INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

REVIEW ARTICLE

A review on state-of-the-art Automatic Speaker verification system from spoofing and anti-spoofing perspective

Ankita Chadha[1]*, **Azween Abdullah**[2], **Lorita Angeline**[2], **Sivakumar Sivanesan**[2]

1 Doctorate Student, School of Computer Science and Engineering, Taylors University, 47500, Selangor, Malaysia
2 School of Computer Science and Engineering, Taylors University, 47500, Selangor, Malaysia

OPEN ACCESS

Received: 09.07.2021

Accepted: 19.10.2021

Published: 29.11.2021

Citation: Chadha A, Abdullah A, Angeline L, Sivanesan S (2021) A review on state-of-the-art Automatic Speaker verification system from spoofing and anti-spoofing perspective. Indian Journal of Science and Technology 14(40): 3026-3050. https://doi.org/10.17485/IJST/v14i40.1279

* **Corresponding author**.

*chadhaankitanaresh@sd.taylors.edu.my

Funding: Taylors University, Lakeside Campus, Selangor, Malaysia

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background/Objectives: The anti-spoofing measures are blooming with an aim to protect the Automatic Speaker Verification systems from susceptible spoofing attacks. This review is an amalgam of the possible attack types, the datasets required, the renowned feature representation techniques, modeling algorithms involving machine learning, and score normalization techniques. **Method/Findings**: A detailed analysis of existing datasets is carried based on the total speaker samples, the number of speakers, and source of availability-open or licensed. This may foster choosing the right dataset for building the anti-spoofing frameworks. Further, the feature extraction schemes are elaborated with an intention to cover the vast span of features existing in various parts of raw speech for obtaining speaker-specific traits. Further, the machine learning algorithms ranging from discriminative to generative to mixed form are explored for seeking the right algorithm in specific attack conditions. On the whole, these analyses of existing features and machine learning algorithms together contribute to classifying the unknown test samples as genuine or spoofed. The score normalization techniques are also considered in this review to avoid any misclassifications and ultimately reduce the False Acceptance Ratios. The performance of any anti-spoofing speaker verification system may be evaluated using standard objective measures such are Equal Error Rate, False positive ratios, and graphical plots. These measures are briefly explained in this review. Overall, the critical analysis of individual methods-feature extraction, machine learning, score normalization, and all the anti-spoofing datasets are also discussed for giving a kick-start to any researcher beginning to explore in this direction. The shortcomings and risks involved in building an enhanced speaker verification system that is robust to almost all the attack types are listed in this article. The review of studies conducted so far has led to vital future directions that are enlisted in the concluding remarks of the article.

Keywords: Automatic Speaker Verification; Spoofed Detection; AntiSpoofing; Voice Conversion; Speech Synthesis; Replay Speech

1 Introduction

The task of permitting pre-enrolled speakers with an intension to disallow unknown ones is called Speaker Verification and that system is an Automatic Speaker Verification (ASV)[1]. The ASV is a crucial part of a Speaker Recognition platform after the Speaker Identification (SI) mechanism[2]. The SI system opts for the most likely speaker among the presented list of speakers while ASV investigates if the claimed identity is true or false. Depending upon the input sequence, the ASV may be text-dependent or independent. The former being preferred in authentication scenarios as it needs higher accuracy[3]. Furthermore, the chances of an ASV being susceptible to spoofing attacks is inevitable as any imposter could mimic or synthesize the target's voice for getting through the intended system. Hence, this study focuses on spoofing and anti-spoofing measures from the system analysis and decision algorithm perspective.

