Face Recognition and Speaker Recognition on the basis of their facial features and skin tones

[1]Brijnath Mangalm (M-tech scholar)
Brij.nath634@gmail.com
[2]Sandeep Dubey (Asst. Prof. RGPM Bhopal)
dubeysandeep7@gmail.com

ABSTRACT
Face Recognition and Speaker Recognition are two technologies which can perform the process of verifying the legitimacy without the consent of the Person in question. Face recognition is the more straightforward of the two which analyses the person on the basis of their facial features and skin tones. Speaker recognition is the lesser documented technology because of the issues involved in correct analysis of different speaker tones with varying dialects and dictions. Also the terms Speaker Recognition and Speech Recognition are frequently confused. There is a difference between the act of certification (commonly referred to as speaker impetration or speaker certification) and recognizance. Finally, there is a difference between speaker recognizance (recognizing who is speaking) and speaker diarisation (recognizing when the same speaker is speaking). Recognizance the speaker can simplify the task of converting speech in systems that have been trained on particular person's voices or it can be used to authenticate or verify the identity of a speaker as part of a security process. This work discusses the Implementation of an Enhanced Speaker Recognition system using MFCC and vector quantization Algorithm. MFCC has been used extensively for purposes of Speaker Recognition. This work has augmented the existing work by using Vector Quantization and Classification using the Linde Buzo Gray Algorithm. A complete test system has been developed in MATLAB which can be used for real time testing as it can take inputs directly from the Microphone. Therefore, the design can be translated into a Hardware having the necessary real time processing Prerequisites. The system has been tested using the VID TIMIT Database and using the Performance metrics of False Acceptance Rate (FAR), True Acceptance Rate (TAR) and False Rejection Rate (FRR). The system has been found to perform better than the existing systems under moderately noisy conditions.

KEYWORDS: Voice recognition, Speech signal, SVM, ML, MFCC

INTRODUCTION

Sound is formed by the vibration of any medium, for example, the vibrations of the computer speaker or simply of air molecules or pressure in the air. These vibrations are generally modelled as two types of layers, interleaved, travelling together through the medium; high pressure layers (molecules compressed more than normal) and low pressure layers (molecules relaxed more than normal). The vibrations affect the ears and that is the underlying theory behind listening. In fact, sound can be perceived as a signal; the amplitude of which corresponds to the pressure change and the length of which corresponds to the distance between two consecutive high (or two consecutive low) pressure layers.

Human speech is one form of sound; which people have developed through time to carry valuable information for communication, such as thoughts and feelings. However, it also carries other derived characteristics such as the speaker's identity, language, diction, dialect, gender and mood. It is based upon the combination of lexical and names that are drawn from very large database (usually about 10,000 different words) . Each spoken word is created out of the phonetic combination of a limited set of vowels and consonant speech sound units. These vocabularies, the syntax which structures them, and their set of speech sound units differ, causing the existence of many thousands of

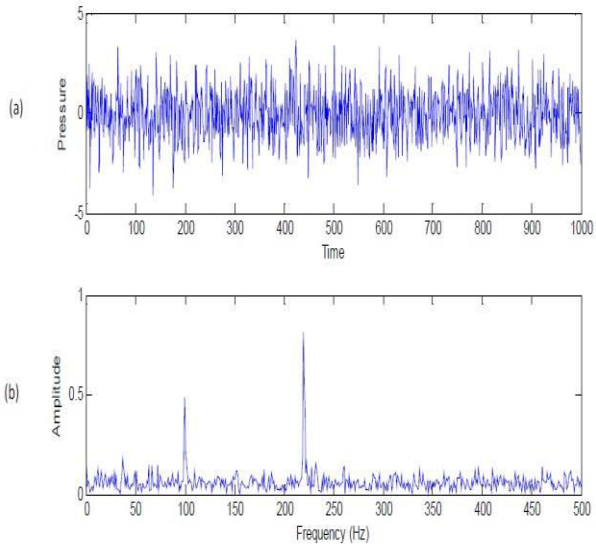
different types of mutually unintelligible human languages. Most human speakers are able to communicate in two or more of the languages, hence called polyglots. The vocal abilities that enable humans to produce speech also provide humans with the ability to sing. Speech is researched in terms of the speech production and speech perception of the sounds used in vocal language. Other research topics concern repetition of speech, the ability to relate heard and spoken words into the vocalizations needed to recreate. This plays a key role in the vocabulary expansion in children and speech errors. Several academic disciplines study these including acoustics, psychology, speech pathology, linguistics, cognitive science, communication studies, and otolaryngology and computer science. Another area of research is how the human brain in its different areas such as the Broca's area and Wernicke's area underlies speech forgotten or lost.

