A Survey on Automatic Speaker Recognition Systems

Conference Paper *in* Communications in Computer and Information Science · January 2010

DOI: 10.1007/978-3-642-17641-8_18 · Source: DBLP

5 authors, including:

Zia U Saquib
Centre for Development of Advanced Compu…
36 PUBLICATIONS **82** CITATIONS
SEE PROFILE

Nirmala Salam
CDAC Mumbai
9 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Nipun Pandey
Centre for Development of Advanced Compu…
3 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Akanksha Joshi
Centre for Development of Advanced Compu…
14 PUBLICATIONS **30** CITATIONS
SEE PROFILE

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2009) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

A Survey on Automatic Speaker Recognition Systems

Zia Saquib , Nirmala Salam*, Rekha P Nair, Nipun Pandey, Akanksha Joshi

CDAC-Mumbai,Gulmohar Cross Road NO.9,Juhu, Mumbai-400021

Abstract

"It's me!" This pronouncement is usually made over the telephone or at an entryway out of sight of the intended hearer. It embodies the expectation that the sound of one's voice is sufficient for the hearer to recognize the speaker. In short, "It's me!" is the original real-world, speaker-recognition challenge. Like human listeners, voice biometrics uses the features of a person's voice to ascertain the speaker's identity. The best-known commercialized forms of voice biometrics are Speaker Recognition. Speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices. Voice signal carries information related to not only the message to be conveyed , but also about speaker , language, emotional status of the speaker, environment and so on. In a speaker recognition task the speech signal is processed to extract speaker-specific information. Speaker recognition system (SRS) is one of the most viable biometric authentication systems. In this paper we have presented a literature survey on SRS. This literature survey paper gives brief introduction on SRS first, and then discusses general architecture of SRS, biometric standards relevant to voice/speech, then existing commercial application of SRS, its market survey, national and international status. We have also surveyed various approaches for SRS. Finally, on the basis of our literature survey we conclude by providing the best method and approach for speaker recognition system.

Keywords: SRS, Speaker Recognition System, Voice, Speech;

1. Introduction

1.1. Brief Overview of Speaker Recognition

Voice biometrics specifically was first developed in 1970, and although it has become a sophisticated security tool only in the past few years, it has been seen as a technology with great potential for much longer. At a Gartner Group IT/Expo event held in 1997, Microsoft founder Bill Gates said, "Biometric technologies, those that use voice, will be one of the most important IT innovations of the next several years."

The most significant difference between voice biometrics and other biometrics is that voice biometrics is the only commercial biometrics that process acoustic information. Most other biometrics is image-based. Another important difference is that most commercial voice biometrics systems are designed for use with virtually any standard telephone or on public telephone networks. The ability to work with standard telephone equipment makes it possible to support broad-based deployments of voice biometrics applications in a variety of settings. In contrast, most other

* Corresponding author. Tel.: +91 022 26201606; fax: +91 022 26210139.
E-mail addresses: nirmala@cdacmumbai.in

biometrics requires proprietary hardware, such as the vendor's fingerprint sensor or iris-scanning equipment. By definition, voice biometrics is always linked to a particular speaker. The best-known commercialized forms of voice biometrics are Speaker Recognition. Speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices.

A speaker's voice is extremely difficult to forge for biometrics comparison purposes, since a myriad of qualities are measured ranging from dialect and speaking style to pitch, spectral magnitudes, and format frequencies. The vibration of a user's vocal chords and the patterns created by the physical components resulting in human speech are as distinctive as fingerprints. Voice Recognition captures the unique characteristics, such as speed and tone and pitch , dialect etc associated with an individual's voice and creates a non-replicable voiceprint which is also known as a speaker model or template. This voiceprint which is derived through mathematical modeling of multiple voice features is nearly impossible to replicate. A voiceprint is a secure method for authenticating an individual's identity that unlike passwords or tokens cannot be stolen, duplicated or forgotten

1.2. Voice Production Mechanism

The origin of differences in voice of different speakers lies in the construction of their articulatory organs, such as the length of the vocal tract, characteristics of the vocal chord and the differences in their speaking habits. An adult vocal tract is approximately 17 cm long and is considered as part of the speech production organs above the vocal folds (earlier called as the vocal chords). As shown in Figure 1.2 (a), the speech production organs includes the laryngeal pharynx (below the epiglottis), oral pharynx (behind the tongue, between the epiglottis and vellum), oral cavity (forward of the velum and bounded by the lips, tongue, and palate), nasal pharynx (above the velum, rear end of nasal cavity) and the nasal cavity (above the palate and extending from the pharynx to the nostrils). The larynx comprises of the vocal folds, the top of the cricoids cartilage, the arytenoids cartilages and the thyroid cartilage. The area between the vocal folds is called the glottis.

The resonance of the vocal tract alters the spectrum of the acoustic as it passes through the vocal tract. Vocal tract resonances are called formants. Therefore the vocal tract shape can be estimated from the spectral shape (e.g., formant location and spectral tilt) of the voice signal. Speaker recognition systems use features generally derived only from the vocal tract. The excitation source of the human vocal also contains speaker specific information. The excitation is generated by the airflow from the lungs, which thereafter passes through the trachea and then through the vocal folds. The excitation is classified as phonation, whispering, frication, compression, vibration or a combination of these. Phonation excitation is caused when airflow is modulated by the vocal folds.

When the vocal folds are closed, pressure builds up underneath them until they blow apart. The folds are drawn back together again by their tension, elasticity and the Bernoulli effect. The oscillation of vocal folds causes pulsed stream excitation of the vocal tract. The frequency of oscillation is called the fundamental frequency and it depends upon the length, mass and the tension of the vocal folds. The fundamental frequency therefore is another distinguishing characteristic for a given speaker.

Whispered excitation is caused by the flow of air rushing through a small triangular opening between the arytenoids cartilages at the rear of the nearly closed vocal folds. A turbulent airflow results after this, which has a wide band noise characteristic. Frication excitation is caused due to the constrictions in the vocal tract. The shape of the broadband noise excitation depends upon the place, shape, and degree of constriction determine. The spectral concentration generally increases in frequency when the constriction moves forward. Sounds that are generated by friction are called fricatives. Frication can occur with or without phonation. Compression excitation is produced from release of a completely closed and pressurized vocal tract. This results in a silence (in the pressure accumulation phase) followed by a short noise burst. If the release is sudden, a stop or plosive is generated. If the release is gradual, an affricate is formed. Vibration excitation is a result of air being forced through a closure other than the vocal folds, especially at the tongue.

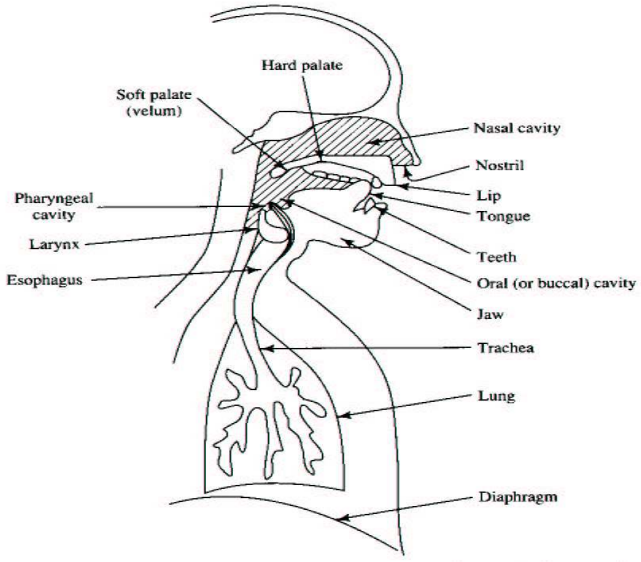

Fig 1 The Speech Production Mechanism

Speech produced by phonated excitation is called voiced, speech produced by phonated excitation plus frication is called mixed voice and speech produced by other types of excitation is called unvoiced. Due to the differences in the manner of production it is reasonable to expect some speech models to be more accurate for certain classes of excitation than the others. Unlike phonation and whispering the places of frication, compression and vibration excitation are actually inside the vocal tract itself. This could cause difficulties for models that assume an excitation at the bottom end of the vocal tract. The respiratory system also plays a role in the resonance properties of the vocal system of an individual. When the vocal folds are in vibration, resonances occur above and below the folds. Sub glottal resonances are largely dependent upon the properties of the trachea, which is typically 12 cm long and 2 cm in diameter , made up of rings of cartilage joined together by connective tissue joining the lungs and the larynx. Due to this physiological dependence, the sub glottal resonances possess speaker dependent properties. Other physiological speaker dependent properties include vital capacity (the maximum volume of air one can blow out after maximum intake), maximum phonation time (the maximum duration a syllable can be sustained), phonation quotient (ratio of vital capacity to maximum phonation time) and glottal airflow (amount of air going through vocal folds). Other aspects of speech production that could be useful for discriminating between speakers are learned characteristics, including speaking rate, prosodic effects and dialect .

1.3. How the Technology Works

The underlying premise for speaker recognition is that each person's voice differs in pitch, tone, and volume enough to make it uniquely distinguishable. Several factors contribute to this uniqueness: size and shape of the mouth, throat, nose, and teeth, which are called the articulators and the size, shape, and tension of the vocal cords. The chance that all of these are exactly the same in any two people is low. The manner of vocalizing further distinguishes a person's speech: how the muscles are used in the lips, tongue and jaw. Speech is produced by air passing from the lungs through the throat and vocal cords, then through the articulators. Different positions of the articulators create different sounds. This produces a vocal pattern that is used in the analysis.

A visual representation of the voice can be made to help the analysis. This is called a spectrogram also known as voiceprint, voice gram, spectral waterfall, and sonogram. A spectrogram displays the time, frequency of vibration of the vocal cords (pitch), and amplitude (volume). Pitch is higher for females than for males.

