This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

Digital Object Identifier not assigned

A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities

MUHAMMAD MOHSIN KABIR [1], M. F. MRIDHA [1], (Senior Member, IEEE), JUNGPIL SHIN [2], (Senior Member, IEEE), ISRAT JAHAN [1], AND ABU QUWSAR OHI [1],

[1]Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh
[2]School of Computer Science and Engineering, University of Aizu, Aizuwakamatsu, Japan

Corresponding author: M. F. Mridha (e-mail: firoz@bubt.edu.bd) and Jungpil Shin (e-mail: jpshin@u-aizu.ac.jp)

The authors would like to thank the Advanced Machine Learning (AML) lab for resource sharing and precious opinions.

ABSTRACT Humans can identify a speaker by listening to their voice, over the telephone, or on any digital devices. Acquiring this congenital human competency, authentication technologies based on voice biometrics, such as automatic speaker recognition (ASR), have been introduced. An ASR recognizes speakers by analyzing speech signals and characteristics extracted from speaker's voices. ASR has recently become an effective research area as an essential aspect of voice biometrics. Specifically, this literature survey gives a concise introduction to ASR and provides an overview of the general architectures dealing with speaker recognition technologies, and upholds the past, present, and future research trends in this area. This paper briefly describes all the main aspects of ASR, such as speaker identification, verification, diarization etc. Further, the performance of current speaker recognition systems are investigated in this survey with the limitations and possible ways of improvement. Finally, a few unsolved challenges of speaker recognition are presented at the closure of this survey.

INDEX TERMS Automatic Speaker Recognition, Feature Extraction, Recognition Techniques, Performance Measures, Challenges.

I. INTRODUCTION

SPEAKER recognition is a biometric scheme applied to authenticate user's individuality using the specific characteristics elicited from their speech utterances. It is the automatic process of acknowledging the speaker depending on the speech signal's characteristic features. The Speaker recognition system uses the speaker's voice utterances to recognize their individuality and control access to services, such as voice dialling, voice mail, security control, etc.

The first automatic speaker recognition (ASR) system came into existence in 1962 through an article by Lawrence G. Kersta, a *Bell Laboratories* physicist designated, *"Voiceprint Identification"* [1], [2]. In 1960, Gunnar Fant developed a physiological model of humans voice production system, which sets a speech analysis base. The speaker recognition system's evolution from the late 1900s to the early 2000s is upheld in Figure 1 [3].

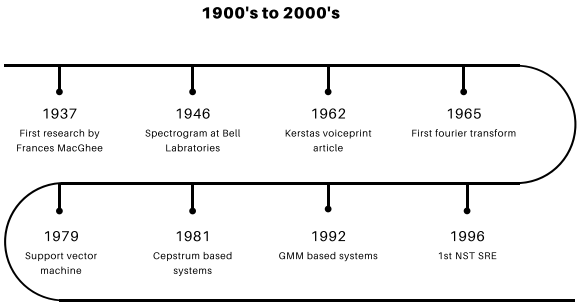

FIGURE 1. Evolution of speaker recognition from the early 1900s to 2000s. The above diagram explains how the evolution of ASR started in the 1900s before the ages of machine learning and deep learning.

A standard speaker recognition system measures the characteristics of a person's voice or speech to assess that person's individuality. Voice or speech is the most logical way to evolve the perceptions of humans. With the advent of human-computer research technologies, ASR systems have advanced in the last six decades. Nowadays, these advanced systems are used in different areas such as person identi-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

fication, person verification, voice dialling, online banking, telephone shopping, security control, and forensic applications as well. A typical speaker recognition system has three primary sections: pre-processing, feature extraction, and speaker modelling. Figure 2 presents a basic structure of speaker recognition systems.

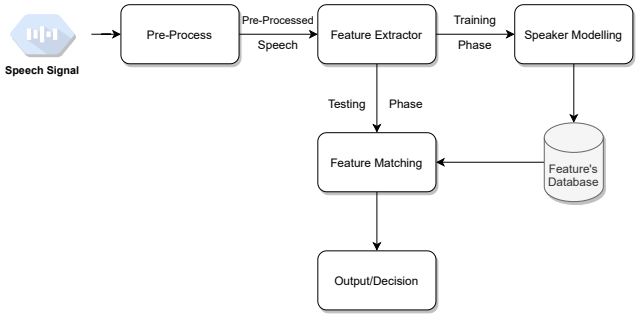

FIGURE 2. Basic structure of a automatic speaker recognition system. The figure illustrates three initial steps where the inputs are pre-processed and generate speaker models after being converted into feature vectors to identify speakers.

Pre-processing: Pre-processing is the initial action in an automated speaker recognition model. It is a crucial process conducted on speech signal input to manifest an effective and dynamic ASR system [4], [5]. In this part of a speaker recognition system, first, the speech signal is cleaned. Then the non-speech portions are removed from the signal. Then preliminary tasks endpoint detection and pre-emphasis are completed.

Feature extraction: Feature extraction, also known as front end pre-processing, is applied to speaker recognition systems training and testing phases. It employs to converts digital speech signal to sets of feature vectors or numerical descriptors. These feature vectors contain the essential characteristics of the speaker's voice [6].

Speaker modelling: The goal of modelling methods is to generate speaker recognition algorithms for feature matching of the speaker's voice. The methods comprising enhanced speaker-specific information with a compressed volume are defined as speaker models [7]. When training or enrolling, state speaker models are generated by practicing the particular features extracted from the contemporary speaker. The speaker model compares with the modern speaker architecture for identification or verification tasks in the recognition state.

A standard automated speaker recognition system has contained these characteristics. Now, the classifications of the ASR are going to be discussed.

A. CLASSIFICATION OF SPEAKER RECOGNITION

An ASR can be split into several classes based on the recognition criteria. Figure 3 presents the different types of speaker recognition approaches. The following subsections extensively outline the illustrated recognition approaches.

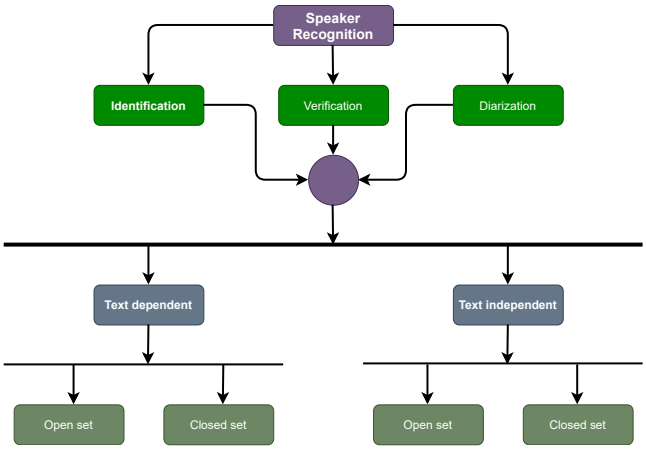

FIGURE 3. Classification of speaker recognition. The speaker recognition encompasses speaker identification (identify unknown speakers), verification (ensure particular speaker's identity), and diarization (identify speakers from speech segments) with the text-dependent and independent criterion.

1) Identification, verification, and diarization

This level of classification is the most prominent among other classification criteria. Automatic speaker identification (SI), speaker verification (SV) and speaker diarization (SD) are often acknowledged as the most fundamental and practical approaches for bypassing illegal access to computer systems. A brief explanation of these subdomains of speaker recognition are described below:

- **Speaker identification (SI):** SI determines an anonymous speaker's identity depending on the speaker's spoken utterances. Speaker identification finds the exact speaker from a set of recognised voices of speakers. It is the way to find a person based on the different utterances contained in the database. This approach is a 1: N match where a particular utterance is compared against N templates.
- **Speaker verification (SV):** SV deals with the voice to authenticate a specific identity asserted by the speaker. The SI system's acquired characteristics are correlated with all the speakers' characteristics composed in a voice model database. In contrast, in SV systems, the acquired characteristics are only linked with the speaker's stored features he or she claimed to be. It is a 1:1 match where one speaker's speech is likened to one template.
- **Speaker diarization (SD):** SD is partitioning a voice with multiple people into homogeneous segments associated with each individual. It is an essential part of speaker recognition systems. It has applications in many critical fields, such as video captioning, understanding the content of any conversations, etc.

SI, SV, and SD system's recognition criteria can be either text-dependent or text-independent, which is discussed in the following part.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

2) Text-dependent and text-independent recognition

Text-dependency is another level of classification of speaker recognition (SR). This classification is based upon the text uttered by the speaker during the classification process. The two subdomains of speaker recognition based on text-dependency are explained below.

- **Text-dependent:** Text-dependent ASR defines a method in which the test utterance is equivalent to the text employed during the enrollment phase [8]–[13]. The test speaker has previous knowledge of the model. The local lexicon posses low enrollment and trial stages to give an authentic result. Still, it faces few scientific and technical challenges. In the 1990s, the first text-dependent speaker recognition introduced the main features of the present state of the art with feature extraction method, speaker modelling, and score normalization using a likelihood ratio score [14]. Since then, numerous architectures have inquired at various times.
- **Text-independent:** In a text-independent ASR task, the speech signal's training and testing are entirely unconstrained [15]–[19]. It usually takes long sections of the speech signal to be developed for both the training and testing phase. Here, in the testing phase, the test speaker does not have any previous knowledge about the samples of the enrollment phase.

However, text-independent speaker recognition is more convenient to text-dependent speaker recognition system (SRS) because the speaker can freely speak to the system. Although, it needs more extended training and testing utterances to obtain better accuracy. These text-dependent and text-independent approaches can be further categorised into the open set and closed set speaker recognition problems, which are discussed in the next portion.

3) Open set and closed set

ASR architectures are classified as an open set or closed set based on the number of qualified speakers available in the system. These two types of ASR methods are described below:

- **Open set:** An open set method is structured with any number of trained speakers. This system is called an open set because the anonymous speech could come from a broad set of unfamiliar speakers.
- **Closed set:** A closed set method has only a designated number of speakers enrolled on the system. In this method, the system tries to discover a speaker's individuality from a set of observed voices.

Additionally, the speech or speaker recognition techniques make a different domain of speaker recognition systems. This domain covers many more methods used to recognize and analyze the speaker's emotions or the emotions expressed in the speech. But this study is extensively focused only on speaker recognition approaches.

However, speaker identification and verification have acquired increased impact and importance in the research field as voice command, speech technology, speech analysis, etc. An eternally progressing obligation to search for voice elements and search based on speaker identity is growing wild. This tremendous visibility and significance of speaker recognition systems imply numerous research work over the years. Besides, a few literature surveys on a different portion of speaker recognition has been done as well. However, mentioning all the excellent research on ASR, a comprehensive survey has become necessary.

In this comprehensive survey, more than 28 literature survey on SR is mentioned and particularly described nine amongst them. But these surveys do not precisely address the current trends and applications, impact, and challenges of the speaker recognition systems. No recent extensive survey has been done on the entire domain of speaker recognition. The summaries and comparison of the recognized literature survey on speaker recognition are shown in Table 1.

In the previous survey, Tirumala et al. [25] carried out a systematic literature review on different significant feature extraction approaches. The paper aimed to identify substantial feature extraction methods in the last six years and provided recommendations based on the investigation. The authors answered three critical questions regarding the speaker recognition domain and investigated the foundations for optimal features with the feature extraction process. They derived feature extraction methods and architectures and discovered the most traditional and prosperous feature extraction methods in the last six years. Finally, they elaborated on a few challenges of the speaker recognition domain. In [24], the authors described a review of various methods for speaker identification using deep learning. They characterised the implementations of deep learning based on speaker identification methods and algorithms. The authors also described major deep learning architectures that achieved the state of art accuracy. They aimed to introduce detailed architectures to speaker recognition researchers to decrease the knowledge gap and enhance the significance of deep learning architectures.

Further, Dicsken et al. [26] completed a tutorial survey on speaker-specific information extraction approaches. The work's theories were categorized into three classes considering their robustness toward channel mismatch, additive noise, and other depravities such as vocal effort, emotion mismatch, etc. The authors mainly focused on the extraction of speaker-specific information in degraded conditions alongside short-time features. In [23], authors upheld a systematic review of the ASR models, particularly with those introduced in the last decade. The authors provided the reader with aspects of how humans also redact speaker recognition. Their review aimed to explain the main domains of speaker recognition, deriving the essential similarities and distinctions. The authors accentuated how spontaneous speaker recognition methods have originated over time toward more present architectures. Togneri et al. [22], represented the actual illustrations for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

TABLE 1. Eight standard literature surveys of the Speaker Recognition domain are demonstrated in this table. In addition to a brief discussion of each paper mentioned in the table, the challenges faced in recognizing speakers are given in detail.

Reference	Year	Main Purpose	Challenges
[20]	2010	A survey of ancient and modern automatic text-independent SR methods with efficiency and marks future scopes of SR to get robust techniques.	There are limited training data, mismatched handsets for training and testing, background noise and unbalanced text.
[21]	2010	An extensive survey of automatic speaker recognition systems.	Unauthorized access, privacy in biometric technologies, the possibility of cracking data encryption, designing long-range features, managing vast quantities of correlated features, odd feature range distributions, original disappeared features and heterogeneous feature types.
[22]	2011	Overview on text-independent, closed-set, speaker identification in the modelling and classification paradigms with key extracted features on both clean and missing data.	The authors only surveyed the speaker identification approaches that are based on missing data methods.
[23]	2015	A comparative study of human versus machine speaker recognition.	The search for alternative compact representations of speakers and audio segments emphasizing the identity relevant parameters while suppressing the nuisance components in system development.
[24]	2016	A comprehensive analysis of deep learning approaches and applications for speaker identification	The paper addressed knowledge in improving SID performance.
[25]	2017	Describes the foundations for optimal characteristics and how feature parameter suitability is determined for the feature extraction method in speaker identification. The feature extraction strategies and architectures are also explained.	Consolidate the capacity to trade with all channel data and noisy data, handling complex pattern recognition problems with deep learning techniques, multi-learner model, feature extraction from unlabeled data alongside incomplete, tampered or damaged data.
[26]	2017	Summarize feature extraction methods under degraded conditions, alongside the short-time features, with few normalization techniques for gaining robustness.	Performance improvement for extracting more robust features in the appearance of the noise and channel mismatch situations.
[27]	2021	A comprehensive survey of automatic speaker recognition based on deep learning.	Most speaker feature extraction techniques require handcraft acoustic features as the input, which is not always optimal. Besides, The state-of-the-art deep models have many parameters, which are difficult to apply to portable devices. This paper presents few more challenges of ASR.
Ours	2021	A systematic survey on speaker recognition where the feature extraction approaches, algorithms, limitations of subdomains of speaker recognition such as speaker identification, speaker verification, diarization is described concisely.	Data-driven dependency; Intra-speaker variability; Speaker-based variability; Conversation-based variability; Technology-based variability; Limited data and constrained lexicon; Aging of speaker models; Forward compatibility.

speaker identification and described the latter study on missing data approaches to enhance robustness. In their study, the feature extraction approaches, various speaker modelling and model classification were explained. In [21], the authors introduced a comprehensive literature survey of ASR systems. They categorized the modules of speaker recognition and demonstrated different models for each module. Besides, they gave a brief explanation of the enormous applications of SR Systems. The authors also elaborated on the issues and challenges of SR Systems. Kinnunen et al. [20], presented a brief introduction of ASR approaches with an emphasis on text-independent recognition from the 1980s until 2009. They emphasized the recent techniques introduced around 2009. Their research served as a short survey of the analytical inquiries and the explications of the speaker recognition domain. Zhongxin Bai et al. [27] reviews various significant speaker recognition subdomains such as speaker identification, verification, diarization etc., focusing on deep-learning-based approaches. Modern and newly published deep learning-based feature extraction approaches, ASR algorithms are extensively explained in this paper. Besides, a few other surveys are introduced in the speaker recognition domain at different times [28]–[47]. As these surveys did not precisely uphold the speaker recognition domain, an extensive study in this domain was necessary.

