Speaker Recognition System and Its Forensic Implications: A Review

Surbhi Mathur*, Sumit K Choudhary, J M Vyas
**Assistant Professor Jr., Gujarat Forensic Sciences University, Gandhinagar, Gujarat*
sue.mathur@gmail.com

Abstract: - **Speaker recognition is the process of automatically recognizing the unknown speaker by extracting the speaker specific information included in his/her speech wave. This paper will help the readers to understand the need of this speaker recognition technique in a much better way. It outlines the basic concepts of speaker recognition along with its diverse applications. It also presents an idea of selecting a robust parameter for the purpose of identification to attain the accurate results, limitations faced and the recent built up advances for identification, so as to provide a technological perspective in this important area of speaker recognition.**

Keywords*:- *Forensic Science, Speaker, Recognition, Identification, Verification, Voice, Speech

I. INTRODUCTION

Speaker recognition comprises all those activities which attempt to link a speech sample to its speaker through its acoustic or perceptual properties [1]. Speech signal is a multidimensional acoustic wave [fig: 1], which provides information regarding speaker characteristics, spoken phrase, speaker emotions, additional noise, channel transformations etc [2; 3]. The human voice is unique personal trait. For indistinguishable voice, the two individuals should have the identical vocal mechanism and identical coordination of their articulators, which is least probable. However, the some variations also occur in the speech exemplars obtained from the same speaker. This is due to the fact that a speaker cannot exactly imitate the same utterance again and again. Even, the signature of an individual also shows variation from trails to trials.

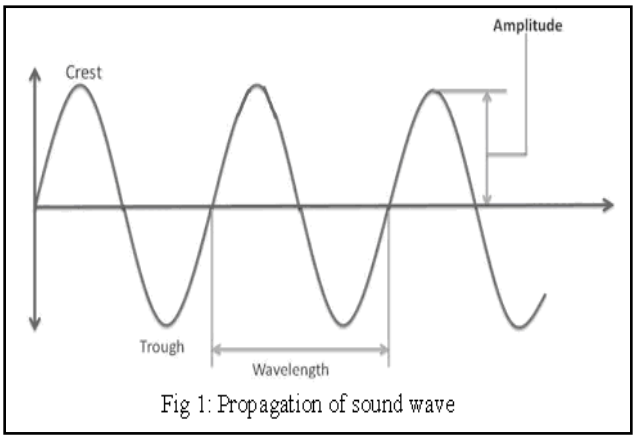

Fig 1: Propagation of sound wave

The process of Speaker recognition has two broad application areas, explicitly, Speaker identification and Speaker verification. Speaker identification deals with identifying a speaker of a given utterance amongst a set of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance [fig: 2]

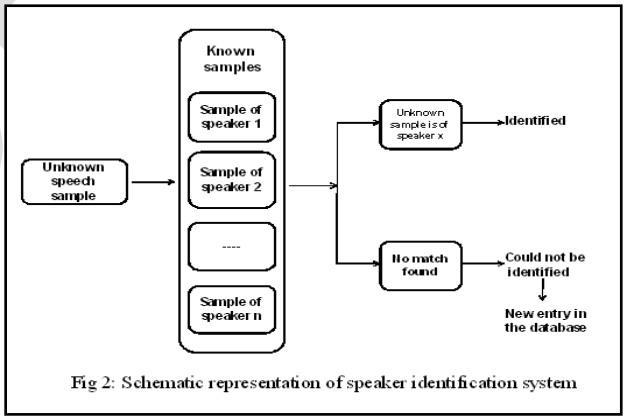

Fig 2: Schematic representation of speaker identification system

There are two modes of operation related to known voices: closed set and open set. The closed set mode is considered as multiple class classification modes. Such system assumes that the voice which has to be determined or identified belongs to a set of known voices. While in open set the speaker which do not belong to a set of known speakers, is referred as an imposter. This task can be used for forensic purposes, in which an offender's is used to reveal his or her identity, among several known suspects.

In contrast, Speaker verification is a more direct and converged effort leading to either acceptance or rejection of the claimed identity of a speaker. To be precise, this investigation reveals whether a speaker is the one who he claims to be [fig: 3] [4; 5; 6]?