Sound & Human Speech

Sound is formed by the vibration of any medium, for example, the vibrations of the computer speaker or simply of air molecules or pressure in the air. These vibrations are generally modelled as two types of layers, interleaved, travelling together through the medium; high pressure layers (molecules compressed more than normal) and low pressure layers (molecules relaxed more than normal). The vibrations affect the ears and that is the underlying theory behind listening. In fact, sound can be perceived as a signal; the amplitude of which corresponds to the pressure change and the length of which corresponds to the distance between two consecutive high (or two consecutive low) pressure layers.

Signal

A signal is the continuous measure of a quantity in terms of time. An example of a signal is the measured voltage of a certain point in an electric circuit. A signal that repeats itself every period T is called a periodic signal; with the value T being its period. The number of times that the signal repeats itself in a time unit, i.e. one second, is called the frequency (mathematically, the inverse of the period).



The time series and (b) frequency spectrum of a signal

Sampling

In signal processing, sampling is the reduction of a continuous signal to a discrete signal. A common example is the conversion of a sound wave (a continuous signal) to a sequence of samples (a discrete-time signal). A sample refers to a value or set of values at a point in time and/or space. A sampler is a subsystem or operation that extracts samples from a continuous signal. A theoretical ideal sampler produces samples equivalent to the instantaneous value of the continuous signal at the desired points.

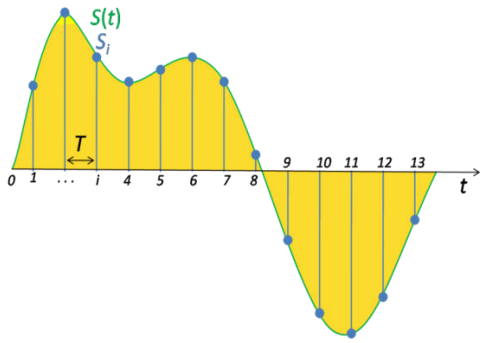


Fig: Signal sampling representation

Nyquist Frequency

The Nyquist frequency, named after electronic engineer Harry Nyquist, is ½ of the sampling rate of a discrete signal processing system.[1][2]

It is sometimes known as the folding frequency of a sampling system.[3] An example of folding is depicted in Figure.

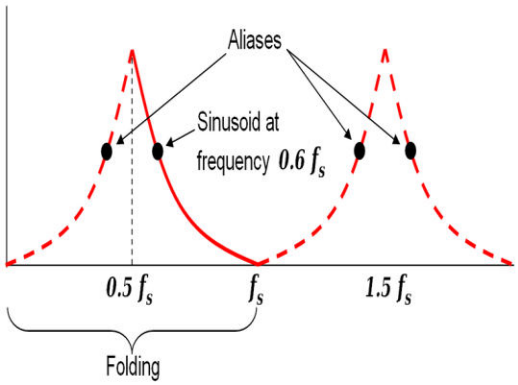


Fig. The black dots are aliases of each other.

The solid red line is an example of adjusting amplitude vs. frequency. The dashed red lines are the corresponding paths of the aliases.

Aliasing

under sampling of the sinusoid at 0.6 fs is what allows there to be a lower-frequency alias, which is a different function that produces the same set of samples. That condition is what's usually called aliasing. The mathematical algorithms that are typically used to recreate a continuous function from its samples will misinterpret the contributions of under sampled frequency components, which causes distortion. Samples of a pure 0.6 fs sinusoid would produce a 0.4 fs sinusoid instead. If the true frequency was 0.4 fs, there would still be aliases at 0.6, 1.4, 1.6, etc., but the reconstructed frequency would be correct. In a typical application of sampling, one first chooses the highest frequency to be preserved and recreated, based on the expected content (voice, music, etc.) and desired fidelity. Then one inserts an anti-aliasing filter ahead of the sampler. Its job is to attenuate the frequencies above that limit. Finally, based on the characteristics of the filter, one chooses a sample-rate (and corresponding Nyquist frequency) that will provide an acceptably small amount of aliasing. In applications where the sample-rate is pre-determined, the filter is chosen based on the Nyquist frequency, rather than vice-versa. For example, audio CDs have a sampling rate of 44100 samples/sec. The Nyquist frequency is therefore 22050 Hz. The anti-aliasing filter must adequately suppress any higher frequencies but negligibly affect the frequencies within the human hearing range. A filter that preserves 0–20 kHz is more than adequate for that.

