a mathematical transformation which is use to transform a signal among time domain and frequency domain. It gives the facility to reversible i.e. from one domain to other. Using Fourier transform a periodic function over time is able to simplify to the computation of a discrete set of complex amplitudes and this is called Fourier series coefficients. When a time domain function is sampled for computer processing it is possible to reconstruct the original Fourier transform as per Poisson summation formula this is also known as Discrete Fourier transform (DFT) [13].

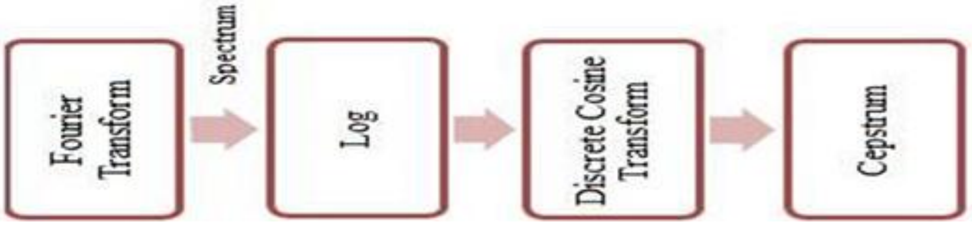

Fig. 4. Conversion of Signal to Cepstrum using FFT

When transform applied in a speech signal it converts it in to frequency domain from time domain. Let y(t) be the speech signal in the time domain and y0, y1, y2…………..yN-1 be the samples of speech signal y(t) in the time domain. The DFT is implemented by using FFT.

$$Y_k = \sum_{n=0}^{N-1} Yn \;\; e^{-f2\pi kn/N} \qquad (3)$$

Where yn =y(nΔt), it is the sampled value of continuous signal y(t). Where k=0, 1, 2………………N-1. And Δt is the sampling interval [14].

FFT also use to increase the speed of computation time. And the sampling value of each speech signal frame should be limited in 2^n times because of some limitation to the FFT [14]. To reduce the computation time, FFT have the benefit of the properties of symmetry and periodicity of the Fourier Transform. FFT is a complex transform due to have performance restrictions in the method and it operate on an imaginary number and special algorithm. FFT has a complex exponential so as to define a complex sinusoid with frequency also it has unchangeable [5].

Discrete Fourier Transform (DFT) The purpose of frequency analysis is to devise a method to extract an estimate of frequency components which are not known a priori. The process is known as the Discrete Fourier Transform [9].

IV. CONCLUSION

The function of transform to speech signal is not only to take out frequency information from a speech signal however also carry on the individual properties of each speaker. For Speaker and speech recognition 'frequency method' is a useful tool of the speech signal analysis. Using MFCCs for speaker identification process, it is not robust in case of noise and telephone degradation. Therefore the feature extraction from the wavelet transform of the degraded attaches more speech features from the estimation and detail components of these signals. And these signals support to achieving higher identification rate.

ACKNOWLEDGEMENT

This work is sponsored by the CST-UP, Lucknow, India, under CST/D-413.

REFERENCES

[1] Kinnunen, Tomi. "Spectral Features for Automatic Text-Independent Speaker Recognition." University of Joensuu Department of Computer Science P.O. Box 111, FIN-80101 Joensuu, Finland. (2003): 1-151. Print.

[2] Ziotko, Bartosz, , et al. "Hybrid Wavelet-Fourier-HMM Speaker Recognition." Department of Electronics, AGH University of Science and Technology Krak. n. page. Print.

[3] singh, Nilu. "A Study on Speech and Speaker Recognition Technology and its Challenges." proceedings of National Conference on Information Security Challenges, DIT, BBAU, Lucknow, INDIA. lucknow: Bharat Book Center, 2014. 34-37. Print.

[4] Bimbot, Frederic, , et al. "A Tutorial on Text-Independent Speaker Verification." EURASIP Journal on Applied Signal Processing. 4. (2004): 430–451. Print.

[5] Ernawan, Ferda, and Nanna Suryana. "SPECTRUM ANALYSIS OF SPEECH RECOGNITION VIA DISCRETE TCHEBICHEF TRANSFORM."International Conference on Graphic and Image Processing (ICGIP 2011),. 8285. (2011): 1-8. Print.

[6] Ziolko, Mariusz, , et al. "WAVELET-FOURIER ANALYSIS FOR SPEAKER RECOGNITION."Department of Electronics, AGH University of Science and Technology, Kraków, Poland al. Mickiewicza 30, 30-059 Kraków. 1-6. Print.

[7] Faraoun, K. M. , and A. Boukelif. "Artificial Immune Systems for text-dependent speaker recognition."Evolutionary Engineering and Distributed Information Département d'informatique, Djillali Liabès University. Systems Laboratory, EEDIS -SBA– Algeria. (2006): 1-8. Print.

[8] N. Do, Minh . "How to Build an Automatic Speaker Recognition System." 1-11. Print.

[9] Ulrike Schild, Angelika B.C. Becker and Claudia K. Friedrich, "Phoneme-free prosodic representations are involved in pre-lexical and lexical neurobiological mechanisms underlying spoken word processing" Elsevier Brain & Language, vol- 136, year 2014, pp. 31–43

[10] GRUBESA1, SANJA, et al. "SPEAKER RECOGNITION METHOD COMBINING FFT, WAVELET FUNCTIONS AND NEURAL NETWORKS." Speaker recognition. 1-4.

[11] A. Reynolds, Douglas . "Automatic Speaker Recognition Using Gaussian Mixture Speaker Models." THE LINCOLN LABORATORY JOURNAL VOlUME 8, NUMBER 2,1995 : 173-192. Print.

[12] Hspice, Star-. "Performing FFT Spectrum Analysis."Star-Hspice Manual, Release 1998.2. . Performing FFT Spectrum Analysis . 1-26. Print.

[13] Sek, Michael. "FREQUENCY ANALYSIS FAST FOURIER TRANSFORM, FREQUENCY SPECTRUM." Victoria university, ,A new school of thought. 1-12. Print.

[14] TANG HSIEH, CHING, , et al. "Robust Speaker Identification System Based on Wavelet Transform and Gaussian Mixture Model." JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 19. 19. ((2003)): 267-282. Print.

[15] Agrawal, Upendra Kumar , Upendra Kumar Agrawal, et al. "FRACTIONAL FOURIER TRANSFORM COMBINATION WITH MFCC BASED SPEAKER IDENTIFICATION IN CLEAN ENVIRONMENT." International Journal of Advanced Science, Engineering and Technology. Vol 1,.1 (2012): 26-28. Print.

[16] Jin, Qin. "Robust Speaker Recognition."Language Technologies Institute School of Computer Science Carnegie Mellon University 5000 Forbes Avenue, Pittsburgh, PA 15213, 1 2007. 1-177. Print.

[17] "Fourier transform." From Wikipedia, the free encyclopedia. 18 May 2014.

[18] Kekre, H B, and Vaishali Kulkarni. "Speaker Identification using Frequency Dsitribution in the Transform Domain." (IJACSA) International Journal of Advanced Computer Science and Applications. 3.2 (2012): 73-78. Print.

[19] "Speech Signal Processing." ee.columbia.edu. N.p.. Web. 29 May 2014.