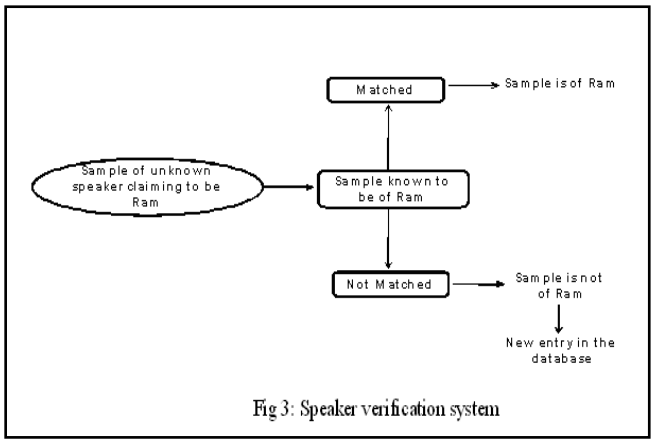

Fig 3: Speaker verification system

It can be considered as a true-or-false binary decision problem. It is sometimes referred to as the open-set problem, because this task requires distinguishing a claimed speaker's voice known to the system from a potentially large group of voices unknown to the system. Today verification is the basis for most speaker recognition applications and the most commercially feasible task.

II. SIGNIFICANCE

A. Security or access control

The voice of a person can be successfully used as a biometric feature as it is well accepted by the users and can be easily recorded using microphones and hardware of low costs [7]. It can provide an unconventional and more secure means of permitting entry without any need of remembering a password, lock combination etc or the use of keys, magnetic card or any other fallible device which can be easily stolen [8; 9].

Although the voice of a person cannot be stolen but it can be copied using some recording devices. Therefore, the voice-based security systems must protect themselves against such flaws. The other concern is voice disguise. An imposter can gain illicit entry by disguising or imitating the voice of a genuine speaker, to access this personal data. Similarly, a valid person may be denied the entry because of some accidental changes in his or her voice due to illness, emotional or physical stress etc.

B. Law enforcement

Voice of a person can plays a vital role in forensic examination. . In the present era, widely available facilities of telephones, mobiles and tape recorders results in the misuse of the device and thus, making them an efficient tool in commission of criminal offences such as kidnapping, extortion, blackmail threats, obscene calls, anonymous calls, harassment calls, ransom calls, terrorist calls, match fixing etc. The criminals nowadays are more frequently misusing these modes of communication, believing that they will remain incognito, and nobody would recognize them. It is fortunately no longer true. The voice of an individual can successfully recognize him and pin the crime on him [10].

The results obtained through speaker recognition analysis are not easily accepted in the court of law. But with advancements made in this field and with the judges understanding the value of statistical findings, the situation is expected to change in the future [11; 12]. But the results in this case also are vulnerable to two types of voice disguise: deliberate and unintentional.

III. COMPONENTS OF SPEAKER RECOGNITION SYSTEM

The main components of speaker recognition system are feature extraction and classification. The classification module is further divided into two parts: pattern matching and decision [fig: 4] [13].

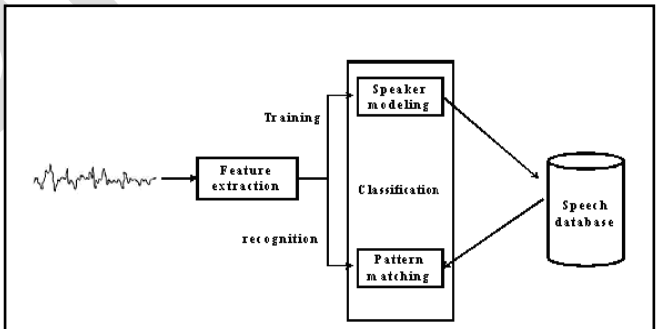

Fig 4: Mechanism of speaker recognition

A. Feature extraction

This is the foremost step in the process of speaker recognition. This segment processes the acquired data, i.e., a set of feature vectors or parameters from the speech signal representing some speaker-specific information, which results from complex transformations occurring at different levels of the speech production: semantic, phonologic, phonetic, and acoustic [14; 15; 16].

