IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Speech and Speaker Recognition Approaches: A Survey

[1]Mr.Sukumar B S, [2]Dr. G N Kodanda Ramaiah, [3]Dr. Sarika Raga, [4]Dr. Lalitha Y S

[1]Research scholar, [2]Professor and HOD, [3]Associated Professor, [4] Professor

[1]Department of ECE,

[1] Dept. of ECE, RRC, Visvesvaraya Technological University, Belagavi, India

[2]Department of ECE, Kuppam Engineering College, Kuppam India.

[3]Department of ECE, Visvesvaraya Technological University PG Centre, Muddenahalli, Chikkaballapura, India

[4]Department of ECE, Don Bosco Institute of Technology, Bangalore, Visvesvaraya Technological University, India.

Abstract: Speech and speaker recognition technologies are one of the emerging technologies in today's world that have obtained rapid advancements. They're becoming integral components in various applications such as virtual assistants, authentication systems, and customer service platforms. The primary purpose of speech recognition is to transcribe spoken language into text and that of speaker recognition is to identify or verify an individual's identity based on vocal characteristics. This paper presents an exhaustive survey of techniques both in this and other fields which include acoustic-phonetic approach, pattern recognition, and artificial intelligence-based models to develop hybrid systems that integrate all the methods for improved accuracy. Also, this paper explores various techniques of speaker recognition. Here, again, one technique considered is the feature extraction method that is necessary to distinguish speakers from one another. Despite these innovations, however, current systems of speech recognition face problems like variability in accents, noisy environments, and data insufficient for training models to lead to deterioration in model performance. Complexity in complexities of model training further hampers progress toward the development of highly accurate adaptive systems for recognition. In discussing these challenges and surveying the current landscape of speech and speaker recognition technologies, insights are provided not only into their potential but also further developments needed for applications over wider spectra and more effectively in the real world.

Index Terms - Speech recognition, Speaker recognition, Hybrid approaches, Challenges, Feature extraction.

1 **Introduction:** The increasing focus on security has led to a surge in the use of biometrics. In addition to facial recognition, distinctive characteristics such as retina patterns, vocal attributes, and iris variations may also differentiate individuals. Biometrics may be categorised into two types: physiological and behavioral, as seen in Fig. 1 [1]. The former methods includes facial recognition, fingerprints, and iris patterns, whilst the latter comprises voice recognition, keystroke dynamics, and signatures. Table 1 delineates the standard attributes of biometric technologies and their efficacy for accuracy, user-friendliness, user acceptability, implementation simplicity, and cost [2] [3]. According to the data shown in the table, it can be concluded that speech technology is very advantageous, since it is user-friendly, readily implementable, and broadly embraced by users owing to its affordability.

Table 1: Comparison of various biometric attributes [2]

Biometric Type	Accuracy	Ease of use	Characteristics of biometric technologies User acceptance	Ease of implementation	Cost
Voice	Medium	High	High	High	Low
Face	Low	Low	High	Medium	Low
Iris	Medium	Medium	Medium	Medium	High
Fingerprint	High	Medium	Low	High	Medium
Retina	High	Low	Low	Low	Medium
Hand geometry	Medium	High	Medium	Medium	High
Signature	Medium	Medium	High	Low	Medium

Speech, as a medium of language expression, is produced by the coordinated motions of articulatory organs, controlled by neural processes in the brain's speech functional regions [5]. Speech is crucial for everyday communication and serves as a fundamental channel for social interactions [6]. In speech processing, speaker identification and speech recognition are the two applications often used by academics to analyse spoken language [7]. Prior to exploring the intricacies of speaker recognition, it is essential to comprehend the distinction between speaker recognition and speech recognition. Speech recognition focusses on the uttered words, while speaker or voice recognition seeks to identify the speaker rather than the words themselves [8].

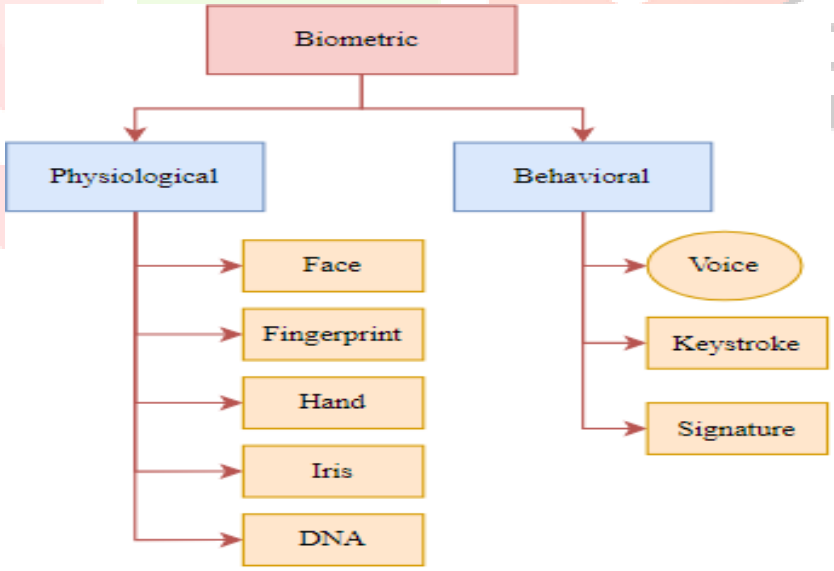

Figure 1 Types of biometrics

1.1 Speech recognition vs Speaker recognition

Speech recognition is beneficial for individuals with diverse disabilities, including those with physical impairments who struggle to type due to difficulty, discomfort, or impossibility, as well as those who have challenges in word recognition and spelling, such as those with dyslexia [9].

Table 2 Speech recognition vs Speaker recognition

Feature	Speech recognition	Speaker recognition
Recognition	Recognises what is spoken and translates it into text.	Identifies the speaker by analysing voice patterns, speaking style, and other verbal characteristics.
Purpose	To recognise as well as digitally record what the speaker is saying.	To recognize the speaker
Focus	The speaker's vocabulary is transformed into digital text.	Biometric characteristics of the speaker, including pitch and in Identification purposes.
Application	Speech to text.	Voice biometrics

The efficacy of voice recognition, which involves transforming audio into text, is significantly influenced by the language and the text corpus. Conversely, speaker recognition involves identifying the individual who is speaking. Pitch, speaking manner, and accent are among the characteristics that account for the variations. Speaker recognition technology has been used in several areas, including biometrics, security, and human-computer interface. Table 2 delineates the distinctions between speaker recognition and speech recognition regarding recognition, purpose, focus, and application [10].

2. Speech Recognition

There has been an increasing focus on modelling raw speech signals for applications such as automated speech recognition (ASR) and automatic speaker recognition (ASpR). Creating voice and speaker identification algorithms that remain impervious to variables such as age, gender, accent, or emotional condition is a significant problem in the domain [11] [12]. De- La-Calle-Silos et al., [13] developed a temporal prototype and application pattern of auditory nerve firings to enhance the efficacy of automated speech recognition systems. A novel feature extraction method for noise elimination, grounded in noise suppression has been developed to enhance the accuracy of speech recognition amidst additive noise.