Quantization

Quantization, in mathematics and digital signal processing, is the process of mapping a large set of input values to a (countable) smaller set – such as rounding values to some unit of precision. A device or algorithmic function that performs quantization is called a quantizer. The round-off error introduced by quantization is referred to as quantization error. In analog-to-digital conversion, the difference between the actual analog value and quantized digital value is called quantization error or quantization distortion. This error is either due to rounding or truncation. The error signal is sometimes modeled as an additional random signal called quantization noise because of its stochastic behavior.

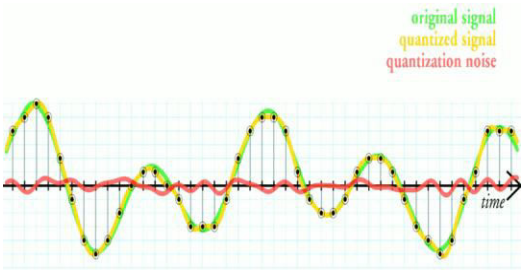


Fig the simplest way to quantize a signal is to choose the digital amplitude value closest to the original analog amplitude.

The quantization error that results from this simple quantization scheme is a deterministic

Function of the input signal.

Windowing

Window functions are used as a temporary bound on the original signal to limit the stream on the interesting range. The most commonly used window function is the rectangle function. The effect of this window on the original signal is to produce a time bounded signal that is similar to the original signal inside the rectangular range and is set to null outside of the

specified limits. In digital signal processing, windowing is thought of as a function affecting the samples of the signal to produce a new series of samples, i.e. a new signal. Figure shows the effects of different windowing functions on a digital flat signal.

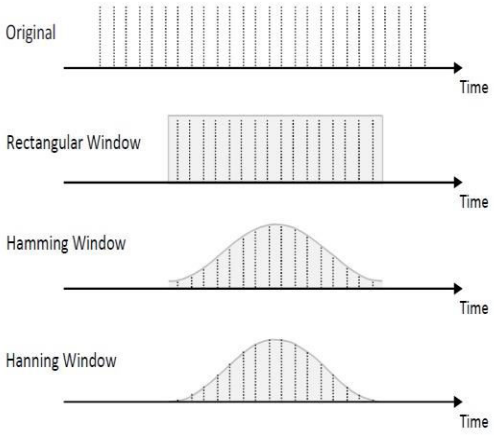
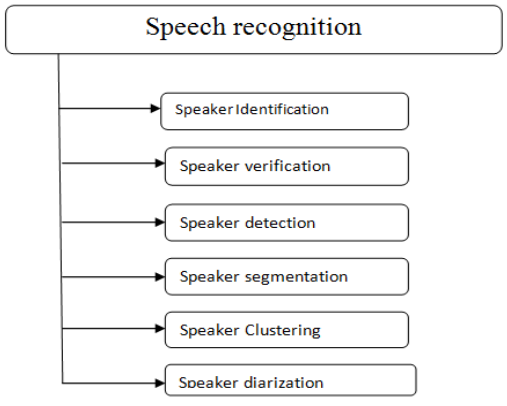


Fig The effects of different window functions on a digital signal

Filtering

Causes for Differences in Speech Signal

SPEAKER AND SPEECH RECOGNITION



Automatic Speaker Recognition

ASR is the process used to identify or verify a person using speech features extracted from an utterance. A typical ASR system consists of a feature extractor followed by a robust speaker modeling technique for generalized representation of extracted features and a classification stage that verifies or identifies the feature vectors with linguistic classes. In the extraction stage of an ASR system, the input speech signal is converted into a series of low-dimensional vectors, the necessary temporal and spectral behavior of a short segment of the acoustical speech input is summarized by each vector (Reynolds, 2002 [7],

SPEAKER RECOGNITION CLASSES

SR is now possible using a range of different approaches each with costs and benefits. As SR is a very important activity research today encompasses the range of difference approaches and for this reason there has been a classification of the approaches into classes. The SR approach classes are:

1. Conventional.

a. Speaker identification

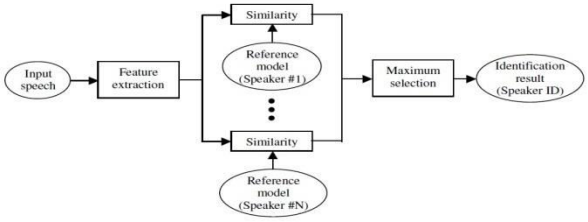
b. Speaker verification

2. Text Conversion.

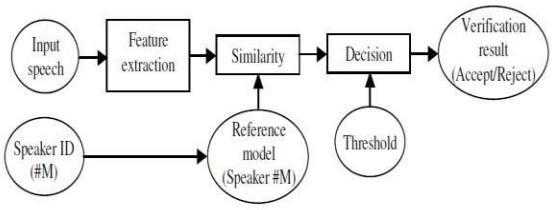
a. Text independent recognition

b. Text dependent recognition

Speaker Identification

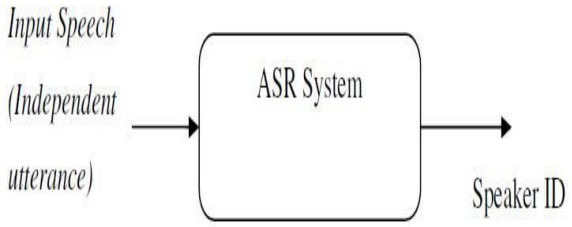


Speaker Verification



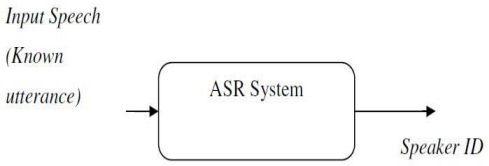
Text-independent recognition

In Figure text-independent SR system is shown where the key feature of the system is speaker identification utilizing random utterance input speech (Chakraborty and Ahmed, 2007)[10].



Text-dependent recognition

In Figure, a text-dependent SR system is shown where recognition of the speaker's identity is based on a match with utterances made by the speaker previously and stored for later comparison. Phrases like passwords, card numbers, PIN codes, etc. made be used (Chakraborty and Ahmed, 2007)[10].



SPEECH FEATURE EXTRACTION

Mel-Frequency Cesptrum Coefficients Processor

Frame Blocking

Windowing

Fast Fourier Transform (FFT)

Mel-frequency wrapping

Cestrum

Pattern Recognition

Vector Quantization

Clustering

K-means clustering

Linde-Buzo-Gray Clustering Technique

Information theoretic based cluste**ring**

Fuzzy C-means Clustering

BACKGROUND

Harrington, J., and Cassidy, S. Techniques in Speech Acoustics. Kluwer Academic Publishers, Dordrecht [34] and Harris, F. On the use of windows for harmonic analysis with the discrete fourier transform[35] have discussed Feature matching which involves the actual procedure to identify the unknown speaker by comparing the extracted features from his/her voice input with the ones that are already stored in our speech database. The speech signal continuously changes due to articulatory movements, and therefore, the signal must be broken down in short frames of about 20-30 milliseconds in duration. Within this interval, the signal is assumed to remain stationary and a spectral feature vector is extracted from each frame. Usually the frame is pre-emphasized and multiplied by a smooth window function prior to further steps. Pre-emphasis boosts the higher frequencies whose intensity would be otherwise very low due to downward sloping spectrum caused by glottal voice source.The window function (usually Hamming), on the other hand, is needed because of the finite-length effects of the discrete Fourier transform (DFT); for details, refer to **Deller, J., Hansen, J., and Proakis, J. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York, 2000[36], Harris, F. On the use of windows for harmonic analysis with the discrete fourier transform. Proceedings of the IEEE 66, 1(January 1978)[35], Oppenheim, A., Schafer, R., and Buck, J. Discrete-Time Signal Processing, second ed. Prentice Hall, 1999**. [13]. The frame length is usually fixed, pitch-synchronous analysis has also been studied in **Nakasone, H., Mimikopoulos, M., Beck, S., and Mathur, S. Pitch synchronized speech processing (PSSP)**

for speaker recognition. In Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004) (Toledo, Spain, May 2004), pp. 251–256.[37], **Zilca, R., Kingsbury, B., Navr´atil, J., and Ramaswamy, G. Pseudo pitch synchronous analysis of speech with applications to speaker recognition. IEEE Trans. Audio, Speech and Language Processing 14, 2 (March 2006), 467–478[38] and Gong, W.-G., Yang, L.-P., and Chen, D. Pitch synchronous based feature extraction for noise-robust speaker verification. In Proc. Image and Signal Processing (CISP 2008) (May 2008), vol. 5, pp. 295–298 [39]**. The experiments in [37, 38] indicate that recognition accuracy reduces with this technique, whereas [39] obtained some improvement in noisy conditions. Pitch-dependent speaker models have also been studied in **Arcienega, M., and Drygajlo, A. Pitch-dependent GMMs for text-independent speaker recognition systems. In Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001) (Aalborg, Denmark, September 2001),**pp. 2821–2824 [40]and Ezzaidi, H., Rouat, J., and O'Shaughnessy, D. Towards combining pitch and MFCC for speaker identication systems. In Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001) (Aalborg, Denmark, September 2001), pp. 2825–2828. [41].