B. Criteria of feature selection

In a scheme for the mechanical recognition of the speakers, it is desirable to use acoustic parameters that are closely related

to voice characteristics that distinguish speakers. It involves selection of those parameters which are motivated by known relations between the voice signal and vocal-tract shapes and gestures. Speaker recognition by and large depends upon both low level and high level information obtained from a person's speech. High level information include values like dialect, accent, the talking style, the subject manner of context, phonetics, prosodic and lexical information [17]. These features are currently recognized and analyzed by humans only. The Low-level features refer to the information like fundamental frequency (F0), formant frequency, pitch, intensity, rhythm, tone, spectral magnitude and bandwidths of an individual's voice [18]. An ideal feature would:

- Have lower intraspeaker variability and high interspeaker variability.
- Be robust against noise and distortion
- Occurs frequently and naturally in speech
- Be easily measured from the speech signal
- Difficult to mimic
- Not be affected by speaker's health or long term variations in voice

There are different ways to categorize the features [19]. From the viewpoint of their physical interpretation, following categories have been proposed:-

a. Short-term spectral features –These features, as the name suggests, are computed from the short frames of about 20 to 30 milliseconds in duration. They are usually the descriptors of the resonance properties of the supralaryngeal vocal tract.
b. Voice source features –These features characterize the glottal excitation signal of voiced sounds such as glottal pulse shape and fundamental frequency, and it is reasonable to assume that they carry speaker-specific information.
c. Spectro-temporal features -It is very much a rational assumption that the spectro temporal. Signal details such as formant transitions and energy modulations contain useful speaker-specific information.
d. Prosodic features - Prosody refers to non-segmental aspects of speech, including syllable stress, intonation patterns, speaking rate and rhythm. These features depends upon the long segments like syllables, words, and utterances and reflects differences in speaking style, language background, sentence type and emotion of the speaker.
e. High level features –These features attempt to capture conversation-level characteristics of speakers, such as characteristic use of words ("uh-huh", "you know", "oh yeah", etc.). Other features are the dialect of any language used in the conversation by the speaker, accent of the speaker and the style of speaking.

C. Pattern matching and decision

The pattern matching module deals with comparison between the estimated features to the speaker models. Some of the pattern matching methods used in speaker recognition include Hidden markov models (HMM), dynamic time warping (DTW), neural networks and vector quantization (VQ) [20]. In case of verification, this module provides an expert with a similarity score between the test sample and the claimed identity. While, in case of identification, the module gives similarity score between the test sample and all the available samples in the database. The evaluation of these scores is done using decision module and the results are accordingly presented.

The effectiveness of a speaker recognition system is measured differently for different tasks. Since the output in identification system is a speaker identity from a set of known speakers, the identification accuracy is used to measure the performance. For the verification systems, two types of error can be observed: false acceptance of an impostor and false rejection of a target speaker [21].

IV. VARIOUS APPROACHES OF SPEAKER RECOGNITION

In the discipline of speaker recognition a wide range of methods and procedures are adopted by the experts for identification.

A. Auditory analysis

Such type of analysis involves a group of trained phoneticians giving their judgement regarding the similarity and dissimilarity between the two speech events, after hearing the samples again and again to find out some similarities in their linguistic, phonetic and acoustic features. Human listeners are robust speaker recognizers when presented with the degraded speech. Listener performance free from all types of limitations like the signal to noise ratio, speech bandwidth, the amount of speech material, distortions occurring in the speech signals as a result of speech coding, transmission systems, etc.

In this technique, different utterances of the speakers are segregated in respect of each speaker by way of repeated listening of recorded conversation. The segregated conversations of each speaker are repeatedly heard to identify linguistic features and phonetic features like articulation rate, flow of speech, degree of vowels and consonant formation, rhythm, striking time, pauses etc. There are cues in voice and speech behaviour, which are individual and thus make it possible to recognize the familiar voices [22].

Experts working in several governments forensic laboratories including laboratories in Germany, Austria,

the Netherlands and Spain, and in private practice in countries like the United Kingdom and Germany, are still practising this phonetic-acoustic technique for identification. However, with any human decision process, it is generally believed that the auditory analysis by a listener leads to a subjective decision [23].

B. Spectrographic approach or voiceprint identification

This involves the semi-automatic measurements of particular acoustic speech parameters such as vowel formants, articulation rate, which is sometimes combined with the results of auditory phonetic analysis by a human expert. In 1941, an electro mechanical acoustic spectrograph was developed by Dr. Raleph Potter, Bell Telephone Laboratory, with an idea to convert sounds into pictures [24].