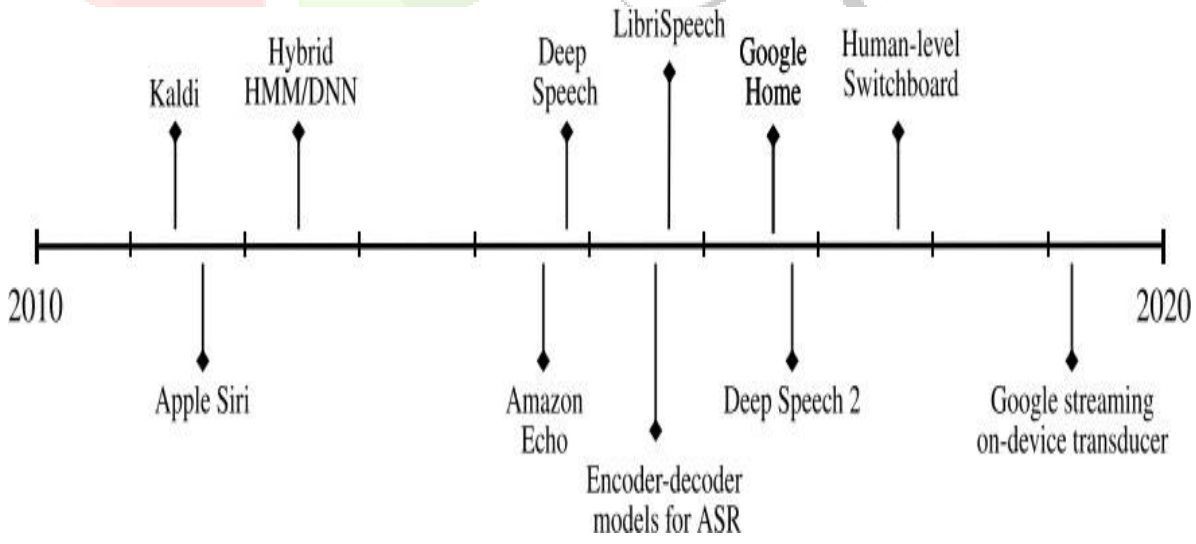

Figure 2a: A timeline of important breakthroughs in speech recognition from 2010 to 2020.

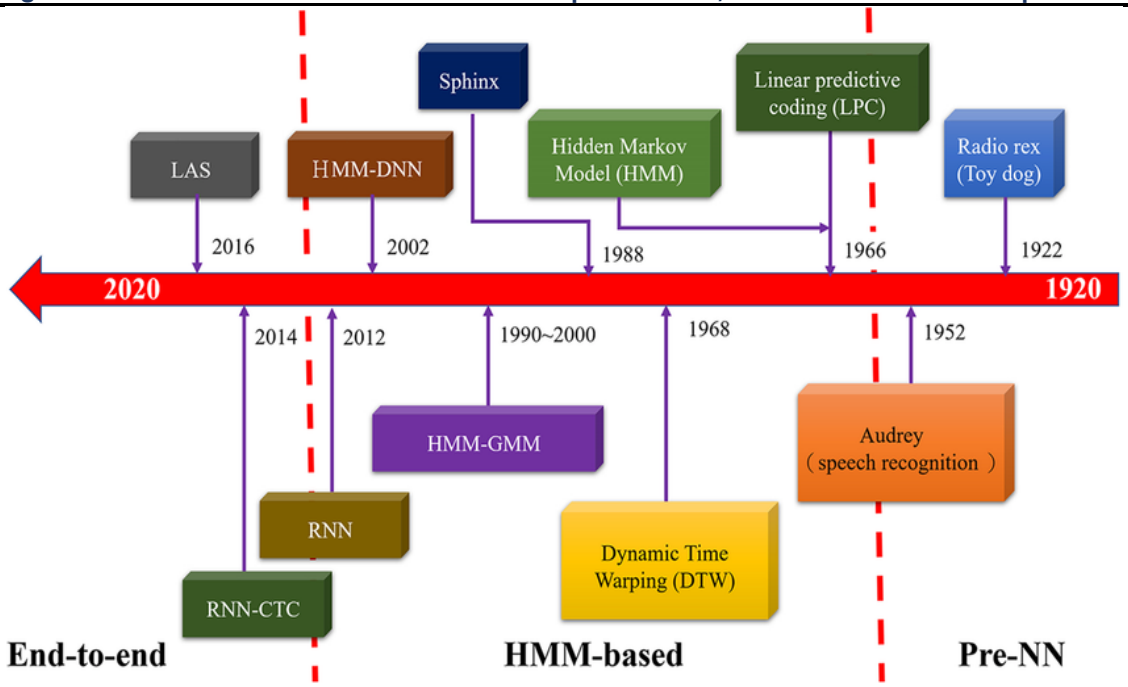

Figure 2b: Evolution of speech recognition from the early 1900s to 2020s

The period from 2010 to 2020 saw significant advancements in voice recognition and associated technologies. Figure 2a presents a history of significant advancements in the research, software, and implementation of voice recognition during the last decade. The decade saw the introduction and proliferation of mobile speech assistants such as Apple Siri. Farfield gadgets like as Amazon Alexa and Google Home were introduced and became widespread [14]. The advancement of these technologies was partially facilitated by the significant improvement in word error rates of automated voice recognition due to the emergence of deep learning. Figure 2b presents the evolution of speech recognition technologies from the early 1900s to 2020 [88].

The decade witnessed the launch of voice-activated gadgets and assistants, the proliferation of open-source speech recognition software like as Kaldi [15], and the establishment of extensive benchmarks like LibriSpeech [16]. Moreover, speech recognition models have evolved from hybrid neural network designs [17] to more comprehensive end-to-end models, including Deep Speech [18], Deep Speech 2 [19], encoder-decoder models incorporating attention [20], and transducer-based speech recognition [21].

2.1 Approaches for Speech Recognition

2.1.1 Acoustic Phonetic Approach:

It is a conventional approach of speech identification that asserts that the spoken language comprises finite distinctive phonetic units, often known as phonemes, and that these units are typically represented by a collection of acoustic characteristics that are formed in the speech signal over time. The acoustic properties of phonetic units are highly variable, both with speakers and with neighboring sounds; this is referred to as the vocalisation effect. The acoustic-phonetic approach assumes that the rules underlying the unpredictability are simple and can be willingly learnt by a machine [22].

2.1.2 Pattern Recognition Approach:

This method relies on the categorisation of input data into identifiable modules by means of the extraction of important data aspects or attributes from unimportant contextual details. Using this approach, the raw data is gathered and statistically analysed. and after that, produce the pattern based on the statistical characteristic. It follows the same basic structural pattern and yields the same outcomes. If there are two distinct scenarios, the system has to be educated or improved to get the intended outcome. Consequently, the speech approach system may provide a range of quantitative and

statistical methodologies to ascertain the needed answers or outcomes. There are four predominant methodologies for speech recognition patterns. In a Statistical or Stochastic Based method, speech modelling is conducted by statistically managing all components of speech variance. Hidden Markov Models (HMM) may be used for modelling. This technique relies on prior assumptions [23] [24]. The system's performance may be constrained by erroneous assumptions or prior modelling [25].

