Speaker Recognition: A Review

*K.Karthika**

Abstract

Human can identify a speaker by hearing their voice over digital device or through phone in addition to in-person. Automatic Speaker Recognition (ASR), a voice biometric authentication technology, has been developed to take advantage of this innate human competency. It is capable of examining the voice signals and traits gleaned from speaker sounds, from which speakers can be identified. As a crucial component of speech biometrics, ASR has lately emerged as a successful study topic. While providing an overview of ASR and the key concepts navigating speaker recognition technology, this literature review highlights the research perspectives in the area of speaker recognition. This paper also illustrates the concepts of ASR with its types, challenges, and its applications.

Keywords—ASR, Speaker Identification, Speaker Verification

I. INTRODUCTION

Voice is the most traditional means of communication since it is the way humans typically express their opinions. In the past few decades and still today, significant advancement has been accomplished via human-computer research aiming at technologies to make human interaction as simple and appropriate as possible[1].

Speaker recognition is a method to validate the identity of user by eliciting certain traits from their voice utterances. It is a mechanism that depends on the distinct features of the speech signal to recognize the speaker. In order to determine the speaker's individuality, the speaker recognition system validate the spoken phrases of the speaker and identify their uniqueness which include applications like security control, voice dialing and voice mail.

The initial ASR system i.e., the physiological model of human voice came into existence in the year 1962 and was developed by Gunnar Fant at Bell Laboratories that is comprised of a speech analysis base. ASR systems have progressed over the past 60 years with the development of human-computer studies. These cutting-edge systems are being utilized in various applications, including forensic applications, voice calls, online banking, telephone shopping, person identification, and person verification. In the past ten years from 2010 to2020, speech recognition and related technologies made significant advancements [2]. The timeline of some of the most significant advancements in speech recognition research, software, and applications over the preceding ten years is shown in the figure below:

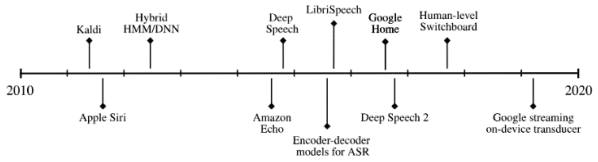

Fig 1 Evolution of Speech Recognition to the year 2030

II. SPEAKER RECOGNITION

Generally, Speaker recognition identifies the person from the voice features. Recognizing the speaker simplifies the task of interpreting the speech in the system or authenticating the identity of the speaker based on the acoustic features. These acoustic features differ for every individual as it depends upon both the learned behavioral and anatomy patterns of every person[3].

Various approaches of speaker recognition are portrayed in Fig:2. A brief explanation of these subdomain are illustrated below:

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
* Corresponding Author

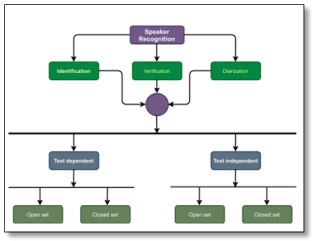

Fig 2: Types of Speaker Recognition

• Speaker Identification (SI)

Based on the utterances of the speaker, SI can detect a fictitious speaker[4]. It is a method of detecting the person by matching the various utterances in the database. By comparing a specific voice against N templates, it executes a 1:N match.

• Speaker Verification (SV)

It utilizes the voice samples to validate the identified speaker. Every speaker's individual patterns are stored in the voice model database and the acquired trait is compared with the one template in a 1:1 match[5].

• Speaker Diarization (SD)

It partitions different voices into homogeneous segments that are associated with every user that includes different applications namely conversational material comprehension and video conferencing [6].

All the above falls either on Text -Dependent or Text Independent which are discussed below:

• Text – Dependent System

It is a method wherein the test utterances are equivalent to the text used during the enrolment step[7]. In this case, the test speaker will be acquainted with the speaker model whose trial and low enrollment stages yields reliable and authenticated results.

• Text – Independent System

In this system, the training speech samples are different from testing and the test speaker in the testing phase will not have any previous knowledge [8] about the voice patterns that are already trained.

Based upon the number of valid speakers, ASR frameworks can be divided either as open set or closed set which are as follows:

• Open – Set

It is a method that is designed and structured with any range of speakers i.e., trained speakers and is said to be open set as the unfamiliar speech samples can be used for validation.

• Closed set

A closed set technique employs only the authorized number of speakers that are registered in the system. It evaluates speaker based upon the samples that are already trained and stored in the database.

III. PROCESS OF SPEAKER RECOGNITION

Pre-processing, feature extraction, and speaker modeling are the three components that make up a standard speaker recognition system. A simple outline of speech recognition systems is shown in Figure 3.

