

Analysis of Automatic Speaker Verification Systems and Challenges

Swathika R

Ph.D. Research Scholar,
Department of Computer Science,
Bharathiar University,
Coimbatore, Tamil Nadu.
Swathi19cs@gmail.com

Dr Geetha K

Assistant Professor,
Department of Computer Science,
Bharathiar University,
Coimbatore, Tamil Nadu.
geethakab@gmail.com

Abstract: The ability to verify uttered words is known as speech recognition, and the ability to verify who is saying them is known as speaker recognition. With the growing relevance of computerised information processing and telecommunications, the utility of validating a person based on his speech characteristics is expanding.. In this paper, speaker recognition, architecture which supports speaker verification, methods for feature extraction and modelling, applications, underlying techniques and some performance metrics used to evaluate speaker verification were analysed. It is also reviewed the some of the fortes and flaws of current speaker recognition technologies and it outlines some potential future trends in research, development and applications. The paper accomplishes with discussions on research opportunities in this area.

Keywords: *Speaker identification, Speaker Verification, Speech Corpus.*

I.INTRODUCTION

In applications like Voice Bout [1, 2], speaker verification (SV) is a method of determining whether an utterance belongs to a specific speaker based on that speaker's known utterances. Speaker verification methods are categorised into two variants based on the constraints of the utterances used for enrolment and verification: text-dependent speaker verification (TD-SV) and text-independent speaker verification (TISV) (TI-SV). In TD-SV, the transcript of both enrolment and verification utterances is phonetically constrained, but in TI-SV, the transcript of enrolment and verification utterances has no lexical limitations, revealing a greater diversity of phonemes and utterance durations [3, 4]. In this paper, we concentrate on TI-SV and a subtask of TD-SV known as global password TD-SV, where the verification is done chevalier [5, 6] "OK Google". In addition to our unique human ability to hear and interpret spoken language, the ear allows us to perform a wide range of functions. Only a few examples include locating goods, appreciating music, and verifying people by their sounds.

Along with efforts to build computer procedures for understanding spoken messages, there is currently a lot of interest in establishing methods for verifying people based on their voices. It would be a pleasant and natural form of communication to be able to speak to your computer and have it recognise and comprehend what you say. It would cut down on the quantity of work you need to do.

The listener receives information from the speech signal at several levels. Speech, at its most basic level, delivers a message through words. However, speech also provides information about the language being uttered, as well as the speaker's emotion, gender, and general veracity. We have been dealing with diverse research efforts in the domain of man-machine interface since the evolution of computers. The input peripherals are although very popular mediums to interact with the computer but has some limitations as keyboard requires a certain amount of skill for effective and fast usage and mouse on the other hand requires a good hand and eye coordination. The physically challenged people find computers difficult to use. Speech which is a natural and very easy way of exchanging the information if used as a medium to interact with the computer and can solve all these problems. Signal processing has made it possible for computers to follow human voice commands and understand human languages as speech can be characterized in terms of signal carrying message information.

Speech signal processing could be divided into three different tasks:

- Analysis
- Recognition
- Coding.

Recognition research fields could be subdivided into three parts:

- Speech
- Speaker

- Language recognition systems.

While discourse acknowledgment targets perceiving the word expressed in discourse, language acknowledgment focuses on the discovery of language spoken and the objective of speaker acknowledgment frameworks is to separate, describe and perceive the data in the discourse signal conveying speaker personality. No two people are indistinguishable in light of the fact that their vocal parcel shapes, larynx sizes, and different pieces of their voice creation organs are unique. Notwithstanding these actual contrasts, every speaker has their trademark way of talking, including the utilization of a specific highlight, beat, inflection style, articulation design, decision of jargon, etc [7].

A. Automatic Speaker Recognition

Automatic Speaker recognition and speech recognition belong to the class of voice signal processing, but speaker recognition emphases on the identity information of the speaker, while speech recognition focuses on the text information equivalent to the voice. Voiceprint can be understood as the pattern of the voice frequency spectrum obtained by the time-frequency analysis technology of the waveform signal of human voice. Due to the inherent differences in the physiological structure of each person, it also causes the diversity of human speech styles, which provides us with a principle basis for automatically identifying the speaker's identity information through machines.