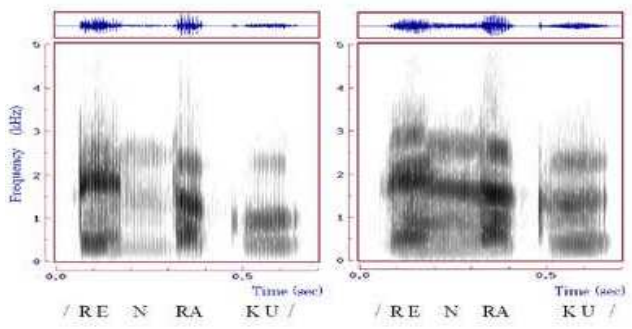

Fig 2. These voiceprints are a visual representation of two different speakers saying “RENRAKU

1.4. Methodology

Each speaker recognition system has two phases: Enrollment and verification. During enrollment, the speaker's voice is recorded and typically a number of features are extracted to form a voice print, template, or model. In the verification phase, a speech sample or "utterance" is compared against a previously created voice print

Speaker recognition **systems fall into two categories: text-dependent and text-independent.**

If the text must be the same for enrollment and verification this is called text-dependent recognition. In a text-dependent system, prompts can either be common across all speakers (e.g.: a common pass phrase) or unique. In addition, the use of shared-secrets (e.g.: passwords and PINs) or knowledge-based information can be employed in order to create a multi-factor authentication scenario.

Text-independent systems are most often used for speaker identification as they require very little if any cooperation by the speaker. In this case the text during enrollment and test is different. In fact, the enrollment may happen without the user's knowledge, as in the case for many forensic applications. As text-independent technologies do not compare what was said at enrollment and verification, verification applications tend to also employ speech recognition to determine what the user is saying at the point of authentication.

2. General Speaker Recognition System Architecture

There are two major commercialized applications of speaker recognition technologies and methodologies: Speaker Identification and Speaker Verification.

2.1. SIS (Speaker Identification System)

Speaker Identification can be thought of as the task of finding who is talking from a set of known voices of speakers. It is the process of determining who has provided a given utterance based on the information contained in speech waves. The unknown voice comes from a fixed set of known speakers, thus the task is referred to as closed set identification. Speaker identification is a 1: N match where the voice is compared against N templates. Error that can occur in speaker identification is the false identification of speaker.

2.2. SVS(Speaker Verification System)

Speaker Verification on the other hand is the process of accepting or rejecting the speaker claiming to be the actual one. Since it is assumed that imposters (those who fake as valid users) are not known to the system, this is referred to as the open set task. Speaker verification is a 1:1 match where one speaker's voice is matched to one template

(also called a "voice print" “speaker model” or "voice model"). Errors in speaker verification can be classified into the following two categories: (1) false rejections: a true speaker is rejected as an imposter, and (2) False acceptances: a false speaker is accepted as a true one.

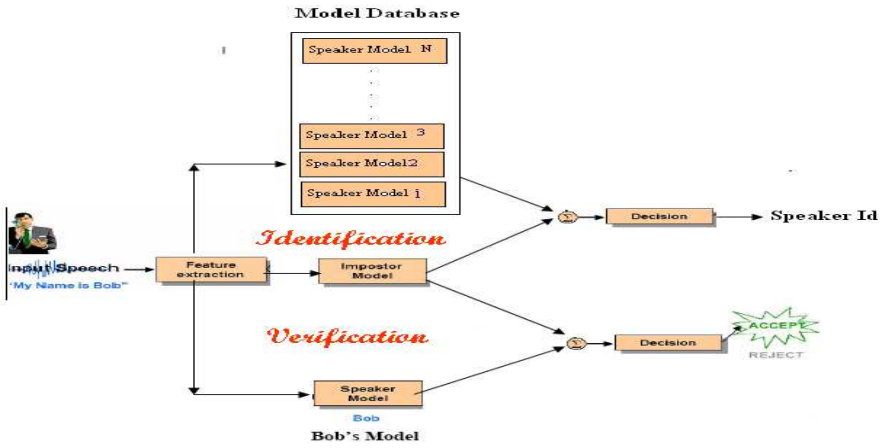

Fig 3. General SIS and SVS Architecture

There are various architectures of Speaker Recognition System provided by researchers and developers. Some of them are:

A. Speaker Identification System

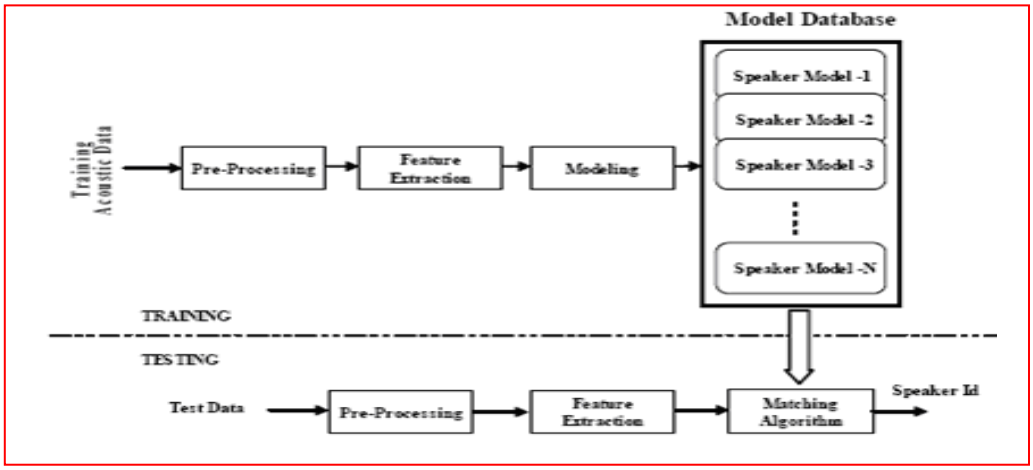

Fig 4. Improved Text-Independent Speaker International Journal of Signal Processing 5;1 2009

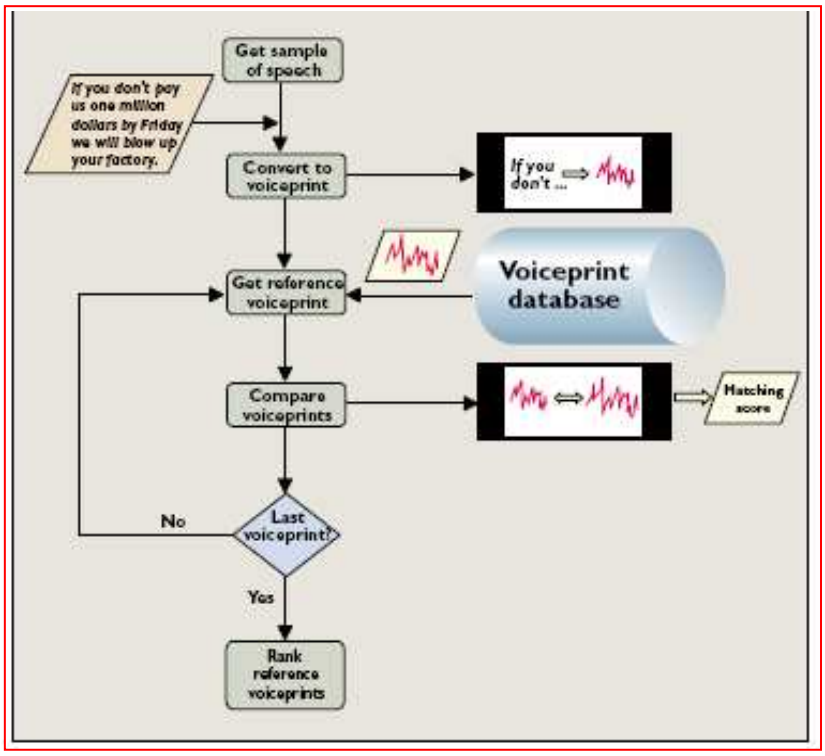

Fig 2.1(b) September 2000/Vol. 43, No. 9 COMMUNICATIONS OF THE ACM

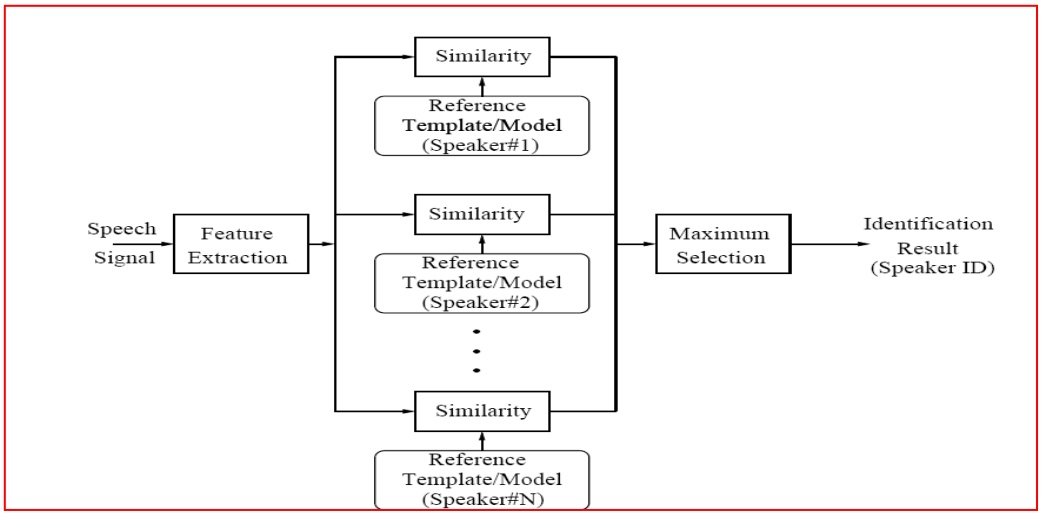

Fig 2.1(c) MS Thesis IIT Madras 2003

B. Speaker Verification System

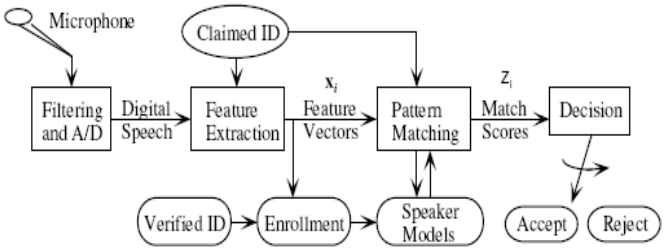

Fig 2.2(a) (Speaker Recognition Proceeding of IEEE 1997)