The developments in reducing channel and noise interference have led to ASV systems being employed in security-based scenarios such as phone banking[4]. However, the primary concern in using ASV is susceptibility to imposters. According to studies conducted in[5,6], there are two basic types of attacks: direct or Physical Access attacks (PA) and indirect or Logical Access attacks (LA). The PA attacks occur at the sensor level while LA attacks are observed after the sensor, that is at the feature representation stage or modelling stage. Also, another way of categorizing the spoofing attacks is through imposter variations which may be impersonated, replayed, voice converted, or synthetic speech. The impersonation or mimicry is the first-ever attack on a speaker verification system because of its ease of production. The only criteria are, the target and imposter's voice must have a similar fundamental frequency which is usually the case for twins. Mimicry is a product of a professional artist, who is trained and dedicated for the sole purpose of copying. The imposter or mimicry artist usually mimics the prosody and timbre of the target speaker[7]. The replayed speech is easy to reproduce for the attacker as it involves playing pre-recorded speech which is certainly captured without the permission of the target. This seems to be a text-dependent scenario where a fixed phrase is used for verifying the speaker's traits. Capturing real-time speech or modified speech of the intended target is tough, but is often considered as a challenge by the imposters and conducted irrespective of the challenges due to no human intervention for re-producing them.

The synthetic speech may be a by-product of a Text-to-Speech (TTS) system (i.e. Speech Synthesis (SS) or a Voice transformation or conversion (VC) speech. Presently, the TTS produces quite intelligible speech as it imitates the human speech production mechanism[8]. On the contrary, the VC speech may be generated from either (or all of these) human speech production, perception, and prosodic models[9]. On the whole, the ASV comprises of two-fold operating conditions: first being the training phrase where a statistical model is a result of extracting appropriate features from a known speaker's voice and trained through convenient machine learning algorithms. The saved model is then used during the testing phase to find if the unknown speaker's speech sample belongs to the known speaker or not. The block schematic of internal blocks during each mode of operation is demonstrated in Figure 1.