A sound spectrograph is an instrument which is able to give a permanent record of changing energy-frequency distribution throughout the time of a speech wave [fig:5] [25]. The spectrograms are the graphic displays of the amplitude as a function of both frequency and time [26].

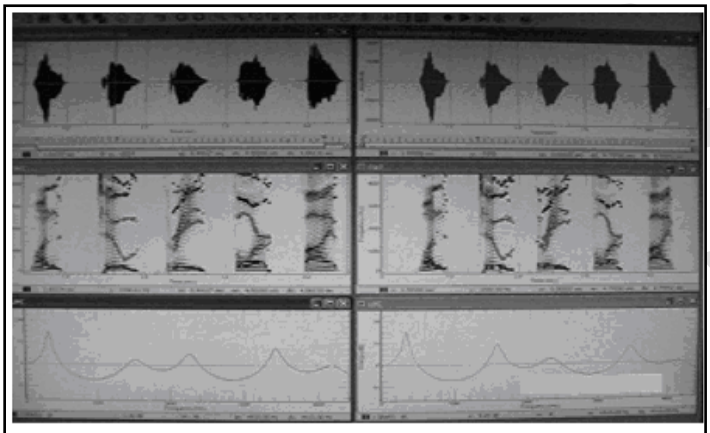

Fig 5: spectrographic analysis of speech samples

Examiners visually inspect and compare similarities or differences of patterns of the energy distribution in the spectrograms. It is generally believed that formant structures and other spectral characteristics which are evident from a spectrogram are unique for each individual. The most widely used features are fundamental frequencies [fig: 6], formant bandwidths, formant frequencies, spectral composition of fricatives and plosives for individual segments, and transitions.

However, the main drawback of this voiceprint analysis is that the spectrograms of the speech signal from same individual will show large intraspeaker variations, because of the fact that no speaker actually is capable of

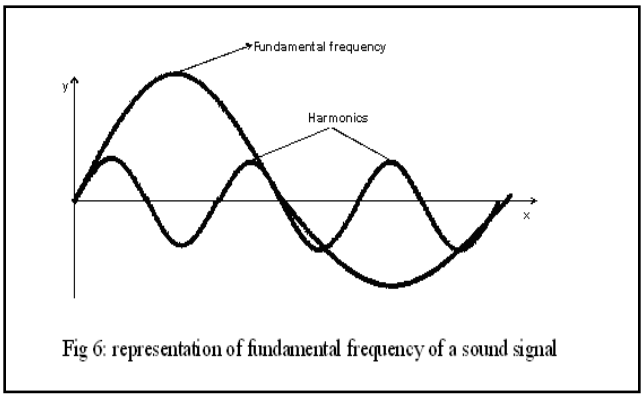

Fig 6: representation of fundamental frequency of a sound signal

producing two identical speech utterances [27]. This method is obviously neither objective nor superior to aural-perceptual methods; it is basically a shifting of subjective judgement to the visual domain. The objectivity, reliability and validity of the method have been discussed controversially. The method has been widely used in the US, parts of Europe and other countries until the 1980s but in the present scenario it has been losing its ground. The FBI are using it for investigative purposes, most U.S. courts do not accept voiceprint evidence. Today voiceprint identification is not used in forensic labs in Europe, but still practised in developing countries like China, Vietnam etc.

C. Recent advances or automatic approach

This approach differs greatly from the earlier methods used for identification as it is both universal as well as automatic. It is considered universal because it does not focus on specific acoustic parameters and consider the speech as a continuously varying complex wave or signal. While, it's automatic nature reduces the subjective evaluation of any speech material to minimum. Most of such automatic identification system today involves techniques like:

1) Gaussian mixture models

These are used to characterise or 'model' the speech of the known speaker (from the database) and that of the unknown speaker (i.e., the perpetrator). In addition to this, a relevant speaker population is defined and a probability-density function of the speech variance of this set is calculated. This technique however faces two types of challenges. The problem of within- and between- speaker variations, may results in overlapping of speaker models. As a result, speakers may not always be reliably distinguished, and the system will produce a certain proportion of false-positives.

The second problem arises due to the extreme sensitivity to transmission channel effects of automatic procedures,

including the effects of different handsets, telephone lines, GSM-coding and perception-based compression techniques [28].