In this work, we present an elaborated survey of ASR. The review is restricted to scholarly work published between 2000 and 2021. The survey aims to discuss the findings of different related research areas such as speaker identification, speaker verification, speaker diarization, etc. Lastly, the paper addresses this field's present challenges and gives recommendations and suggestions for future research directions. The overall contributions of the survey include:

- The paper presents a systematic review of the speaker recognition systems, along with the historical backgrounds.
- The paper introduces the feature extraction procedures, architectural procedures, dataset inspections, and performance comparisons of speaker recognition architectures.
- The paper summarizes speaker recognition procedures based on the existing systems, datasets, feature extraction techniques. Further, the paper exploits the limitations of such systems.
- Finally, the survey concludes by identifying the present challenges of speaker recognition systems, with future research directions.

The rest of the paper is outlined as follows. Section II explains the survey methodology. Section III demonstrates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

TABLE 2. The table explains the inclusion and exclusion criteria maintained to select review articles.

Inclusion/Exclusion	Criteria
Inclusion	IC1: Research articles are written in English.
	IC2: Articles that have been published between 2000 to 2021 [Few old papers are used in the survey for specific purpose.]
	IC3: Papers published in Academic Journals, Conference/Workshop Proceedings, Book Chapters, and thesis dissertations.
Exclusion	EC1: Duplicate articles
	EC2: Conflicting with the theme of the review
	EC3: Lack of sufficient information

the dataset used in ASR. The feature extraction techniques are explained in Section IV. Section V describes famous ASR techniques and algorithms. Section VI investigates the papers regarding speaker recognition systems. Section VII analyzes the performance evaluation methods of ASR. Section VIII addresses the challenges of ASR with future research scopes. Finally, Section IX concludes the article.

II. SURVEY METHODOLOGY

This survey is processed through a systematic literature review (SLR) approach proposed by Kitchenham [48], [49]. In this paper, SLR steps are described in three phases: planning, conducting and reporting the review. In the following subsections, the steps are elaborated.

A. PLANNING THE REVIEW

This sub-section briefly describes the following things. The i. research question, ii. sources of review materials; and iii. inclusion and exclusion criteria.

1) **Research question:** The primary research questions were:

RQ1: How does ASR systems contribute to the field of SV, SI and SD?
RQ2: Which datasets are universally used for ASR systems?
RQ3: Which feature extraction approaches are widely used for ASR systems?
RQ4: Which techniques are broadly used for ASR systems?
RQ5: What kind of evaluation matrices are used for ASR systems evaluation?
RQ6: What are the challenges and future research possibilities on ASR systems?

2) **Sources of review materials:** The survey is restricted to excellent academic articles published via the *ScienceDirect*, *SpringerLink*, *MDPI*, *Hindawi*, *ACM Digital Library*, *IEEE Xplore*, etc., and also different famous conferences.
3) **Inclusion and exclusion criteria:** This survey's essential materials are gathered using *PRISMA* (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) described in Figure 4. Moreover, the inclusion and exclusion criteria in *PRISMA* are presented in Table 2. This table expresses the paper selection criteria and standard. On which standard the paper is selected for review or rejected.

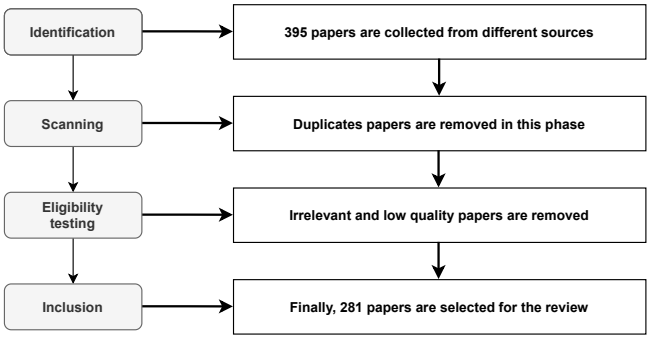

FIGURE 4. The image explains the PRISMA workflow for this literature review. Number of papers collected during the systematic literature review are mentioned in the PRISMA statement.

B. CONDUCTING THE REVIEW

This phase discusses how the necessary information is extracted from the articles. Extracting essential information and performing the literature review in a structured way, five subphases considered as described below.

1) **Topical relationship:** This phase describes how the articles are correlated with others selected for this survey. Word cloud in Figure 5 is created with the keywords of the papers and principal words from the papers' titles explaining how much the chosen articles are correlated.

FIGURE 5. Word cloud for the titles and keywords of the selected articles.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

TABLE 3. Popular datasets used in speaker recognition domain. Datasets are included for the results of speaker recognition under different environmental conditions during the training and testing phase. Different sizes and conditions of the datasets differ the effectiveness of speaker identities from voices.

Dataset Name	No. of Speakers	No. of. Utterances	Condition	Used in
TIMIT [50]	530	6300 sentences	Clean Speech	[51]–[56]
LIBRISPEECH [57]	44	1000 hours	Segmented English read speech	[58]–[61]
MIT Mobile [62]	88	7884 sentences	Mobile Devices	[63]–[66]
Switchboard [67]	500	2500 conversations	Telephony	[68]–[72]
POLYCOST [73]	133	1285 conversations	Telephony	[74]–[77]
ICSI Meeting Corpus [78]	53	75 meetings	Meetings	[79], [80]
Forensic Comparison [81]	552	1264 conversations	Telephony	[82]
RSR2015 [13]	300	151 hours	Mobile device	[83]–[85]
ANDOSL [86]	204	33900 utterances	Clean Speech	[56], [87]
RedDots [88]	45	104,000 sentences	Phonetic variation	[89]–[91]
SITW [92]	299	8 session/speaker	Multi-media	[93]–[96]
ELSDSR [97]	22	154 sentences	Clean Speech	[98]–[101]
YOHO [28]	138	1380 sentences	NA	[56], [77], [102]–[106]
CMU [107]	74	9487130 utterances	NA	[106]
VOICES [108]	300	1440 hours	Noisy room	[109]
NIST-SRE [110]	2000+	Varies year by year	Clean Speech	[77], [95], [96], [111]–[122]
CN-CELEB [123]	1000	274 hours	Unconstrained	[124]
Hi-MIA [125]	340	1561 hours	Multiple microphone	[126]
VoxCeleb1 [127]	1251	153516 utterances	Multi-media	[96], [128]–[130]

2) **Aims and outcomes:** Objective, contribution, challenges of different useful articles are presented in Section VI and VIII.
3) **Evaluation metrics:** All the evaluation measurement techniques are explained in Section IX.
4) **Research type:** It shows the paper type, such as academic journals, conference/workshop proceedings, book chapters, or thesis work.
5) **Publication year and type:** At the beginning of this work, 395 papers were collected from various resources, and 281 papers were finalized for the survey. Amongst them, more than 90% of articles are published between 2010 to 2021. We have worked with more recent articles to make this review more advanced. Figure 6 present the number of papers by year.

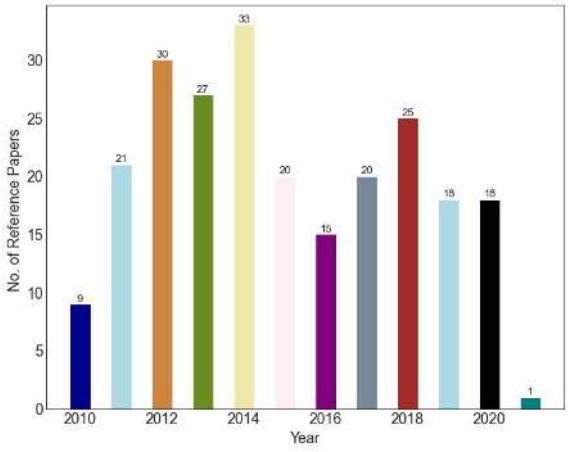

FIGURE 6. The figure illustrates the statistics of papers collected from the year 2010 to 2021.

C. OUTCOME:

Finally, the gathered information has been analyzed, addressed current issues and challenges, and provided future research opportunities.

III. DATASETS USED IN SPEAKER RECOGNITION

With the advent of voice-related applications, the essentiality of verified speaker/speech databases has radically increased in the speaker recognition domain. The shared dataset available for speaker recognition lets the researchers demonstrate, evaluate, and compare model's performance with the existing system. To create these databases, the data can be collected from various sources according to the application, such as speech recorded from acoustic laboratories [86], [131], speech recorded from mobile device [62], [132], telephone calls [73], forensic data from police, [133], etc. The mentioned datasets are acquired from single-speaker systems and have no audience noise or overlapping speech. Researchers also worked with the multi-speaker environment datasets with audience noise and overlapping speech, such as recorded meeting data [78], [134], and audio broadcast [135]. Few datasets have artificial degradation to mimic real-world noise, such as the TIMIT dataset [50]. The different audio formats like uncompressed, lossless compressed, and lossy compressed are used to record a speech. Therein, waveform (.wav) format is mostly used. This audio format is arranged into three general chunk types, the RIFF chunk, the FORMAT chunk and the DATA chunk, where the DATA chunk contains the real sample data. Selecting the proper dataset reduces the pre-processing time and makes the systems more efficient. These speaker recognition datasets can be classified into two categories:

i **Clean speech dataset:** Clean speech refers to the condition of the dataset where there is zero presence of noise in the dataset. Most of the dataset we use are clean speech dataset [50], [67], [110].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

ii **In the wild dataset:** In the wild dataset, data do not collect beneath controlled situations, and hence it carries actual noise, vibration, intra-speaker variability and squeezing artefacts. Popular in the wild datasets are SITW [92] and VoxCeleb [127].

Mostly popular datasets in speaker recognition domain are presented in Table 3.

IV. FEATURE EXTRACTION APPROACHES

A series of feature vectors can represent the speech signal to apply mathematical tools without losing generality. The goal of feature extraction is to interpret a speech signal by preset the signal's number of components. The reason is every piece of data in the acoustic signal is weighty to work, and some of the data is inappropriate [198], [199]. In practical, real-life systems, several features are used in combinations for speaker recognition tasks. Every speaker has identical guttural raw features, mainly learned/behavioural based and physiological/natural features. Below, such features are briefly explained:

i **Behaviour-based speech features:** This feature is extracted by measuring the number of parameters like experience, personality, education, familial connections, community, and communication media. High-level feature and prosodic & spectro-temporal features are two categories of learned based features. High-level characteristics comprise phones, idiolect, semantics, accent, and pronunciation. On the other side, prosodic & spectro-temporal characteristics include pitch, energy, rhythm, duration, and temporal features. Prosodic characteristics are the non-segmental appearance of speech formed in lengthy utterances such as prosodic properties accent is an associative technicality used to describe pitch variations, stress, rhythm, loudness, rhythm. Table 4 gives a chronology of learned based feature extraction methods.

ii **Physiological based speech features:** These types of features are affected by the vocal tract's length, dimension and fold size. The short-term spectral characteristics are the types of physiological features that can be measured from small speech utterances; These features are applied to explain the short-term spectral container connected to the supralaryngeal vocal tract's timbre and resonance characteristics. Voice source features are characteristics of the verbal flow. The short term spectral features are additionally categorised into two types: Spectrum and Gammatone pulse features. Short term physiological based feature extraction approaches are given in Table 5.

However, the Mel-frequency Cepstral Coefficients (MFCC) [200], Linear Predictive Coefficients (LPC) [201], and Linear Predictive Cepstral Coefficients (LPCC) [202], Perceptual Linear Prediction (PLP) [203]-based feature extraction strategies are recognized as the most effective, economical, and universally adopted feature extraction technique in the speaker recognition domain. These extensively accepted models for feature extraction are categorised into two types based on the coefficients, such as filterbank coefficients and predictive coefficients. MFCC and LFCC (Linear Frequency Cepstral Coefficients) use filterbank coefficients and LPC, LPCC use predictive coefficients for the feature extraction procedure. Besides, PLP is used in combination with cepstral and autocorrelation coefficient. Hence, in this section, a few renowned feature extraction approaches are demonstrated.

a: Mel-frequency cepstral coefficients (MFCC)

MFCCs is the extensively employed feature extraction method in speech and speaker recognition tasks. In the 1980s, the MFFCs was proposed by Davis and Mermelstein and have remained state-of-the-art from then. The standard MFCC feature extraction procedure is illustrated in Figure 7.

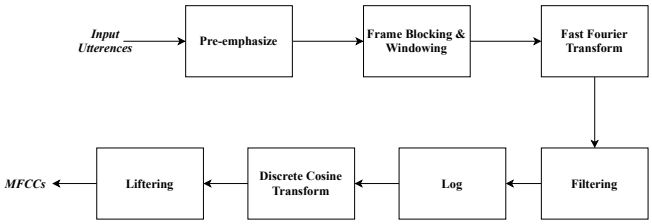

FIGURE 7. The figure represents a block diagram explaining the steps to generate MFCC features. The seven computational steps extracts parameterized representation of input signals where the critical frequency bandwidth of signals is evenly spaced in the mel-frequency cepstrum. The feature extraction affects its response with efficiency in recognition performance.

The first action is to employ a pre-emphasis filter on the input signal to expand the high frequencies. This filter is implemented to a signal x applying the first-order filter in the subsequent equation:

$$y(t) = x(t) - \alpha x(t-1) \tag{1}$$

where, α is the filter coefficient.

After pre-emphasis, framing is applied to partition the signal into low-time frames. After splitting the signal into short-time frames, a window function like the Hamming window is used for every frame.

$$w[n] = 0.54 - 0.46cos(\frac{2\pi n}{N-1}) \tag{2}$$

where, $0 \leq n \leq N-1$, N is the window length.

Then the fast fourier transform (FFT) is applied using the next formula:

$$P = \frac{|FFT(x_i)|^2}{N} \tag{3}$$

where, x_i is the i^{th} frame of signal x.

Then the mapping from linear frequency to MFCC is defined, and the following equation determines filter bank coefficients where f is physical frequency and $Mel(f)$ is the approximation of the Mel from physical frequency.

$$Mel(f) = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

TABLE 4. The table aggregates the behaviour-based feature extraction approaches in speaker recognition.

Learned-based approaches			
High level	**References**	**Prosodic**	**References**
PMFC (Phoneme Mean F-ratio Coefficient)	[136]	Sub-band Auto Correlation Classification (SACC)	[137]–[143]
PMFFCC (Phoneme Mean F-ratio Frequency Cepstrum Coefficient)	[144]	PROSACC	[145]
Polish vowel	[146]	Shifted Delta Cepstrum (SDC) and additional temporal information	[147]
Vowel phonemes	[148]	DPE to present pitch and energy by using twelve DCT coeffs	[147]
Maximum-Likelihood Linear Regression (MLLR)	[149]	Empirical Mode Decomposition (EMD) features extraction method	NA

TABLE 5. The table aggregates the physiological-based feature extraction approaches in speaker recognition.

Physiological based speech features			
Spectrum	**References**	**Gammatone pulse**	**References**
MFCC (Mel-Frequency Cepstral Coefficients)	[137], [139], [141], [142], [150]–[184]	Gammatone Feature	[185]–[187]
Linear Predictive Cepstral Coefficients (LPCC)	[138], [188]–[191]	Gammatone Frequency Cepstral Coefficients (GFCC)	[136], [162], [184], [186], [187], [192]
Linear Predictive Coefficients (LPC)	[189], [193], [194]	Hilbert Envelope of Gammatone Filterbank	[195]
Linear Predictive Residual (LPR)	[138], [196]	Perceptual Linear Prediction (PLP)	[141], [145], [197]

TABLE 6. The table demonstrates a correlation between the feature extraction methods based on four signal processing criteria.

Name	Type of Filter	Shape of Filter	Speed of Computation	Type of Coefficient	Noise Resistence	Sensitivity to Quantization
MFCC	Mel	Triangular	High	Cepstral	Medium	Medium
LPC	Linear prediction	Linear	High	Auto-correlation coefficient	High	High
LPCC	Linear prediction	Linear	Medium	Cepstral	High	High
PLP	Bark	Trapezoidal	Medium	Cepstral and Auto-correlation	Medium	Medium

$$f = 700(10^{m/2595} - 1) \tag{5}$$

Finally, the MFCC computed using the following discrete cosine transformation formula:

$$MFCC_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{6}$$

where, N is the number of bandpass filters, m_j is the log bandpass filter output amplitudes. The main drawbacks of MFCC is that the features are not precisely correct in background sound [204], [205], and it does not work better for all case [206].

b: Linear prediction coefficients (LPC)

LPC architecture is practised to get the filter coefficients coequal to the uttered speech by decrementing the mean square error (MSE) within the input signal and estimated signal [207]. LPC analyze the speech features by the prognosis of any input signal at a particular time as a linear weighted aggregation of including examples. The LPC architecture is developed based on the following equation for a given sample at time n, [208], [209]:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{7}$$

where $\hat{s}$ is the predicted speech, s is the input speech, and $a_k = 1, 2, 3, .., p$ are the predictor coefficients.

$$e(n) = s(n) - \hat{s}(n) \tag{8}$$

where, $e(n)$ is prediction error, which is calculated as in [209], [210].