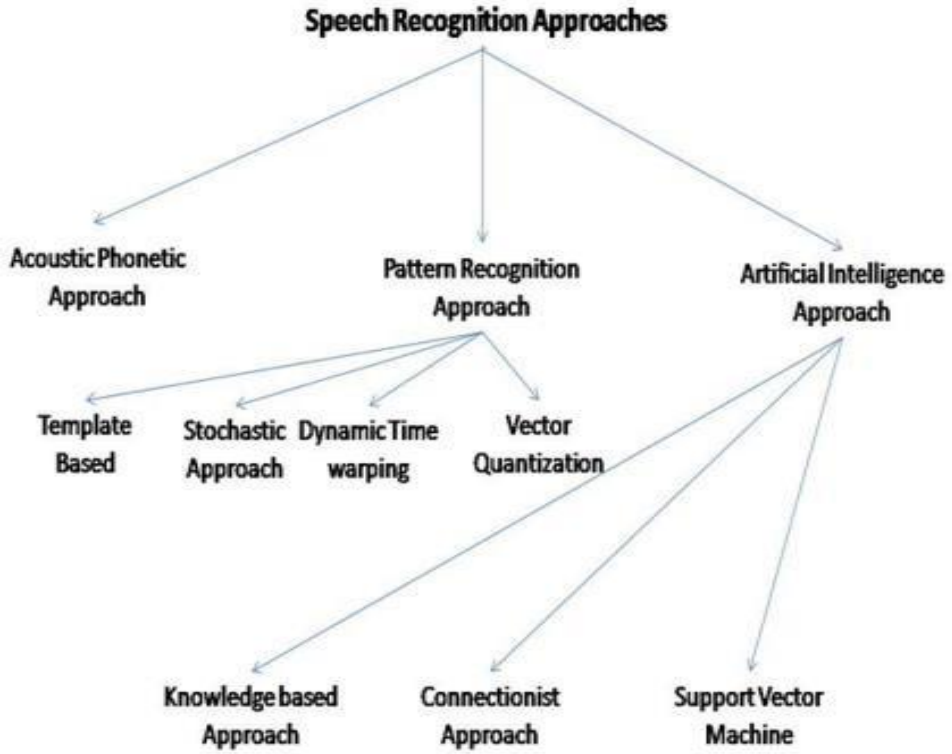

Figure 3 Speech recognition approaches

2.1.3 Artificial Intelligence approach

The AI technique for recognising speech is based on neural networks. The development of natural human interfaces for communication with computers is now one of the foremost issues in contemporary research. Computer simulation demonstrates the efficacy of outcomes. Three methodologies exist for the creation of artificial neural networks [25] [26] as shown in Figure 3.

2.2 Hybrid Speech Recognition Systems

ASR that leverages the robust representation learning capabilities of DNNs with the sequential modelling proficiency of HMMs [27]. This model is referred to as the DNN-HMM hybrid system. This framework models the dynamics of the speech signal using Hidden Markov Models (HMMs), while the observation probabilities are computed using Deep Neural Networks (DNNs). Each output neuron of the DNN is trained to predict the posterior probability of the continuous density HMMs' state based on the acoustic observations. Figure 4 illustrates the design of this framework.

The HMM models the sequential characteristics of the speech signal, while the DNN determines the scaled observation probability for all senones. This structure has two primary advantages: training may be executed with the Viterbi method, which identifies the route that maximises the likelihood of observation between the values of the HMM represented in matrix form, and decoding is often very efficient. [29] conducted a comparative analysis of GMM- HMM and DNN-HMM pronunciation verification methodologies, revealing that the hybrid DNN-HMM consistently surpasses the traditional GMM-HMM across all tests including both normal and disturbed speech.

The DNN-HMM demonstrates an accuracy over 85% when applied to disordered speech. [30] proposed a comparison between the hybrid GMM-HMM model and the DNN-HMM model using the

GlobalPhone (GP) multilingual text and voice database. The experimental findings indicate that the hybrid DNN-HMM frameworks surpass the GMM-HMM frameworks, irrespective of the training speech size used. The attained relative enhancement varies from 7.14% to 59.43%. [31] introduced a hybrid SGMM-HMM framework for acoustic modelling, which was evaluated against baseline GMM-HMM methods. Initially, GFCC and MFCC feature vectors [32] are extracted; subsequently, these features are used to train context-dependent (triphone) and context-independent (monophone) states using the given acoustic observations.

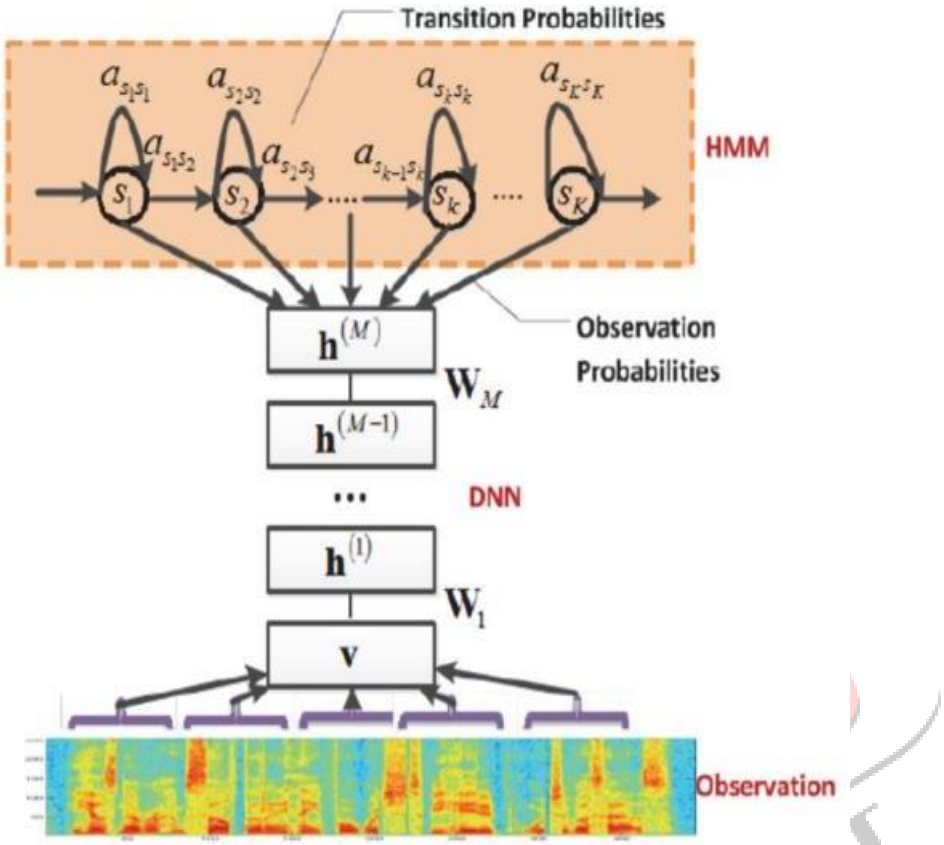

Figure 4 Architecture of the DNN-HMM hybrid system [28]

The monophone state is first taught to facilitate the training of the triphone state, which aids in the development of the GMM-HMM model. This information state is also used by the SGMM-HMM methodology. The suggested system is assessed using medium vocabulary in the Punjabi language. The sanitised data is used for the training and evaluation of the proposed system using two hybrid classifiers: SGMM-HMM and GMM- HMM. System testing is conducted with two front-end methodologies: MFCC and GFCC. The experimental findings indicate that the SGMM-HMM framework achieves an enhancement of 3–4% compared to the GMM-HMM approach.