2) Long-term averaging

The long- term speech spectrum is used as an important cue of determining the voice quality [29]. In this technique, large number of feature vectors is collected for each known speaker. The average and variance of each component of the feature vector are calculated, and vector of mean value, and vector of the variances, is used to model each speaker. A similar model is made for the unknown speaker.

This technique is most useful for text independent recognition, where large amount of data is required for construction of the speaker's model. This method will not be beneficial if the utterances are too short and if contains the insufficient amount of data.

The major disadvantage of long-term averaging is that each speaker's model consists of a single cluster of data represented by an average and variance vector. If the data contain multiple clusters of vectors, the variance will be very high. Since human speech is composed primarily of vowels, it is natural to expect feature vectors to form clusters, each one based on the pronunciation of a specific vowel.

3) Vector quantization

This is a technique in which each speaker's model is prepared which consists of several clusters of data, along with their centroids. VQ reduces these sets of vectors to a codebook, which provides an efficient way of building and comparing models of speakers [30]. VQ is used in several ways in speaker recognition. In some systems it is used simply to compress data. In other systems, VQ is a preprocessing step for other methods such as HMMs.

For text-dependent identification and verification several codebooks are created or "trained" for each speaker, who speaks a prescribed text several times. These codebooks are considered as the speaker's template. During the operational phase the same prescribed text is spoken by the unknown person. The comparison is done on the basis of observed differences or similarities between the unknown person's template, and each trained template, after removing the variations in the speaking rate.

For text-independent speaker recognition a single codebook is created for each speaker. The codebook is considered as an accurate model of the speaker because it is formed from a much larger amount of speech than in the text-dependent case.

This method introduces a new factor affecting the performance of the system, which is code-book size. Larger codebooks will perform a better job of characterizing a speaker's voice, but these results in increased computational expenses and the danger of not producing results in real time, which is a significant factor for verification. The advantage of this method is that it requires only a small amount of data to create a speaker's model without causing any loss to the accuracy.

4) Hidden Markov Models

These models are which are useful for modelling the stationary as well as the transient properties of speech. These made it possible to deal with the time sequential data and can provide the time scale invariability in recognition [31]. These are appropriate for speech sounds as it contains both vowel and consonants. HMMs are able to represent signals that exhibit diverse behaviour because of their probabilistic nature [32] and can be efficiently used for both text-dependent and text-independent speaker recognition system.

In case of text-dependent SR system, a single HMM is trained for each individual uttering the prescribed text. When an unknown person speaks the same speech, an HMM is created on the spot and compared with all the others. Commonly, the feature vectors used with HMMs are averaged through vector quantization and expressed as codebook values. While in case of text-independent recognition, the "states" are trained to represent each person's pronunciation of the different phonetic classes, such as rounded vowels or nasal consonants.

During training, the parameters of the HMM are adjusted to best represent the significant features of each person's speech. During the operational phase, it is determined mathematically which model is most consistent with the unknown input speech, and that model determines the unknown speaker, or confirms or denies the verification.

5) Neural networks

These are computational models that attempt to imitate the human brain through interconnected nodes that behave like simple nerve cells [33]. These are versatile devices and can be used for variety of purposes.

In a typical system, a neural network is created for each speaker and trained to be active (i.e., to give an output near 1.0) when the input belongs to the speaker, and inactive (an output near 0) for some other speaker. Example: In a population including 3 speakers X, Y, and Z, we would have three binary networks: one trained to distinguish between (X, Y), one between (X, Z), and one between (Y, Z). If an unknown, A, is to be determined

from among X, Y and Z the procedure would be as follows: Feed A into the (X, Y) network and record the output score for X and the output score for Y. Repeat the process with the (X, Z) network and the (Y, Z) network. For example, suppose A was actually X. Then X would have a high score when A is fed into (X, Y) and (X, Z). In those cases both Y and Z would have low scores, when fed into (Y, Z), both Y and Z would have low scores. In total, X would be the winner.

However, the use of a single large network works well for small populations of speakers, but it has two disadvantages. One is that if the size of the population exceeds a few dozen, the training times go way up, and the performance goes way down. The second is when new speakers are added to the population, the entire network must be retrained.