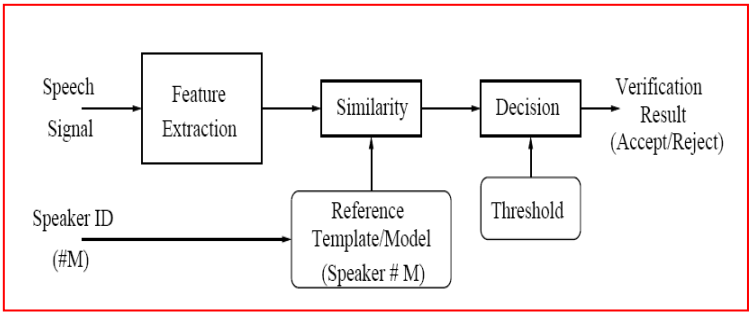

Fig 2.2(b) MS Thesis IIT Madras

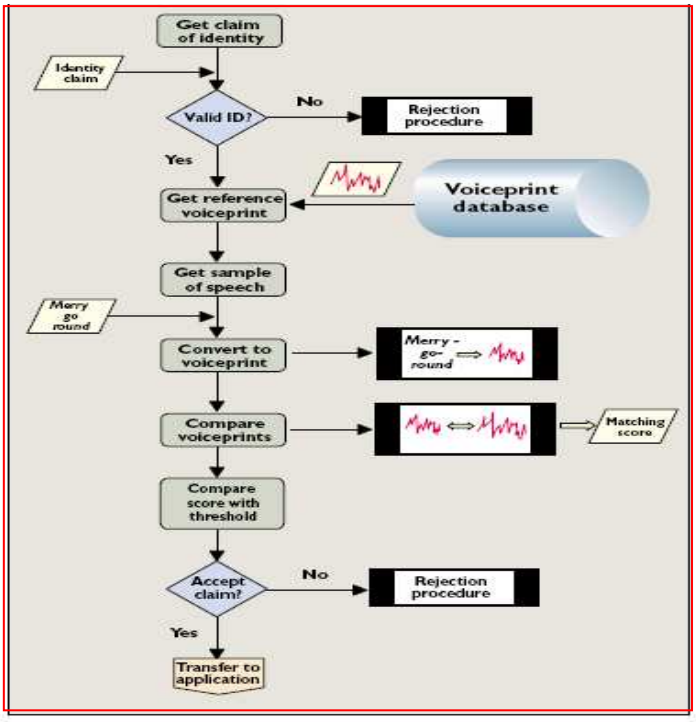

Fig 2.2(c) September 2000/Vol. 43, No. 9 COMMUNICATIONS OF THE ACM

2.3. Descriptions of the key terminology used in the above systems

- Training Data

The process of capturing sample biometric (here voice) data for the purpose of extracting features and generating template/model/reference data is known as Training .The data captured for the above process is training data.

- Test Data

The input biometric data captured by the biometric system for verification or identification is called test data.

- Pre-processing

The Pre-processing of voice data sample involves modulation, removing unwanted noise (denoising) from speech, dividing sounds into voiced and unvoiced sounds and channel compensation. Pre-processing is basically done for speech /voice enhancement.

- Feature Extraction

The process of converting a captured biometric (here voice) sample into biometric data (representing characteristic of the sample) so that it can be compared to a reference template.

- Modeling

The process of creating template/speaker model from the features vectors extracted from the voice data sample is called Modeling.

- Model Database

The repository or database of templates/speaker models/reference model which are later used for recognition by biometric system is known as a Model database.

- Matching

The process of comparing a biometric (here voice) sample against a previously stored template and scoring the level of similarity. An accept or reject is then based upon whether this score exceeds the given threshold

- Voiceprint

A sample of speech that has been converted to a form that a voice biometrics system can analyze. A voice print is not a recording, set of words or a wave pattern of a voice. It cannot be played back or used for any other purpose than a comparison with subsequent voice prints.

- Filtering and A/D

The process of converting voice data sample which are originally in wave/analog form to digital form is called Filtering and A/D.

- Enrolment

The process involved when a person enrolls their voice data onto a voice biometric system to create a reference template for future verification / identification

- Pattern Matching

The process of comparing a biometric (here voice) sample against a previously stored template and scoring the level of similarity. An accept or reject is then based upon whether this score exceeds the given threshold

- Digital Speech

The modulated speech obtained using filtering is a digital speech.

- Reference Template

Reference templates are the template generated using feature vectors of a voice data sample .It is also known as a reference model or speaker model.

- Threshold

The point that must be reached for a speech sample to be considered a match with a previously enrolled voiceprint. The threshold level can be changed depending on the security requirements of a customer

- Encryption

Encryption is the conversion of biometric data into a form that cannot be understood or manipulated by unauthorized people.

3. Voice Biometric Standards

Standards play an important role in the development and sustainability of technology, and work in the international and national standards arena will facilitate the improvement of biometrics. The major standards work in the area of speaker recognition involves:

- Speaker Verification Application Program Interface(SVAPI)
- Biometric Application Program Interface (BioAPI)
- Media Resource Control Protocol (MRCP)
- Voice Extensible Markup Language (VoiceXML)
- Voice Browser (W3C)

Of these, BioAPI has been cited as the one truly organic standard stemming from the BioAPI Consortium, founded by over 120 companies and organizations with a common interest in promoting the growth of the biometrics market.

4. Commercial Application of SRS

The applications of speaker recognition technology are quite varied and continually growing. Voice biometric systems are mostly used for telephony-based applications. Voice verification is used for government, healthcare, call centers, electronic commerce, financial services, and customer authentication for service calls, and for house arrest and probation-related authentication. Below is an outline of some broad areas where speaker recognition technology has been or is currently used.

- Union Pacific Railroad

Union Pacific moves railcars back and forth across the United States every day. The railcars travel loaded in one direction and empty on the way back. When the loaded railcar arrives, the customer is notified to come pick up the contents. Once emptied, the customer needs to alert Union Pacific to put the railcar back to work. Union Pacific now has an automated system that utilizes voice authentication to allow a customer to release empty railcars. Customers enroll in the voice authentication system over the phone. When they call back to release an empty railcar, the system authenticates them and allows them to release their railcars. In this case, voice authentication has allowed customers to get off the phone faster, and Union Pacific to guarantee that a customer is not releasing a railcar that doesn't belong to him.

- New York Town Manor

New York Town Manor is a residential community in Pennsylvania designed for senior citizens with technologically advanced features. The residents no longer have to remember passwords. They do carry ID cards that are used in conjunction with voice authentication to allow access to the complex.

To enter their apartments, they speak for a few seconds while the system authenticates them. With this approach, voice authentication provides an extra measure of security.

- Bell Canada

Technicians for Bell Canada used to have to carry laptops on the job with them. A technician would dial up using a modem to report the current job as finished and to get the next job. Bell Canada has rolled out a new system that uses voice authentication to verify the identity of the technician through a phone call and give him access to the data. This eliminates the need for a laptop.

- Password Journal

Anyone who has ever had a diary has probably worried that someone would read it without permission. One company has solved this problem by adding voice authentication as a privacy measure to their Password Journal product. The journal has its own speaker, raises an alarm if an unauthorized person attempts to access it, and keeps track of how many failed attempts there have been.

- Password Reset

Some companies are allowing users to reset passwords themselves. Users dial an automated system. The system asks questions. When the user answers, the system authenticates his voice and allows him to reset his own password. This saves companies time and money in support costs, and users need not spend time on hold waiting for the next available support person.

- Banking

Reducing crime at Automated Teller Machines is an ongoing struggle. Banks have started using biometrics to authenticate users before allowing ATM transactions. Users generally must provide a pin number and a voice sample to be allowed access. Royal Canadian Bank is using voice authentication to allow access to telephone banking.

- US Social Security Administration

The United States Social Security Administration is using voice authentication to allow employers to report W-2 wages online. Used in combination with a pin number, the voice authentication provides system security and user convenience.

- Law Enforcement

In Louisiana, criminals are kept on a short leash with voice biometrics. This inexpensive approach allows law enforcement to check in with offenders at © SANS Institute 2004, Author retains full rights. Key fingerprint = AF19 FA27 2F94 998D FDB5 DE3D F8B5 06E4 A169 4E46 © SANS Institute 2004, As part of the Information Security Reading Room Author retains full rights. Lisa Myers Page 12 7/24/2004 random times of the day. The offender must answer the phone and speak a phrase that is used for authentication. This system guarantees that they are where they are supposed to be! Voice authentication has also been used in criminal cases, such as rape and murder cases, to verify the identity of an individual in a recorded conversation. There is a terrorism application also. Voice authentication is frequently used to validate the identity of terrorists such as Osama Bin Laden on recorded conversations. Hopefully these clues will one day assist in his capture.

- AHM (Australia Health Management)

Since 2007, Australia private health insurer AHM has successfully managed one of the largest public-facing deployments of speaker verification. With more than 400,000 yearly calls into its main contact center, ahm has implemented an automated voice verification system to provide quick, accurate authentication of callers enhancing member security and improving the customer experience.

- VoiceCash

VoiceCash, an enabler of mobile payment solutions, has made some noise this week with a couple of press releases formally launching its prepaid MasterCard twin card for the German market and announcing a new office and regional CEO in Dubai.

Based in Germany, VoiceCash is targeting consumers interested in cross-border money transfers offering pre-paid payment cards that can be managed online or via SMS communications. The transfers can be authenticated utilizing voice verification technology supplied by VoiceTrust.

- SIMAH

The Saudi Arabia Credit Bureau is deploying a voice biometric solution provided by Agnitio and IST, a contact center system integrator. The technology is part of IST's iSecure product and will be deployed through SIMAH's new Cisco contact center.

- Vodafone Turkey

Offers Customers Voice Biometric-Based Authentication. The adoption of voiceprints to authenticate wireless subscribers is accelerating, thanks to a new installation of PerSay's VocalPassword(TM) developed and installed by Turkish speech application specialist SPEECHOUSE at Vodafone Turkey. With roughly 16 million subscribers, Vodafone Turkey is the second largest mobile carrier in the country. Yet, if past is prologue, the

incorporation of voice authentication into the customer care fabric of any Vodafone subsidiary is bound to have implications across all of its properties - an empire of over 300 million customers.