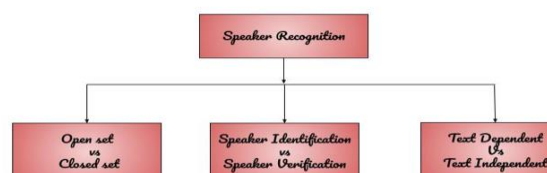


Fig.1 Automatic speaker recognition system

Fig.1 shows the components of an automatic speaker recognition system. In speaker acknowledgment the underlying system includes extraction. In highlight extraction module the crude signs are changed into include vectors in Enrolment Front end Pattern Matching Background Modeling Score Normalization. Include extraction input discourse choice. Which speaker explicit properties are accentuated and measurable redundancies is hidden. In the enrolment mode, a speaker model is prepared utilizing highlight vectors of the objective speaker. In acknowledgment mode, the element vectors

separated from the obscure individual's expression of the individual are with the framework data set and a closeness score is produced. Ultimate conclusion is gone with by the choice module in light of the closeness score. Basically all cutting edge speaker acknowledgment frameworks utilize a bunch of foundation speakers or partner speakers. This is done to upgrade the power and computational effectiveness of the recognizer. In the enrolment stage, foundation speakers are utilized as the negative models in the preparation of a discriminative model, or in preparing stage an all inclusive foundation model from which the objective speaker models are adjusted. In the acknowledgment stage, foundation speakers are utilized in the standardization of the speaker match score [8].

B. Speaker Verification System

Speaker check is utilized to decide if an individual professes to be as per his/her voice facts. This errand is otherwise called voice check and speaker location. Speaker check is a 1:1 match where one speaker's voice is matched to one format or in other sense Pattern Matching between the guaranteed speaker model enrolled in the information base and the faker model will be performed then, at that point (shows figure.2) assuming the match is over a specific limit, the personality guarantee is confirmed. Utilizing a high limit, framework gets high wellbeing and forestalls frauds to be acknowledged, yet in the mean while it likewise faces the challenge of dismissing the authentic individual, as well as the other way around.

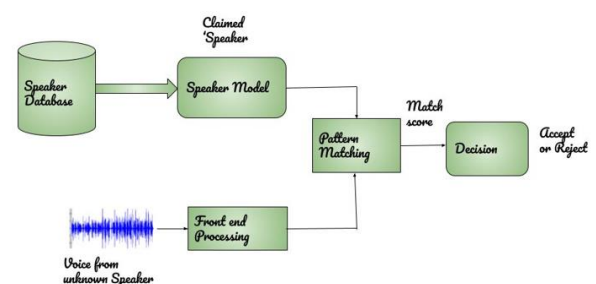


Fig.2 Speaker Verification system

C. Speaker Identification System

Speaker ID is the most common way of tracking down the character of an obscure speaker by looking at his/her voice with voices of enlisted speakers in the information base. It is a one to-numerous correlations (1: N match where the voice is analyzed against N layouts) [9].

In Speaker ID System, N speaker models are scored in equal and the most one of the speaker's ID in the information base, or will be nothing from what was just mentioned if and provided that the matching score is underneath a few edge and it's on account of an open in all probability one is accounted for, and thus choice will be open-set Speaker ID framework. Fig.3 shows the Speaker Identification framework.

The positive highlights for Security and Intelligence Services (SIS) ought to have the accompanying credits:

- Easy to extract, easy to measure, occur frequently and naturally in speech
- Not be affected by speaker physical state (e.g. illness)
- Not change over time and utterance variations (fast talking vs. slow talking rates)
- Not be affected by ambient noise
- Not subject to mimicry

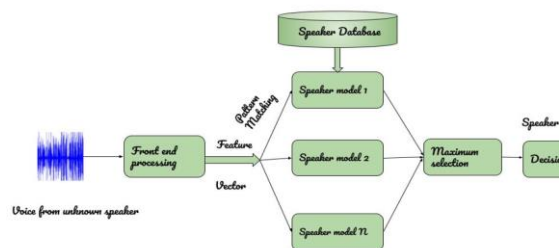


Fig.3 Speaker Identification System.

