**Underlying Text Independent Speaker Recognition**

Nilu Singh1, R. A. Khan2

1,2SIST-DIT, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, UP, India

1, 2[nilu.chouhan@hotmail.com](mailto:nilu.chouhan@hotmail.com)

**Abstract**

This paper discusses the evaluation of automatic speaker recognition from numerous perspectives. A general discussion of the Speaker recognition task, challenges and issues involved in its evaluation is discussed. Speaker recognition have many factors that found to have an impact on performance of recognition rate such as pitch, frequency, handset type and back ground noise etc. We start with the fundamentals of automatic speaker recognition, features type concerning feature extraction technique and speaker modeling. Automatic Speaker Recognition (ASR) is the procedure to automatically recognizing a person based on their speech waves including speaker specific information. It is a procedure to Identification and verification of a speaker. ASR supports access control for different voice based services such as voice dialing, voice mail, telephone shopping, database access services, banking services by telephone, information and reservation services, security control for confidential information, remote access to computers and the another important application of speaker recognition technology is in forensics.

**Keywords**

Automatic Speaker Recognition (ASR), Identification, Verification, High- Level Features, Technical Challenges of ASR, Evaluation Factors for Performance of ASR.

1. **Elementary of Speaker Recognition Technology**

Security is the need of human being in many references, to protect their data/information people generally use password or PIN but there are chances to lose security because if someone lost their PIN or somehow password known by someone else so that data is not secured as much as necessary. There are lots of cases of scam of data and stealing of PIN so to overcome from this problem there is need of a new technology which is more secure and not easy to fraud and theft the technology is Biometric Identification Techniques (BIT).In this technology biometric characteristic are uses as security factor since it is unique for individual therefore these can be used to authenticate the user for access control [1]. There are many Biometric Identification Techniques such as Face recognition, finger print, Iris, voice recognition and speaker recognition etc. In recent speaker recognition is more suitable and secure technology for recognizing a people [2][3]. The vocabulary used for testing speaker recognition may be some words or it may be digit, but digit is use very often due to its use in much security purpose. Such as to access a system there is a need to speak a word or users have to articulate a Personal Identification Number (PIN) when she/he want access to the laboratory/entrance door etc. Such type of applications required automatic speaker recognition system for checking the voice characteristics of the given utterance. Speaker recognition is still a vibrant research area since it is a complex task to recognize a people based on his/her voice. Also since speech signals are highly variants so it is challenging task. Except than these some other factors also challenge the speaker recognition technology for example noise and recording environments [4] [5].



Figure 1: Steps in the development of Speaker Recognition System

Speaker recognition has many practical applications in terms of security systems, human voice use for security because it is not easy to loss or forget and also no body steels it. There are many real time examples such as voice is used to identify criminals in familiar places. Since speech easily transmitted by wired or wireless medium so it can used for access control such as Banking by telephone, credit card purchases etc. users voice help to improve his/her identity and security by storing speakers voice model in a small chip and used as a replacement for a PIN code. We can listen many times that someone ATM card is stolen and make a transaction (it is very easy to find PIN) but if transaction is done by speaking PIN code so it is not possible to produce particular persons voice hence transaction is not made and ATM card is useless for theft, so here we see that data and information are more secure [own thought] [6][7].

In Figure 1 speaker recognition system procedure are described that how to build ASR. Voice biometric can be classified as text-dependent and text-independent, the speaker recognition technology shows that text-dependent systems are easy to implement than text-independent systems. Use of text-independent verification systems are huge such as telephone banking, it is not an easy task because huge amount of voice database models [8].

1. **PRINCIPAL OF SPEAKER VERIFICATION AND IDENTIFICATION**

Speaker recognition can be categorized as speaker identification and speaker verification the fundamental difference between these is the number of decision alternatives. In case of identification the number of decision alternatives is equivalent to the number of speaker voice database while in case of verification claimed identity either accepted or rejected i.e. it is not depend on speaker voice database. Thus the performance of speaker identification decreases as the size of voice database increases and the performance of speaker verification is constant since it is independent from the size of voice database [9].

Again speaker recognition system can be categories as text-dependent and text-independent methods. In text-dependent method fixed utterance to be spoken while in text- independent no fixed utterance to be speaks. Text-dependent technique based on model/template matching techniques in which comparisons are done between voice samples of training and testing utterances. In general text-dependent system performance of recognition is higher as compare to text-independent systems since this method achieves voice unique features associated with their phoneme [9] [10].