Alternatives to FFT-based signal decomposition such as non-harmonic bases, aperiodic functions and data-driven bases derived from independent component analysis (ICA) have been studied in **Gopalan, K., Anderson, T., and Cupples, E. A comparison of speaker identication results using features based on cepstrum and Fourier-Bessel expansion. IEEE Trans. on Speech and Audio Processing 7, 3 (May 1999), 289–294**, **Imperl[42], B., Kacic, Z., and Horvat, B. A study of harmonic features for the speaker recognition. Speech Communication 22, 4 (September 1997), 385–402[43] , Jang, G.-J., Lee, T.-W., and Oh, Y.-H. Learning statistically efficient features for speaker recognition. Neurocomputing 49 (December 2002), 329–348** [44]

The **mel-frequency cepstral coefficients** (MFCCs) are popular features in speech and audio processing. MFCCs were introduced in early 1980s for speech recognition and then adopted in speaker recognition. Even though various alternative features, such as spectral subband centroids (SSCs) discussed in **Kinnunen, T., Zhang, B., Zhu, J., and Wang, Y. Speaker verification with adaptive spectral subband centroids. In Proc. International Conference on Biometrics (ICB 2007) (Seoul, Korea, August 2007), pp. 58–66[45], Thian, N., Sanderson, C., and Bengio, S. Spectral subband centroids as complementary features for speaker authentication. In Proc. First Int. Conf. Biometric Authentication (ICBA 2004) (Hong Kong, China, July 2004), pp. 631–639.** [46] have been studied, the MFCCs seem to be difficult to beat in practice. Lesser used features such as linear predictive cepstral coefficients (LPCCs) and line spectral frequencies have been discussed in **Huang, X., Acero, A., and Hon, H.-W. Spoken Language Processing: a Guide to Theory, Algorithm, and System Development. Prentice-Hall, New Jersey, 2001** [47] , perceptual linear prediction (PLP) coefficients in **Hermansky, H. Perceptual linear prediction (PLP) analysis for speech. Journal of the Acoustic Society of America 87 (1990), 1738–1752.** [48] and partial correlation coefficients (PARCORs), log area ratios (LARs) and formant frequencies and bandwidths in **Rabiner, L, and Juang, B.-H. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, New Jersey, 1993**. [49]

2.2 Recent works

Speaker recognition has been an active research area and Some of the more recent worksin the field have been mentioned here.

"Including human expertise in speaker recognition systems: report on a pilot evaluation" by CS Greenberg, AF Martin, IEEE Journal of Speech and Signal, 2011[50] discusses Speaker Recognition Evaluation (SRE10) included a test of Human Assisted Speaker Recognition (HASR) in which systems based in whole or in part on human expertise were evaluated on limited sets of trials.

"Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors"by M McLaren, D Van Leeuwen - Acoustics,

Speech and Signal, 2011[51]. The recently developed i-vector framework for speaker recognition has set a new performance standard in the research field. An i-vector is a compact representation of a speaker utterance extracted from a low-dimensional total variability subspace

Multi-variability speech database for robust speaker recognition by BC Haris, G Pradhan, A Misra, S Shukla NCC), 2011 National, 2011[52]. In this paper, the authors have presented an initial study with the recently collected speech database for developing robust speaker recognition systems in Indian context. The database contains the speech data collected across different sensors, languages, speaking styles etc.

In the paper Performance **Comparison of Speaker Recognition using Vector Quantization by LBG and KFCG by H. B. Kekre and Vaishali Kulkarni [53]**, two approaches for speaker Recognition based on Vector quantization are proposed and their performances are compared. Vector Quantization (VQ) is used for feature extraction in both the training and testing phases. Two methods for codebook generation have been used. In the 1st method, codebooks are generated from the speech samples by using the Linde-Buzo-Gray (LBG) algorithm. In the 2nd method, the codebooks are generated using the Kekre‟s Fast Codebook Generation (KFCG) algorithm.