The speaker ID issue might additionally be partitioned into shut set and open set. The shut set alludes to a situation where the speaker is known and has a place with a bunch of M speakers. In the open set case, the speaker might be out of the set and subsequently, a nothing from what was just mentioned classification is vital. Another distinctive part of speaker ID frameworks is that they can either be text-free or message subordinate contingent upon the application. In the message autonomous case, there is no limitation on the sentence or expression to be spoken, while in the message subordinate case, the information sentence or expression is fixed for every speaker. A text-subordinate is normally utilized in speaker confirmation frameworks in which an individual's secret word is basic for checking his/her character Table.1 shows the standard discourse corpus [10].

Table. I standard speech corpus

Name of the speech corpus	No of Speakers
TIMIT speech corpus	630 speakers (438 male and 192 female)
SIVA speech corpus(Italian speech corpus)	840 speakers
POLYVAR Speech Corpus	143 speakers (85 male and 58 female)
POLYCOST Speech Corpus	133 speakers (74 male and 59 female)
KING Speech Corpus	51 male speakers
YOHO Speech Corpus	138 speakers (106 male and 32 female)
SWITCHBOARD Speech Corpus	Switchboard I consists of 543 speakers and Switchboard II consists of 657 speakers
NIST 2001 SRE Speech Corpus	174 speakers (74 male and 100 female)
NIST 2002 SRE Speech Corpus	330 speakers (139 male and 191 female)
NIST 2004 SRE Speech Corpus	616 speakers (248 male and 368 female)
TSID Speech Corpus	35 speakers (four female, 31 male)
Speaker Recognition Corpus (OGI)	100 speakers consisting of 47 male and 53 female speakers

ANDOSL	129 speakers 67 female and 62 male
Digit-SPL	68 males and 19 females
t-DCF	107 Speakers (46 males,61 females)

Table. II standard speech corpus

D. Countermeasure for Speaker Verification

In speaker verification process performance is evaluated based on False Acceptance Rate (FAR) and False Rejection Rate (FRR) with the following formulation

$$\text{False Rejection Rate} = \frac{\text{no of rejective true speaker}}{\text{total number of true speaker}}$$

$$\text{FalseAcceptance Rate} = \frac{\text{no of accepted imposter}}{\text{total number of imposter}}$$

EER = False Reject Rate = False Acceptance Rate

False Acceptance:

Also called as missing probability ratio in which the number of verified identities for which the test speaker varies from the target speaker normalized against the total number of acceptances.

False Rejection rate:

Also called as false alarm probability in which the number of identities which were not verified for which the test speaker was the same as the target speaker normalized against the total number of rejections.

E. Challenges & Applications of Speaker Verification

The table.II shows the challenges and applications of speaker verification.

S. No	CHALLENGES	APPLICAT IONS
1	Challenging audio	Transaction authenticatio n

2	Treatment of whispered speech	Personalizati on of IVR dialogue
3	Different styles of phonation	Information Retrieval
4	Speech under stress	Access control
5	Multiple sources of speech and far-field audio capture	Remote digital time and attendance logging
6	Channel mismatch	Audio mining of data

Table. II Challenged and Applications

II.STRENGTHS AND WEAKNESSES

Speaker Verification has its different assets and shortcomings and the accompanying standards are utilized to assess the appropriateness of speaker confirmation as biometrics: Voice accounts are not difficult to acquire and don't need costly equipment. The genuine benefit of voice check is that it tends to be done over phone lines or utilizing PC mouthpieces, with variable recording and transmission quality. Design matching calculations should have the option to deal with surrounding commotion and varying nature of the accounts [11]. Table.III shows the qualities and shortcomings of the speaker confirmation.

Strength and Weakness	Description
Portability	Speaker verification is easy to use, has low computation requirements (can be ported to cards and handhelds) and, given appropriate constraints, has high accuracy.

Acceptability	Speaker check is inconspicuous; talking is a characteristic cycle so no uncommon activities are required. It is utilized for observation applications or overall when the subject doesn't know about it than the normal security worries of distinguishing ignorant subjects apply. In addition speaker data can be gotten effectively from anyplace utilizing the recognizable phone organization (or web) with no extraordinary client hardware or preparing.
Circumvention	A major issue with speaker verification is spoofing using voice recordings. The risk of spoofing with voice recordings can be mitigated if the system requests a random generated phrase to be repeated, an impostor cannot anticipate the random phrase that will be required and therefore cannot attempt a playback spoofing attack.
Performance	Robustness is very dependent on the setup, when telephone lines or computer microphones are used the algorithms will have to compensate for noise and issues with room acoustics. Furthermore speaker recognition is, because the voice is a behavioural biometric, impacted by errors of the individual such as misreading and mispronunciations.
Mobility	Mobility of system means that people are using verification systems from more uncontrolled and harsh acoustic environments like cars, crowded airports which can stress accuracy