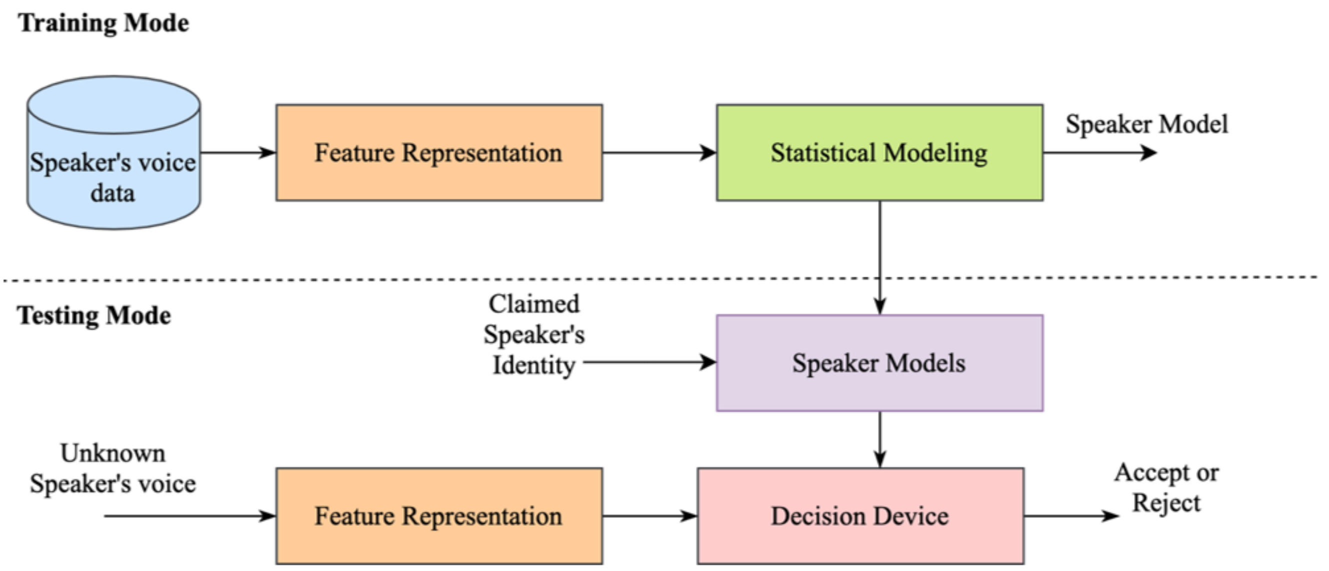

Fig 1. A block schematic of Automatic Speaker Verification framework

The literature offers quite significant reviews of developments in the ASV domain and anti-spoofing measures. The work[10] describes the basics of biometrics along with spoofing attacks while[11,12] describes a survey of techniques existing in the speaker verification system from the ASV Spoof challenge perspective that includes protocols, databases, and future directions. Another review covers the detection of speech synthesis and replay attacks[13]. The review of ASV based on short utterances is presented in[14] which explains challenges including the trends in this field. On the other hand in this article, a thorough analysis of state-of-art features, training algorithms along with popular databases and evaluation metrics are presented. Unlike past reviews, this work concentrates on popular methods addressing in-depth function of the blocks belonging to the ASV system. Such an intensive review of feature extraction and machine learning algorithms employed in anti-spoofing frameworks has not been conducted according to the best of the authors' knowledge. In fact, this article also presents a critical evaluation of these internal blocks for providing a base in developing anti-spoofing frameworks. Thus the main objective of this article are three-fold:

1. Investigating the available datasets for developing anti-spoofing measures in order to analyse the nature and types of attacks. This will promote a deciding criteria for selecting the right kind of dataset.
2. Investigating the feature representation techniques that help reducing the raw data redundancies and represent the spoofed and genuine speakers efficiently. This will a kick start for researchers looking out for existing feature extraction techniques in addition to their pros and cons.
3. Exploring existing machine learning and score normalization algorithms for accurate categorization of input test sample (as genuine or spoofed). This will keep a track of evolution of machine learning algorithms right from discriminative models to generative models to the artificial neural networks.

The article is structured as follows: Section 2 describes the ASV system for spoofed speech while section 3 identifies popular databases employed in studying ASV. Section 4 covers extensive literature related to feature representation techniques, section 5 describes the machine learning algorithms while the score normalization techniques are explained in section 6. Furthermore, section 7 elaborates evaluation metrics and lastly, section 8 summarizes the article review with hints for future work.

2 ASV system for spoofing attack

The mode of operation for an ASV for imposter speech is nearly identical to the standard ASV only with additional requirements to detect such mimicked or synthetic speech. Also in the testing mode, the test sample is matched with the trained model to obtain a score signifying the speaker belongs to a known or unknown class as depicted in Figure 2.

To decrypt the process of spoofing attacks on the ASV and taking necessary actions to prevent it, the characteristics of natural speech as against artificial or mimicked speech must be investigated. The human behavioural traits such as huskiness,