The deployment is a milestone in a couple of respects. The mobile market holds huge potential for speaker authentication for customer care and electronic payments. In addition, Vodafone Turkey clearly sees subscriber authentication as a source of differentiation in a long-standing battle for share versus Turkcell. In that pursuit, SPEECHOUSE has successfully integrated speaker authentication into the IVR-based Vodafone Voice Portal Platform. The immediate result is the use of a spoken password for secures self service applications the mobile equivalent of password reset, GSM PUK (Personal Unlocking Key) reset.

PerSay VocalPassword is a biometric speaker verification system that verifies a speaker during an interaction with an IVR or a voice application.

Vodafone Turkey has integrated PerSay VocalPassword with Avaya Voice Portal Platform to enable secure self-service applications such as GSM Personal Unlocking Key reset and access to Vodafone Call Centers.

5. Market Survey

Tabel 1.List of Vendors

S.No	Country	Company/Website	Product
1.	USA,New York	Persay http://www.persay.com/	Free Speech
2.	Spain	Agnito http://www.agnitio.es	Voice / Speech Recognition, Physical Access Control, Justice / Law Enforcement, Time and Attendance, HealthCare Biometrics, Financial and Transactional, Other Uses of Biometrics
3.	Slovenia, Europe	TAB Systems Inc. http://www.tab-systems.com/	Product & Services: Facial Recognition, Time & Attendance, Voice Recognition, Physical Access Control. .
4.	Washington DC	DAON http://www.daon.com	Fingerprint Readers, Iris Recognition, Facial Recognition, Voice / Speech Recognition, Smart Cards, Signature / Keystroke, Middleware / Software, Logical Access Control, Border Control / Airports.
5.	USA	Smartmatic http://www.smartmatic.com	Facial Recognition, Fingerprint Readers, Iris Recognition, Voice/Speaker, Border Control/Airports, Justice/Law Enforcement and Physical Access Control.
6.	Russia	SPEECH TECHNOLOGY CENTER (STC) http://www.speechpro.com/	Voice / Speech Recognition, Financial and Transactional, Justice / Law Enforcement, Logical Access Control, Other Uses of Biometrics.
7.	Italy	Loquendo http://www.loquendo.com/en/	Biometric Sensors and Detectors, Voice / Speech Recognition, Financial and Transactional, Justice / Law Enforcement, Logical Access Control, Other Uses of Biometrics, Physical Access Control.

8.	Barcelona	SeMarket http://www.semarket.com	Smart cards, Voice / Speech Recognition Biometrics
9.	Newyork	Recognition Technologies Ltd. http://www.speakeridentification .com/	RecoMadeEasy® Speaker Recognition (SPKR) -- SIV System.

6. Current Status of SRS

6.1. National Status

Table 2. R&D in India

S.No	Institute/Organization	Project
1	Indian Institute of Technology, Guwahati (http://www.iitg.ac.in/ece/)	1.1) Study of Source Features for Speech Synthesis and Speaker Recognition Sponsoring Agency: UK-INDIA Education and Research Initiative (UKIERI) 1.2) Keyword spotting in continuous speech Sponsoring Agency: DST, NewDelhi It's a Indo-Swiss Joint Research Project between IIT Guwahati, IIT Madras & IDIAP Switzerland 1.3) Development of Person Authentication System based on Speaker Verification in Uncontrolled Environment Sponsoring Agency: MIT (http://www.mit.gov.in/content/list-ongoing-projects)
2	Indian Institute of Technology, Kharaghpur	2.1) Development of speaker verification software for single to three registered user(s) 2.2) Development of robust speaker verification system to increase security in limited user Environment 2.3) Development of a Lung Sound Analyzer 2.4) FPGA based Automatic Speaker Recognition 2.5) Complex Biomedical Signal Analysis 2.6) Development of Speaker Recognition Software for Telephone Speech 2.7) Development of speech database for speaker recognition application

3.	Indian Institute of Technology, Madras Donlab: http://lantana.tenet.res.in/projects.php	TATA Power Project involves Speaker Identification in Radio Communications Channel.
4.	CFSL, Chandigarh	CFSL is the first Forensic Laboratory in the Country to develop text independent speaker identification system indigenously. A number of important cases related to corruption, threatening calls and identification of individuals through their voice have been solved by CFSL, Chandigarh. CFSL Chandigarh has the technique to match voices irrespective of the language used by the person.
5.	Speech Processing Lab Language Technology Research Centre, IIIT Hyderabad	The speech processing lab at Language Technologies Research Center, IIITH is actively involved in research and imparting training at the graduate and under-graduate level through semester long courses in the fields of speech processing which includes speech signal analysis, speech recognition, speech synthesis, speaker recognition, speech enhancement and spoken dialog systems. Two research projects with IIT Guwahati.

6.2. International Status

Table 3. R&D Worldwide

S.No	Institute /Organization	Project/Description
1.	Speech Technology and Research Laboratory, SRI International,CA	**Speaker Recognition and Talk Printing** The goal of our project is to discover "TalkPrint" features -- features that capture these habitual variations in speaking style, and to model them in conjunction with standard features to improve automatic speaker recognition.
2.	Speech and Speaker Modeling Group, University Of Texas at Dallas	**2.1) Dialect / Accent Classification** This project is focused on developing new algorithms for automatic detection and classification of Dialect/Accent in natural speech communication. **2.2) In-Set Speaker Recognition** This project focuses on small enrollment and test data sizes (2-8 sec), with open domain content. **2.3) Speaker Normalization** In Normalization, focus is on making speech systems robust under differing operational conditions. These include variation in environment, language, bandwidth, noise, and speaker differences.
3.	The Centre for Speech Technology Research, University of Edinburgh, United Kingdom	**Voice transformation** Transforming the quality and intonation of the speech of one speaker so that it sounds like another speaker

4.	Human Language Technology Group, Lincoln Laboratory, Massachusetts Institute of Technology	**Forensic Speaker Recognition Project** The Science and Technology Directorate Command, Control and Interoperability Division's (CID) forensic speaker recognition technology is reducing voice analysis completion rates by half. CID is partnering with the United States Secret Service and the Massachusetts Institute of Technology's Lincoln Laboratory to develop a cutting-edge suite of software tools that automate labor-intensive components of speech analysis and compare speaker language and dialect features. The technology identifies how common a speech feature is by comparing the feature between speech samples and the U.S. population at large. Previously, it could take up to ten hours to analyze five minutes of speech evidence. The software tools under development have reduced this analysis time to less than one hour. This project is part of the Knowledge Management Tools program area of CID.

7. Publicly Available Speech Database

1. TIMIT

TIMIT is available from the LDC for $100
http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1
A 3% sample is distributed for free with NLTK [http://nltk.sourceforge.net

2. NIST (Performance Evaluation)

http://itl.nist.gov/iad/mig/tests/sre

3. NOIZEUS database

http://www.utdallas.edu/~loizou/speech/noizeus/

4. NTIMIT

http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S2

5. YOHO database

http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S16

6. Switchboard (SWBI) conversational speech

8. Performance Metrics

Biometric systems are not perfect. There are two important types of errors associated with biometric system, namely a false accept rate (FAR) and a false reject rate (FRR). The FAR is the probability of wrongfully accepting an impostor user, while the FRR is the probability of wrongfully rejecting a genuine user. System decisions (i.e. accept/reject) is based on so-called thresholds. By changing the threshold value, one can produce various pairs of (FAR,FRR). For reporting performance of biometric system in verification mode, researchers often

use a decision error trade-off (DET) curve. The DET curve is a plot of FAR versus FRR and shows the performance of the system under different decision thresholds [54], see Figure 8(a). Using machine learning terminology, FAR and FRR are analogues to False Negative and False Positive, respectively. A modified version of the DET curve is a ROC (Receiver Operating Characteristic) curve, which is widely used in the machine learning community. The difference between DET and ROC curves is in ordinate axis. In the DET curve the ordinate axis is FRR, while in the ROC curve it is 1-FRR (i.e. probability of correct verification). Usually, to indicate the performance of biometric system by a single value in verification mode, an equal error rate (EER) is used. The EER is the point on the DET curve, where FAR=FRR, see Figure 8(a). To evaluate the performance of a biometric system in identification mode, a cumulative match characteristics (CMC) curve can be used. The CMC curve is a plot of rank versus identification probability and shows the probability of a sample being in the top closest matches [55], see Figure 8(b). In identification mode, to indicate performance of the system by a single number, the recognition rate (i.e. identification probability at rank 1) is used. In the next sections, when performance of the method is referred to the recognition rate the system is evaluated in the identification mode, and when it is referred to the EER the system is evaluated in the verification mode.

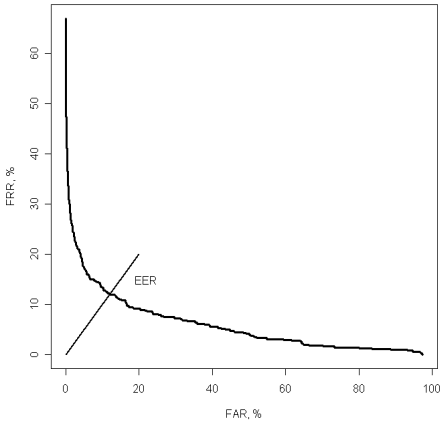

Fig 8(a) Example of DET Curve

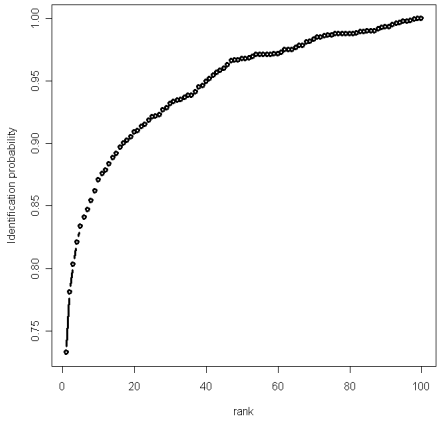

Fig 8(b) Example of CMC Curve

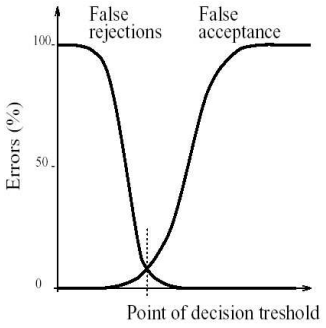

Fig 8(c)

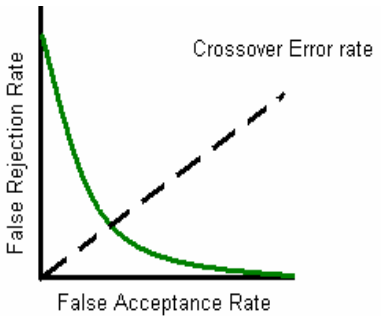

Fig 8(d)

If you plot FAR and FRR against each other, the point at which they intersect is called the crossover error rate (CER). The lower the CER, the better the system is performing

9. Issues Pertaining to SRS

Speaker recognition and verification has been an area of research for more than four decades and thus have many challenges that are needed to overcome. We have discussed many approaches previously and there are some limitations with the approach one follows and thus poses few challenges in the area. We can discuss some of the issues which forms current area of research in speaker recognition now a days.