Fully automatic systems are generally introduced on a small scale, in forensic casework. At present countries like France [34] and Switzerland [35; 36] are using such methods, which are also being tested in Spain [37] and the United States of America [38]. The FBI recently completed an evaluation project in which four automatic speaker recognition systems were tested on a specially designed forensic database compiled by the FBI. The results confirmed that the performance levels of automatic systems can be quite high when text and transmission conditions are controlled. Deterioration is usually encountered in the conditions related to forensic realm.

V. EXPRESSING RESULTS IN FORENSIC SPEAKER RECOGNITION

Like in other disciplines in the forensic field, a voice expert generally renders his or her opinion in terms of probability of evidence under two rival assumptions:

- Prosecution hypothesis: the unknown or the test sample is originated from the given source
- Defence hypothesis: the unknown or the test sample originates from some other member of a potential suspect population, like the adult male population of a town or a particular region.

The ratio between these two probabilities is known as likelihood ratio, which assumes some numerical value.

Likelihood ratio= Hp/Hd

Where, Hp= Prosecution hypothesis

Hd=Defence hypothesis

Yet, even these high numbers do not indicate how likely the questioned voice sample is to have originated from the suspect. It only expresses the relative strength of the evidence.

VI. PROBLEMS OR LIMITATIONS IN SPEAKER RECOGNITION

Short duration samples are more demanding and should be carefully analysed.

- Dissimilarity in the language of questioned and specimen voice samples
- Emotion Variability in questioned and specimen samples [39]
- Misspoken or misread prompted phrases
- Poorly recorded/noisy samples are difficult to analyse
- Insufficient number of comparable words
- Disguise in speech samples poses a problem in speaker recognition and/or the degree of disguise is decided by the expert
- Extreme emotional states (e.g. stress or duress) [40]
- Change in physical state of the speaker (e.g. eating, effect of ethanol, etc.) [41]
- The attitude of the how the speech is said by the speaker
- Channel mismatch or mismatch in recording conditions (e.g. using different microphones for enrolment and verification)
- Different pronunciation speed of the test data compared with the training data.
- Speaker's health [42; 43]
- Aging (the vocal tract can drift away from models with age)

CONCLUSION

In lieu of the above discussion, it can be inferred that the comparison of voice samples is quite complicated but absolutely possible. The skill of an examiner itself along with chosen parameters and selection of appropriate technique for identification is largely decisive and can facilitate accurate and conclusive results. There have been many advancements and success made in this field, however, much remains to be done in order to overpower the daunting limitations which still prevails and limits the process. If we successfully overcome all such limitations, this technique with its promising features will have an obvious advantage over the pre-existing ones for establishing individual identity.

REFERENCES

1. Ramachandran RP, Farrell KR, Ramachandran R, Mammone RJ. Speaker recognition-general classifier approaches and data fusion methods. Pattern recognition 35 (2002) 2801 – 2821
2. Arslan L, Gorgun S, Naci U. Handset normalization for voice authentication. (levent@gvs.com.tr)