**Example** to differentiate between identification and verification here take an example. To present your identity card to control authority is verification process i.e. in this case control authority compare your face with identity card photo and the only possible result is either true or false that means the identity card either is your or not here not any third option. That’s why speaker recognition is called 1 to 1 (1:1) process or the result either true or false. But in case of identification process control authority try to match a photo or sketch from previously found criminals and try to get the closet match to someone and there are several matching processes done and then get closet match it is identification procedure. So identification can be described as to take a identity card and match its photo with many people’s then find some closet to it i.e. 1 to many (1 : n) process [12] [11].

1. **High- Level Features**

Automatic Speaker Recognition (ASR) use to recognizing people by their voice, as it is well known that no two individual voice same/identical due to the differences in their physical structure of sound production system such as larynx size, vocal tract shape etc. every human being have their speaking style as well as the use of particular accent, intonation, rhythm, type of pronunciation and so on. So to develop a more accurate ASR system it is required to use a number of these features. A speaker can be differentiating by their voice timber and tone of voice/accent as well as the kind of words (lexicon) that speaker use in their conversation. High-level features use speakers characteristic vocabulary supposed idiolect (for e.g. a speaker uses a phrase like ya - ya, you know, oh yeah, no-no etc. can be used for recognition) use to describe speakers. The concept of high-level modeling is to use tokens of sequence instead of utterance [13] [14][15].

High level features used for text-independent speaker verification like prosody (rising/falling of pitch/energy), pronunciation, idiolect, behavior and place of articulation etc.[13] [9]. Speech production system of each and every speaker can be related to physiological and behavioral characteristics of speaker. These characteristic come from vocal tract characteristics (Spectral feature) and voice source characteristics i.e. supra-segmental features of speech signal [9]. For text-independent speaker verification high-level features have been successfully used such as word idiolect, prosody, pronunciation, phone usage etc.

1. **TECHNICAL CHALLENGES OF ASR**

It is well known that every people have distinct voice so characteristic of voice also distinct for every people. When people interact to each other it is happens many times that one person’s voice is similar to some one other but in real it is not possible. The people’s voice changes time to time i.e. voice is unpredictable because it changes over time. For example when someone have cold or some disease or many other factors which are easily affected the voice. Hence similarity between individual voices is not a challenge but the technical challenge in speaker recognition is changeability of voice for a person. Some other factors are also affects the voice such as mismatch of handsets, background noise, voice transmission etc. There are discussing some functions which is responsible for variation in voice and affect the recognition performance [10] [15]-

* **Speakers:** It is too difficult to recognize a speaker because every time speaker’s voice is unique so speakers variability is also make problem difficult.
* **Time period:** session variability is also a problem for recognition. Through experiments it is found that speaker recognition performance better when training and testing voice sample taken in the same session. But for some cases it is not possible to take voice sample in same session. For example a speaker voice before 1-2 year or more and speaker’s voice present time.
* **Speakers Circumstance:** lung infections or breathing problem are the definitive health problem for speaker recognition since voice affected frequently. Another case also affected the voice such as speaker sad, emotional or happy etc.
* **Speaker’s surroundings:** it may be playing an important role that is speaker’s native place, speakers education level these function also affect the voice. In case of supportive system where speaker control can have an advantageous effect on recognition performance.
* **Speaking style and speaking rate:** these factors are controlled by speaker but speakers are not usually aware about this. These factors are depend on the place and time for example if someone give a speech or talking to their family member or speak loudly such changes create complex causes in the speech signal/voice.
* **Acoustical noise:** acoustical noise or some background at the time of taking voice sample this is also a problem. Noise and distortion affect the recognition performance.
* **Different microphone:** mismatch of microphone is a big problem for speaker recognition. Also headphone distance from mouth or voice recording through different distances from the mouth using different microphone/headphone etc.

1. **EVALUATION FACTORS FOR PERFORMANCE OF ASR**

To evaluate a system it is necessary to measure the performance of that system. Evaluation factors are those that impact on the system performance. Performance of any system depends on many factors, for speaker recognition here discussed some important functions which affected more the system performance [10]:

* **Number of Training data**

Speaker’s voice training database is a factor which effect on system performance such as from large training voice data a system acquires more voice characteristics of a speaker. Since system has more voice characteristics hence its performance better. Since voice characteristic changes time to time hence for system performance training data acquired in different sessions.

* D**uration of voice sample**

As it is known that a speech signal contains more characteristics, but for recognition for training and testing voice sample contains a segment from the whole speech, this is the duration of speech signal. If speech segment is longer then it acquires more features and hence system performance is better. So duration of speech signal is also a factor which affected the system performance.

* **Background Noise (Distortion)**

Distortion/noise in speech signal also affects the system performance. Noise affects the speaker recognition system performance. When a speaker voice recorded by some type of microphone then maybe there is channel distortion voice, CPU fan voice, or any background noise such as room fan, door knock voice etc. these all noise/distortion affect the system performance.