Standard approaches to automatic speaker recognition use spectrum-related features based on very short time slices of speech. Models based on such information suffer from a lack of robustness to channel mismatches, and fail to capture longer-range characteristics of how a person talks, including the speaker's word patterns, and patterns in speech prosody (the timing, pausing, and intonation of speech). One core technical challenge in this work is to design long-range features (which by definition occur less frequently than very short-range features) that provide robust additional information even for short (e.g., 30 seconds) training and test spurts of speech.

A second crucial challenge area is to develop methods for feature selection and model combination at the feature level, that can cope with large numbers of interrelated features, odd feature space distributions, inherent missing features (such as pitch when a person is not voicing), and heterogeneous feature types.

A third issue is how to employ Talk Print features successfully for a new language and across languages, since traditional speech recognition and derived features are inherently language-dependent.

The “realistic conditions” in which a voice is recorded impose on speech signals a high degree of variability. All these sources of variability can be classified as follows:

(i) peculiar intra speaker variability: type of speech, gender, time separation, aging, dialect, sociolect, jargon, emotional state, use of narcotics, and so forth

(ii) forced intra speaker variability: Lombard effect, external-influenced stress, and cocktail-party effect

(iii) channel-dependent external variability: type of handset and/or microphone, landline/mobile phone, communication channel, bandwidth, dynamic range, electrical and acoustical noise, reverberation, distortion, and so forth

The tuning of decision thresholds is very troublesome in speaker verification. If the choice of its numerical value remains an open issue in the domain (usually fixed empirically), its reliability cannot be ensured while the system is running. This uncertainty is mainly due to the score variability between trials, a fact well known in the domain. This score variability comes from different sources.

First, the nature of the enrollment material can vary between the speakers. The differences can also come from the phonetic content, the duration, the environment noise, as well as the quality of the speaker model training.

Secondly, the possible mismatch between enrollment data (used for speaker modeling) and test data is the main remaining problem in speaker recognition. Two main factors may contribute to this mismatch: the speaker him/herself through the intra speaker variability (variation in speaker voice due to emotion, health state, and age) and some environment condition changes in transmission channel, recording material, or acoustical environment.

On the other hand, the inter speaker variability (variation in voices between speakers), which is a particular issue in the case of speaker-independent threshold-based system, has to be also considered as a potential factor affecting the reliability of decision boundaries. Indeed, as this inter speaker variability is not directly measurable, it is not straightforward to protect the speaker verification system (through the decision making process) against all potential impostor. Lastly, as for the training material, the nature and the quality of test segments influence the value of the scores for client and impostor trials.

9.1. Some security related issues with voice biometrics are:

Hackers might attempt to gain unauthorized access to a voice authenticated system by playing back a pre-recorded voice sample from an authorized user. One way to thwart this sort of attack is to use a challenge response system. The system can prompt the user to repeat a random set of words or phrases in a specified order. Then the system verifies that the voice sample matches, and that the sample contains the requested words and phrases in the correct order. This makes it difficult for anyone to use a prerecorded voice sample for authentication.

A major issue facing all biometric technologies that store data is maintaining the privacy of that data. As soon as a user registers with a voice biometric system, that voiceprint is stored somewhere just like an address or a phone number. What if companies decide to sell voiceprints like addresses? Will we need a public "opt out" registry like we now have for telephone numbers to prevent the sharing of biometric data?

If the data is encrypted in storage and in transport, there is always the possibility of cracking the encryption and stealing the data. Biometric data is unique in that once it has been compromised; a user cannot merely request a new one like one can with a password reset or a new credit card number. Each person only has one voice.

10. SRS Modules

Speaker Recognition Modules mainly comprises of :

i) Preprocessing
- Speech Enhancement(Denoising)
- Channel Compensation

ii) Feature Extraction
- Low Level Feature Extraction
- High Level Feature Extraction

iii) Modeling
- Speaker Model Generation
- Imposter Model Generation
- Ciphering

iv) Matching/Decision Logic
- Score Normalization

10.1. Preprocessing

The captured voice may contain unwanted background noise, unvoiced sound, and there can be a device mismatch, environmental mismatch between training and testing voice data which subsequently leads to degradation in the performance of Speaker Recognition System. The process of removal of this unwanted noise, dividing sounds into voiced and unvoiced sounds and channel compensation etc for the enhancement of speech/voice is called pre-processing.

10.1.1. Speech Enhancement (Denoising)

Most published works in the areas of speech recognition and speaker recognition focus on speech under the noiseless environments. The quality of a speech signal is judged, depending on the application, by one or more of

the following factors—intelligibility, perceptual quality, listener fatigue, signal-to-noise ratio (SNR), speech distortion, and (occasionally) recognition accuracy of an automatic speech recognizer. Numerous schemes have been proposed and implemented that perform speech enhancement under various constraints/assumptions and deal with different issues and applications.

The principal degradations in the speech signal are:

1. Additive acoustic noise – such as the noise added to the speech signal when recorded in an environment with noticeable background noise, like in an aircraft cockpit.

2. Acoustic reverberation – results from the additive effect of multiple reflections of an acoustic signal.
3. Convolutive channel effects - resulting in an uneven or band limited response, can result when the communication channel is not modeled effectively for the channel equalizer to remove the channel impulse response.
4. Non-linear distortion such as arises from clipping – such as when inappropriate gain is applied at the signal input stage.
5. Additive broadband electronic noise
6. Electrical interference
7. Codec distortion– distortion caused by the coding algorithm due to compression
8. Distortion introduced by recording apparatus – poor response of microphone

The aims of speech enhancement vary according to the application and may include:

1. Improvements in the intelligibility of speech to human listeners.
2. Improvement in the quality of speech that make it more acceptable to human listeners.
3. Modifications to the speech that lead to improved performance of automatic speech or speaker recognition systems.
4. Modifications to the speech so that it may be encoded more effectively for storage or transmission.

Table 4. Various approaches for Speech Enhancement

S. No.	Approach	Characteristics
1.	Perceptually weighted multi-band spectral subtraction speech enhancement technique [2008][1]	1.1) Estimates a suitable factor that will subtract just the necessary amount of the noise spectrum from each frequency bin (ideally) to prevent destructive subtraction of the speech. 1.2) Improvement over the conventional power spectral subtraction method. 1.3) This technique takes into consideration that the noise elements are masked by the speech power in the formant regions and conversely unmasked in the valleys between the formants.
2.	Spectral Subtraction Speech Enhancement Technology Based on Fast Noise Estimation [2009][2]	2.1) Based on fast noise estimation 2.2) This algorithm can significantly improve the performance compared to the basic spectral subtraction algorithm.
3.	Speech Enhancement Based on Generalized Minimum Mean Square Error Estimators and Masking Properties of the Auditory System [2006][3]	3.1) A Generalized MMSE estimator (GMMSE) is formulated after study of different methods of MMSE family. 3.2) GMMSE auditory masking threshold (AMT) enhancement method is combination of GMMSE and auditory enhancement scheme using the masking threshold of the human auditory system.
4.	MMSE estimator for speech enhancement considering the constructive and destructive interference of noise [2010][4]	This method considers the constructive and destructive interference of noise in the speech signal.

5.	Extension of the signal subspace speech enhancement approach to colored noise [2003][5]	This algorithm is an extension to signal subspace approach for speech enhancement to colored-noise processes.
6.	Signal subspace approach for speech enhancement in non-stationary noises [2007] [6]	6.1) This approach is based on Signal Subspace Approach combined with RL noise estimation for non-stationary noise. 6.2) Signal/ Noise Karhunen-Loeve transform is used to implement signal subspace decomposition for noisy speech signal. 6.3) RL noise estimation is used for non-stationary noise.
7.	Speech enhancement using the multistage Wiener filter [2009][7]	7.1) A subspace speech enhancement approach for estimating a signal which has been degraded by additive uncorrelated noise. 7.2) This approach utilizes the multistage Wiener filter (MWF).
8.	A Modified Speech Enhancement Algorithm Based on the Subspace [2009] [8]	8.1) Modified subspace speech enhancement Algorithm. 8.2) Reduces residual noise caused by wrongly estimating noise Eigen value matrix and speech Eigen value matrix. 8.3) This method tracks real time noise Eigen value matrix in the subspace domain by applying statistical information in the whole time, and corrects speech Eigen value matrix making use of the principle of winner filtering.
9.	Noise Reduction System Based on LPEF and System Identification with Variable Step Size [2007][9]	9.1) Noise reduction method is based on a linear prediction error filter (LPEF) 9.2) Background noise is estimated by system Identification. 9.3) Variable step size based noise reconstruction system is proposed.
10.	Speech Enhancement Using Harmonic Emphasis and Adaptive Comb Filtering [2010][10]	10.1) A spectral weighting function is derived by constrained optimization to suppress noise in the frequency domain. 10.2) Two design parameters are included in the suppression gain, namely, the frequency-dependent noise-flooring parameter (FDNFP) and the gain factor. 10.3) Further enhancement of the harmonics is achieved by adaptive comb filtering derived using the gain factor with a peak-picking algorithm.
11.	An Effective Approach for Speech Enhancement by Multi-band MMSE Spectral Subtraction [2007][11]	11.1) Single channel speech enhancement algorithm that combines the merits of multi-band analysis and MMSE spectral subtraction. 11.2) A scale factor is introduced which reflects the difference of noise influence in different bands to reduce the color noise.
12.	Improvement of speech enhancement techniques for robust speaker identification in noise (Specifically addressed for speaker identification) [2009][12]	12.1) In this approach start-end points detection, silence part removal, frame segmentation and windowing technique have been used to pre-process and wiener filter has been used to remove the silence parts from the speech utterances. 12.2) To measure the performance of the proposed speech enhancement techniques, genetic algorithm has been used as a classifier for the noise robust automated speaker identification system