LPC analyzes the vocal tract signal from a given speech [210] efficiently. It provides accurate results of speech parameters and is relatively better for computation [204], [211]. However, LPC suffers from high sensitivity to quantization noise [212].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

c: Linear prediction cepstrum coefficients (LPCC)

LPCC is more robust and reliable than LPC, and it has been broadly used for a few decades. LPCCs are equivalent to LPC when organized in the cepstrum domain. LPCC is solely the coefficients of all-pole filter and is equal to the flattened container of the log spectrum of the speech signal. The following equation is calculated the LPCC:

$$LPCC_i = LPC_i + \sum_{k=1}^{i-1} \frac{k-i}{i} LPCC_{i-k} LPC_k \quad (9)$$

LPCC features produce a lower error rate (ER) than LPC [213]. However, LPCC results are shady for having massive sensitivity to quantization noise [214].

d: Perceptual linear prediction (PLP):

PLP architect human speech depending on the notion of the psychophysics of sound [203]. PLP removes unnecessary information of the signal and therefore enhance speech recognition rate. The PLP architecture is divided into three stages: the critical-band resolution curves, the equal-loudness curve, and the intensity-loudness power-law relation. Figure 8 demonstrates the PLP architecture.

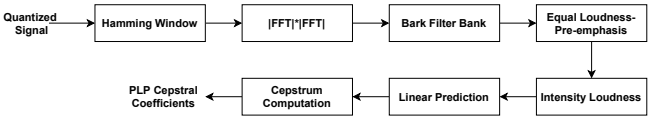

FIGURE 8. The figure represents a block diagram explaining the steps to generate PLP features. The input signal converts frequency to bark that integrates into Bark Filter Bank using hamming window and critical band analysis. An equal-loudness pre-emphasis observes sensitivity of hearing by the filter-banks. Then the equalized values are processed to linear prediction and obtain cepstral coefficients.

However, the adoption of the feature extraction strategy depends on the work and individual. Table 6 represents a comparison between the widely used feature extraction architectures described above, based on various dimensions of speech signal processing. The table will be beneficial to select between these algorithms for future work.

V. SPEAKER RECOGNITION TECHNIQUES

Based on the working criteria, speaker recognition architectures are divided into stage-wise and end-to-end architecture. A stage-wise ASR algorithm commonly compositions of a front-end for the feature extraction and a back-end for the speaker features similarity computation. On the other hand, the end-to-end speaker recognition uses speech utterances as the enrollment signal and instantly returns the similarity rate. Figure 9 represents the algorithm used in stage-wise and end to end speaker recognition modelling. This section broadly discussed the algorithms are shown in figure 9.

A. STAGEWISE SPEAKER RECOGNITION

In a stagewise speaker recognition systems, the recognition tasks such as speaker identification, speaker verification or speaker diarization are processed in two stages: front-end

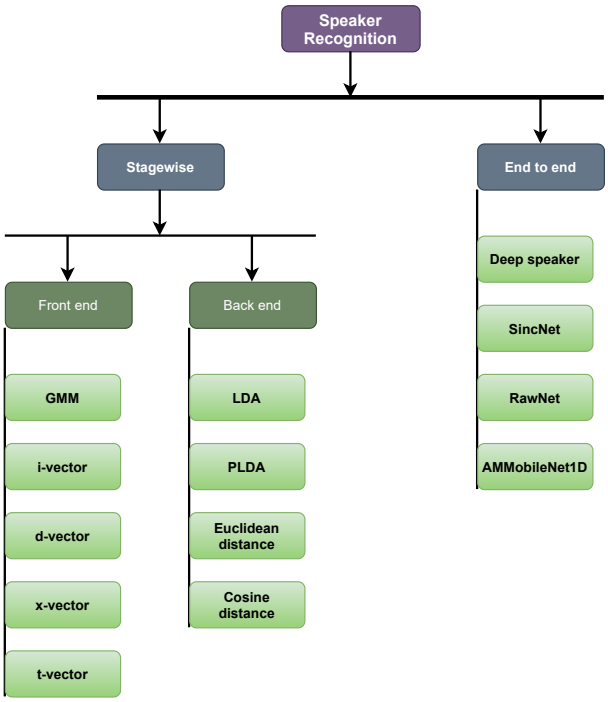

FIGURE 9. A taxonomy of the speaker recognition approaches based on the architectural constraint.

and back-end [27]. Various algorithms are employed in the front end and back end to complete the speaker recognition task. A standard stage-wise speaker recognition architecture is demonstrated in Figure 10. In the following portion, the front-end and back-end architectures are extensively investigated.

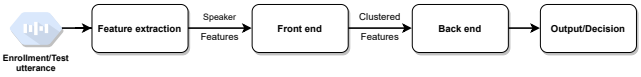

FIGURE 10. The figure demonstrates a block diagram of the stage-wise speaker recognition architecture. Front-end extracts low dimensional speaker utterances to high dimensional feature vectors. Back-end evaluated the result from input and test features similarity using threshold values.

1) Front-end

The front-end of stagewise speaker recognition architecture changes an utterance in the time-frequency domain or time domain into a high-dimensional feature vector. The widespread algorithms use in the front end nowadays are Gaussian mixture model (GMM), i-vector, x-vector, t-vector, and d-vector. These algorithms are briefly explained below:

a: Gaussian mixture model (GMM)

The GMM is a probabilistic model [215] in which datasets are assumed to be formed by a mixture of a fixed number of Gaussian distributions with uncertain variables [216]. Mixture models can be thought of as making generalizations of k-means clustering to provide details about the data's covariance structure and the centres of the undiscovered Gaussian

distributions. It is a function made up of many Gaussian distributions, each defined by $k\epsilon\{1, 2, ..., K\}$ where K is the number of clusters in the dataset. The following parameters characterize each Gaussian k in the mixture:

- μ – mean with a specified centre.
- Σ – a covariance defines its width. In a multivariate case, this will be analogous to the measurements of an ellipsoid.
- A combining probability that determines the size of the Gaussian function.

The combining factors are probabilities which must satisfy the following condition:

$$\sum_{k=1}^{K} \pi_k = 1 \tag{10}$$

To do so, we must ensure that each Gaussian matches the data sets in each cluster. The matching is precisely what maximizing probability achieves. The Gaussian density function is expressed by:

$$N(x|\mu, \Sigma) = \frac{exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))}{(2\pi)^{D/2}\,|\Sigma|^{1/2}} \tag{11}$$

where, x represents data points, D is the number of dimensions of each data point. μ and Σ are the mean and covariance, respectively. The log calculation of equation 11 found significant. The mathematical derivation can be given as:

$$lnN(x|\mu, \Sigma) = -\frac{D}{2}ln2\pi - \frac{1}{2}ln\Sigma - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \tag{12}$$

GMM is a beneficial method that is commonly used for a variety of clustering based tasks. The most common probability function for text-independent ASR using continuous features, in which there is no advance awareness of what the speaker would say, has also been Gaussian mixture models.

b: i-vector (identity vector)

i-vector [217] is applied to reduce high dimensional space to low dimensional space for speaker and channel variations by simple factor analysis. Instead of using two different spaces, the approach uses only a single space with both speaker and channel named "total variability space". The new GMM supervector is defined by $M = m + Tw$, where m is the speaker, and channel-independent supervector and w is the total factor with standard normal distributed vectors named i-vectors. For a given utterance, i-vectors used to represent the speech signal by posterior distribution. Then, the Baum-Welch statistics are extracted using UBM to estimate i-vectors by following statistics:

$$N_c = \sum_{t=1}^{L} P(c \mid y_t, \Omega) \tag{13}$$

$$F_c = \sum_{t=1}^{L} P(c \mid y_t, \Omega) y_t \tag{14}$$

$$\tilde{F}_c = \sum_{t=1}^{L} P(c \mid y_t, \Omega)(y_t - m_c) \tag{15}$$

here, $[y_1, y_2, ..., y_L]$ represents a sequence of L frames, Ω formed by C mixture components in F feature space. $P(c \mid y_t, \Omega)$ denotes posterior probability of vector y_t where $c = 1, 2, ..., C$ is Gaussian index. Then the centralized first order Baum-Welch statistics computed on the UBM mean mixture components and estimating i-vectors.

c: d-vector (deep vector)

d-vector framework [10] first preprocess data by extracting acoustic features. A DNN model is trained and acoustic feature is concatenated with context frames in hidden layer because of the inefficient information of single speakers. d-vector receives the output activation of every frame from the last hidden layer using feedforward propagation and is represented by averaging all frames' deep embedding features from an utterance. The output layer is removed to decrease the DNN model size for runtime. Finally, the decision derives by computing the distance between the target and test d-vectors. The reason to choose the last hidden layer is to observe well generalized unseen speakers. Figure 11 shows the architecture of d-vector.

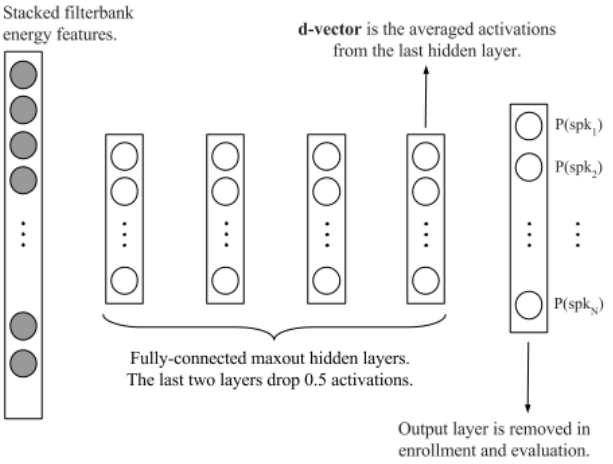

FIGURE 11. The figure illustrates d-vector architecture implemented in the front-end of a multi-stage speaker recognition system (from [10]).

d: x-vector

x-vector system is developed on DNN embeddings where neural networks trained to discriminate between speakers

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

[96]. The approach [218] follows an end-to-end system [219] that produces embeddings united with similarity metric using time-delayed DNN and compare them by a separated trained classifier like PLDA. First, short term temporal frame-level context is extracted by time delay. Then, a statistic pooling layer aggregates over the input segment and computes mean and standard deviation. The computation finally classifies the segment level feature to the speaker by DNN. The produced segment level speaker embeddings are known as x-vectors. Figure 12 represents the network structure of the x-vector architecture.

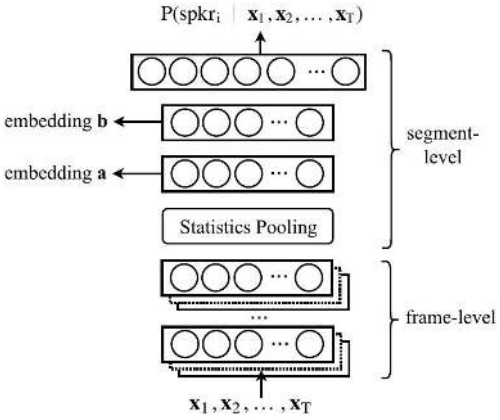

FIGURE 12. Network structure of x-vector (from [218])

.

e: t-vector

The t-vector system as triplet speaker embeddings [220] is also trained discriminatively like x-vector in speaker diarization or verification. High-resolution filter bank features are extracted to modify the system [71] to deal with the long duration data samples without overlapping. Also, the null utterance part is emitted by energy-based voice activity detection. Using an end-to-end system, segment level embedding generated by averaging the output in a sequential order to estimate the t-vector. Then the Inception-ResNet-V1 network specified in [71] is applied for discriminative speaker training. The speaker embeddings combined with Cosine Distance Scoring (CDS) and PLDA classifier can improve the performance against channel and noise variabilities. The t-vector architecture developed for the speaker recognition system is transformed from [71], with loss function changes and acoustic features. Figure 13 presents the triplet loss based speaker verification architecture which is further converted into t-vector.

2) Back-end

The back-end specifies the relationship evaluation between input and test speaker features and then conforms the result with a threshold. Then, the decision took by the threshold values, and the speaker recognition processed completed. In

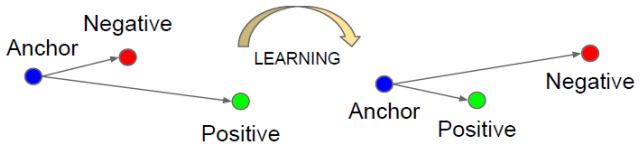

FIGURE 13. Triplet loss based speaker verification system (from [71])

.

speaker recognition tasks, the back-end aims to compensate for the channel variability and reduce interfaces. Linear discriminant analysis (LDA), probabilistic linear discriminant analysis (PLDA), cosine distance, Euclidean distance are the popular algorithms for the back-end. In the below sections, an extensive description of these back-end architectures is given.

a: Linear discriminant analysis (LDA)

LDA is a technique for reducing dimensions. As the name suggests, dimensionality reduction methods minimize the number of parameters in a dataset while preserving as much information as available. It is a standard statistical method for reducing dimensionality in classification and pattern recognition concerns [221]. As each class has a Gaussian distribution and a general covariance matrix, it considers the best optimum linear transformation. The speech class differentiation parameter in the context of A is defined by LDA as follows [23]:

$$\lambda = \frac{A^T S_b A}{A^T S_w A} \tag{16}$$

here, S_b and S_w represent between and within-class covariance matrices. A is a projection matrix that contains the k eigenvectors corresponding to the k largest eigenvalues of $S-1$. wS_b is the answer to the LDA optimal solution. Overall class disperses for feature vectors x are determined by [222]:

$$S_b = \sum_{c=1}^{C} n_c(\mu_c - \mu)(\mu_c - \mu)^T \tag{17}$$

$$S_b = \sum_{c=1}^{C} \sum_{k \epsilon c} (x_k - \mu_c)(x_k - \mu_c)^T \tag{18}$$

here, C is the number of different speaker classes, n_c is the number of various samples in class c, μ is the total mean of all samples and μ_c is the mean of samples in class c.

b: Probabilistic linear discriminant analysis (PLDA)

PLDA is a probabilistic variant of linear discriminant analysis that can accommodate more complex data sets. PLDA has a wide range of applications in several fields of study, namely computer vision, speech recognition, etc. Even for a single example of an unknown class, PLDA will generate a class centre using discrete non-linear parameters. Researchers consider different instances of a previously unknown class in statistical analysis to see if they relate to the same category—it also clusters studies from previously unseen groups. PLDA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

is a generative theory that includes given data sets are drawn from a distribution. In PLDA, the model parameters that best represent the training data must be determined. Two factors determine the representation where the data is presumed to be obtained: It should reflect various data types, and parameter processing should be easy and swift. Gaussian is the most common representation that meets these requirements. A typical Gaussian PLDA implies that an i-vector w is constructed as follows [223]:

$$w = m + V_y + z \tag{19}$$

here, m is the mean of i-vectors, y denotes the speaker latent variable with standard typical prior and the residual. Z is normally distributed with zero mean and full covariance matrix Σz. PLDA uses the expectation-maximization (EM) algorithm to estimate the model parameters $(V, \Sigma z)$.