Variability	The varied microphones and channels that people use can cause difficulties since most speaker verification systems rely on low-level spectrum features susceptible to transducer/channel effects.
Universality	Clearly for individuals who are quiet or disliking their voice because of serious ailment this biometric arrangement isn't useable.
Permanence	Speech signal used for speaker verification is a behavioural signal that may not be consistently reproduced by a speaker so an issue with speaker recognition is that the voice changes with ageing, and is also influenced by factors such as sickness, tiredness, stress, etc.

Table. III Challenges and Applications

III. CONCLUSION AND FUTURE TRENDS

This paper analysed an analysis of speaker verification. Focus is also flashed on the various applications and factors affecting the system. The speaker verification system still has various drawbacks which can be further reduced by carrying out research in sub-domains and merging of other biometrics system with speaker verification. The main application for the technology is in the area of access control, where the speakers are required to be authenticated before they can be allowed access to certain facilities or some other restricted services in various domains which is having some secured information. The future trend in access control is to integrate speaker verification technology into a multi-level and a hybrid authentication approach, where results from different biometric technology like finger print, face, iris and speaker verification could be fused together to achieve better reliability in authentication. However, the biggest advantage of speech based biometrics is the ability perform authentication where a direct physical or visual contact with the subject is not feasible. Thus the technology has a clear advantage for

authenticating transactions that occur over the voice channel like telebanking. A more controversial application of speaker verification technology is in the area of forensics where the results of the technique could be offered as evidence in judicial trials. Compared to fingerprinting and DNA based authentication technology, the existing speaker verification techniques have their drawbacks and limitations due their sensitivity to corruption by noise and the ability to masquerade the signal using voice recording devices. We believe that there is an enormous potential for speaker verification and recognition technology in multimedia and biometric applications. However, key challenges still remain to be solved and are currently limiting the wide-scale deployment of the technology.

replayed speech”, *Computer Speech & Language*, 64, November 2020.

11. http://www.biometricsolutions.com/solutions/index.php?story=speaker_recognition

REFERENCES

1. Yury Pinsky, “Tomato, tomahto. Google Home now supports multiple users,” <https://www.blog.google/products/assistant/tomato-tomahtoogle-home-now-supports-multiple-users>, 2017.
2. Mihai Matei, “Voice match will allow Google Home to recognize your voice,” <https://www.androidheadlines.com/2017/10/voice-matchwill-allow-google-home-to-recognize-your-voice.html>, 2017.
3. Tomi Kinnunen and Haizhou Li, “An overview of textindependent speaker recognition: From features to supervectors,” *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.
4. Frederic Bimbot, Jean-Francois Bonastre, Corinne Frenouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-Garcia, Dijana Petrovska-Delacretaz, and Douglas A Reynolds, “A tutorial on text-independent speaker verification,” *EURASIP journal on applied signal processing*, vol. 2004, pp. 430–451, 2004.
5. Guoguo Chen, Carolina Parada, and Georg Heigold, “Small footprint keyword spotting using deep neural networks,” in *Acoustics, Speech and Signal Processing (ICASSP)*, 2014 IEEE International Conference on. IEEE, 2014, pp. 4087–4091.
6. Rohit Prabhavalkar, Raziell Alvarez, Carolina Parada, Preetum Nakkiran, and Tara N Sainath, “Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks,” in *Acoustics, Speech and Signal Processing (ICASSP)*, 2015 IEEE International Conference on. IEEE, 2015, pp. 4704–4708.
7. <http://www.jmarkowitz.com/glossary.html>.
8. Ing. Milan Sigmund, CSc. “Speaker Recognition, Identifying People by their Voices”, Brno University of Technology, Czech Republic, Habilitation Thesis, 2000
9. Tomi kinnunen, Haizhou Li, “An overview of text independent speaker recognition: From features to supervectors”, *Speech Communication*.
10. X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K.A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J. -F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, Z.-H. Ling, “ASVspoof 2019: a large-scale public database of synthetic, converted and