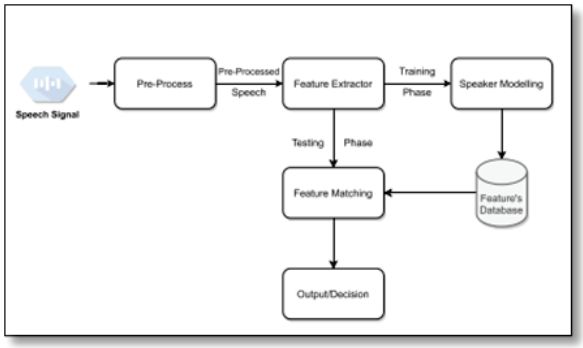

Fig 3: Structure of ASR

• Feature Extraction

It is a process by which the speech signals are analysed and more discriminative features are extracted by converting them into parametric values[9]. Different methods are used for extraction by applying mathematical tools for creating the models for the speaker. A feature extraction technique should possess the following characteristics:

- It should be robust.
- It should be capable of maintaining low intra-speaker variation and high inter-speaker variation
- Simple and easy to measure.
- Robust for channel distortion, noise, and mimicry.

Commonly used methods include:

- Mel Frequency Cepstral Coefficients (MFCC)
- Linear Prediction Cepstral Coefficients (LPCC), and
- Perceptual Linear Prediction Cepstral Coefficients (PLPC

- **Speaker Modelling**

Modelling of speaker is used for retrieving the speaker features. The type of modelling can be selected based upon the storage and computation requirement, model training and its performance [10]. It is classified as:

- Generative Models
- Discriminative Models

- **Generative Models** form the model based upon the features that are estimated from the targeted speaker. Vector Quantization (VQ) ,Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) are its various types.
- **Discriminative Models** formulate the boundary between the speakers by training the data for all targeted and non-targeted speakers. Support Vector Machine (SVM) and Artificial Neural Network (ANN) fall under this category[11].

IV. DATA SETS

The demand of validated databases for speaker/speech databases has grown along with the development of voice-based technologies. These datasets aid the researcher in demonstrating, assessing, and contrasting the effectiveness of the current system[12]. Datasets for speaker recognition can be divided into:

- Cleaned Speech Dataset: These datasets do have zero presence of noises[13].
- Wild Datasets: It carries noise as it is collected beyond the controlled environments. The noises include vibrations, background sounds, squeezing, and inter-speaker variability[14].

Data Set Name	No.of Speakers	Type of Speech	No of Utterance
TIMIT	530	Clean	6300 sentences
VoxCeleb1	1251	Multi-Media	1535136 utterances
VOICES	300	Noisy Room	1440 hours
CMU	74	NA	9487130 utterances
ELSDSR	22	Clean	154 sentences
SwitchBoard	500	Telephony	2500 conversations
POLYCOST	133	Telephony	1285 conversations
NIST-SRE	2000+	Clean	Varies year by year
YOHO	138	NA	1380 conversations
MIT Mobile	88	Mobile Devices	7884 sentences
LIBRISPEECH	44	Segmented English Read Speech	1000 hours

Table 1: Popular datasets used in speaker recognition domain [15]

V. CHALLENGES

Numerous challenges that are present in speaker recognition methods applicable to both text-dependent and text-independent tasks include:

- **Limited Data**

The amount of data for training and testing may not be abundant which may degrade the recognition of speaker.

- **Aging of Speaker Models**

Several models may experience aging source in terms of channel usage, natural ageing and changes in behaviour. Behavioural alterations happens when the user faces vulnerability in their voices. Biological ageing may make the changes in recognition.

- **Deployment cost**

The cost of deployment is high for estimating the conventional positive, false rejection and false acceptance rates. This is due to implementation of huge volume of dialogues which requires threshold for protection.

- **Low language resource**

The language of low resource was applied for retrieving the result which make the range down the average. Therefore, unsatisfactory results on the state-of-art approaches were profound.

- **Speaker – based variability**

The differences in the way of speaker speaks will affects the accuracy of the system in ASR. The variability include determinants namely: emotion, style and psychological etc.,

- **Conversation-based variability**

The variability in vocal communication with the system or different person's accent may fall under this category. It incorporates monologue, two-way conversation and dialect spoken etc.,

- **Intra-speaker variability**

There can be a substantial amount of diversity since the same person doesn't always deliver the precise speeches in the same manner at all times. The constraints posed by this heterogeneity make speaker recognition tasks very difficult

VI. EFFICIENCY OF SPEAKER RECOGNITION

The efficiency of the ASR is evaluated based upon the parameters such as speed and accuracy. The accuracy of the system is tested with False Acceptance Rate(FAR) and False Rejection Rate (FRR).Some commonly used performance metrics include Equal Error Rate (EER) , Detection Error Tradeoff (DET) and Receiver Operating Characteristics (ROC). Evaluation of the system can be performed using the following:

- Receiver Operating Characteristic (ROC)

It is a probability curve that illustrate degree of separating the classes. It is said to be positive, if the threshold value is minimum and when increases it is negative.