Following parameterization, the verification score for each of the two trial i-vectors w_1 and w_2 will be calculated using the log-likelihood ratio of the hypothesis H_s, that both i-vectors are accurate from the same speaker, and the H_d assumption that both are two distinct speakers [224], that mathematically represented as:

$$verification\_score = log\frac{p(w_1, w_2)|H_s}{p(w_1, w_2)|H_d} \tag{20}$$

PLDA score can be calculated by:

$$\begin{aligned} PLDA\_score = logN(\begin{bmatrix} w1 \\ w2 \end{bmatrix}; \begin{bmatrix} m \\ m \end{bmatrix}, \begin{bmatrix} S_T & S_B \\ S_B & S_T \end{bmatrix}) \\ -logN(\begin{bmatrix} w1 \\ w2 \end{bmatrix}; \begin{bmatrix} m \\ m \end{bmatrix}, \begin{bmatrix} S_T & 0 \\ 0 & S_T \end{bmatrix}) \end{aligned} \tag{21}$$

where $S_B = W_T$ and $S_T = S_B + \Sigma z$.

c: Euclidean distance

The Euclidean distance $||x - y||_2$ between two vectors $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ can be computed as [225]:

$$||x - y||_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{22}$$

where the calculation measured in Euclidean vector space $\mathbb{R}^n$, and $n \in \mathbb{N}$.

d: Cosine distance

Cosine distance is calculated from cosine similarity [226]. Cosine similarity is applied to define the similarity between two non zero vectors. It calculates the cosine of the angle between two vectors in a multi-dimensional space. The relationship between cosine similarity and cosine distance is disproportionate. The cosine similarity raises when the distance between two vectors reduces and vice versa. The following equations calculate the cosine similarity and cosine distance, respectively. The functions can be mathematically presented as below:

$$cos(\theta) = \frac{A.B}{||A||\,.\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{23}$$

$$Cosine_{distance} = 1 - cos(\theta) \tag{24}$$

here, A, B are two non-zero vectors and $cos(\theta)$ refers to the cosine similarity.

B. END TO END SPEAKER RECOGNITION

End-to-end speaker recognition is a modern technique that receives a set of speech utterances as the input and returns their affinity score immediately. Here, both the front-end and back-end task is done by a single architecture. Few standard end-to-end speaker recognition architectures are deep speaker, raw-net, AM-MobileNet, Sinc-net etc. This subsection extensively explains these architectures. A general end-to-end speaker recognition architecture is upheld in the Figure 14. In this section, some of the standard end-to-end systems are demonstrated.

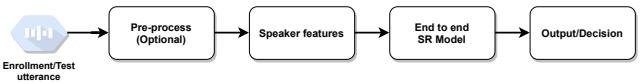

FIGURE 14. The figure demonstrates the end-to-end speaker recognition architecture.The preprocessed speaker embedding from input signal generates the output results by deep speaker modeling. Following the steps, the system produces single similarity score from the directly mapped input utterances to identify speakers.

a: Deep speaker

Deep Speaker algorithm [227] produces utterance-level speaker embedding using a deep neural network where speaker similarity is measured by cosine similarity. First, the raw audio is converted to a minibatch size of Fbank coefficients. Then, a feed-forward DNN is used to extract features over the preprocessed audio. In deep training, ResCNN architecture containing ResBlock [228] is used to remove computational complexity and help the frequency dimension immutable when the channel increases. Also, Deep speech 2 (DS2) [229] style GRU architecture extracts frame-level acoustic feature like ResCNN for faster training and less chance of divergence. After both layers, the average sentence layer, affine and length normalization layers are applied to convert the frame-level input to speaker embeddings. Finally, the triplet loss layer based on cosine similarities is conducted over an entire batch for negative selection across GPUs. This algorithm improves accuracy in both speaker verification and identification tasks. Figure 15 demonstrates the deep speaker architecture.

b: SincNet

SincNet [230] is a CNN based architecture that deals with high dimensional inputs, noisy and inconsistent multi-band shapes in the first convolutional layer. SincNet reduces the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

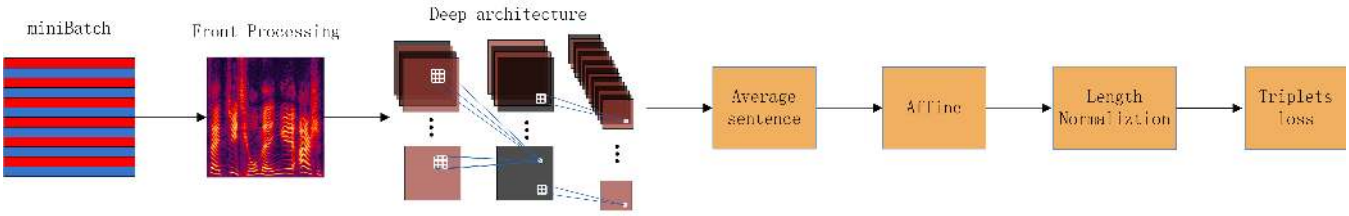

FIGURE 15. Diagram of the Deep Speaker architecture (from [227])

filter parameters and helps to converge faster by strengthening the network to focus only on the parameters. SincNet implicates the waveform by getting started a function g using a set of parameters:

$$y[n] = x[n] * g[n, \theta] \tag{25}$$

here, $y[n]$ is the filtered output, and $x[n]$ is a portion of the speech signal. The function implements rectangular band-pass filters as the difference between two low-pass filters using low f_1 and high f_2 cutoff frequencies.

$$g[n, f_1, f_2] = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \tag{26}$$

The generated filters with cutoff frequencies can directly allocate more filters where significant speaker identity clues are lying. Then, the high-frequency selection attains by Hamming window to get the ideal band-pass filters. The filters obtained by SincNet are interpretable and readable than any other techniques of speaker identification and verification.

c: RawNet

RawNet [231] is a reformed architecture of the CNN-LSTM model that enhances the pre-training scheme with additional objective functions of speaker embeddings. In the architecture, residual blocks connected to a global average pooling layer are constructed first to process Input features for frame-level embeddings extraction using leaky ReLU. A gated recurrent unit (GRU) layer then aggregates the features to utterance level embedding instead of the LSTM layer.

Center loss L_c and speaker-basis loss L_{BS} as an additional objective function employs in RawNet to reduce intra-class and enhance inter-class covariance while embeddings are discriminate.

$$L_C = \frac{1}{2} \sum_{i=1}^{N} \|x_i - c_{y_i}\|_2^2 \tag{27}$$

$$L_{BS} = \sum_{i=1}^{M} \sum_{j=1, j \neq i}^{M} cos(w_i, w_j) \tag{28}$$

DNN is trained using the final objective function besides categorical cross-entropy loss L_{ce} for feature enhancement. The fully connected layer initiates the utterance level to speaker embeddings that decrease the number of frames' parameters and increase efficiency.

d: Additive Margin MobileNet1D

Additive Margin MobileNet1D (AM-MobileNet1D) [66] directly process waveform audio on mobile devices that reduce storage size, energy-consuming and processing memory. To fit the audio signal and tackle the speaker recognition problems on mobile, the MobileNetV2 [232] architecture is modified from a 2D convolutional neural network to 1D. The modification reduces complexity and model size by faster speed. The MobileNet1D uses additive margin softmax (AM-Softmax) layer to extract features that include each linear surface of separation with the additional region. The AM-Softmax equation defined as:

$$Loss = -\frac{1}{n} \sum_{i=1}^{n} log \frac{\phi_i}{\phi_i + \sum_{j=1, j \neq y_i}^{c} exp(s(W_j^T f_i))} \tag{29}$$

$$\phi_i = exp(s(W_{y_i}^T f_i - m)) \tag{30}$$

here, W is the weight matrix, and f_i is the input and $W_{y_i}^T f_i$ is the target logit from the i_{th} sample for the last fully connected layer. It introduces two new parameters scaling factor s and additive margin size m in softmax function. The model forces the distance between two same samples to be closer and different samples to be more distant, reducing frame-level error rate seven times faster than SincNet.

VI. SUMMARY OF PAPERS REGARDING SPEAKER RECOGNITION

In the last decades, the ASR approaches gained a broad interest from many researchers that driven this system to become an effective identity authentication means. Speaker recognition is related to physiological and behavioural features of the speech utterance method of an individual voice. This section first analyses the papers of general speaker recognition systems in Table 7 and then explains the sub-domains that are presented in Table 8, 9, and 10.

Ahilan et al. [233] explored how the current factor analysis approaches perform when utterances length are reduced. The paper provided a comparative analysis of Joint Factor Analysis (JFA) and i-vector based systems, including various compensation techniques using the dataset of 2008 NIST SRE. The experiment showed that all the techniques' performance difference was very narrow at short utterance (<10s). In [234], a weighted feature extraction method is proposed to improve the effectiveness of the feature parameter. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

TABLE 7. A summary of some significant papers of speaker recognition are mentioned in the table.

Reference	Task	Method	Dataset	Feature Extract	Accuracy	Limitation
[233]	Short utterance effect investigation	Joint vector analysis(JFA) and i-vectors including various compensation techniques	NIST 2008 SRE (telephone based)	13 feature-warped MFCC with appended delta coefficients	Overall results show that as the utterance length decreases, performance degrades at an increasing rate	The analysis approaches have not provided any clear differences in performance for short speech.
[234]	Improves the effectiveness of feature parameter	Weighted feature extraction method	Voice sample of 20 speakers	Weighted LPCC	94.67% for non-sequential weighting coefficients	Hard to determine weighting coefficients.
[235]	Feature extraction in front-end ASR system	Bottleneck Neural Network approach	2010 NIST SR	Bottleneck features	Combined MFCC with Bottleneck approach optimizes a recording-level criterion	Proposed bottleneck feature extraction method slightly worse than MFCCs.
[236]	Speaker adaptation	DNN acoustic model	Switchboard English conversational telephone speech	Speaker identity vectors along with the regular ASR features	i-vectors can be used to adapt neural network models	Result may change with the increasing number of speakers' data.
[237]	Single DNN training	DNN BNF i-vector system	NIST Switchboard data	DNN bottleneck features	On DAC13: 55% reduction in EER for out-of-domain.	NA
[238]	Deep speaker feature learning	Convolutional Time Delay Deep Neural Network structure (CT-DNN)	Train: Fisher database Test: CSLT-COUGH100	Highly speaker sensitive features	Best in d-vector system with PLDA scoring	Implication of the experimental results for the acoustic and linguistic research.
[239]	Enhance end-to-end system for SR	CNN based end-to-end system	Voxceleb and NIST LRE 07	LR task: SDC features, SR task: MFCC features	CNN with LDE-ASoftmax performs best	NA
[240]	Speaker recognition under noisy and unconstrained conditions	VGGVox (CNN trunk based embedded system)	VoxCeleb1, VoxCeleb2 (train only)	DNN	VoxCeleb2 addresses the lack of VoxCeleb1 alongside achieves significant margin over previous works	Very low label error in VoxCeleb1 is detected and solved in VoxCeleb2.
[130]	Improve text-independent SR system	LM(Logistic Margin) and AM-Softmax with dropout method	VoxCeleb dataset	ResNet-20	Reduces prediction errors by up to 18%	NA
[241]	Short utterance speaker recognition	Self-adaptive GMM-MAP-UBM method	Self-recording voice	24-dimensions MFCC	Decrease the equal error rate from 4.9% to 2.5%.	Improve richness of feature extraction for large data
[242]	How the speaker recognition model extracts discriminative embeddings	CNN	Voxceleb1, TIMIT	MFCCs	The networks are better at discriminating broad phonetic classes than individual phonemes	Exploring a larger speaker dataset, a different loss function, such as the angular softmax loss, or adding an attention layer
[243]	Identify speakers for short utterances with imbalance length pairs	Meta-learning framework: ResNet34	VoxCeleb	40-dimensional log mel-filterbank (MFB) features	state-of-the-arts performance on short utterances	NA
[244]	Combine two features to enhance degraded audio signals performance	1D-Triplet-CNN	TIMIT, Fisher, Nist SRE 2008 and 2010	MFCC-LPC features	Perform in a substantial margin	Fails to verify some of sample data of each datasets.

paper analyzed each component of traditional LPCC and generated the weighted LPCC where vector quantization is used for feature matching. The experiment then evaluated that weighted LPCC has high accuracy and also learned the method non-sequentially is better than the one sequentially. Further, the study concluded that the different distribution of weighting coefficients is essential to the system's accuracy. Yaman et al. [235] presented an approach of using a bottleneck neural network to provide features in an SR system. A network re-training technique was described in this paper that optimizes a recording-level criterion with significant gains. The experiments on the same and different microphone tasks of the NIST 2010 SRE dataset showed that the bottleneck features benefit from the combined MFCC. In [236], authors presented a DNN acoustic model consisting of i-vectors and ASR features to perform speaker adaptation. The experimental results on a Switchboard, 300 hours corpus, showed that the DNN with i-vectors inputs is better than those trained on speaker-independent features only and also provide additional gains with normalized speaker features. The authors observed that i-vector extraction requires a single decoding pass but performance similar to other speaker adaptation techniques.

Richardson et al. [237] introduced the utilisation of a single DNN for speaker and language recognition and explained the construction of a DNN BNF i-vector method. The paper illustrated considerable achievement obtains when practising the DAC13 SR and LRE11 LR benchmarks technique. The evaluation showed a more significant reduction in error rate for both out-of-domain and in-domain condition using tandem features. In [238], the authors used a convolutional time-delay DNN (CT-DNN) architecture to extract speaker-sensitive features. The paper investigated that when the test utterances are short, the learned feature is highly preferential and can be used to obtain high accuracy. The experiment evaluated in the Fisher database explained that CT-DNN could generate high-quality speaker features. The study showed that the speaker trait is primarily a deterministic short-time feature rather than a long-time distributional pattern. In [239], a unified and interpretable end-to-end system for both SR and LR was developed. Besides the time average pooling layer, authors introduced self-attentive pooling (SAP) and learnable dictionary encoding (LDE) layers that aggregated the variable-length input sequence into an utterance level representation. Center loss and angular softmax loss introduced in the system to get more discriminative speaker embedding. Experimental results on Voxceleb and NIST LRE 07 datasets showed that the proposed encoding layer and loss function could significantly improve the system. Chung et al. [240] introduced VoxCeleb2, a truly large-scale audio-visual speaker recognition dataset gathered from open-source tools. The paper developed CNN models and training strategies to recognize identities from voice under noisy and unconstrained conditions effectively.

Hajibabaei et al. [130] investigated different methods to advance the efficiency of a text-independent ASR system without using additional data and more extensive criteria by expanding the training and testing data. The results showed the sufficient dimensionality of embedding space, and the usage of more discriminative loss functions increases accuracies. Also declared that the dropout method of a fully connected layer improves the verification accuracy. In [241], a hybrid model of adaptive GMM and CNN for short utterance was presented. The CNN method was trained to process spectrograms to extract the deep features of the entire frequency spectrum of short utterance. The experimental results showed the improvement of accuracy and richness of feature extraction in short utterance and speech. Shon et al. [242], proposed a Convolutional Neural Network (CNN) based ASR model for extracting robust speaker embeddings. The paper modified the embedding structure to extract frame-level speaker embeddings from each hidden layer. In [243], the authors introduced a meta-learning framework for imbalance length pairs which solved the poor performance of models with short utterances. Further, the authors proposed prototypical networks with different speech length of two meta-learning schemes as support and query sets. The experimental results of combined two schemes on short utterance (1-2 seconds) outperformed speaker verification models and unseen speaker identification on the VoxCeleb datasets. 1D-Triplet CNN was presented in [244] with a combination of MFCC and LPC features to improve the quality of the input speech signal. Experiments on the TIMIT dataset, Fisher dataset, NIST SRE 2008 and 2010 datasets observed robust performance to a wide range of audio degradations hence still failed to verify all the data correctly.

Besides, a few other speaker recognition models are also introduced in this domain at different times [227], [249]–[263]. In paper [20], authors presented an overview of Classical approaches like vector quantization, Gaussian mixture model, support vector machine (SVM), and Supervector methods of automatic text-independent SR. The authors also elaborated the normalization and adaption methods to handle mismatch of training and testing, unbalanced text, limited training data, background noise, and non-cooperative users. The study showed the NIST database's performance and mentioned recent methodological difficulties such as text dependency, channel impacts, speech durations, and cross-talk speech.

Singh et al. [264] provided information about three specific application areas as authentication, surveillance and forensic speaker recognition technologies. The authors discussed irrelevant information in automatic speaker recognition applications that may degrade the system accuracy. Also added relevant information of applications such as linguistic information. In [265], the MFCC technique for feature extraction and vector quantization for feature matching were used for designing a speaker recognition system. The paper suggested some modifications of existing MFCC that can be used to improve the performance of SR. Ferrer et al. [266] presented SRI's submission along with an analysis of the approaches that provided significant gains for the evaluation. The paper

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

TABLE 8. Some papers of speaker identification have been summarized in this table.