In the paper Speaker recognition using **Vector Quantization by MFCC and KMCG clustering algorithm by HB Kekre, VA Bharadi, AR Sawant IEEE, 2012[54]** authors have implemented a speaker recognition system using a combination of Mel Frequency Capestral Coefficients (MFCC) &Kekre's MCG clustering algorithm.

As is evident from the listed works in the chapter, it can quite clearly be seen that the field has been an active research area in the past few years by the no of manuscripts that have been published on the topic. However, the primary metrics on which the system is evaluated for the performance i.e TAR, FAR and FRR have been found to be different in different works.

CONCLUSION

The system has been found to perform satisfactorily under noisy conditions as well however has been found prone to increase in FAR if user inputs are from microphone under noisy conditions. The testing has been done by using standard Microphones in acoustically silent environments and then additional hum has been added for noise simulations. The GUI developed for the purpose has capabilities of real time speaker recognition, making it a significant contribution to the work. The work has been simulated and tested using MATLAB R 2012. Although the GUI that has been developed takes inputs in real time, however the performance of the system needs to be tested on a Hardware platform F2812 Floating point Processor for its actual real time performance to be verified. That will require further optimization of the LBG algorithm for it to match to the hardware needs. The platform should be such that its internal mathematical operations should not be affected due to exchange of data to and from real world. System can be implemented using other combination of other quantization methods as well.

REFERENCES.

1) Grenander, Ulf (1959). Probability and Statistics: The Harald Cramér Volume. Wiley. "The Nyquist frequency is that frequency whose period is two sampling intervals."
2) Harry L. Stiltz (1961). Aerospace Telemetry. Prentice-Hall. "the existence of power in the continuous signal spectrum at frequencies higher than the Nyquist frequency is the cause of aliasing error"
3) Thomas Zawistowski, Paras Shah. "An Introduction to Sampling Theory". Retrieved 17 April 2010. "Frequencies "fold" around half the sampling frequency - which is why [the Nyquist] frequency is often referred to as the folding frequency."
4) Campbell Jr., J.P., 1997. Speaker recognition: A tutorial. Proceedings of the IEEE, 85(9), pp.1437–1462
5) Daniel Jurafsky& James H. Martin “Automatic Speech Recognition” Speech and Language Processing: An

introduction to natural language processing, computational linguistics, and speech recognition, 2007

6) He´bert, M., 2008. Text-dependent speaker recognition. In: Benesty, J.,Sondhi, M., Huang, Y. (Eds.), Springer Handbook of Speech Processing. Springer-Verlag, Heidelberg, pp. 743–762
7) Douglas A Reynolds "An Overview of Speaker Recognition Technology", MIT Lincoln Laboratory, MA 2002
8) Speaker segmentation and clustering (2008) M Kotti, V Moschou, C Kotropoulos., 2008"
9) Digital Speech Processing: Synthesis, and Recognition, Second Edition, SadaokiFurui, 2000 CRC Press
10) An Automatic Speaker Recognition System. P. Chakraborty, F. Ahmed, Md. MonirulKabir, M. Shahjahan, and Kazuyuki Murase. ICONIP 1, volume 4984 of Lecture Notes in Computer Science, page 517-526. Springer, (2007).
11) . R. Rabiner,B.-H. Juang, C.-H. Lee ,An Overview of Automatic Speech Recognition Automatic Speech and Speaker Recognition, The Kluwer International Series in Engineering and Computer Science Volume 355, 1996, pp 1-30
12) PhaophakSirisuk, Fearghal Morgan, Tarek El-Ghazawi, Hideharu Amano Reconfigurable Computing: Architectures, Tools and Applications: 6th Edition page 359, Bangkok, Thailand, March 17-19, 2010, Proceedings
13) From Frequency to Quefrency: A History of the Cepstrum Alan V. Oppenheim and Ronald W. Schafer, IEEE Signal Processing Magazine Reprinted 2004
14) Atal, B. S., and Hanauer, S. L. (1971) "Speech analysis and synthesis by linear prediction of the speech wave," J. AcoustoSoco Am. 50, 637-655..
15) . R. Sambur A, . E. RosenbergL, . R. Rabinera, nd C. A. McGonegal "On reducing the buzz in LPC synthesis", M Bell Laboratories
16) Davis, S. ; Mermelstein, P Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences Acoustics, Speech and Sig Processing, IEEE Transactions on (Volume:28 , Issue: 4)Aug 1980, Page(s): 357 - 366
17) TodorGanchev , Nikos Fakotakis , George Kokkinakis Comparative evaluation of various MFCC implementations on the speaker verification task (2005),