10.1.2. Channel Compensation

Channel effects, are major causes of errors in speaker recognition and verification systems.The main measures to improving channel robustness of speaker recognition system are channel compensation and channel robust features. Some common channel effects known to cause Corruption to speech signal in speaker recognition and verification systems are :

i) Band limiting where a signal can be distorted by a filtering effect from the medium the signal is being sent through.

ii) Additive White Gaussian Noise (AWGN) which is distortion caused by electrical, thermal and/or environmental factors where random signal distortion is added to a signal during transmission through a vulnerable channel.

iii) Linear Time–Invariant and Linear Time-Variant filtering which are filtering effects that cause convolution distortion to a signal being transmitted

Usually, channel compensation includes feature domain compensation, model domain compensation and score domain compensation.

i)Feature domain compensation: aims to remove channel mismatch when feature vectors are being extracted. These include well-known and widely used techniques such as cepstral mean subtraction, RASTA filtering and cepstral subtraction. MAP (Maximum A Posterior Probability)

ii) Model domain compensation: modifies models to minimize channel mismatch. An example is SMS (Speaker Model Synthesis) , which learns how model parameters change between different channels and applies a transformation to synthesize speaker models under unseen enrollment conditions.

iii)Score domain compensation: attempts to remove model score scales and shifts caused by channel mismatch. Examples of score domain compensation technique are Hnorm and Tnorm.

High level features have the potentiality to improve the channel robustness

Table 5. Various approaches for Channel Compensation

S.No	Approach	Characteristic
1.	Cohort-Based Speaker Model Synthesis For Channel Robust Speaker Recognition [2006][28]	1.1) Mel cepstral coefficients plus delta, which were computed with 20ms frame length every 10ms 1.2) Utilizes channel-dependent UBMs as a priori knowledge of channels for speaker model synthesis. 1.3) Speaker similarity is measured through Kullback-Leibler distance between the speaker model and cohort speaker model
2.	Text-independent Speaker Identification Based on MAP Channel Compensation and Pitch -dependent Features [2010] [51]	2.1) In the first aspect, this paper applies MAP channel compensation technique, which was used in speech recognition, to speaker recognition system. 2.2) In the second aspect, this paper introduces pitch-dependent features and pitch-dependent speaker model to recognize again, and then uses ANN to combine the three pitch-dependent results and one GMM score for getting a fusion result.
3.	Performance of a Text-Independent Remote Speaker Recognition Algorithm over Communication Channels with Blind Equalisation [2005][57]	3.1) Examine the reliability of the MFCC algorithm when applied on speech transmitted through a communication channel with channel equalisation for remote speaker identification 3.2) Blind equalisation techniques with QPSK modulation

10.2. Feature Extraction

The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate).The heart of any speaker recognition system is to extract speaker dependent featuresfrom the speech which should ideally have following characteristics

• have large between-speaker and small within-speaker variability
• be difficult to impersonate/mimic
• not be affected by the speaker's health or long-term variations in voice

• occur frequently and naturally in speech
• Easily measurable
• Not be affected by background noise nor depend on the specific transmission medium
• Occur naturally and frequently in speech.

Types of Features: They are basically categorized into two types

i) Low Level features
ii) High Level Features

10.2.1. Low Level Features

Table 6. Various approaches for extracting low level features of voice signal

S.No	Approach	Characteristics
1	Significance of Formants from Difference Spectrum for Speaker Identification [2007][13]	Auto-associative neural network (AANN) and formant features
2	Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter [2009][14]	Gaussian filter based mel (MFCC)and inverted mel (IMFCC)scaled filter bank is proposed in this paper
3	Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach [2009][15]	3.1) Perceptual Features 3.2) Iterative clustering approach for both speech and speaker recognition
4	Combination of Pitch and MFCC GMM Supervectors for Speaker Verification [2008][16]	Fusion of pitch and MFCC GMM supervectors Systems on score level
5	A Supervised Text-Independent Speaker Recognition Approach [2007][17]	5.1) A text-independent voice recognition system representation of the vocal feature vectors as truncated acoustic matrices with DDMFCC coefficients 5.2) Hausdorff-based metric, used in the speech feature vector classification process
6	Speaker identification using Warped MVDR Cepstral Features [2009][18]	Replaces the widely used Mel-frequency cepstral coefficients by warped minimum variance distortion less response cepstral coefficients for speaker Identification
7	Feature Extraction and Test Algorithm for Speaker Verification [2006][19]	7.1) After MFCC extraction, both Cepstral Mean Subtraction (CMS)and RASTA filtering are used to remove linear channel convolutional effect on the cepstral features 7.2) The voiced vectors and unvoiced vectors are transformed independently in the framework Gaussianization 7.3) The test sequence is adapted to a new model via the UBM
8	Speaker Identification Using Admissible Wavelet Packet Based Decomposition [2010][20]	8.1) Proposed an admissible wavelet packet based filter structure for speaker identification. 8.2) Multiresolution capabilities of wavelet packet transform are used to derive the new features.

10.2.2. High Level Features

Higher level features are long range features of voice that have attracted attention in automatic speaker recognition in recent years. The short term spectral features of voice fails to capture a wealth of longer-range and linguistic information that also resides in the signal. Higher-level information can significantly improve performance when combined with lower-level cepstral information. Higher-level information also offers the possibility of increased robustness to channel variation, since features such as lexical usage or temporal patterns do not change with changes in acoustic conditions. And finally, higher-level features can provide useful metadata about a speaker, such as what topic is being discussed, how a speaker is interacting with another talker, whether the speaker is emotional or disfluent, and so on.

Table 7. Various approaches for extraction of High Level Feature

S.No	Approach	Characteristic
1.	Speaker Verification using Support Vector Machines and High-Level Feature [2007][26]	1.1) Features: word n grams, phones, pitch gestures 1.2) Speaker Model: SVM and new kernel based upon linearizing a log likelihood ratio scoring system
2.	Gaussian Mixture Modelling of Broad Phonetic and Syllabic Events for Text-Independent Speaker Verification [2005][27]	2.1) Presegmentation of utterances at word level using a state-of-the-art English ASR system 2.2) Words are modeled using Hidden Markov Models (HMMs). 2.3) Features: Syllabic event, four broad phonetic classes are used in order to limit the total number of possible syllables. 2.3) Speaker Model: GMM
3.	Extraction and representation of prosodic features for language and speaker recognition [2008][56]	3.1) Syllable-like unit is chosen as the basic unit for representing the prosodic characteristics. 3.2) Approximate segmentation of continuous speech into syllable-like units is obtained by locating the vowel onset points (VOP) automatically 3.3) Quantitative parameters are used to represent F 0 and energy contour in each region between two consecutive VOPs
4.	Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification [2007][59]	4.1) Introduces the use of continuous prosodic features for speaker recognition and we show how they can be modeled using joint factor analysis 4.2) Prosodic feature extracted using a basis consisting of Legendre polynomials

10.3. Modeling

10.3.1. Speaker Model Generation

The feature vectors of speech are used to create a speaker's model. The numbers of reference templates that are required for efficient speaker recognition depend upon the kind of features or techniques that the system uses for recognizing the speaker. In the recognition phase, features similar to the ones that are used in the reference template are extracted from an input utterance of the speaker whose identity is required to be determined. The recognition

decision depends upon the computed distance between the reference template and the template devised from the input utterance.

Table 8. Various approaches for Speaker Model Generation

S.No	Approach	Characteristic
1.	Efficient Speaker Identification and Retrieval [2005][21]	1.1) GMM Simulation algorithm 1.2) GMM compression algorithm
2.	Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter [2009][14]	2.1) GMM based Modeling 2.2) Fusion of two GMM for each speaker one for MFCC and other for IMFCC feature sets
3.	Feature Dimensionality Reduction Through Genetic Algorithms For Faster Speaker Recognition [2008][22]	3.1) GA(Genetic Algorithm) 3.2) Comparison with LDA and PCA
4.	Efficient Speaker Recognition Using Approximated Cross Entropy (ACE) [2007][24]	4.1) Based on approximating Gaussian mixture modeling (GMM) likelihood scoring using approximated cross entropy (ACE) 4.2) GMM compression algorithm
5.	Speaker Model Clustering for Efficient Speaker Identification in Large Population Applications [2009][25]	GMM-based speaker models are clustered using a simple k-means algorithm.

10.3.2. Imposter Modeling

Table 8. Various approaches for Imposter Modeling

S. No	Approach	Characteristics
1.	An experimental comparison of modeling techniques for speaker recognition under limited data condition [2009][62]	1.1) This paper gives an experimental evaluation of the modeling techniques like Crisp Vector Quantization (CVQ), Fuzzy Vector Quantization (FVQ), Self-Organizing Map (SOM), Learning Vector Quantization (LVQ), and Gaussian Mixture Model (GMM) classifiers. 1.2) Gaussian Mixture Model–Universal Background Model (GMM–UBM) is used 1.3) The experimental knowledge is then used to select a subset of classifiers for obtaining the combined classifiers.
2.	Impostor Detection in Speaker Recognition Using Confusion-Based Confidence Measures [2006][63]	Introduces confusion-based confidence measures for detecting an impostor in speaker recognition, which does not require an alternative hypothesis.
3.	Impostor Modeling Techniques For Speaker Verification Based On Probabilistic Neural Networks [2003][64]	The impact of two different impostor modeling techniques on the performance of a Probabilistic Neural Networks (PNNs)-based text-independent speaker verification system is studied.

10.3.3. Ciphering/Encryption

Encryption is the conversion of voice biometric data into a form that cannot be understood or manipulated by unauthorized people.