Reference	Task	Method	Dataset	Feature Extract	Accuracy	Limitation
[245]	Identify unknown speaker	BPNN based approach	text-dependent dataset spoken by 5 different female speakers	MFCC	85% at the filter number 32	NA
[246]	Personal authentication based on neural network	ASR model of MFCC and ANN classifier	Sample of 10 phrases for each 50 users and the phrase	16 MFCC features	92%	Text-dependent phrase only.
[247]	Performance analysis in noise presence or degradation for identification.	SVM (support vector machine) technique	Sample of Arabic clause iterates 15 times each of 80 speakers with degradation	MKMFCC (Multiple Kernel Weighted MFCC)	Higher identification rate	Result differ in other languages.
[248]	Low complexity solution for short utterance speaker identification	Gaussian mixture model	TIMIT database	combined MFCC and CMN feature vectors	Overall performance reduce both time and complexity	NA

included a multiclass voice-activity detection (VAD) system and the fusion of subsystems trained with clean and noisy data, a fusion strategy for acoustic characterization using metadata. In [267], a new MFCC and VAD based approach was presented for the SR system where VAD removes the background noise. Also, a new criterion for voice detection was proposed. Experimental results were presented on the dataset of three speakers' sample voices and showed 90% accuracy with no false recognition. In [183] the authors designed speaker identification experiments to analyze noise robustness of MFCC and GFCC. This study revealed that the nonlinear rectification accounts for the noise robustness differences primarily, suggested how to enhance MFCC robustness, and further improved GFCC robustness by adopting a different time-frequency representation. The paper [268] applied a CNN/i-vector approach to identify speakers in noisy conditions for Automatic speech recognition (ASR). Datasets used for the experiments was supplied under the DARPA RATS program with an additional two latest data collection. The performance evaluated that CNN/i-vector front end is better than the UBM/i-vector on heavily degraded speech data fusion with UBM/i-vector system using different features considerably outperforms with 26% in miss rate. The paper also highlighted a future scope of the CNN/i-vector approach by analyzing the language and channel dependency.

As speaker recognition has different applications, the following subsections summarize the papers of speaker identification, speaker verification and speaker diarization that are presented in Table 8, 9, and 10.

A. SUMMARY OF PAPERS REGARDING SPEAKER IDENTIFICATION

In this section, we represent the speaker identification systems evolution from the early implementations to current trends. The summary of the papers are uphold in table 8.

In the paper [245], a BPNN based system was presented for speaker identification where the MFCC feature was used with some modifications. The paper investigated the extracted features to identify the unknown speaker using five unknown female speakers dataset, and the results showed that at 32 number of filters, the efficiency is 85%. The paper used neural networks for training and testing to find the error values that defined the clarity of the algorithm. In [246], an automatic text-dependent speaker recognition model was presented. The paper used MFCC features to get hybrid features and then a multilayer feed-forward backpropagation ANN classifier used to recognize speakers. The experimental results evaluated on a fixed spoken phrase, and the CIR performance is 92% but only for ten users. The paper [247] presented a robust noise MKMFCC–SVM based on the Multiple Kernel Weighted Mel Frequency Cepstral Coefficient (MKMFCC) and SVM for SI. The paper used MKMFCC to extract features from degraded audio using multiple kernels like exponential and tangential and categorized with the SVM approach. The comparative analysis with MFCC-SVM SI in different transforms and SNR on a database includes 80 speakers (each iterating Arabic clauses 15 times) provided a higher identification rate in noise presence or degradation. The paper [248] proposed GMM based on a low complexity solution with new feature vectors in a text-independent speaker recognition system for short utterances. The paper used Cepstral Mean Normalization (CMN) to reduce the effect of the extracted features' variability and combined with MFCC to enhance the performance. The experimental results on the TIMIT database demonstrated remarkable results of the system's effectiveness without incorporating lengthy, extra-data and complicated calculations to handle short utterances data.

TABLE 9. Some papers of speaker verification have been summarized in this table.

Reference	Task	Method	Dataset	Feature Extract	Accuracy	Limitation
[269]	Speaker verification with short speech utterance	Probabilistic linear discriminant analysis (PLDA) approach	NIST 2008	MFCC	Heavy-tailed PLDA achieve better performance than Gaussian PLDA	NA
[270]	Improve speaker embedding quality in short utterances text independent speaker verification	CNN based end-to end SV framework and an i-vector SV framework with deep discriminant analysis compensation	NIST SRE corpus	I-vector system based MFCCs, A-softmax loss based fbank features	24.4% and 13.9% respectively outperform by i-vector/PLDA framework	Proposed deep discriminant analysis compensation isn't compatible with PLDA.
[271]	Generate likelihood ratios in an end-to-end speaker verification system	Two Tied Factor Deep Neural Network (TF2-DNN) model and autoencoder model	RSR 2015 database I	MFCCs	Competitive result respect to DNN	NA
[272]	Extract self attentive speaker embeddings	DNN with PLDA and extension of the x-vector architecture	NIST SRE 2016	MFCCs	consistent with both short and long testing utterances.	Small dataset lead mismatch during train and test segmentation.
[218]	DNN embedding and fusion with i-vectors investigation for text independent speaker verification	A feed-forward DNN embedding system	NIST SRE 2010 and 2016	MFCCs	Embeddings are better on short utterances and fusing are complementary	More appropriate symmetric metric can be used.

B. SUMMARY OF PAPERS REGARDING SPEAKER VERIFICATION

This section presents a concise overview of the speaker verification domain by elaborating its modelling methodologies and various feature extraction approaches.The summary of the papers are presented in table 9.

Kanagsundaram et al. [269] investigated the effect of short utterances available for development, enrolment, and verification using the PLDA approaches. Applying the NIST 2008 SR evaluation database, the study showed that heavy-tailed PLDA (HTPLDA) deliver greater accuracy than Gaussian PLDA (GPLDA) as evaluation sentence dimensions are minimised. The paper presented mismatched and matched evaluation performance by pooling the advancement data, preferably concatenating two independently trained total-variability spaces from every channel. In [270], the application of angular softmax is presented to increase speaker embedding performance. In the paper, a CNN based end-to-end SV framework and a deep discriminant analysis for compensation in i-vector space are two a-softmax loss-based SV frameworks investigated. An end-to-end text-dependent speaker recognition model was proposed in the paper [271] based on deep neural networks using tied hidden factors. A two-step backpropagation algorithm used to train model parameters and hidden variables in the paper. The authors calculated the gradients of tied hidden variables by aggregating all session frames for JFA. For effectiveness against overfitting and uncertainty, the dropout Bernoulli distribution and likelihood ratios used in enrolment and trial evaluation. The evaluation on the RSR2015 part I database provided well-calibrated scores in results. A self-attention mechanism into DNN embeddings was proposed in [272] to extract speaker embeddings for text-independent SV. The speaker embeddings used as a weighted mean of a speaker's frame-level hidden vectors in the paper. A PLDA classifier was used to compare combinations of embeddings. The experimental results on NIST SRE16 showed that the improvement by the mechanism was compatible with both short and long testing utterances. The paper [218] investigated deep neural network embeddings that had been replaced with i-vectors for text-independent speaker verification. The experiments on NIST SRE 2010 and 2016 showed that the feed-forward DNN embedding system outperformed i-vectors for short speech segments and are ambitious on long-duration test circumstances. The paper also evaluated equivalent performance using a fusion of i-vector and embeddings.

C. SUMMARY OF PAPERS REGARDING SPEAKER DIARIZATION

This section demonstrates a few speaker diarization papers. The summary of the articles is upheld in table 10. In [273], Yusuke et al. introduced an end-to-end neural-network-based speaker diarization technique. This technique does not have individual sections for feature extraction and clustering. In-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

TABLE 10. Some papers of speaker diarization have been summarized in this table.

Reference	Task	Method	Dataset	Feature Extract	Accuracy	Limitation
[273]	End-to-end neural-network based SD method	Neural probabilistic model of SD	Switchboard 2, Switchboard cellular, NIST SRE	23-dimensional log-Mel-filterbank features	The proposed method achieved diarization error rate (DER) of 12.28%	Better simulation techniques can improve the error rate
[274]	Overlap detection system for improved meeting diarization	Integrated overlap detector and diarization system	AMI Meeting Corpus	MFCC	Relative improvement of about 7.4% DER	NA
[275]	Improving the short-term spectral feature based overlap detector	HMM/GMM overlap detector	AMI meeting corpus	Long-term conversational features	DER is reduced by 5%	Exploring prior probabilities obtained from the conversational features
[276]	Discriminatively trained DNN for SD	DNN	CALLHOME conversational telephone speech corpus	Feed forward DNN	DER 9.9	Does not respond as well to current unsupervised calibration strategy
[277]	Speaker diarization system	LSTM-based d-vector	NIST SRE 2000 CALLHOME	LSTM	12.0% diarization error rate	NA
[278]	Supervised speaker diarization	Unbounded interleaved-state recurrent neural networks (UIS-RNN)	NIST SRE 2000 CALLHOME	RNN	7.6% diarization error rate	NA
[279]	Speaker diarization via unsupervised PLDA i-vector	Incorporates probabilistic linear discriminant analysis (PLDA) for i-vector scoring	CALLHOME conversational telephone speech corpus	i-vector	DER 13.7%	Error rate can be reduce further

stead, a single neural network outputs SD results directly. The authors developed a permutation-free objective function to depreciate diarization errors instantly. This SD architecture obtains a diarization error rate (DER) of 12.28%. In [274], the authors developed an overlapped speech detector in the SD. Kefi et al. propose an overlap detector and diarization system that trained on AMI meeting corpus dataset and achieved relative-improvement of about 7.4% diarization error rate. The paper [275] represents an architecture to enhance the short-term spectral feature-based overlap detector by incorporating information from long-term conversational features. HMM/GMM based technique applied on AMI meeting corpus finds that the diarization error rate is minimized by 5% pertinent to the baseline overlap detector. In [276], the authors introduced a substitute technique for learning feature affirmation by deep neural networks to dispel the i-vector model from this speaker diarization system. Wang et al. [277] proposed a novel d-vector based model for SD. The authors merged LSTM-based d-vector audio embeddings with modern nonparametric clustering to significantly result in the speaker diarization system. This architecture experimented on NIST SRE 2000 CALLHOME dataset and found a significant result of 12% DER. Zhang et al. [278] introduced a diarization approach fully supervised denoted unbounded interleaved-state recurrent neural networks (UIS-RNN). This unsupervised method was evaluated on NIST SRE 2000 CALLHOME dataset and obtained a 7.6% diarization error rate. In [279], the authors proposed a supervised diarization system i-vector and PLDA. The model evaluated on CALLHOME conversational telephone speech corpus and obtained a diarization error rate of 13.7%. The paper [280] describes an unsupervised speaker diarization method that creates massive progress on unsupervised SD.

VII. PERFORMANCE OF SPEAKER RECOGNITION

The SR system's performance can be evaluated depending on parameters such as accuracy and the system's speed. The accuracy of architecture can be calculated from the false acceptance ratio (FAR) and false rejection ratio (FRR). FAR is the percentage of negative inputs which are deter-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

mined as positive. FRR is the percentage of positive inputs, which are destined as negative. Some widely used SR performance measures systems are receiver operating characteristic (ROC), equal error rate (EER), detection error tradeoff (DET). The following sections broadly describe the evaluation systems.

A. RECEIVER OPERATING CHARACTERISTIC (ROC)

The area under the curve (AUC) and receiver operating characteristic (ROC) curve is a performance measurement for recognition problems at numerous thresholds. The receiver operating characteristic is a probability curve, and the area under the curve demonstrates the degree or measure of separability. It represents how much the architecture is efficient to separate classes. If the threshold value is minimized, it classifies more items as positive, which increases both true positive and false positive. Figure 16 demonstrates a standard ROC-AUC curve. To better explain the ROC-AUC curve, the true positive and false positive rates are discussed below:

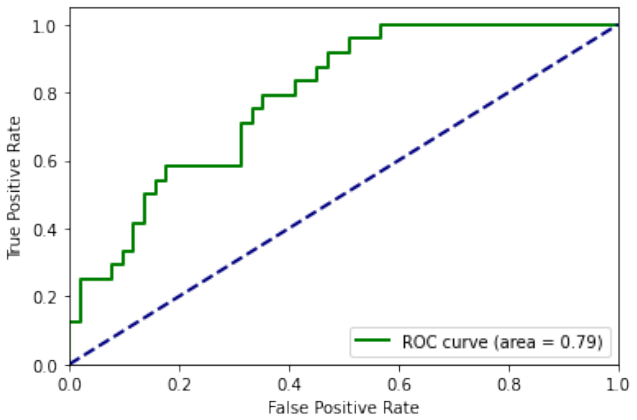

FIGURE 16. The figure illustrates TPR vs FPR at different classification thresholds, which is the condition of the ROC-AUC curve.

True positive rate (TPR): TPR means true positive rate or sensitivity that tells us what proportion of the positive items got correctly classified. The mathematical formula of TPR is:

$$TPR = \frac{TP}{TP + FN} \tag{31}$$

here, TP, FN refers to true positive and false negative respectively.

False positive rate (FPR): FPR means the false positive rate that tells us what proportion of the negative items got incorrectly classified. The mathematical formula of FPR is:

$$FPR = \frac{FP}{TN + FP} \tag{32}$$

where, FP, TN refers to false positive and true negative respectively.

B. EQUAL ERROR RATE

The equal error rate (ERR) is the algorithm that is used to accurately predict the thresholds for its false acceptance and false rejection rates. The standard value is referred to as the equal error rate whenever the rates are similar. The value demonstrates that the percentage of false acceptances equals the portion of false rejections. Fundamentally, it is a mathematical method of detecting errors and error margins in inaccurate results. There are also two possible conditions: Y, which means that the expression is acknowledged as the speaker's, and N, meaning that the expression is refused. From these expressions, the conditional probability can be summarized as [28]:

$P(Y|y)$ is the probability of correct acceptance,
$P(Y|n)$ the probability of false acceptance (FA),
$P(N|y)$ the probability of false rejection (FR),
$P(N|n)$ the probability of correct rejection.

The relationships between these parameters: $P(Y|y) + P(N|y) = 1$ and $P(Y|n) + P(N|n) = 1$. SR systems can be measured using the two probabilities $P(Y|y)$ and $P(Y|n)$.

C. DETECTION ERROR TRADEOFF

A detection error tradeoff (DET) is a visual representation of error margins for binary classifications that plots the false rejection rate versus the false acceptance rate [281]. It is now common practice to map the error curve on a regular deviate scale, in which situation the curve is referred to as the DET curve. The x and y-axes in the curve are measured non-linearly by the regular natural deviates, resulting in more linear tradeoff curves than ROC curves. It uses most sequential values to emphasize the significance of the difference in the critical operating zone. The DET curve representation is simple to understand and compares the system's output over a wide range of operating situations. The probit function (used to model binary outcome variables) gives the normal deviate mapping, such that:

$x = probit(P_{fa})$
$y = probit(P_{fr})$

where P_{fa} and P_{fa} are the false-accept and false-reject rates. Probabilities are mapped from the unit interval $[0, 1]$ to the extended real line $[-\infty, +\infty]$.

VIII. CHALLENGES AND FUTURE SCOPES OF ASR SYSTEMS

Speaker recognition approaches face numerous challenges. In data-driven strategies, intra-speaker variability is the most common challenge. These challenges are common for both text-dependent and text-independent speaker recognition tasks. This section first analyzed speaker recognition's general challenges and then explained the technology and deployment challenges.

a: General challenges

The difficulties of speaker recognition tasks for both text-dependent and text-independent models are:

- **Data-driven dependency:** Even though the speaker recognition approaches are practical, these strategies are incredibly data-driven. A massive quantity of knowledge is required to train the background methods. The

database got to be structured and arranged in a very controlled way requiring notable human efforts.