3. Astuti W. Text dependent speaker identification using PLP & SLV techniques. MSc dissertation report, international islamic university, Malaysia, 2007
4. Hannani AE, Delacretaz DP, Fauve B, Mayoue A, Mason J, Bonastre JF et al. Text-independent speaker verification: state of the art and challenges. Springer Berlin / Heidelberg, Vol. 4391/2007
5. Markowitz J. The many roles of speaker classification in speaker verification and identification. Springer Berlin / Heidelberg, Vol. 4343/2007
6. Pawlewski M, Jones J. Speaker verification: part 1, Biometric technology today. Vol. 14, Issue 6, June 2006
7. Chenafa M, Istrate D, Vrabie V, Herbin M. Biometric system based on voice recognition using multiclassifiers. Springer Berlin / Heidelberg, Volume 5372/2008
8. Skosan M, Mashao D. An overview of speaker recognition technology. Proceedings of CHI-SA 2005
9. Karpov E. Real time speaker identification. department of computer science, university of Joensuu, 2003
10. Sharma BR. Scientific criminal investigation. universal law publishing company
11. French P. An overview of forensic phonetics with particular reference to speaker identification. Forensic linguistics, 1(2): 169-181. 1994
12. Majewski W, Basztura C. Integrated approach to speaker recognition in forensic applications. Forensic linguistics, 3(1): 50-64. 1996
13. Tazi EB, Benabbou A, Harti M. Design of an automated speaker recognition system based on adapted MFCC and GMM methods for Arabic speech. IJCSNS International Journal of Computer Science and Network Security, vol.10 no.1, 2010
14. Campbell JP. Speaker recognition: A tutorial. Proceeding of the IEEE, 85:1437–1462, September 1997
15. Jin Q. Robust speaker recognition. School of computer science, carnegie mellon university, 2007
16. Chenafa M, Istrate D, Vrabie V, Herbin M. Speaker recognition using decision fusion. Biosignals-international conference on bio-inspired systems and signal processing, 2008
17. Shriberg E. Higher-level features in speaker recognition. Speaker classification I, Springer Berlin/Heidelberg, Volume 4343/2007
18. Graevenitz GAV. About speaker recognition technology. Bergdata biometrics GmbH , Bonn, Germany
19. Kinnunen T, Li H. An overview of text-independent speaker recognition: from features to supervectors. Speech communication, July 2009
20. Cepeda LF. Semi-automatic forensic speaker identification. research thesis, North Carolina State University, 2007
21. Richiardi J, Drygajlo A, Prodanov P. Confidence and reliability measures in speaker verification. Journal of the franklin institute 343 (2006) 574–595
22. Zetterholm E. Detection of speaker characteristics using voice imitation. Springer Berlin / Heidelberg, Volume 4441/2007
23. Braun A, Kunzel HJ. Is forensic speaker identification unethical - or can it be unethical not to do it?. forensic linguistics 5(1), 10-21, 1998
24. Kent RD, Read C. The acoustic analysis of speech. university of Wisconsin- Madison, A.I.T.B.S Publishers and distributors, Delhi
25. Samudravijaya K. Speech and speaker recognition: a tutorial. Tata institute of fundamental research, Mumbai
26. Kvistedal YA. A research paper in forensic science. the university of Auckland, New Zealand,2000
27. Gfroerer S. Auditory-instrumental forensic speaker recognition. Eurospeech, Geneva, 2003
28. Broeders APA. Forensic speech and audio analysis forensic linguistics. 13th Interpol forensic science symposium, France, october 16-19 2001
29. Harmegnies B, Landercy A. Intra-speaker variability of the long term speech pattern. Speech communication 7 (1988) 81-86, North-Holland
30. Kekre HB, Sarode TK. Speech data compression using vector quantization. International journal of computer and information science and engineering 2;4 2008
31. Yamato J, Ohya J, Ishii K. Recognizing human action in time sequential images using hidden markov model. IEEE, 1992
32. Abdulla WH, Kasabov NK, The concepts of hidden markov model in speech recognition. technical report TR99/09, university of Otago, New Zealand
33. Bennani Y, Gallinari P. Neural networks for discrimination and modelization of speakers. Speech communication 17 (1995)
34. Marescal F. The forensic speaker recognition method used by the french gendarmerie. internal publication, IRCGN: Paris, 1999
35. Meuwly D. Reconnaissance de Locuteurs: l'Apport d'une Approche Automatique. PhD Thesis, University of Lausanne, 2001
36. Pfister B. Personenidentifikation anhand der Stimme'. Kriminalistik 55(4), 287-292, 2001
37. Gonzalez-Rodriguez J, Ortega-Garcia J, Lucena-Molina J. On the application of the bayesian framework to real forensic conditions with GMM-based systems. paper presented at- a speaker odyssey, Crete, Greece, 2001
38. Nakasone H, Beck SD. Forensic automatic speaker identification. paper presented at- a speaker odyssey, Crete, Greece, 2001
39. Shan Z, Yang Y. Scores selection for emotional speaker recognition. Springer Berlin / Heidelberg, Volume 5558/2009
40. Resa CO, Moren IL, Ramos D, Rodriguez JG. Anchor model fusion for emotion recognition in speech. BioID multicomm2009, LNCS 5707, pp. 49–56, 2009.
41. Effects of ethanol intoxication on speech suprasegmentals. J. Acoust Soc Am, Vol 110, Dec 2001
42. Tull RG, Rutledge JC. Cold speech for automatic speaker recognition. Acoustical society of america,131st meeting lay language papers,1996
43. Singh S, Bucks R, Cuerden JM. Speech in alzheimer disease. SST 1996 proceedings, (www.asta.org/sst/Abstract-SST-1996.html)