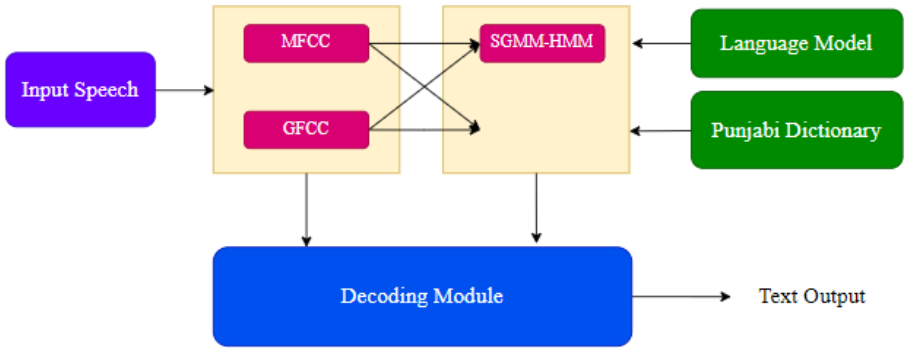

Figure 5 illustrates the design of the SGMM-HMM framework.

2.3 Speaker Recognition A conventional speaker recognition system evaluates the attributes of an individual's voice or speech to determine their identity. Vocalisation is the most rational method to enhance human senses. In the last sixty years, automatic speech recognition systems have progressed

significantly due to the emergence of human-computer research technologies. Currently, these sophisticated systems are used in several domains like personal identity, authentication, voice dialing, online banking, telephone commerce, security management, and forensic applications.

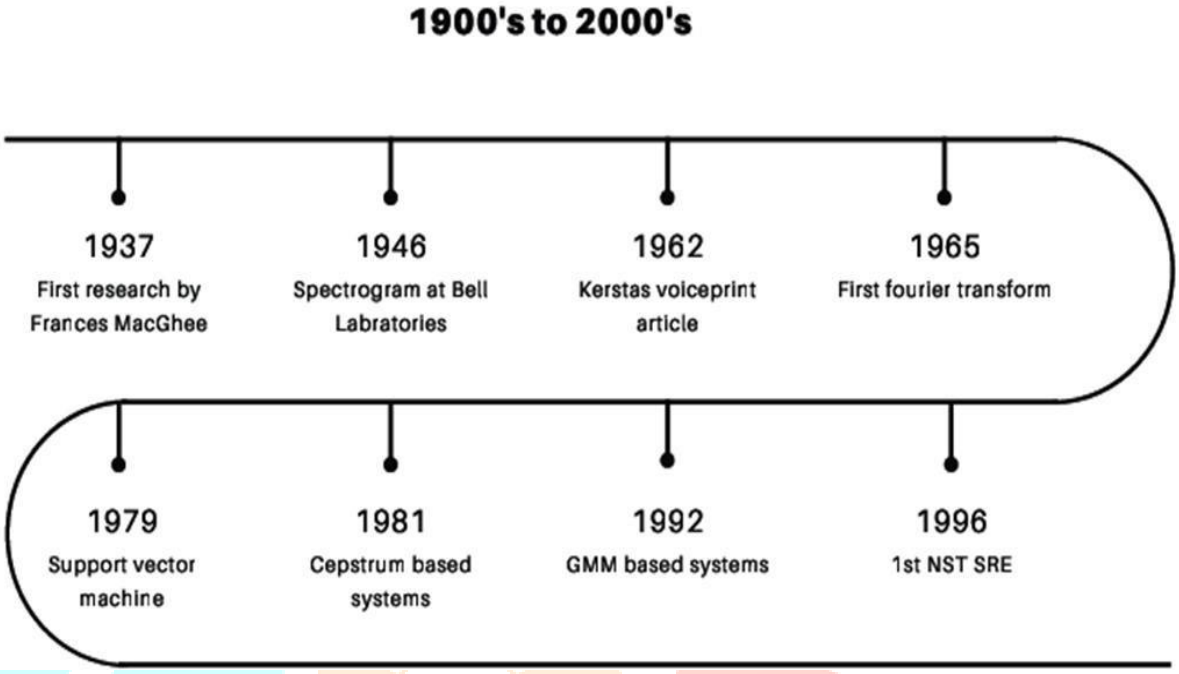

Figure 6 Evolution of speaker recognition from the early 1900s to 2000s

A conventional speaker identification system has three main components: pre-processing, feature extraction, and speaker modelling [35].

A Speaker Recognition (SR) system evaluates the characteristics of an individual's voice or speech to ascertain their identification. Although recognising a voice on the phone is a routine activity for humans, autonomous speech recognition tasks are challenging. Autonomous SR systems have seen both notable accomplishments and disappointments throughout the years. Significant progress over the last fifty years has facilitated the resolution of several substantial issues for SR systems. Contemporary systems provide an effective method for validating user access privileges, recognising individuals within a group, and facilitating restricted applications in forensics. The initial studies in SR focused on human talents. Later, wartime research resulted in substantial improvements in autonomous systems, including the development of a tool for visual speech assessment. The development of autonomous systems was enabled by advancements in signal processing and computer technology. Despite certain restrictions, several applications have made significant progress towards commercial systems [36].

Speaker recognition allows a conversational robot to authenticate users for access control, identify conversing parties, personalise the device, and tailor the device and its applications to people. Speaker-recognition technologies enable systems to autonomously identify the speaker or, more specifically, to assess if the incoming speech matches that of an enrolled user. This decision may thereafter be used for user authentication in access control, identification of communicating entities, and the personalisation and modification of the radio and its applications. Figure 7 illustrates the fundamental processes of a speaker-recognition system, including its two operational phases: enrolment and verification [37].