- **Intra-speaker variability:** Sometimes, the same person does not speak the exact speeches in the same style every interval, which causes a significant degree of variability. We have discussed three types of variability that create excessive challenges for speaker recognition tasks.
- **Speaker-based variability:** It displays differences in how a speaker speaks and will affect system accuracy for ASR. Those can be regarded as the inherent variability of the speaker, including the following determinants: vocal style, emotion, phycological, etc.
- **Conversation-based variability:** It reflects different scenarios concerning the vocal communication with either a different person or any system or disputes concerning the particular language or accent. It incorporates the human-to-human conversation of the dialect spoken, monologue, two-way conversation, etc.
- **Technology-based variability:** It involves the time and place of the audio capturing and subsequent issues: electromechanical, environmental, data quality—duration, sampling rate, recording quality, and audio compression.
- **Low resource languages:** State-of-the-art ASR approaches give satisfactory results on the dataset described in Section III. Still, when the language of the low resources is applied to these approaches, the result becomes more down than average. So, there is quite a notable quantity of effort demanded in the field of low-resource language ASR systems.

b: Technology challenges

The technology challenges of speaker recognition architectures are intensely correlated to the core algorithms. Few technological challenges are:

- **Limited data and constrained lexicon:** Modern industrial purposes apply training periods that generally composed of repetitious duplications of the enrollment lexicon. The trial period falls from a unique replication of a sub-portion of the recorded lexicon for an entire speech input of 4-5 seconds. Certain obligations are constrained by studies that illustrate, end customers best observe shorter enrollment and testing sessions. In the most prominent circumstances, the testing lexicon is preferred to match the enrollment lexicon precisely.
- **Channel usage:** This is not surprising to observe that top clients in real applications use numerous types of phones: landline phones, payphones, cordless phones, cell phones, etc. This advances the effect of their influence on the efficiency of channel usage. A cross-channel endeavour is designated as a measuring interval introducing from a separate channel than throughout the training period. It is a vital region where the speaker recognition architectures must be improved.
- **Aging of speaker models:** Several speaker models ageing sources exist natural ageing, channel usage, and behavioural changes. Biological ageing is associated with the physiological developments that happen to the phonatory device over extended times. Channel usage shifts across time can let the speaker model enhance antiquated concerning the contemporary channel usage. Subsequently, behavioural alterations happen when users acquire extra vulnerability to the voice interface and reconstruct how they communicate. In sum, these constituents influence the models and score and therefore are reflected in the efficiency.

c: Deployment challenges

The deployment challenges of speaker recognition approaches are encountered when transferring the speaker recognition architecture into an actual application include:

- **Cost of deployment:** For the speaker recognition task, a broad range of dialogues can be implemented. These dialogues need their individual collection of thresholds based on the significant stage of protection. Lately, an attempt to develop off-the-shelf security settings into results has occurred. This method does not need any information and is quite reliable for small- to medium-scale methods or basic security settings for an experiment. However, maximum developers desire to have a more precise understanding of their protection layer's accuracy. They crave to estimate original data of the conventional false accept, false reject, and prompt rates. Thus the cost of the deployment becomes high.
- **Forward compatibility:** Here, the principal purpose is that the enrollee database should be forward congruous with revising the application and its underlying architectures. Indeed, an application produced using a security layer based on a first name and last name lexicon is restricted to using this lexicon. The explanation of these features is an indispensable part of the speaker model, and any modification to this will harm efficiency. It also regulates what research can contribute to an existent algorithm.

Besides, challenging audio datasets, different phonation styles, speech under stress in the record, channel mismatch, and speech modality make today's speaker recognition approaches challenging. However, solving these challenges involves the huge advanced research possibilities of ASR.

IX. CONCLUSION

Speaker recognition is an eminent research domain widely investigated and integrated into numerous systems to identify or verify individuals. However, limited investigation has been conducted in the vast research domain, and most of them are currently outdated. Therefore, the paper focuses on the renowned research area and explores various dimensions of such an exciting research field. The article targets the research domain in multiple aspects, including the fundamentals, feature extraction methodologies, datasets, architectural

constructions, performance measurements, and challenges. Primitively, the study explains the fundamental theories and existing review work of the ASR domain. The following describes the research methodologies and the datasets used in this domain. Then the paper introduces feature extraction approaches of ASR systems extensively. Further, the paper aggregates the current implementations of the ASR systems, explains and provides a comparative review based on the aspects mentioned earlier. Later the performance measurement methods of ASR systems present. Finally, the paper exploits some of the limitations of the current ASR systems and addresses future directives. The article significantly upholds the existing defaults and variations of ASR systems that would help new researchers quickly adapt the research domain concepts. Moreover, the comparisons and future directives would help explore a broader perception of speaker recognition technologies, including architectural and feature extraction terminologies.

REFERENCES

[1] N. Singh, A. Agrawal, and R. Khan, "The development of speaker recognition technology," *International Journal of Advanced Research in Engineering & Technology (IJARET)*, vol. 9, no. 3, pp. 8–16, 2018.

[2] 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.

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

[4] H. Jayanna and S. M. Prasanna, "Analysis, feature extraction, modeling and testing techniques for speaker recognition," *IETE Technical Review*, vol. 26, no. 3, pp. 181–190, 2009.

[5] N. Singh, R. Khan, and R. Shree, "Mfcc and prosodic feature extraction techniques: a comparative study," *International Journal of Computer Applications*, vol. 54, no. 1, 2012.

[6] M. R. Hasan, M. Jamil, M. Rahman *et al.*, "Speaker identification using mel frequency cepstral coefficients," *variations*, vol. 1, no. 4, 2004.

[7] F. K. Soong, A. E. Rosenberg, B.-H. Juang, and L. R. Rabiner, "Report: A vector quantization approach to speaker recognition," *AT&T technical journal*, vol. 66, no. 2, pp. 14–26, 1987.

[8] M. Hébert, "Text-dependent speaker recognition," in *Springer handbook of speech processing*. Springer, 2008, pp. 743–762.

[9] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent speaker recognition using plda with uncertainty propagation," *matrix*, vol. 500, no. 1, 2013.

[10] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[11] S. Hayakawa and F. Itakura, "Text-dependent speaker recognition using the information in the higher frequency band," in *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. IEEE, 1994, pp. I–137.

[12] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[13] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[14] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital signal processing*, vol. 1, no. 2, pp. 89–106, 1991.

[15] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[16] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[17] 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 text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, pp. 1–22, 2004.

[18] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[19] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE signal processing magazine*, vol. 11, no. 4, pp. 18–32, 1994.

[20] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[21] Z. Saquib, N. Salam, R. P. Nair, N. Pandey, and A. Joshi, "A survey on automatic speaker recognition systems," *Signal Processing and Multimedia*, pp. 134–145, 2010.

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

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

[24] S. S. Tirumala and S. R. Shahamiri, "A review on deep learning approaches in speaker identification," in *Proceedings of the 8th international conference on signal processing systems*, 2016, pp. 142–147.

[25] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Systems with Applications*, vol. 90, pp. 250–271, 2017.

[26] G. Dişken, Z. Tüfekçi, L. Saribulut, and U. Çevik, "A review on feature extraction for speaker recognition under degraded conditions," *IETE Technical Review*, vol. 34, no. 3, pp. 321–332, 2017.

[27] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, 2021.

[28] R. V. Pawar, R. M. Jalnekar, and J. S. Chitode, "Review of various stages in speaker recognition system, performance measures and recognition toolkits," *Analog Integrated Circuits and Signal Processing*, vol. 94, no. 2, pp. 247–257, 2018.

[29] J. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part ii: Review of speaker recognition techniques and study on the effects of different variability factors," *Biomedical Signal Processing and Control*, vol. 48, pp. 128–143, 2019.

[30] M. Hamidi, H. Satori, N. Laaidi, and K. Satori, "Conception of speaker recognition methods: A review," in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. IEEE, 2020, pp. 1–6.

[31] D. Sztahó, G. Szaszák, and A. Beke, "Deep learning methods in speaker recognition: a review," *arXiv preprint arXiv:1911.06615*, 2019.

[32] A. Irum and A. Salman, "Speaker verification using deep neural networks: A," *International Journal of Machine Learning and Computing*, vol. 9, no. 1, 2019.

[33] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, and T. Kinnunen, "Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction," *Speech Communication*, vol. 99, pp. 62–79, 2018.

[34] A. P. Singh, R. Nath, and S. Kumar, "A survey: Speech recognition approaches and techniques," in *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE, 2018, pp. 1–4.

[35] S. A. Imam, P. Bansal, and V. Singh, "Speaker recognition using automated systems," *AGU International Journal of Engineering and Technology (AGUIJET)*, vol. 5, pp. 31–39, 2017.

[36] 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.

[37] M. Manjutha, J. Gracy, P. Subashini, and M. Krishnaveni, "Automated speech recognition system–a literature review," *COMPUTATIONAL METHODS, COMMUNICATION TECHNIQUES AND INFORMATICS*, p. 205, 2017.

[38] M. Farrús, "Voice disguise in automatic speaker recognition," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–22, 2018.

[39] G. Chaudhary, S. Srivastava, and S. Bhardwaj, "Feature extraction methods for speaker recognition: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 12, p. 1750041, 2017.

[40] H. Bouraoui, C. Jerad, A. Chattopadhyay, and N. B. Hadj-Alouane, "Hardware architectures for embedded speaker recognition applications: a survey," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 16, no. 3, pp. 1–28, 2017.

[41] N. G. Desai and N. V. Tahilramani, "Speaker recognition system using watermark technology for anti-spoofing attack: A review," *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, vol. 4, no. 4, pp. 152–156, 2016.

[42] S. Dhonde and S. Jagade, "Feature extraction techniques in speaker recognition: A review," *International Journal on Recent Technologies in Mechanical and Electrical Engineering (IJRMEE)*, vol. 2, no. 5, pp. 104–106, 2015.

[43] T. F. Zheng, Q. Jin, L. Li, J. Wang, and F. Bie, "An overview of robustness related issues in speaker recognition," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–10.

[44] V. Sharma and P. Bansal, "A review on speaker recognition approaches and challenges," *International Journal of Engineering Research and Technology (IJERT)*, vol. 2, no. 5, pp. 1581–1588, 2013.

[45] A. Lawson, P. Vabishchevich, M. Huggins, P. Ardis, B. Battles, and A. Stauffer, "Survey and evaluation of acoustic features for speaker recognition," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5444–5447.

[46] S. Furui, "40 years of progress in automatic speaker recognition," in *International Conference on Biometrics*. Springer, 2009, pp. 1050–1059.

[47] P. Premakanthan and W. Mikhael, "Speaker verification/recognition and the importance of selective feature extraction," in *Proceedings of the 44th IEEE 2001 Midwest Symposium on Circuits and Systems. MWSCAS 2001 (Cat. No. 01CH37257)*, vol. 1. IEEE, 2001, pp. 57–61.

[48] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[49] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.

[50] W. M. Fisher, "Ther darpa speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition, Feb. 1986*, 1986, pp. 93–99.

[51] A. Roy, M. M. Doss, and S. Marcel, "A fast parts-based approach to speaker verification using boosted slice classifiers," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 241–254, 2011.

[52] A. Venturini, L. Zao, and R. Coelho, "On speech features fusion, α-integration gaussian modeling and multi-style training for noise robust speaker classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1951–1964, 2014.

[53] S. Hyon, J. Dang, H. Feng, H. Wang, and K. Honda, "Detection of speaker individual information using a phoneme effect suppression method," *Speech Communication*, vol. 57, pp. 87–100, 2014.

[54] P. K. Ajmera and R. S. Holambe, "Fractional fourier transform based features for speaker recognition using support vector machine," *Computers & Electrical Engineering*, vol. 39, no. 2, pp. 550–557, 2013.

[55] S. M. Govindan, P. Duraisamy, and X. Yuan, "Adaptive wavelet shrinkage for noise robust speaker recognition," *Digital Signal Processing*, vol. 33, pp. 180–190, 2014.

[56] B. R. Wildermoth and K. K. Paliwal, "Gmm based speaker recognition on readily available databases," in *Microelectronic Engineering Research Conference, Brisbane, Australia*, vol. 7, 2003, p. 55.

[57] 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)*. IEEE, 2015, pp. 5206–5210.

[58] A. C. Kocabiyikoglu, L. Besacier, and O. Kraif, "Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation," *arXiv preprint arXiv:1802.03142*, 2018.

[59] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[60] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "Rwth asr systems for librispeech: Hybrid vs attention–w/o data augmentation," *arXiv preprint arXiv:1905.03072*, 2019.

[61] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6224–6228.

[62] R. H. Woo, A. Park, and T. J. Hazen, "The mit mobile device speaker verification corpus: data collection and preliminary experiments," in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*. IEEE, 2006, pp. 1–6.

[63] T. J. Hazen, E. Weinstein, and A. Park, "Towards robust person recognition on handheld devices using face and speaker identification technologies," in *Proceedings of the 5th international conference on Multimodal interfaces*, 2003, pp. 289–292.

[64] T. I. Ren, G. D. Cavalcanti, D. Gabriel, and H. N. Pinheiro, "A hybrid gmm speaker verification system for mobile devices in variable environments," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2012, pp. 451–458.

[65] F. Curelaru, "Evaluation of the generative and discriminative text-independent speaker verification approaches on handheld devices," in *2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2015, pp. 1–10.

[66] J. A. C. Nunes, D. Macêdo, and C. Zanchettin, "Am-mobilenet1d: A portable model for speaker recognition," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[67] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[68] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, "Improving speaker recognition performance in the domain adaptation challenge using deep neural networks," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 378–383.

[69] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "The ibm 2016 speaker recognition system," *arXiv preprint arXiv:1602.07291*, 2016.

[70] S. Shon, S. Mun, W. Kim, and H. Ko, "Autoencoder based domain adaptation for speaker recognition under insufficient channel information," *arXiv preprint arXiv:1708.01227*, 2017.

[71] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1633–1644, 2018.

[72] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4889–4893.

[73] D. Petrovska, J. Hennebert, H. Melin, and D. Genoud, "Polycost: a telephone-speech database for speaker recognition," *RLA2C, Avignon, France*, pp. 211–214, 1998.

[74] M. Katz, S. E. Krüger, M. Schafföner, E. Andelic, and A. Wendemuth, "Speaker identification and verification using support vector machines and sparse kernel logistic regression," in *International Workshop on Intelligent Computing in Pattern Analysis and Synthesis*. Springer, 2006, pp. 176–184.

[75] G. Sarkar and G. Saha, "Real time implementation of speaker identification system with frame picking algorithm," *Procedia Computer Science*, vol. 2, pp. 173–180, 2010.

[76] D. Sharma and I. Ali, "A modified mfcc feature extraction technique for robust speaker recognition," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2015, pp. 1052–1057.

[77] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition," *Speech communication*, vol. 54, no. 4, pp. 543–565, 2012.

[78] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[79] Q. Jin, "Robust speaker recognition," Ph.D. dissertation, Carnegie Mellon University, Language Technologies Institute, School of . . . , 2007.

[80] Z. Chen, R. Yang, Z. Zhao, D. Cai, and X. He, "Dialogue act recognition via crf-attentive structured network," in *The 41st international acm sigir conference on research & development in information retrieval*, 2018, pp. 225–234.

[81] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, and D. Chow, "Forensic database of voice recordings of 500+ australian english speakers," 2015.

[82] A. Drygajlo, D. Meuwly, and A. Alexander, "Statistical methods and bayesian interpretation of evidence in forensic automatic speaker recognition," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[83] J. Alam, P. Kenny, P. Ouellet, T. Stafylakis, and P. Dumouchel, "Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the rsr2015 corpus," in *Odyssey speaker and language recognition workshop*, 2014, pp. 123–130.

[84] A. Larcher, P.-M. Bousquet, K. A. Lee, D. Matrouf, H. Li, and J.-F. Bonastre, "I-vectors in the context of phonetically-constrained short utterances for speaker verification," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4773–4776.

[85] G. Wang, K.-A. Lee, T. H. Nguyen, H. Sun, and B. Ma, "Joint speaker and lexical modeling for short-term characterization of speaker." in *Interspeech*, 2016, pp. 415–419.