Table 9. Various approaches for Encryption of Voice Signal

S.No	Approach	Characteristic
1.	Securing Telecommunication based on Speaker Voice as the Public Key [2007][23]	1.1) Proposes a technique to generate a public cryptographic key from user's voice while speaking over a handheld device. 1.2) RSA and DH are used to generate the secret shared key
2.	Speech Encryption Using Circulant Transformations [2002][60]	2.1) A new analog encryption technique using unitary circulant transformationof the sampled analog signal is proposed 2.2) Introduces primarily a phase distortion 2.3) optimal speaker-specific key generation scheme is then developed

10.4. Matching /Decision Logic

10.4.1. Score Normalization

Table 10. Various approaches for Score Normalization

Sr.No	Approach	Characteristics
1.	Score Normalization for Multimodal Recognition Systems [2010][65]	1.1) The paper shows the results obtained using a number of fusion algorithms (Neural Networks, SVM, Weighted Sum, etc.) on the scores generated with three independent monomodal biometric systems (Iris, Signature and Voice). 1.2) Also shows the behavior of the most popular score normalization techniques (z-norm, tanh, MAD, etc). 1.3) A new normalization algorithm (DLin) is proposed.
2.	Double Gauss Based Unsupervised Score Normalization In Speaker Verification [2008][66]	2.1) An unsupervised score normalization is proposed. 2.2) A target speaker score Gauss and an impostor score Gauss are set up as a prior; the parameters of the impostor score model are updated using the test score. 2.3) Then the test score is normalized by the new impostor score Model.
3.	Speaker Verification using Speaker and Test Dependent Fast Score Normalization [2007][67]	3.1) Novel score normalization technique based on test-normalization method (Tnorm) is presented. 3.2) Selects a speaker dependent subset of impostor models from the fixed cohort using distance-based criterion. 3.3) Selection of the sub cohort is made using a distance measure based on a fast approximation of Kullback-Leibler (KL) divergence for Gaussian Mixture Models (GMM). 3.4) The proposed technique is known as KL-Tnorm.

4.	A Cohort Methods for Score Normalization in Speaker Verification System, Acceleration of On-line Cohort Methods [2007] [68]	4.1) A new normalization technique, unconstraint cohort extrapolated normalization, is introduced. 4.2) The world, cohort, and unconstraint cohort normalization techniques are also presented.
5.	Speaker Adaptive Cohort Selection For Tnorm In Text-Independent Speaker Verification [2005][69]	5.1) A new method of speaker Adaptive-Tnorm that offers advantages over the standard Tnorm by adjusting the speaker set to the target model is presented. 5.2) Proposes an approach for speaker dependent Tnorm selection to help improve verification performance.
6	A Unified Framework for Score Normalization Techniques Applied to Text-Independent Speaker Verification [2005][70]	6.1) This paper proposes to unify several of the state-of-the-art score normalization techniques applied to text-independent speaker verification systems. 6.2) A new framework is proposed in which the two well-known Z- and T-normalization techniques can be easily interpreted as different ways to estimate score distributions.
7	Feature And Score Normalization For Speaker Verification Of Cellular Data [2003][71]	7.1) This paper presents some experiments with feature and score normalization for text-independent speaker verification of cellular data. 7.2) The following methods, which have been proposed for feature and score normalization, are reviewed and evaluated on cellular data: cepstral mean subtraction (CMS), variance normalization, feature warping, T-norm, Z-norm and the cohort method.

11. Conclusion

In this paper, we have presented an extensive survey of automatic speaker recognition systems. We have categorized the modules in speaker recognition and discussed the characteristics of different approaches for each module. In addition to this, we have presented a study of the various commercial applications, list of vendors worldwide and the current international and national status of research being carried out in the field of speaker recognition. We have also discussed issues and challenges pertaining to the speaker verification systems.

12. References

1. F.A.; Alam, J.; Alam, F.; O'Shaughnessy, D.;Perceptually weighted multi-band spectral subtraction speech enhancement technique ChowdhuryElectrical and Computer Engineering, 2008. ICECE 2008. International Conference on Digital Object Identifier: 10.1109/ICECE.2008.4769239 Publication Year: 2008 , Page(s): 395 – 399
2. Luo Jun; Zhiming He;Spectral Subtraction Speech Enhancement Technology Based on Fast Noise Estimation;International Conference on Information Engineering and Computer Science, 2009. ICIECS 2009. Digital Object Identifier: 10.1109/ICIECS.2009.5362621, 2009
3. Hansen, J.H.L.; Radhakrishnan, V.; Arehart, K.H; Speech Enhancement Based on Generalized Minimum Mean Square Error Estimators and Masking Properties of the Auditory System ; Audio, Speech, and Language Processing, IEEE Transactions on Volume: 14 , Issue: 6 Digital Object Identifier: 10.1109/TASL.2006.876883 Publication Year: 2006 , Page(s): 2049 – 2063
4. Hasan, T.; Hasan, M.K;MMSE estimator for speech enhancement considering the constructive and destructive interference of noise; Signal Processing, IETVolume: 4 , Issue: 1 Digital Object Identifier: 10.1049/iet-spr.2008.0114 Publication Year: 2010 , Page(s): 1 – 11
5. Lev-Ari, H.; Ephraim, Y;Extension of the signal subspace speech enhancement approach to colored noise; Signal Processing Letters, IEEE Volume: 10 , Issue: 4 Digital Object identifier: 10.1109/LSP.2003.808544 Publication Year: 2003 , Page(s): 104 – 106
6. Chiung-Wen Li; Sheau-Fang Lei;Signal subspace approach for speech enhancement in nonstationary noises ; Communications and Information Technologies, 2007. ISCIT '07. International Symposium on Digital Object Identifier: 10.1109/ISCIT.2007.4392269 Publication Year: 2007 , Page(s): 1580 – 1585
7. Tinston, M; Ephraim, Y;Speech enhancement using the multistage Wiener filter;Information Sciences and Systems, 2009. CISS 2009. 43rd Annual Conference n Digital Object Identifier: 10.1109/CISS.2009.5054690 Publication Year: 2009 , Page(s): 55 – 60
8. Hairong Jia; Xueying Zhang; Chensheng Jin; A Modified Speech Enhancement Algorithm Based on the Subspace ; Knowledge Acquisition and Modeling, 2009. KAM '09. Second International Symposium on Volume: 3 Digital Object Identifier: 10.1109/KAM.2009.19 Publication Year: 2009 , Page(s): 344 – 347
9. Sasaoka, N.; Watanabe, M.; Itoh, Y.; Fujii, K; Noise Reduction System Based on LPEF and System Identification with Variable Step Size ;Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on Digital Object Identifier: 10.1109/ISCAS.2007.378850 Publication Year: 2007 , Page(s): 2311 – 2314

10. Wen Jin; Xin Liu; Scordilis, M.S.; Lu Han; Speech Enhancement Using Harmonic Emphasis and Adaptive Comb Filtering; Audio, Speech, and Language Processing, IEEE Transactions on Volume: 18 , Issue: 2 Digital Object Identifier: 10.1109/TASL.2009.2028916 Publication Year: 2010 , Page(s): 356 – 368
11. Ning Cheng; Wen-Ju Liu; Bo Xu; An Effective Approach for Speech Enhancement by Multi-band MMSE Spectral Subtraction; Natural Language Processing and Knowledge Engineering, 2007. NLP-KE 2007. International Conference on Digital Object Identifier: 10.1109/NLPKE.2007.4368027 Publication Year: 2007 , Page(s): 157 – 161
12. Islam, M.R.; Rahman, M.F.; Khan, M.A.G; Improvement of speech enhancement techniques for robust speaker identification in noise; Computers and Information Technology, 2009. ICCIT '09. 12th International Conference on Digital Object Identifier: 10.1109/ICCIT.2009.5407130 Publication Year: 2009 , Page(s): 255 – 260
13. Kishore Prahallad*+, Sudhakar Varanasi, Ranganatham Veluru, Bharat Krishna M, Debashish S Roy ; Significance of Formants from Difference Spectrum for Speaker Identification ; INTERSPEECH-2006, paper 1583-Tue1CaP.1.
14. Sandipan Chakroborty* and Goutam Saha ; Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter ;International Journal of Signal Processing 5;1 2009
15. A.Revathi1, R.Ganapathy and Y.Venkataramani ; Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach ;International Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 2, November 2009
16. Wei Huang1, Jianshu Chao2, Yaxin Zhang1; Combination of Pitch and MFCC GMM Supervectors for Speaker ;IEEE 2008 ICALIP2008
17. Tudor Barbu ; A Supervised Text-Independent Speaker Recognition Approach ;World Academy of Science, Engineering and Technology 33 2007
18. Matthias W¨olfel, Qian Yang, Qin Jin, Tanja Schultz ; Speaker identification using Warped MVDR Cepstral Features ;ISCA 2009 Brighton U.K
19. Wu Guo, Renhua Wang and Lirong Dai ; Feature Extraction and Test Algorithm for Speaker Verification ; International Symposium on Chinese Spoken Language Processing 2006 Singapore
20. Mangesh S. Deshpande and Raghunath S. Holambe; Speaker Identification Using Admissible Wavelet Packet Based Decomposition ;International Journal of Signal Processing 6:1 2010
21. Hagai Aronowitz, David Burshtein ; Efficient Speaker Identification and Retrieval ; in Proc. Interspeech, 2005, pp. 2433–2436.
22. M. Zamalloayz, L. J. Rodriguez-Fuentesy, M. Penagarikanoy, G. Bordely, J. P. Uribez ; Feature Dimensionality Reduction Through Genetic Algorithms For Faster Speaker Recognition ;16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP
23. Monther Rateb Enayah and Azman Samsudin ; Securing Telecommunication based on Speaker Voice as the Public Key; IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.3, March 2007
24. Hagai Aronowitz and David Burshtein, Senior Member, IEEE ; Efficient Speaker Recognition Using Approximated Cross Entropy (ACE; IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 7, SEPTEMBER 2007
25. Vijendra Raj Apsingekar and Phillip L. De Leon, Senior Member, IEEE ;Speaker Model Clustering for Efficient Speaker Identification in Large Population Applications ; IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 17, NO. 4, MAY 2009
26. William M. Campbell, Member, IEEE, Joseph P. Campbell, Fellow, IEEE, Terry P. Gleason, Douglas A. Reynolds, Senior Member, IEEE, and Wade Shen; Speaker Verification using Support Vector Machines and High-Level Feature ;IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 7, SEPTEMBER 2007
27. Brendan Baker, Robbie Vogt and Sridha Sridharan ;Gaussian Mixture Modelling of Broad Phonetic and Syllabic Events for Text-Independent Speaker Verification ; INTERSPEECH 2005 Lisbon
28. Wei Wu, Thomas Fang Zheng, and Mingxing Xu; COHORT-BASED SPEAKER MODEL SYNTHESIS FOR CHANNEL ROBUST SPEAKER RECOGNITION ; Center for Speech Technology, State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China Proc ICASSP (2006)
29. J. Pelecanos and S.Sridharan, “Feature warping for robust speaker verifi cation,” Proc. Speaker Odyssey , 2001.
30. Bing Xiang, Upendra V. Chaudhari, Jiri Navratil, et al., “Short-time gaussianization for robust speaker verifi cation,” ICASSP , 2002.
31. Hynek Hermansky and Nelson Morgan, “Rasta processing of speech,” IEEE Transactions on Speech and Audio Processing , vol. 2, no. 4, pp. 578–589, Oct 1994.
32. Remco Teunen, Ben Shahshahani, and Larry Heck, “A model-based transformational approach to robust speaker recogntion,” ICSLP , 2000.
33. Douglas A. Reynolds, “Channel robust speaker verifi cation via feature mapping,” ICASSP , 2003.
34. Douglas A. Reynolds, “Comparison of background normalizations for text-independent speaker verifi cation,” EuroSpeech , 1997.
35. Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, “Score normalization for textindependent speaker verifi cation system,” Digital Signal Processing , vol. 10, pp. 42–54, Jan 2000.
36. D.E.Sturim and D.A.Reynolds, “Speaker adaptive cohort selection for tnorm in text-independent speaker verifi cation,” ICASSP , 2005
37. Sturim, D.E., Reynolds, D.A., Dunn, R.B., Quatieri, T.F.: Speaker Verification Using Text-Constrained Gaussian Mixture Models. In: Proc. ICASSP. vol. 1., Orlando, pp. 677–680 (2002)
38. Gauvain, J.L., Lamel, L.F., Prouts, B.: Experiments with Speaker Verification Over the Telephone. In: Pardo, J.M., Enr´ıquez, E., Ortega, J., Ferreiros, J., Mac´ıas, J., Valverde, F.J. (eds.) Proc. EUROSPEECH, Madrid (1995)
39. Newman, M., Gillick, L., Ito, Y., McAllaster, D., Peskin, B.: Speaker Verification ThroughLarge Vocabulary Continuous Speech Recognition. In: Bunnell, H.T., Idsardi, W. (eds.) Proc. ICSLP. vol. 4, Philadelphia, pp. 2419–2422 (1996)
40. Boakye, K., Peskin, B.: Text-Constrained Speaker Recognition on a Text-Independent Task. In: Proceedings Odyssey-04 Speaker and Language Recognition Workshop, Toledo, Spain (2004)
41. Aronowitz, H., Burshtein, D., Amir, A.: Text Independent Speaker Recognition Using Speaker Dependent Word Spotting. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP ’04), Jeju Island, South Korea, pp. 1789–1792 (2004)
42. Andrews,W.D., Kohler, M.A., Campbell, J.P., Godfrey, J.J., Hernandez-Cordero, J.: Gender- Dependent Phonetic Refraction for Speaker Recognition. In: Proc. ICASSP. Orlando, vol. 1,pp. 149–152 (2002) 17. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Jones, D.A.,