* **Different handsets/ microphones for voice recording**

Using different microphones are one of the most severe problems for speaker recognition. When voice acquire by telephone then this problem affected more due to differences in telephone systems. To achieve better performance it is necessary to obtained training and testing data by same/single type of microphone.

1. **VOICE DATA COLLECTION**

To build a voice database there are lots of medium available to record the voice/speech. For speaker recognition two types of data required training data and testing data. Testing data should be captured by in different sessions. Except than Computer/Laptop there is also some external audio/video device need to store voice database. To find paramount value of speech signal this is need to performing several testing in speech collected in same session, since in same session data speech characteristics have strong correlation [10].

1. **SPEECH FEATURES**

Speech signal acquire many characteristics of speaker which are sufficient for recognizing a speaker. To extract features from a speech signal the speech would be [9][10][13]-

* Speech produced naturally as speaker speak habitually
* Speech signal robust hence does not affected by noise and distortion
* Easy to measure features
* Speech not affected by variation in time
* Difficult to copycat
* Independent from speaker variability

In the Table1 describe the properties of speech feature, how to extract features form speech signal etc. To extract features from speech signal there are many techniques available such as short term spectral features, high level features, voice source feature, prosodic features, spectro-temporal features etc. To extract short term features Mel Frequency Cepstrum Coefficients techniques used and in this procedure frame size about 20-30 ms. this technique work on resonance properties of vocal tract. To use prosodic and spectro-temporal features with intonation and rhythm, duration is tens/hundreds of milliseconds. And to extract high-level features need to detention of speakers conversation characteristics [13] [21].

|  |  |  |  |
| --- | --- | --- | --- |
| Feature type | Extracted by | Properties | Dependency factor |
| Short-term spectral feature and voice source features | Spectrum, glottal features | * Required small amount of data * Easy to extract * Affected by background noise and channel mismatch * Text- independent | * Size of the vocal fold * Length of the vocal tract |
| High level features | phonemes, idiolect, accent, articulation, semantics | * Required lots of training data * Difficult to extract * Robust against channel and noise effect | * Native place * Education and language * Influenced by local language |
| Prosodic and Spectro- temporal | Pitch, duration, energy, temporal features | * more robust,   but less discriminative   * easier to impersonate | * Emotional expression * Vocal fold vibration |

Table1: Instant view of speech features perspective of their physical understanding

To extract short term spectral features are simple and most discriminative while prosodic and high-level features have received much attention at high computational charge. Selection of speech features is depending on their discrimination, robustness, and practicality. Selection of speech features also depend on proposed application, available computing resource, and availability of speech data and behavior of speaker i.e. speaker are cooperative or not [16]. Speech production is the method of to travel sound wave in air with speedy vibration, and periodic air rhythms transient through vibrating the vocal chords to produce voiced sound[7] [17].

1. **APPLICATIONS OF SPEAKER RECOGNITION**

Speaker recognition is come in the area of voice biometric. And voice biometric can be defined as, it is a human being generated voice signal/speech which is use for authenticating a human identity. Automatic Speaker Recognition can be used in many areas such as [6] [10]-

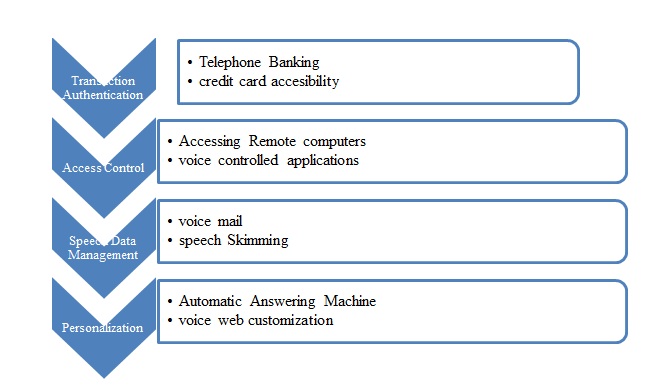
****

Figure 2: Applications of ASR

Speaker recognition applications can be defined as that helps the speaker as security systems that make available control of physical entrance and information access. Other examples of speaker recognition systems in forensic applications but in such cases these systems is fewer encumbered by design [7]. And performance also inferior due to speaker may be using a different microphone or may be speaker emotional or angry or sad due to lots of variety of reasons voice recording quality is low. Difference between forensic applications and security applications is, in forensic required more speech data forensic applicationsmay have minutes of speech data to use while in security where time is of the principle and speech fragments are required in a few seconds [18] [19].