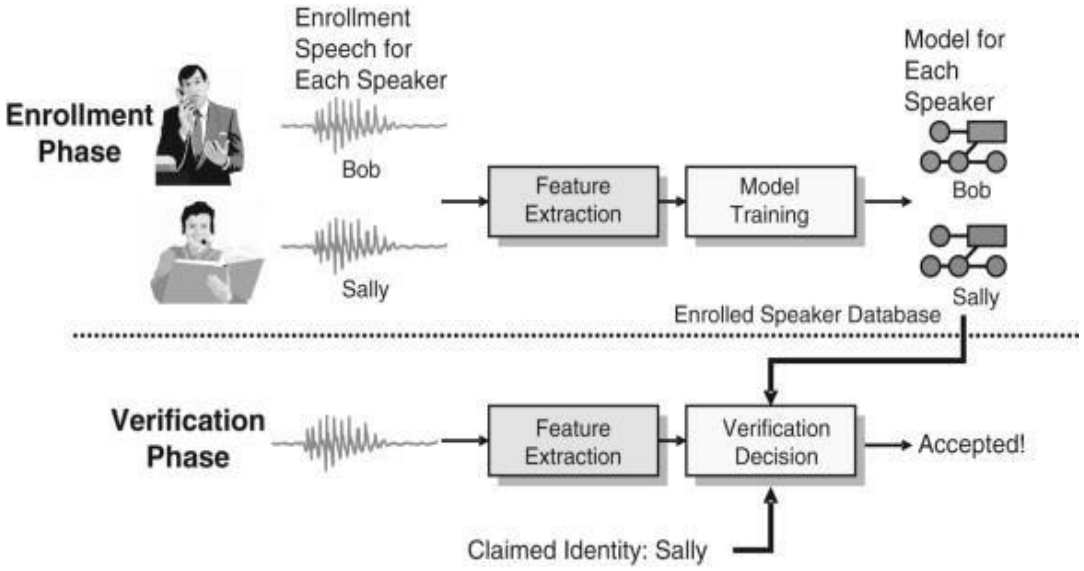

Figure 7 Basic Process of Speaker Recognition System

2.3.1 Types of Speaker Recognition

Speaker recognition may be categorised as speaker identification, speaker verification, speaker detection, speaker segmentation, speaker clustering, and speaker diarization [38]. A concise overview of these categories is provided in Table 3 to provide a clearer comprehension of their distinctions.

Table 3 Types of Speaker Recognition

Speaker Recognition Types	Description
Speaker Identification	Speaker identification involves categorizing an unidentified voice, delivered anonymously, as belonging to one of a predetermined set of N reference speakers [39].
Speaker Verification	Speaker verification involves determining if an unknown voice corresponds to a designated reference speaker, resulting in two potential outcomes: acceptance of the reference speaker or rejection of the imposter [40].
Speaker Segmentation	Speaker segmentation involves identifying the moments at which the speaker transitions during the presentation of an audio stream to the system [42].
Speaker Clustering	Speaker clustering is the process of accurately grouping a large number of utterances supplied to the system, which is commonly done online [43].
Speaker Diarization	Speaker diarization involves the automated segmentation of audio into distinct speaker segments and the identification of segments uttered by the same speaker [44].

2.3.2 Standard literature survey on Speaker Recognition

Reference	Year	Aim	Challenge
[45]	2010	An examination of both ancient and contemporary automated text- independent speech recognition systems, emphasising efficiency and identifying future prospects for the development of robust approaches in speech recognition.	There exists a lack of training data, incongruent devices for training and evaluation, ambient noise, and an imbalance in the text.
[46]	2010	A comprehensive examination of automated speaker identification systems.	Unauthorized access, privacy in biometric technology, the possibility of breaking data encryption, developing long-range features, handling enormous numbers of linked features, odd feature range distributions, original disappearing features and heterogeneous feature types.
[47]	2011	Overview of text-independent, closed-set speaker identification in modelling and classification paradigms, including significant extracted characteristics from clean and missing data.	The authors exclusively examined speaker identification systems that rely on missing data methodologies.
[48]	2015	A comparison between human and computer speaker recognition.	In system development, the search for alternative compact representations of speakers and audio segments that emphasise the identification of key attributes while reducing unwanted components.
[49]	2016	A detailed examination of deep learning methods and applications for speaker recognition.	The study addressed expertise on enhancing SID performance.
[50]	2017	Describes the fundamentals of optimum characteristics and how feature parameter appropriateness is evaluated for the feature extraction approach in speaker identification. The feature extraction methodologies and architectures are also discussed.	Consolidate the ability to trade with all channel data and noisy data, handle complicated pattern recognition issues using deep learning methods and a multi-learner model, and extract features from unlabelled data as well as incomplete, manipulated, or destroyed data.
[51]	2017	Summarise feature extraction methods in degraded circumstances, together with short-time features, and use a few normalization strategies to improve resilience.	Improved performance in extracting robust features from noise and channel mismatch scenarios.

[52]	2021	A survey of deep learning-based automated speaker identification systems.	This paper presents additional challenges in Automated Speaker Recognition (ASR) techniques, which often require handcrafted acoustic features and have numerous parameters, making them unsuitable for portable devices.
[35]	2021	This review provides a comprehensive overview of feature extraction methodologies, algorithms, and limits for subdomains of speaker recognition, including identification, verification, and diarization.	Data-driven dependence; inter-speaker variability; speaker-based variability; conversation-based variability; technology- based variability; Limited data and limited lexicon; Ageing speaker models; Forward compatibility.
[53]	2024	It provides a historical context, highlights key developments, and discusses the evolution of speaker recognition based on AI, such as the x-vector system and deep neural networks.	

2.4 Approaches for Speaker Recognition

2.4.1 Feature extraction approaches

A collection of feature vectors may encapsulate the voice signal, enabling the use of mathematical techniques without sacrificing generality. The objective of feature extraction is to analyse a voice signal by determining the number of its components. The rationale is that some data points in the acoustic signal is significant for analysis, and some data is unsuitable. In practical applications, many elements are used in conjunction for speech identification tasks. All speakers possess similar guttural characteristics, mostly derived from learnt behaviour and physiological attributes. The following aspects are briefly elucidated:

Behaviour-based speech features:

This feature is derived by quantifying aspects such as experience, personality, education, family ties, community, and communication channels. High-level features and prosodic and spectro- temporal characteristics are two kinds of learnt features. Prominent traits include phones, idiolect, semantics, accent, and pronunciation. Conversely, prosodic and spectro-temporal traits include pitch, energy, rhythm, length, and temporal attributes. Prosodic qualities refer to the non-segmental features of speech shown in extended utterances, including prosodic traits. Accent is an associative technicality used to delineate differences in pitch, stress, rhythm, and loudness. Table 4 presents a timeline of learned-based feature extraction techniques.