[86] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody, "The australian national database of spoken language," in *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. IEEE, 1994, pp. I–97.

[87] P. Nguyen, D. Tran, X. Huang, and D. Sharma, "Automatic classification of speaker characteristics," in *International Conference on Communications and Electronics 2010*. IEEE, 2010, pp. 147–152.

[88] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. v. Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The reddots data collection for speaker recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[89] T. Kinnunen, M. Sahidullah, I. Kukanov, H. Delgado, M. Todisco, A. Sarkar, N. B. Thomsen, V. Hautamäki, N. Evans, and Z.-H. Tan, "Utterance verification for text-dependent speaker recognition: a comparative assessment using the reddots corpus," 2016.

[90] H. Zeinali, H. Sameti, L. Burget, J. Cernockỳ, N. Maghsoodi, and P. Matejka, "i-vector/hmm based text-dependent speaker verification system for reddots challenge." in *InterSpeech*, 2016, pp. 440–444.

[91] M. J. Alam, P. Kenny, and V. Gupta, "Tandem features for text-dependent speaker verification on the reddots corpus." in *Interspeech*, 2016, pp. 420–424.

[92] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database." in *Interspeech*, 2016, pp. 818–822.

[93] Y. Liu, Y. Tian, L. He, and J. Liu, "Investigating various diarization algorithms for speaker in the wild (sitw) speaker recognition challenge." in *Interspeech*, 2016, pp. 853–857.

[94] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.

[95] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak *et al.*, "State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations," *Computer Speech & Language*, vol. 60, p. 101026, 2020.

[96] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[97] L. Feng and L. K. Hansen, "A new database for speaker recognition," Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, 2005. [Online]. Available: http://www2.compute.dtu.dk/pubdb/pubs/3662-full.html

[98] M. Soleymanpour and H. Marvi, "Text-independent speaker identification based on selection of the most similar feature vectors," *International Journal of Speech Technology*, vol. 20, no. 1, pp. 99–108, 2017.

[99] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, "Speaker recognition using mel frequency cepstral coefficients (mfcc) and vector quantization (vq) techniques," in *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*. IEEE, 2012, pp. 248–251.

[100] P. Dhakal, P. Damacharla, A. Y. Javaid, and V. Devabhaktuni, "A near real-time automatic speaker recognition architecture for voice-based user interface," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 504–520, 2019.

[101] Y. Lan, Z. Hu, Y. C. Soh, and G.-B. Huang, "An extreme learning machine approach for speaker recognition," *Neural Computing and Applications*, vol. 22, no. 3, pp. 417–425, 2013.

[102] P. Angkititrakul and J. H. Hansen, "Discriminative in-set/out-of-set speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 498–508, 2007.

[103] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry." in *Interspeech*, 2013, pp. 930–934.

[104] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. De Leon, "Speaker recognition anti-spoofing," in *Handbook of biometric anti-spoofing*. Springer, 2014, pp. 125–146.

[105] S. Ozaydin, "Design of a text independent speaker recognition system," in *2017 international conference on electrical and computing technologies and applications (ICECTA)*. IEEE, 2017, pp. 1–5.

[106] R. Jahangir, Y. W. Teh, N. A. Memon, G. Mujtaba, M. Zareei, U. Ishtiaq, M. Z. Akhtar, and I. Ali, "Text-independent speaker identification through feature fusion and deep neural network," *IEEE Access*, vol. 8, pp. 32 187–32 202, 2020.

[107] R. Hokking and K. Woraratpanya, "A hybrid of fractal code descriptor and harmonic pattern generator for improving speech recognition of different sampling rates," in *International Conference on Computing and Information Technology*. Springer, 2017, pp. 32–42.

[108] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout *et al.*, "Voices obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

[109] D. Snyder, J. Villalba, N. Chen, D. Povey, G. Sell, N. Dehak, and S. Khudanpur, "The jhu speaker recognition system for the voices 2019 challenge." in *INTERSPEECH*, 2019, pp. 2468–2472.

[110] C. S. Greenberg, V. M. Stanford, A. F. Martin, M. Yadagiri, G. R. Doddington, J. J. Godfrey, and J. Hernandez-Cordero, "The 2012 nist speaker recognition evaluation." in *INTERSPEECH*, 2013, pp. 1971–1975.

[111] C. Hanilçi, T. Kinnunen, F. Ertas, R. Saeidi, J. Pohjalainen, and P. Alku, "Regularized all-pole models for speaker verification under noisy environments," *IEEE Signal Processing Letters*, vol. 19, no. 3, pp. 163–166, 2012.

[112] T. Kinnunen, R. Saeidi, F. Sedlák, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-variance multitaper mfcc features: a case study in robust speaker verification," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 7, pp. 1990–2001, 2012.

[113] M. Sahidullah and G. Saha, "A novel windowing technique for efficient computation of mfcc for speaker recognition," *IEEE signal processing letters*, vol. 20, no. 2, pp. 149–152, 2012.

[114] S. Sarkar and K. S. Rao, "Stochastic feature compensation methods for speaker verification in noisy environments," *Applied Soft Computing*, vol. 19, pp. 198–214, 2014.

[115] R. M. Hecht, E. Noor, G. Dobry, Y. Zigel, A. Bar-Hillel, and N. Tishby, "Effective model representation by information bottleneck principle," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 8, pp. 1755–1759, 2013.

[116] J. Pohjalainen, C. Hanilçi, T. Kinnunen, and P. Alku, "Mixture linear prediction in speaker verification under vocal effort mismatch," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1516–1520, 2014.

[117] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Low-variance multitaper mel-frequency cepstral coefficient features for speech and speaker recognition systems," *cognitive computation*, vol. 5, no. 4, pp. 533–544, 2013.

[118] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper mfcc and plp features for speaker verification using i-vectors," *Speech communication*, vol. 55, no. 2, pp. 237–251, 2013.

[119] S. Ganapathy, S. H. Mallidi, and H. Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1285–1295, 2014.

[120] A. K. Sarkar, C.-T. Do, V.-B. Le, and C. Barras, "Combination of cepstral and phonetically discriminative features for speaker verification," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1040–1044, 2014.

[121] K. K. George, C. S. Kumar, K. Ramachandran, and A. Panda, "Cosine distance features for improved speaker verification," *Electronics Letters*, vol. 51, no. 12, pp. 939–941, 2015.

[122] Z.-Y. Li, W.-Q. Zhang, and J. Liu, "Multi-resolution time frequency feature and complementary combination for short utterance speaker

recognition," *Multimedia Tools and Applications*, vol. 74, no. 3, pp. 937–953, 2015.

[123] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[124] D. Zhou, L. Wang, K. A. Lee, M. Liu, and J. Dang, "Deep discriminative embedding with ranked weight for speaker verification," in *International Conference on Neural Information Processing*. Springer, 2020, pp. 79–86.

[125] X. Qin, H. Bu, and M. Li, "Hi-mia: A far-field text-dependent speaker verification database and the baselines," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7609–7613.

[126] Y. Zhang, H. Yu, and Z. Ma, "Speaker verification system based on deformable cnn and time-frequency attention," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 1689–1692.

[127] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[128] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[129] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.

[130] M. Hajibabaei and D. Dai, "Unified hypersphere embedding for speaker recognition," *arXiv preprint arXiv:1807.08312*, 2018.

[131] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[132] C. McCool and S. Marcel, "Mobio database for the icpr 2010 face and speech competition," Idiap, Tech. Rep., 2009.

[133] D. v. Vloed, J. Bouten, and D. A. van Leeuwen, "Nfi-frits: A forensic speaker recognition database and some first experiments," 2014.

[134] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[135] P. Bell, M. J. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester *et al.*, "The mgb challenge: Evaluating multi-genre broadcast media recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 687–693.

[136] 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*. IEEE, 2012, pp. 345–348.

[137] 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)*. IEEE, 2012, pp. 4377–4380.

[138] Y. Kawakami, L. Wang, A. Kai, and S. Nakagawa, "Speaker identification by combining various vocal tract and vocal source features," in *International conference on text, speech, and dialogue*. Springer, 2014, pp. 382–389.

[139] H. Lu, A. B. Brush, B. Priyantha, A. K. Karlson, and J. Liu, "Speakersense: Energy efficient unobtrusive speaker identification on mobile phones," in *International conference on pervasive computing*. Springer, 2011, pp. 188–205.

[140] B. Nagaraja and H. Jayanna, "Multilingual speaker identification with the constraint of limited data using multitaper mfcc," in *International conference on security in computer networks and distributed systems*. Springer, 2012, pp. 127–134.

[141] O. Plchot, S. Matsoukas, P. Matějka, N. Dehak, J. Ma, S. Cumani, O. Glembek, H. Hermansky, S. Mallidi, N. Mesgarani *et al.*, "Developing a speaker identification system for the darpa rats project," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6768–6772.

[142] 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)*. IEEE, 2015, pp. 4265–4269.

[143] J.-D. Wu and Y.-J. Tsai, "Speaker identification system using empirical mode decomposition and an artificial neural network," *Expert Systems with Applications*, vol. 38, no. 5, pp. 6112–6117, 2011.

[144] 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)*. IEEE, 2012, pp. 4365–4368.

[145] 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*. IEEE, 2013, pp. 6773–6777.

[146] 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.

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

[148] M. Sarma and K. K. Sarma, "Vowel phoneme segmentation for speaker identification using an ann-based framework," *Journal of Intelligent Systems*, vol. 22, no. 2, pp. 111–130, 2013.

[149] R. Saeidi, A. Hurmalainen, T. Virtanen, and D. A. v. Leeuwen, "Exemplar-based sparse representation and sparse discrimination for noise robust speaker identification," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[150] G. Biagetti, P. Crippa, A. Curzi, S. Orcioni, and C. Turchetti, "Speaker identification with short sequences of speech frames." in *ICPRAM (2)*, 2015, pp. 178–185.

[151] 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*. Springer, 2016, pp. 465–477.

[152] Y.-H. Chao, "Speaker identification using pairwise log-likelihood ratio measures," in *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*. IEEE, 2012, pp. 1248–1251.

[153] E. Esmi, P. Sussner, M. E. Valle, F. Sakuray, and L. Barros, "Fuzzy associative memories based on subsethood and similarity measures with applications to speaker identification," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2012, pp. 479–490.

[154] X. Fan and J. H. Hansen, "Speaker identification within whispered speech audio streams," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 5, pp. 1408–1421, 2010.

[155] E. Fang and J. N. Gowdy, "New algorithms for improved speaker identification," *International Journal of Biometrics*, vol. 5, no. 3-4, pp. 360–369, 2013.

[156] N. Fazakis, S. Karlos, S. Kotsiantis, and K. Sgarbas, "Speaker identification using semi-supervised learning," in *International Conference on Speech and Computer*. Springer, 2015, pp. 389–396.

[157] M. Gabrea, "Two microphones speech enhancement systems based on instrumental variable algorithm for speaker identification," in *2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2011, pp. 000 569–000 572.

[158] M. V. Ghiurcau, C. Rusu, and J. Astola, "A study of the effect of emotional state upon text-independent speaker identification," in *2011 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 4944–4947.

[159] G. Chenghui, Z. Heming, and T. Zhi, "Speaker identification of whispered speech with perceptible mood." *Journal of Multimedia*, vol. 9, no. 4, 2014.

[160] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertaş, "Speaker identification from shouted speech: Analysis and compensation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8027–8031.

[161] M.-J. Kim, I.-H. Yang, and H.-J. Yu, "Histogram equalization using centroids of fuzzy c-means of background speakers' utterances for speaker identification," in *International Conference on Statistical Language and Speech Processing*. Springer, 2013, pp. 143–151.

[162] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 6, pp. 1791–1801, 2010.

[163] G. Liu, Y. Lei, and J. H. Hansen, "Robust feature front-end for speaker identification," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 4233–4236.

[164] Y. Michalevsky, R. Talmon, and I. Cohen, "Speaker identification using diffusion maps," in *2011 19th European signal processing conference*. IEEE, 2011, pp. 1299–1302.

[165] V. Mitra, M. McLaren, H. Franco, M. Graciarena, and N. Scheffer, "Modulation features for noise robust speaker identification." in *INTERSPEECH*, 2013, pp. 3703–3707.

[166] A. A. Nugraha, K. Yamamoto, and S. Nakagawa, "Single-channel dereverberation by feature mapping using cascade neural networks for robust distant speaker identification and speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–31, 2014.

[167] A. Pal, S. Bose, G. K. Basak, and A. Mukhopadhyay, "Speaker identification by aggregating gaussian mixture models (gmms) based on uncorrelated mfcc-derived features," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 04, p. 1456006, 2014.

[168] M. A. Pathak and B. Raj, "Privacy-preserving speaker verification and identification using gaussian mixture models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 397–406, 2012.

[169] S. Prasad, Z.-H. Tan, and R. Prasad, "Multi-frame rate based multiple-model training for robust speaker identification of disguised voice," in *2013 16th international symposium on wireless personal multimedia communications (WPMC)*. IEEE, 2013, pp. 1–4.

[170] S. O. Sadjadi and J. H. Hansen, "Robust front-end processing for speaker identification over extremely degraded communication channels," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7214–7218.

[171] S. Safavi, A. Hanani, M. Russell, P. Jancovic, and M. J. Carey, "Contrasting the effects of different frequency bands on speaker and accent identification," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 829–832, 2012.

[172] A. K. Sarkar and S. Umesh, "Eigen-voice based anchor modeling system for speaker identification using mllr super-vector," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[173] A. K. Sarkar, S. Umesh, and J.-F. Bonastre, "Computationally efficient speaker identification using fast-mllr based anchor modeling," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4357–4360.

[174] M. Sidorov, A. Schmitt, S. Zablotskiy, and W. Minker, "Survey of automated speaker identification methods," in *2013 9th International Conference on Intelligent Environments*. IEEE, 2013, pp. 236–239.

[175] J. Taghia, Z. Ma, and A. Leijon, "On von-mises fisher mixture model in text-independent speaker identification." in *INTERSPEECH*, 2013, pp. 2499–2503.

[176] I. Trabelsi and D. B. Ayed, "A multi level data fusion approach for speaker identification on telephone speech," *arXiv preprint arXiv:1407.0380*, 2014.

[177] 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 Transactions on Automation Science and Engineering*, vol. 12, no. 4, pp. 1191–1199, 2015.

[178] L. Wang, Z. Zhang, and A. Kai, "Hands-free speaker identification based on spectral subtraction using a multi-channel least mean square approach," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7224–7228.

[179] Y. Xing, H. Li, and P. Tan, "Hierarchical fuzzy speaker identification based on fcm and fsvm," in *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*. IEEE, 2012, pp. 311–315.

[180] Y. Yang and J. Liu, "Dereverberation for speaker identification in meeting," in *International conference on human-computer interaction*. Springer, 2014, pp. 594–599.

[181] Y. Yang, L. Chen, and W. Wang, "Emotional speaker identification by humans and machines," in *Chinese Conference on Biometric Recognition*. Springer, 2011, pp. 167–173.

[182] L. Zao and R. Coelho, "Colored noise based multicondition training technique for robust speaker identification," *IEEE Signal Processing Letters*, vol. 18, no. 11, pp. 675–678, 2011.

[183] X. Zhao and D. Wang, "Analyzing noise robustness of mfcc and gfcc features in speaker identification," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7204–7208.

[184] Z. Zhang, L. Wang, and A. Kai, "Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–12, 2014.

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

[186] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836–845, 2014.

[187] X. Zhao, Y. Shao, and D. Wang, "Robust speaker identification using a casa front-end," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 5468–5471.

[188] A. Malegaonkar and A. Ariyaeeinia, "Performance evaluation in open-set speaker identification," in *European workshop on biometrics and identity management*. Springer, 2011, pp. 106–112.

[189] P. Qi and L. Wang, "Experiments of gmm based speaker identification," in *2011 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*. IEEE, 2011, pp. 26–31.

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

[191] J.-F. Wang, J.-S. Peng, J.-C. Wang, P.-C. Lin, and T.-W. Kuan, "Hardware/software co-design for fast-trainable speaker identification system based on smo," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2011, pp. 1621–1625.

[192] 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.