**CONCLUSION**

This paper introduces text-dependent speaker recognition systems that have been competent for a specific user. In this paper we have presented Automatic Speaker Recognition technology their technical challenges, speech factors etc. now days ASR technology has newly been executed in large figure of commercial areas successfully. ASR technology is voiced based biometrics, and this technology used in many where such as voice controlled applications, security control for confidential information, remote access to computers, voice based banking, law enforcement etc. All Automatic Speaker Recognition systems have two main components one is feature extraction and second is feature matching. Also Automatic Speaker Recognition systems have two important tasks one is training phase and other one is testing phase.

**ACKNOWLEDGEMENT**

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

**REFERENCES**

1. Nilu, singh. "A Study on Speech and Speaker Recognition Technology and its Challenges." *proceedings of National Conference on Information Security Challenges, DIT, BBAU, Lucknow, INDIA*. Lucknow: Bharat Book Center, ISBN-978-81-7678-220-3, 2013. 34-37. Print.
2. ABDUL SYAFIQ B ABDULL SUKOR, “SPEAKER IDENTIFICATION SYSTEM USING MFCC PROCEDURE AND NOISE REDUCTION METHOD”, University Tun Hussein Onn Malaysia, JANUARY 20 12, pp. 1-28.
3. Sridhar ChandramohanIyer, “Speaker Recognition System using Coefficients and Correlation Approaches in MATLAB”, International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 5, May – 2014, pp.112-115.
4. Minh N. Do, Digital Signal Processing Mini-Project “An Automatic Speaker Recognition System”, Audio Visual Communications Laboratory Swiss Federal Institute of Technology, Lausanne, Switzerland, pp. 1-14.
5. SeyedOmidSadjadi, Malcolm Slaney, and Larry Heck, “MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker-Recognition Research ”,in IEEE Signal processing society newsletter, Microsoft Research SLTC Newsletter, November 2013.
6. Douglas A. Reynolds, “Overview of Automatic Speaker Recognition”, in JHU WS2008 Summer School@2008MIT, M.I.T. Lincoln Laboratory.
7. J. P. Campbell, “Speaker recognition: A tutorial'' Proceedings of the IEEE, vol. 85, pp. 1437--1462, September 1997.
8. Li CHEN, Ying-chun YANG, Zhao-hui WU,” Mismatched feature detection with finer granularity for emotional speaker recognition”, Journal of Zhejiang University-SCIENCE C (Computers & Electronics), ISSN 1869-1951, May 20, 2014, pp.1-25
9. GintarasBarisevicius, “TEXT-INDEPENDENT SPEAKER VERIFICATION”, Kaunas University of Technology, Kaunas, Lithuania, pp. 1-9.
10. [SadaokiFurui](http://www.scholarpedia.org/article/User:Sadaoki_Furui), “Speaker recognition”, ScholarpediaSadaokiFurui (2008), Scholarpedia, 3(4):3715, [doi:10.4249/scholarpedia.3715](http://dx.doi.org/10.4249/scholarpedia.3715).
11. George R. Doddington, Mark A. Przybocki, Alvin F. Martin, Douglas A. Reynolds, “The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective” , In Elsevier Speech Communication 31 (2000) pp. 225 -254.
12. Wikipedia, the free encyclopedia, “Speaker recognition”
13. Nilusingh, R. A. Khan, et al. "Applications of Speaker Recognition." Elsevier- ScienceDirect. (2012): 3122-3126.
14. TomiKinnunen, Haizhou Li, “An Overview of Text-Independent Speaker Recognition: from Features to Supervectors” in Speech Communication http://www.pcworld.com/article/106142/
15. visa\_gets\_behind\_voice\_recognition.html, july 1, 2009.
16. Elizabeth Shriberg, “Higher-Level Features in Speaker Recognition”, Speaker Classification I, LNAI 4343, pp. 241–259, 2007. Springer-Verlag Berlin Heidelberg 2007
17. [HiraShaukat](http://www.slideshare.net/HiraShaukat?utm_campaign=profiletracking&utm_medium=sssite&utm_source=ssslideview), “Speaker recognition using MFCC”, Dec 19, 2013
18. ZaidiRazak, Noor Jamaliah Ibrahim, EmranMohd Tamil, Mohd Yamani IdnaIdris, Mohd. Yakub @ Zulkifli Bin MohdYusoff, “Quranic Verse Recitation Feature Extraction Using
19. Mel-Frequency Cepstral Coefficient (MFCC)”, 4th International Colloquium on Signal Processing and its Applications, March 7-9, 2008, Kuala Lumpur, Malaysia. © Faculty of Electrical Engineering, UiTM Shah Alam, Malaysia. ISBN: 978-983-42747-9-5
20. S. Furui. Recent advances in speaker recognition. AVBPA97, pp 237--251, 1997
21. B. S. Atal, “Automatic Recognition of Speakers from their Voices”, Proceedings of the IEEE, vol 64, 1976, pp 460 – 475.