Table 4 The table summarizes the behavior-based feature extraction techniques in speaker recognition

Learned-Based Approaches			
High level	References	Prosodic	References
PMFC (Phoneme Mean F-ratio Coefficient)	[54]	Sub-band Auto Correlation Classification (SACC)	[55] [56] [57] [58]
Polish vowel and PMFFCC (Phoneme Mean F-ratio Frequency Cepstrum Coefficient)	[59], [60]	PROSACC	[61]

Vowel phonemes	[62]	Shifted Delta Cepstrum (SDC) and additional temporal information	[63]
Maximum-Likelihood Linear Regression (MLLR)	[64]	DPE to present pitch and energy by using twelve DCT coefficients	[63]
Seven Shannon entropy wavelet packet and five formants were extracted from vowelM	[65]		

Physiologically related speech features:

These features are influenced by the length, dimensions, and size of the vocal tract folds. The short-term spectral characteristics refer to physiological variables measurable from brief spoken utterances, these qualities elucidate the short-term spectral envelope associated with the timbre and resonance properties of the supralaryngeal vocal tract. Voice source features are attributes of vocal expression. The short-term spectral properties are further classified into two categories: Spectrum and Gammatone pulse features. Table 5 presents short-term, physiology- based feature extraction approaches.

3. Challenges in Speech and Speaker Recognition

Despite much work on ASR, the absence of homogeneity in speech systems remains a substantial problem. Several difficulties are prevalent in contemporary systems [79]:

3.1 Variation in Accents

Speech recognition systems often fail to address the diverse accents in every language, leading to a domain adaptation bias. This bias is gaining traction, causing discussions on the need for more generalized and inclusive technology.

Table 5 lists physiological-based feature extraction techniques for speaker identification.

Physiological based speech features			
Spectrum	References	Gammatone pulse	References
MFCC (Mel-Frequency Cepstral Coefficients)	[66] [67] [68] [69]	Gammatone Feature	[70] [71]
Linear Predictive Cepstral Coefficients (LPCC)	[72]	Gammatone Frequency Cepstral Coefficients (GFCC)	[73]
Linear Predictive Coefficients (LPC)	[74] [75]	Hilbert Envelope of Gammatone Filter-bank	[76]
Linear Predictive Residual (LPR)	[77]	Perceptual Linear Prediction (PLP)	[78]

The fluctuating syllables and phonetics of the same word make it difficult for machines to process. Common challenges with voice recognition include differences in pronunciation, accent, and intonation [80]. To solve these disparities, exponentially growing amounts of data from people of different regions and dialects can be used in training models. However, many people must adapt to their first language, often leading to an identity crisis for marginalized communities like people with disabilities who rely heavily on voice recognition and speech-to-text tools [81]. Therefore, technology must no longer be disconnected from the complexity of human languages worldwide.

3.2 Insufficient data

Insufficient data denotes the lack of enough data necessary to build representative models for making informed decisions. This signifies a significant and prevalent issue, since the majority of applications need systems that function with the minimal feasible quantity of training data collected throughout the least number of enrolment sessions [82]. The study in [83] emphasised that the difficulty lies in the assumption that the test data distribution aligns with the distribution depicted by the speaker model. The system functions well when the speaker model is thoroughly trained using extensive enrolment speech and sufficient test speech samples. In [3], Singh asserted that the system's performance is contingent upon the volume of training data. As per [84], there is no evidence indicating that the duration of the voice sample yields superior outcomes. Conversely, [83] asserted that an ample quantity of speech data is essential for model training, while also acknowledging the practical challenges in acquiring extended utterances, since users often do not talk for prolonged durations. Consequently, this prompted speaker identification to become a prominent subject for future exploration.

3.3 Model Training

Deep learning is useful in cutting-edge automatic speech recognition systems. However, overfitting is a serious issue with ANNs that only becomes obvious when evaluating fresh data. Overfitting leads a model learns the information and noise in training data to the point that it degrades its performance on fresh data. This implies that the model detects noise or random oscillations in the training data and interprets them as ideas. The difficulty is that these concepts are not applicable to fresh data, restricting the models capacity to generalise. Dropout training may help to address the issue. However, it seems to have been used for word recognition under difficult circumstances. ANNs are less effective in voice signals with extended time sequences. Furthermore, neural networks are more accurate and precise, but they need high compositionality, the longest processing time [85], and are unable to discern between highly similar texts [86].

3.4 Noise

Another issue with noise is that it often results in significant data loss, causing more harm than intended. One technique to dramatically enhance noise-robust voice recognition is to create and operate with speech signals in real-time situations. A few of such examples include using standard datasets in the noise section, such as the CHiME challenge series [78] that provides speech signals in unique noisy environments like dinner parties, domestic environments, etc., the NOISEX, which is a database of recording of various noises like voice babble and factory noises, and the SpEAR Database (Speech Enhancement and Assessment Resource), which provides noise-corrupted speech samples paired with clean speech references. Background noise is another issue that [83] identifies as troublesome since the speaker often talks in a clean setting when training. In contrast, during testing, the speaker talks in a loud environment. Background noise has a major influence on the accuracy of speaker identification. Accuracy is high for clean samples and low for noisy samples with no effect from babbling noise [87]. The recorded voice wave with background disturbances such as white noise, music, etc. has the greatest impact on speaker modelling. It disrupts the assessment test and lowers the performance of the speech identification system.

4. CONCLUSION

The speech and speaker recognition survey identifies some of the prominent differences between these two technologies. The approaches and challenges of speaker recognition and speech recognition have also been discussed. Speech recognition has been focused upon the accurate transformation of a spoken language to text, using different techniques like the Acoustic Phonetic Approach, Pattern Recognition, and Artificial Intelligence, they use hybrid systems to yield best performance. It is on speaker identification or verification, with which it is also concerned-thus identified, and divided into text-dependent and text-independent types based on different feature extraction approaches. The speaker recognition literature presents significant developments but challenges of accent variation, insufficient data, model training complications, and noise still stand out. Overcoming these bottlenecks will be an essential step to allow robust application and usage in a wide variety of real-world environments. Advances in future AI and machine learning research might find applications in overcoming these challenges, thereby improving the robustness of speech and speaker recognition technologies.

REFERENCES

[1] O. Adenuga and A. Oduroye, "Biometric Authentication: A Review," Aug. 2023.

[2]R. de Luis-García, C. Alberola-Lopez, O. Aghzout, and J. Ruiz-Alzola, "Biometric identification systems," *Signal Processing*, vol. 83, no. 12, pp. 2539–2557, 2003.

[3]N. Singh, "A study on speech and speaker recognition technology and its challenges," in *Proceedings of National Conference on Information Security Challenges. Lucknow: DIT, BBAU*, 2014, pp. 34–37.

[4]A. M. Sharma, "Speaker recognition using machine learning techniques," 2019.

[5]L. Rampello, L. Rampello, F. Patti, and M. Zappia, "When the word doesn't come out: A synthetic overview of dysarthria," *J. Neurol. Sci.*, vol. 369, pp. 354–360, 2016.

[6]Z. Qian, K. Xiao, and C. Yu, "A survey of technologies for automatic Dysarthric speech recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2, 2023, doi: 10.1186/s13636-023-00318-2.

[7]J. Keshet and S. Bengio, "Automatic speech and speaker recognition," *Large Margin Kernel Methods, John Willy Sons*, 2009.

[8]C. Kikel, "Difference between voice recognition and speech recognition," *Blog] Total Voice Technol.*, 2019.

[9]R. M. Hanifa, K. Isa, and S. Mohamad, "Malay speech recognition for different ethnic speakers: an exploratory study," in *2017 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2017, pp. 91–96.