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

[194] H. Do, I. Tashev, and A. Acero, "A new speaker identification algorithm for gaming scenarios," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 5436–5439.

[195] S. O. Sadjadi and J. H. 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)*. IEEE, 2011, pp. 5448–5451.

[196] 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*. Springer, 2012, pp. 358–366.

[197] 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," *International journal of multimedia information retrieval*, vol. 3, no. 3, pp. 161–175, 2014.

[198] G. S. Mosa and A. A. Ali, "Arabic phoneme recognition using hierarchical neural fuzzy petri net and lpc feature extraction," *Signal Processing: An International Journal (SPIJ)*, vol. 3, no. 5, p. 161, 2009.

[199] N. Yousefian and M. Analoui, "Using radial basis probabilistic neural network for speech recognition," in *Proceeding of 3rd international conference on information and knowledge (IKT07), Mashhad, Iran*, 2007.

[200] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *Ismir*, vol. 270. Citeseer, 2000, pp. 1–11.

[201] B. S. Atal, "The history of linear prediction," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 154–161, 2006.

[202] M. Hariharan, L. S. Chee, and S. Yaacob, "Analysis of infant cry through weighted linear prediction cepstral coefficients and probabilistic neural network," *Journal of medical systems*, vol. 36, no. 3, pp. 1309–1315, 2012.

[203] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[204] S. Narang and M. D. Gupta, "Speech feature extraction techniques: a review," *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 3, pp. 107–114, 2015.

[205] T. B. Rao, P. P. Reddy, and A. Prasad, "Recognition and a panoramic view of raaga emotions of singers-application gaussian mixture model,"

International Journal of Research and Reviews in Computer Science, vol. 2, no. 1, p. 201, 2011.

[206] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition using mp-based features," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 1–4.

[207] P. Kumar and M. Chandra, "Speaker identification using gaussian mixture models," *MIT International Journal of Electronics and Communication Engineering*, vol. 1, no. 1, pp. 27–30, 2011.

[208] R. Kumar, R. Ranjan, S. K. Singh, R. Kala, A. Shukla, and R. Tiwari, "Multilingual speaker recognition using neural network," *Proceedings of the Frontiers of Research on Speech and Music, FRSM*, pp. 1–8, 2009.

[209] S. Agrawal, A. Shruti, and C. R. Krishna, "Prosodic feature based text dependent speaker recognition using machine learning algorithms," *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5150–5157, 2010.

[210] K. T. Al-Sarayreh, R. E. Al-Qutaish, and B. M. Al-Kasasbeh, "Using the sound recognition techniques to reduce the electricity consumption in highways," *Journal of American Science*, vol. 5, no. 2, pp. 1–12, 2009.

[211] M. Paulraj, Y. Sazali, A. Nazri, and S. Kumar, "A speech recognition system for malaysian english pronunciation using neural network," 2009.

[212] M. El Choubassi, H. El Khoury, C. J. Alagha, J. Skaf, and M. Al-Alaoui, "Arabic speech recognition using recurrent neural networks," in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 03EX795)*. IEEE, 2003, pp. 543–547.

[213] Q.-Z. Wu, I.-C. Jou, and S.-Y. Lee, "On-line signature verification using lpc cepstrum and neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 27, no. 1, pp. 148–153, 1997.

[214] A. M. Ahmad, S. Ismail, and D. Samaon, "Recurrent neural network with backpropagation through time for speech recognition," in *IEEE International Symposium on Communications and Information Technology, 2004. ISCIT 2004.*, vol. 1. IEEE, 2004, pp. 98–102.

[215] D. A. Reynolds, "Gaussian mixture models." *Encyclopedia of biometrics*, vol. 741, pp. 659–663, 2009.

[216] G. S. Kumar, K. P. Raju, M. R. CPVNJ, and P. Satheesh, "Speaker recognition using gmm," *International Journal of Engineering Science and Technology*, vol. 2, no. 6, pp. 2428–2436, 2010.

[217] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[218] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[219] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[220] C. Zhang, F. Bahmaninezhad, S. Ranjan, H. Dubey, W. Xia, and J. H. Hansen, "Utd-crss systems for 2018 nist speaker recognition evaluation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5776–5780.

[221] Q. Jin and A. Waibel, "Application of lda to speaker recognition," in *Sixth International Conference on Spoken Language Processing*, 2000.

[222] G. Liu and J. H. Hansen, "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1978–1992, 2014.

[223] A. Khosravani and M. M. Homayounpour, "Linearly constrained minimum variance for robust i-vector based speaker recognition." in *Odyssey*. Citeseer, 2014.

[224] S. Novoselov, T. Pekhovsky, and K. Simonchik, "Stc speaker recognition system for the nist i-vector challenge." in *Odyssey*, 2014.

[225] P.-E. Danielsson, "Euclidean distance mapping," *Computer Graphics and image processing*, vol. 14, no. 3, pp. 227–248, 1980.

[226] G. Qian, S. Sural, Y. Gu, and S. Pramanik, "Similarity between euclidean and cosine angle distance for nearest neighbor queries," in *Proceedings of the 2004 ACM symposium on Applied computing*, 2004, pp. 1232–1237.

[227] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, "Deep speaker feature learning for text-independent speaker verification," *arXiv preprint arXiv:1705.03670*, 2017.

[228] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[229] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[230] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[231] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[232] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[233] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "I-vector based speaker recognition on short utterances," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2011, pp. 2341–2344.

[234] L. Zhu and Q. Yang, "Speaker recognition system based on weighted feature parameter," *Physics Procedia*, vol. 25, pp. 1515–1522, 2012.

[235] S. Yaman, J. Pelecanos, and R. Sarikaya, "Bottleneck features for speaker recognition," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[236] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 55–59.

[237] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE signal processing letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[238] M. Zhang, Y. Chen, L. Li, and D. Wang, "Speaker recognition with cough, laugh and" wei"," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 497–501.

[239] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.

[240] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[241] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "Gmm and cnn hybrid method for short utterance speaker recognition," *IEEE Transactions on Industrial informatics*, vol. 14, no. 7, pp. 3244–3252, 2018.

[242] S. Shon, H. Tang, and J. Glass, "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1007–1013.

[243] S. M. Kye, Y. Jung, H. B. Lee, S. J. Hwang, and H. Kim, "Meta-learning for short utterance speaker recognition with imbalance length pairs," *arXiv preprint arXiv:2004.02863*, 2020.

[244] A. Chowdhury and A. Ross, "Fusing mfcc and lpc features using 1d triplet cnn for speaker recognition in severely degraded audio signals," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616–1629, 2019.

[245] K. Dash, D. Padhi, B. Panda, and S. Mohanty, "Speaker identification using mel frequency cepstralcoefficient and bpnn," *International Journal of Advanced Research in Computer Science and Software Engineering Research Paper*, vol. 2, 2012.

[246] S. Nandyal, S. Wali, and S. Hatture, "Mfcc based text-dependent speaker identification using bpnn," *International Journal of Signal Processing Systems*, vol. 3, no. 1, pp. 30–34, 2015.

[247] O. S. Faragallah, "Robust noise mkmfcc–svm automatic speaker identification," *International Journal of Speech Technology*, vol. 21, no. 2, pp. 185–192, 2018.

[248] R. Chakroun and M. Frikha, "Robust text-independent speaker recognition with short utterances using gaussian mixture models," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 2020, pp. 2204–2209.

[249] R. Li, J.-Y. Jiang, J. L. Li, C.-C. Hsieh, and W. Wang, "Automatic speaker recognition with limited data," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 340–348.

[250] A. D. Mengistu, "Automatic text independent amharic language speaker recognition in noisy environment using hybrid approaches of lpcc, mfcc and gfcc," *International Journal of Advanced Studies in Computers, Science and Engineering*, vol. 6, no. 5, p. 8, 2017.

[251] Y. Wang and B. Lawlor, "Speaker recognition based on mfcc and bp neural networks," in *2017 28th Irish Signals and Systems Conference (ISSC)*. IEEE, 2017, pp. 1–4.

[252] I. Bisio, C. Garibotto, A. Grattarola, F. Lavagetto, and A. Sciarrone, "Smart and robust speaker recognition for context-aware in-vehicle applications," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8808–8821, 2018.

[253] S. Madikeri, P. Motlicek, and S. Dey, "A bayesian approach to inter-task fusion for speaker recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5786–5790.

[254] J. A. C. Nunes, D. Macêdo, and C. Zanchettin, "Additive margin sincnet for speaker recognition," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–5.

[255] K. S. Prasad, G. K. Ramaiah, and M. Manjunatha, "Speech features extraction techniques for robust emotional speech analysis/recognition," *Indian Journal of Science and Technology*, vol. 10, no. 3, 2017.

[256] K. Saakshara, K. Pranathi, R. Gomathi, A. Sivasangari, P. Ajitha, and T. Anandhi, "Speaker recognition system using gaussian mixture model," in *2020 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2020, pp. 1041–1044.

[257] S. Biswas and S. S. Solanki, "Speaker recognition: an enhanced approach to identify singer voice using neural network," *International Journal of Speech Technology*, pp. 1–13, 2020.

[258] S. Abd El-Moneim, M. Nassar, M. I. Dessouky, N. A. Ismail, A. S. El-Fishawy, and F. E. Abd El-Samie, "Text-independent speaker recognition using lstm-rnn and speech enhancement," *Multimedia Tools and Applications*, vol. 79, no. 33, pp. 24 013–24 028, 2020.

[259] M. Tripathi, D. Singh, and S. Susan, "Speaker recognition using sincnet and x-vector fusion," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2020, pp. 252–260.

[260] N. Tawara, A. Ogawa, T. Iwata, M. Delcroix, and T. Ogawa, "Frame-level phoneme-invariant speaker embedding for text-independent speaker recognition on extremely short utterances," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6799–6803.

[261] R. Peri, M. Pal, A. Jati, K. Somandepalli, and S. Narayanan, "Robust speaker recognition using unsupervised adversarial invariance," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6614–6618.

[262] M. Gupta, S. S. Bharti, and S. Agarwal, "Gender-based speaker recognition from speech signals using gmm model," *Modern Physics Letters B*, vol. 33, no. 35, p. 1950438, 2019.

[263] L. Lu, L. Liu, M. J. Hussain, and Y. Liu, "I sense you by breath: Speaker recognition via breath biometrics," *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 2, pp. 306–319, 2017.

[264] N. Singh, R. Khan, and R. Shree, "Applications of speaker recognition," *Procedia engineering*, vol. 38, pp. 3122–3126, 2012.

[265] A. V. Bhalla, S. Khaparkar, and M. R. Bhalla, "Performance improvement of speaker recognition system," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 3, pp. 138–143, 2012.

[266] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for nist 2012 speaker recognition evaluation," SRI INTERNATIONAL MENLO PARK CA SPEECH TECHNOLOGY AND RESEARCH LAB, Tech. Rep., 2013.

[267] G. Nijhawan and M. Soni, "A new design approach for speaker recognition using mfcc and vad," *International Journal of Image, Graphics and Signal Processing*, vol. 5, no. 9, p. 43, 2013.

[268] M. McLaren, Y. Lei, N. Scheffer, and L. Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[269] A. Kanagasundaram, R. Vogt, D. Dean, and S. Sridharan, "Plda based speaker recognition on short utterances," in *Proceedings of The Speaker and Language Recognition Workshop: Odyssey 2012*. International Speech Communication Association, 2012, pp. 28–33.

[270] Z. Huang, S. Wang, and K. Yu, "Angular softmax for short-duration text-independent speaker verification." in *Interspeech*, 2018, pp. 3623–3627.

[271] A. Miguel, J. Llombart, A. Ortega, and E. Lleida, "Tied hidden factors in neural networks for end-to-end speaker recognition," *arXiv preprint arXiv:1812.11946*, 2018.

[272] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification." in *Interspeech*, vol. 2018, 2018, pp. 3573–3577.

[273] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[274] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4353–4356.

[275] S. H. Yella and H. Bourlard, "Improved overlap speech diarization of meeting recordings using long-term conversational features," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7746–7750.

[276] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4930–4934.

[277] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5239–5243.

[278] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

[279] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 413–417.

[280] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[281] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," National Inst of Standards and Technology Gaithersburg MD, Tech. Rep., 1997.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084299, IEEE Access

MUHAMMAD MOHSIN KABIR is a Computer Science graduate from Bangladesh University of Business and Technology. Currently, he is working as an associate researcher in the Advanced Machine Learning lab. He was born and raised in Dhaka, Bangladesh. He is always willing to learn new things with full enthusiasm and passion. He is particularly interested in deep learning, pattern recognition, and computer vision. He has experienced working in Python, Keras, TensorFlow, Sklearn, Scipy, etc. He is currently researching Content-Based Image Retrieval, Facial Expression Recognition, Automatic Speaker Recognition, and Writer Identification. He is a very dedicated and promising Deep Learning researcher.

M. F. MRIDHA (Senior Member, IEEE) is currently working as an Associate Professor in the Department of Computer Science and Engineering of the Bangladesh University of Business and Technology. He also worked as a CSE department faculty member at the University of Asia Pacific and as a graduate coordinator from 2012 to 2019. He received his Ph.D. in AI/ML from Jahangirnagar University in the year 2017. He joined as a lecturer at the Department of Computer Science and Engineering, Stamford University Bangladesh, in June 2007. He was promoted as a Senior Lecturer at the same department in October 2010 and promoted as an Assistant Professor at the same department in October 2011. Then he joined UAP in May 2012 as an assistant professor. His research experience, within both academia and industry, results in over 80 journal and conference publications. His research interests include artificial intelligence (AI), machine learning, deep learning, and natural language processing (NLP). For more than 10 (Ten) years, he has been with the masters and undergraduate students as a supervisor of their thesis work. His research interests include artificial intelligence (AI), machine learning, natural language processing (NLP), big data analysis, etc. He has served as a program committee member in several international conferences/workshops. He served as an associate editor of several journals.

JUNGPIL SHIN (Senior Member, IEEE) received a B.Sc. in Computer Science and Statistics and an M.Sc. in Computer Science from Pusan National University, Korea in 1990 and 1994, respectively. He received the Ph.D. degree in computer science and communication engineering from Kyushu University, Japan, in 1999, under the scholarship from the Japanese Government (MEXT). He was an Associate Professor, a Senior Associate Professor, and a Professor with the School of Computer Science and Engineering, The University of Aizu, Japan, in 1999, 2004, and 2019, respectively. He has coauthored more than 250 publications published in widely cited journals and conferences. His research interests include pattern recognition, image processing, computer vision, and machine learning, human–computer interaction, non-touch interface, human gesture recognition, automatic control, Parkinson's disease diagnosis, ADHD diagnosis, user authentication, machine intelligence, handwriting analysis, recognition, and synthesis. He is a member of ACM, IEICE, IPSJ, KISS, and KIPS. He served several conferences as a program chair and program committee member for numerous international conferences. He serves as an Editor of journals of IEEE and MDPI Sensors. He serves as a Reviewer for several IEEE and SCI major journals.

ISRAT JAHAN is a B.Sc. student of Computer Science and Engineering at Bangladesh University of Business and Technology (BUBT). She is enthusiastic, determined, a quick learner, and a good communicator. She is a competitive programmer. She has attained more than five Inter-University Programming Contest in Bangladesh and BUBT Intra-University Programming Contest with commendable positions. She has also participated in the National Girls Programming Contest (NGPC). She has experience working with Tensorflow, Keras, Matplotlib, etc., and is interested in deep learning, data mining research fields.

ABU QUWSAR OHI is a computer science graduate and currently working as a lecturer and research assistant at Bangladesh University of Business and Technology in the Department of Computer Science and Engineering. His research focuses on deep learning algorithms, pattern recognition, computer vision, and reinforcement learning. Moreover, he is proficient in prominent frameworks, including TensorFlow, Keras, NumPy, Matplotlib, etc. Apart from his research works, he is a Competitive Programmer and has attained more than 10 Bangladeshi national programming contests, including National Collegiate Programming Contest (NCPC) and International Collegiate Programming Contest (ICPC). He is passionate about learning fundamental algorithms, including data structures, dynamic programming, and game theory. He has experience working in famous languages, including Python and C++.