Leek, T.R.: Phonetic Speaker Recognition with Support Vector Machines. Advances in Neural Information Processing Systems 16, 1377–1384 (2004)
43. J. Ortega-Garcia, J. Gonzalez-Rodriguez, and S. Cruz-Llanas, "Speech variability in automatic speaker recognition systems for commercial and forensic purposes," IEEE Trans. On Aerospace and Electronics Systems, vol. 15, no. 11, pp. 27–32, 2000.
44. Hatch, A.O., Peskin, B., Stolcke, A.: Improved Phonetic Speaker Recognition Using Lattice Decoding. In: Proc. ICASSP. Philadelphia, vol. 1, pp. 169–172 (2005)
45. Titze, I.: Principles of Voice Production. Prentice Hall, Englewood Cliffs (1994)
46. Atal, B.: Automatic Speaker Recognition Based on Pitch Contours. Journal of the Acoustical Society of America 52(6), 1687–1697 (1972)
47. Doddington, G.: Speaker Recognition Based on Idiolectal Differences Between Speakers. In: Dalsgaard, P., Lindberg, B., Benner, H., Tan, Z. (eds.) Proc. EUROSPEECH, Aalborg, Denmark, pp. 2521–2524 (2001)
48. Heck, L.: Integrating High-Level Information for Robust Speaker Recognition (2002), http://www.clsp.jhu.edu/ws2002/groups/supersid/
49. W. M. Campbell, D. E. Sturim, D. A. Reynolds, MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420 Support Vector Machines using GMM Supervectors for Speaker Verication IEEE SIGNAL PROCESSING LETTERS, VOL. 13, NO. 5, MAY 2006
50. D. E. Sturim, W. M. Campbell, Z. N. Karam, D. A. Reynolds, F. S. Richardson.MIT Lincoln Laboratory, THE MIT LINCOLN LABORATORY 2008 SPEAKER RECOGNITION SYSTEM (2009)
51. Wei Wu, Thomas Fang Zheng, and Mingxing Xu . COHORT-BASED SPEAKER MODEL SYNTHESIS FOR CHANNEL ROBUST SPEAKER RECOGNITION, Center for Speech Technology, State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University.
52. Jiqing Han, Rongchun Gao. Text-independent Speaker Identification Based on MAP Channel Compensation and Pitch-dependent Features (2010).
53. Brendan Baker, Robbie Vogt and Sridha Sridharan. Gaussian Mixture Modelling of Broad Phonetic and Syllabic Events for Text-Independent Speaker Verification, Speech and Audio Research Laboratory, Queensland University of Technology. Interspeech 2005
54. Andre G. Adami, Radu Mihaescu, Douglas A. Reynolds, John J. Godfrey. MODELING PROSODIC DYNAMICS FOR SPEAKER RECOGNITION, GI School of Science and Engineering, Oregon Health and Science
55. A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In Eurospeech'97, pages 1895– 898, 1997
56. P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10):1090–1104, 2000
57. Leena Mary ,B. Yegnanarayana, Extraction and representation of prosodic features for language and speaker recognition ELSEVIER Speech Communication 50 (2008) 782–796
58. Katrina Neville, Jusak Jusak, Student Member, IEEE, Zahir M. Hussain, Senior Member, IEEE and Margaret Lech Performance of a Text-Independent Remote Speaker Recognition Algorithm over Communication Channels with Blind Equalisation 2005
59. Patricia Melin, Jerica Urias, Daniel Solano, Miguel Soto, Miguel Lopez, and Oscar Castillo Voice Recognition with Neural Networks, Type-2 Fuzzy Logic and Genetic Algorithms 2006
60. Najim Dehak, Pierre Dumouchel, and Patrick Kenny, Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification 2007
61. G. Manjunath and G. V Anand SPEECH ENCRYPTION USING CIRCULANT TRANSFORMATIONS 2002
62. Kyuhong Kim, Hoirin Kim, and Minsoo Hahn, "Impostor Detection in Speaker Recognition Using Confusion-Based Confidence Measures", Electronics And Telecommunications Research ETRI Journal, Volume 28, Number 6, December 2006.
63. Todor Ganchev, Nikos Fakotakis, George Kokkinakis, "Impostor Modelling Techniques For Speaker Verification Based On Probabilistic Neural Networks", From Proceeding Signal Processing, Pattern Recognition, and Applications – 2003 (SPPRA 2003), Rhodes, Greece.
64. H S Jayanna And S R Mahadeva Prasanna, "An experimental comparison of modelling techniques for speaker recognition under limited data condition", Vol. 34, Part 3, October 2009, pp. 717–728. © Indian Academy of Sciences.
65. Luis Puente, María-Jesús Poza, Belén Ruiz and Ángel García-Crespo, "Score Normalization for Multimodal Recognition Systems", Journal of Information Assurance and Security 5 (2010).
66. Wu Guo, Li-Rong Dai, Ren-Hua Wang, "Double Gauss Based Unsupervised Score Normalization In Speaker Verification", International Symposium on Chinese Spoken Language Processing (ISCSLP 2008), Kunming, China, pp. 165-168.
67. Daniel Ramos Castro, Julian Fierrez-Aguilar, Joaquin Gonzalez-Rodriguez, Javier Ortega-Garcia, "Speaker Verification using Speaker and Test Dependent Fast Score Normalization", Pattern Recognition Letters, vol. 28, 2007, pp. 90-98.
68. Zbyněk Zajíc, Jan Vaněk, Lukáš Machlica, Aleš Padrta, "A Cohort Methods for Score Normalization in Speaker Verification System, Acceleration of On-line Cohort Methods", SPECOM'2007
69. D. E. Sturim and D. A. Reynolds, "Speaker Adaptive Cohort Selection For Tnorm In Text-Independent Speaker Verification", Proceedings of ICASSP, 2005.
70. Johnny Mariéthoz and Samy Bengio, "A Unified Framework for Score Normalization Techniques Applied to Text-Independent Speaker Verification", IEEE Signal Processing Letters, Vol. 12, No. 7, July 2005.
71. Claude Barras and Jean-Luc Gauvain, "Feature And Score Normalization For Speaker Verification Of Cellular Data", Proceedings of ICASSP 2003, pp. 49-52
72. S. K. Singh," FEATURES AND TECHNIQUES FOR SPEAKER RECOGNITION" M. Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay submitted Nov 03
73. Cheedella S Gupta ."Significance of Source Feature for Speaker Recognition", A M.S Thesis IIIT Madras 2003
74. Tomi H. Kinnunen," Optimizing Spectral Feature Based Text-Independent Speaker Recognition" A Phd Thesis UNIVERSITY OF JOENSUU 2005
75. http://www.biometrics.gov/Documents/SpeakerRec.pdf