[10] B. Today, "Differences between Voice and Speech Recognition.," 2018.

[11] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, 2015.

[12] K. Radha, M. Bansal, and R. Bilas, "Speech and speaker recognition using raw waveform modeling for adult and children ' s speech : A comprehensive review," *Eng. Appl. Artif. Intell.*, vol. 131, no. December 2023, p. 107661, 2024, doi: 10.1016/j.engappai.2023.107661

[13] F. De-La-Calle-Silos and R. M. Stern, "Synchrony-based feature extraction for robust automatic speech recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1158– 1162, 2017.

[14] A. Hannun, "The History of Speech Recognition to the Year 2030," 2021, [Online].
Available: http://arxiv.org/abs/2108.00084

[15] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 5206–5210.

[17] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[18] A. Hannun, "Deep Speech: Scaling up end-to-end speech recognition," *arXiv Prepr. arXiv1412.5567*, 2014.

[19] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[20] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[21] Y. He *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6381–6385.

[22] R. D. N. Kaur, "Speech Recognition Using Stochastic Approach: A Review".

[23] A. P. Singh, R. Nath, and S. Kumar, "HANDWRITTEN ENGLISH WORD RECOGNITION USING HMM, BAUM-WELCH AND GENETIC ALGORITHM".

[24] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Readings speech Recognit.*, pp. 267–296, 1990.

[25] S. K. Saksamudre, P. P. Shrishrimal, and R. R. Deshmukh, "A review on different approaches for speech recognition system," *Int. J. Comput. Appl.*, vol. 115, no. 22, 2015.

[26] D. O'Shaughnessy, "Automatic speech recognition: History, methods and challenges,"
Pattern Recognit., vol. 41, no. 10, pp. 2965–2979, 2008.

[27] D. Yu and L. Deng, *Automatic speech recognition*, vol. 1. Springer, 2016.

[28] S. M. Abdou and A. M. Moussa, "Arabic Speech Recognition : Challenges and State of the Art," pp. 1–27.

[29] M. Shahin, B. Ahmed, J. McKechnie, K. Ballard, and R. Gutierrez-Osuna, "A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[30] M. Y. Tachbelie, A. Abulimiti, S. T. Abate, and T. Schultz, "Dnn-based speech recognition for globalphone languages," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8269– 8273.

[31] V. Kadyan and M. Kaur, "SGMM-based modeling classifier for punjabi automatic speech recognition system," in *Smart Computing Paradigms: New Progresses and Challenges: Proceedings of ICACNI 2018, Volume 2*, 2020, pp. 149–155.

[32] Z. Wu and Z. Cao, "Improved MFCC-based feature for robust speaker identification," *Tsinghua Sci. Technol.*, vol. 10, no. 2, pp. 158–161, 2005.

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

[34] C. D. Shaver and J. M. Acken, "A brief review of speaker recognition technology," 2016.

[35] M. M. Kabir *et al.*, "A Survey of Speaker Recognition : Fundamental Theories , Recognition Methods and Opportunities," vol. 9, 2021, doi: 10.1109/ACCESS.2021.3084299.

[36] C. D. Shaver, J. M. Acken, C. D. Shaver, and J. M. Acken, "A Brief Review of Speaker Recognition Technology," 2016.

[37] J. P. Campbell, W. M. Campbell, A. V McCree, C. J. Weinstein, and S. M. Lewandowski, "Cognitive Services for the User," B. A. B. T.-C. R. T. (Second E. Fette, Ed. Oxford: Academic Press, 2009, pp. 305–324. doi: https://doi.org/10.1016/B978-0-12-374535- 4.00010-2.

[38] U. SANDOUK, "Speaker recognition, speaker diarization and identification," *Univ. Manchester Sch.*, 2012.

[39] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE circuits Syst. Mag.*, vol. 11, no. 2, pp. 23–61, 2011.

[40] R. Naika, "An overview of automatic speaker verification system," in *Intelligent Computing and Information and Communication: Proceedings of 2nd International Conference, ICICC 2017*, 2018, pp. 603–610.

[41] A. Vishwakarma and M. Soni, "Speaker Recognition System–A Review," *Int. J. Emerg. Technol. Eng. Res.*, pp. 85–89.

[42] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," *Signal Processing*, vol. 88, no. 5, pp. 1091–1124, 2008.

[43] V. V Patil and P. M. Gaud, "Improved speaker segmentation using clustering technique," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017, pp. 1554–1558.

[44] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[45] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors." 2010.

[46] Z. Saquib, N. Salam, R. P. Nair, N. Pandey, and A. Joshi, "A survey on automatic speaker recognition systems," *Commun. Comput. Inf. Sci.*, vol. 123 CCIS, pp. 134–145, 2010, doi: 10.1007/978-3-642-17641-8_18.

[47] R. Togneri and D. Pullella, "An Overview of Speaker Identification: Accuracy and Robustness Issues," no. May, pp. 23–61, 2011.

[48] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans," no. November, pp. 74–99, 2015.

[49] S. S. Tirumala and S. R. Shahamiri, "A review on deep learning approaches in speaker identification," *ACM Int. Conf. Proceeding Ser.*, no. November 2016, pp. 142–147, 2016, doi: 10.1145/3015166.3015210.

[50] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Syst. Appl.*, vol. 90, pp. 250– 271, 2017, doi: https://doi.org/10.1016/j.eswa.2017.08.015.

[51] G. Dişken, Z. Tüfekçi, L. Saribulut, and U. Çevik, "A Review on Feature Extraction for Speaker Recognition under Degraded Conditions," *IETE Tech. Rev. (Institution Electron. Telecommun. Eng. India)*, vol. 34, no. 3, pp. 321–332, 2017, doi: 10.1080/02564602.2016.1185976.

[52] Z. Bai and X. L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021, doi: 10.1016/j.neunet.2021.03.004.

[53] R. Sharma, D. Govind, J. Mishra, A. K. Dubey, K. T. Deepak, and S. R. M. Prasanna, *Milestones in speaker recognition*, vol. 57, no. 3. Springer Netherlands, 2024. doi: 10.1007/s10462-023-10688-w.

[54] C. Zhao, H. Wang, S. Hyon, J. Wei, and J. Dang, "Efficient feature extraction of speaker identification using phoneme mean F-ratio for Chinese," in *2012 8th International Symposium on Chinese Spoken Language Processing*, 2012, pp. 345–348.

[55] E. El Khoury, A. Laurent, S. Meignier, and S. Petitrenaud, "Combining transcription- based and acoustic-based speaker identifications for broadcast news," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2012, pp. 4377–4380.

[56] Y. Kawakami, L. Wang, A. Kai, and S. Nakagawa, "Speaker identification by combining various vocal tract and vocal source features," in *Text, Speech and Dialogue: 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings 17*, 2014, pp. 382–389.

[57] A. Prasad, V. Periyasamy, and P. K. Ghosh, "Estimation of the invariant and variant characteristics in speech articulation and its application to speaker identification," in *2015 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 4265–4269.