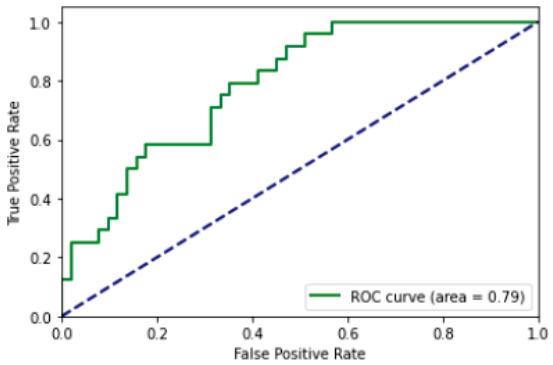

Fig 4: ROC Curve

- **EQUAL ERROR RATE**

This algorithm is used to make accurate predictions of the threshold for its false rejection and false acceptance rate. It tells the amount of percentage of false rejections and false acceptances. The conditional probability is:

- $P(X|x)$ = correct acceptance,
- $P(X|n)$ = False Acceptance (FA),
- $P(N|x)$ = False Rejection (FR),
- $P(N|n)$ = correct rejection.

The following two probabilities P(X|x) and P(X|n) can be used to measure the SR system. The relationships among these parameters can be illustrated as:

$P(X|x) + P(N|x) = 1$ and $P(X|n) + P(N|n) = 1$ ------ [1]

- DETECTION ERROR TRADEOFF

It is the depiction of error margins for binary classification that plots false acceptance and false

rejection rates. To model the binary outcome variable can be represented as:

$x = probit(Pfa)$ -----[2]

$y = probit(Pfr)$ -------[3]

VII. CONCLUSION

Speaker Recognition is an eminent domain which are integrated with other domains for validating the speaker. This paper focuses on exploring the various dimensions of the research field. It also targets on types, feature extraction, modelling approaches and data sets used for evaluation. Moreover, the performance metrics used for validation helps to elaborate the future directives in a broader perception of speaker recognition technologies.

REFERENCES

[1] N. Singh, A. Agrawal, and R. Khan, "The development of speaker recognition technology," Int. J. Adv. Res. Eng. Technol., vol. 9, no. 3, pp. 8–16, 2018.

[2] N. Singh, ‘‘A study on speech and speaker recognition technology and its challenges,’’ in Proc. Nat. Conf. Inf. Secur. Challenges. Lucknow, India: DIT, BBAU, 2014, pp. 34–37.

[3] C D. Shaver and John M. Acken, “The Development of Text-Independent Speaker Recognition Technology”, Journal of Electrical Engineering, pp. 1-8, 2014.

[4] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993

[5] W. Yutai, J. Xiaoqing, L. Feng,"Speaker Recognition Based on Dynamic MFCC Parameters, "School of Information Science and Engineering, University of Jina , 2002.

[6] Z. Saquib, N. Salam, R. P. Nair, N. Pandey, and A. Joshi, ‘‘A survey on automatic speaker recognition systems,’’ Signal Process. Multimedia, vol. 123, pp. 134–145, Dec. 2010.

[7] M. Hébert, ‘‘Text-dependent speaker recognition,’’ in Springer Handbook of Speech Processing. Berlin, Germany: Springer, 2008, pp. 743–762.

[8] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, ‘‘A tutorial on textindependent speaker verification,’’ EURASIP J. Adv. Signal Process., vol. 2004, no. 4, pp. 1–22, 2004.

[9] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, “Spoofing Detection from a Feature Representation Perspective,” In Processing of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2016.

[10] H. Jayanna and S. M. Prasanna, ‘‘Analysis, feature extraction, modeling and testing techniques for speaker recognition,’’ IETE Tech. Rev., vol. 26, no. 3, pp. 181–190, 2009

[11] S. Sujiya and E. Chandra, “A Review on Speaker Recognition,” International Journal of Engineering and Technology (IJET) , vol. 9, no. 3, pp. 1592–1598, 2017, doi: 10.21817/ijet/2017/v9i3/170903513.

[12] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody, ‘‘The Australian national database of spoken language,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), vol. 1, Apr. 1994, p. I-97.

[13] W. M. Fisher, ‘‘Ther DARPA speech recognition research database: Specifications and status,’’ in Proc. DARPA Workshop Speech Recognit., Feb. 1986, pp. 93–99.

[14] A. Nagrani, J. S. Chung, and A. Zisserman, ‘‘VoxCeleb: A large-scale speaker identification dataset,’’ 2017, arXiv:1706.08612. [Online]. Available: http://arxiv.org/abs/1706.08612.

[15] M. M. Kabir et al.,”Survey of SR: Fundamental Theories, Recognition Methods and Opportunities”IEEE Access,2021, 10.1109/ACCESS .2021.3084299.