[58] O. Plchot *et al.*, "Developing a speaker identification system for the DARPA RATS project," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6768–6772.

[59] K. Salapa, A. Trawińska, I. Roterman, and R. Tadeusiewicz, "Speaker identification based on artificial neural networks. Case study: the Polish vowel (pilot study)," *Bio- Algorithms and Med-Systems*, vol. 10, no. 2, pp. 91–99, 2014.

[60] S. Cumani, O. Plchot, and M. Karafiát, "Independent component analysis and MLLR transforms for speaker identification," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4365–4368.

[61] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer, and Y. Lei, "Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 6773–6777.

[62] M. Sarma and K. K. Sarma, "Vowel phoneme segmentation for speaker identification using an ANN-based framework," *J. Intell. Syst.*, vol. 22, no. 2, pp. 111–130, 2013.

[63] M. Kockmann and L. Burget, "Application of speaker-and language identification state- of-the-art techniques for emotion recognition," *Speech Commun.*, vol. 53, no. 9–10, pp. 1172–1185, 2011

[64] R. Saeidi, A. Hurmalainen, T. Virtanen, and D. A. van Leeuwen, "Exemplar-based sparse representation and sparse discrimination for noise robust speaker identification," 2012.

[65] K. Daqrouq and T. A. Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers," *Appl. Soft Comput.*, vol. 27, pp. 231–239, 2015, doi: https://doi.org/10.1016/j.asoc.2014.11.016.

[66] H. Lu, A. J. Bernheim Brush, B. Priyantha, A. K. Karlson, and J. Liu, "Speakersense: Energy efficient unobtrusive speaker identification on mobile phones," in *Pervasive Computing: 9th International Conference, Pervasive 2011, San Francisco, USA, June 12-15, 2011. Proceedings 9*, 2011, pp. 188–205.

[67] G. Biagetti, P. Crippa, A. Curzi, S. Orcioni, and C. Turchetti, "Speaker identification with short sequences of speech frames," in *International Conference on Pattern Recognition Applications and Methods*, 2015, vol. 2, pp. 178–185.

[68] G. Biagetti, P. Crippa, L. Falaschetti, S. Orcioni, and C. Turchetti, "Robust speaker identification in a meeting with short audio segments," in *Intelligent decision technologies 2016: proceedings of the 8th kes international conference on intelligent decision technologies (KES-IDT 2016)–Part II*, 2016, pp. 465–477.

[69] J.-C. Wang, Y.-H. Chin, W.-C. Hsieh, C.-H. Lin, Y.-R. Chen, and E. Siahaan, "Speaker identification with whispered speech for the access control system," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 4, pp. 1191–1199, 2015.

[70] X. Zhang, H. Zhang, and G. Gao, "Missing feature reconstruction methods for robust speaker identification," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1482–1486.

[71] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 836– 845, 2014.

[72] M. Rossi, O. Amft, and G. Tröster, "Collaborative personal speaker identification: A generalized approach," *Pervasive Mob. Comput.*, vol. 8, no. 3, pp. 415–428, 2012.

[73] N. P. Jawarkar, R. S. Holambe, and T. K. Basu, "Effect of nonlinear compression function on the performance of the speaker identification system under noisy conditions," in *Proceedings of the 2nd international conference on perception and machine intelligence*, 2015, pp. 137–144.

[74] M. Chandra, P. Nandi, A. Kumari, and S. Mishra, "Spectral-subtraction based features for speaker identification," in *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014: Volume 2*, 2015, pp. 529–536.

[75] J. B. Ramgire and S. M.Jagdale, "A Survey on Speaker Recognition with Various Feature Extraction Techniques," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 2, pp. 884–887, 2019, doi: 10.26438/ijcse/v7i2.884887.

[76] S. O. Sadjadi and J. H. L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *2011 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 2011, pp. 5448–5451.

[77] S. Khan, J. Basu, and M. S. Bepari, "Performance evaluation of PBDP based real-time speaker identification system with normal MFCC vs MFCC of LP residual features," in *Indo-Japanese Conference on Perception and Machine Intelligence*, 2012, pp. 358–366.

[78] H. Bredin, A. Roy, V.-B. Le, and C. Barras, "Person instance graphs for mono-, cross- and multi-modal person recognition in multimedia data: application to speaker identification in TV broadcast," *Int. J. Multimed. Inf. Retr.*, vol. 3, pp. 161–175, 2014.

[79] S. Basak *et al.*, "Challenges and Limitations in Speech Recognition Technology: A Critical Review of Speech Signal Processing Algorithms, Tools and Systems," *C. - Comput. Model. Eng. Sci.*, vol. 135, no. 2, pp. 1053–1089, 2023, doi: 10.32604/cmes.2022.021755.

[80] S. D. Deshmukh and M. R. Bachute, "Automatic speech and speaker recognition by mfcc, hmm and vector quantization," *Int. J. Eng. Innov. Technol.*, vol. 3, no. 1, pp. 109– 113, 2013.

[81] C. L. Lloreda, "Speech recognition tech is yet another example of bias," *Sci. Am.*, vol. 5, 2020.

S. Furui, "40 years of progress in automatic speaker recognition," in *Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009.* T. F. Zheng and L. Li, *Robustness-related issues in speaker recognition*, vol. 2. Springer, 2017.

[82] V. Sharma and P. K. Bansal, "A review on speaker recognition approaches and challenges," *Int. J. Eng. Res. Technol.*, vol. 2, no. 5, pp. 1581–1588, 2013. *Proceedings 3*, 2009, pp. 1050–1059.

[83] L. Eljawad *et al.*, "Arabic voice recognition using fuzzy logic and neural network," *ELJAWAD, L., ALJAMAEEN, R., ALSMADI, MK, AL-MARASHDEH, I., ABOUELMAGD, H., ALSMADI, S., HADDAD, F., ALKHASAWNEH, RA, ALZUGHOUL, M. ALAZZAM, MB*, pp. 651–662, 2019.

[84] M. A. Al-Alaoui, L. Al-Kanj, J. Azar, and E. Yaacoub, "Speech recognition using artificial neural networks and hidden Markov models," *IEEE Multidiscip. Eng. Educ. Mag.*, vol. 3, no. 3, pp. 77–86, 2008.

[85] D. P. Varun Sharma, "A Review On Speaker Recognition Approaches And Challenges," 2013.

[86] Wen-Tsai Sung, Hao-Wei Kang and Sung-Jung Hsiao "Speech Recognition via CTC-CNN Model" 2023, CMC, 2023, vol.76, no.3, DOI: 10.32604/cmc.2023.040024 PP:3833-3858.

[87] D.P. Varun Sharma, "A Review On Speaker Recognition Approaches And Challenges," 2013.

[88] Wen-Tsai Sung, Hao-Wei Kang and Sung-Jung Hsiao "Speech Recognition via CTC-CNN Model" 2023, CMC, 2023, vol.76, no.3, DOI: 10.32604/cmc.2023.040024 PP:3833-3858.