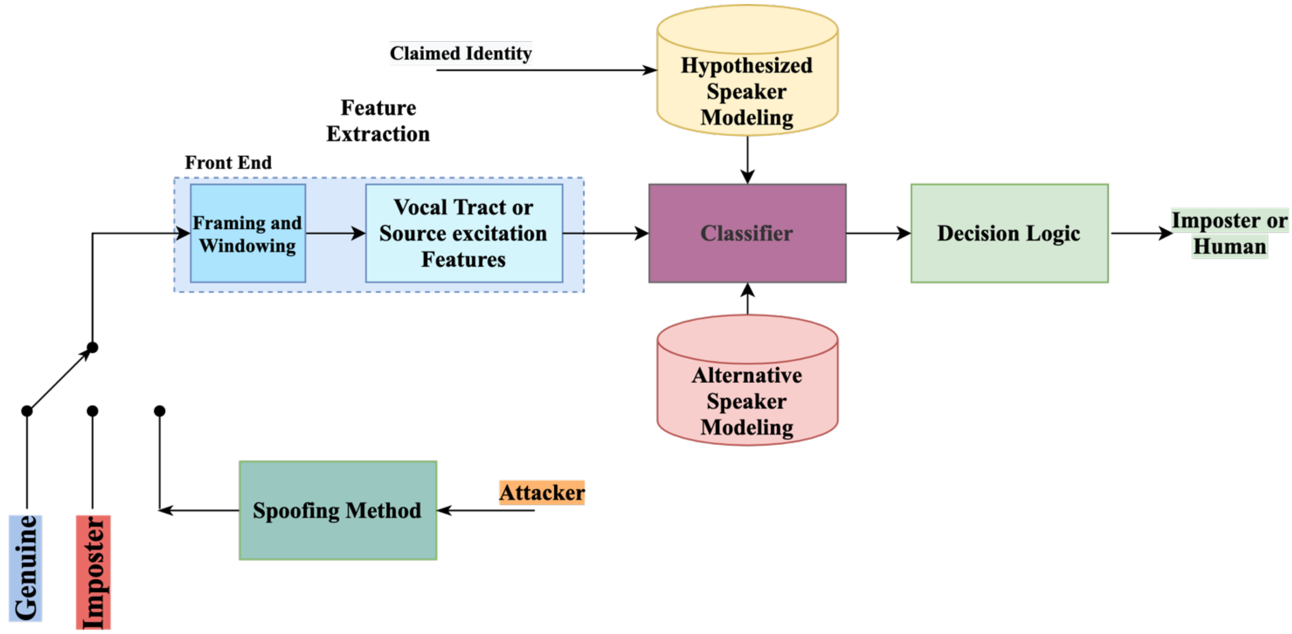

Fig 2. Automatic Speaker Verification with spoofing attack

breathlessness, and speaking rate are utterly impossible to mimic or synthesize individually[11]. Furthermore, the high-level features like pitch and duration are not consistent considering inter-speaker and intra-speaker variations. Timbre is also a potential feature considering natural versus artificial speech.

2.1 Speech-based Spoofing Attacks

The combined effort by individual steps of training and testing are vulnerable including links between each component like a microphone to input feature representation, features to classifiers, and classifiers to decision algorithm[15]. These attacks may be sub-divided into two basic types: Direct or the general spoofing attacks that occur at the microphone and during transmission to the first input component (feature extraction block); while indirect attacks occur inside the ASV itself which usually needs access to the system say at the feature level or classifier or decision logic end. The attacker would replace or modify the contents of these components. Direct access attacks are considered potential risks as opposed to indirect attacks as they don't require system-level access. Additionally, four broad categories of representation attacks are natural impersonation, speech synthesis, VC, and replay speech as described below:

2.1.1 Natural Impersonation

The Impersonation is performed by a professional mimicry artist or imposter who holds the ability to produce similar voice traits and behaviour or even twins with identical spectral characteristics. Through studies, it is inferred that the imposter does not depend on the prior knowledge of machines for copying the target speaker. All the imposter needs is a target speaker's voice sample and a nearly similar spectral pattern would authorize the imposter[16]. In fact, the impersonator tries to reproduce the prosodic parameters of the target[17]. Along with this, the imitator adapts to the target speaker's accent, pronunciation, lexicon, and various high-level features. Thereby the voice produced by impersonation could deceive the human ear perception. However, the practicality of this attack is negligible or extremely low as most often the anti-spoofing ASV considers spectral parameter traits as the base feature technique.

2.1.2 Speech Synthesis

Speech Synthesis (SS) refers to the conventional TTS system but with a target-specific speech that sounds intelligible, natural, and yet it is machine-generated speech from prompted text. Some common applications of SS in the younger generation such as audiobooks, in-the-car navigation, speech translation, etc[18,19]. The speech synthesis involves two basic steps: analyze the text (front-end) and generate waveform or speech (back-end). When analyzing the text, the words and sentences are broken into a less complex linguistic unit called a phoneme. On the contrary, the speech generation utilizes these linguistic parameters to build up a waveform. The careful analysis of literature suggested, four prime waveform generation techniques evolved to date. The first being the acoustic features specially formants representing every phoneme[20]. Following this, the second approach was rooted in diphones which comprised the second half of the first phoneme till the first half of consecutive phoneme. These diphones were further represented using a linear prediction algorithm. The third technique was based on selecting the right speech units and then concatenating them into a single speech sample which is termed as unit selection approach[21]. And lastly, statistical parametric-based synthesis techniques such as Hidden Markov Model (HMM) have shown promising results in the domain of SS[22,23]. Additionally, the DNN is also proposed for SS[23,24].

2.1.3 Voice Conversion

The speech signal from the source speaker is modified statistically/acoustically to sound identical to the target speaker's speech. This is a basic parametric difference between speech synthesis and voice conversion algorithms. To modify a source speaker's speech, the spectral characteristics and prosody of the imposter are mapped to find no audible change in speech. So the voice timbre, prosodic parameters like pitch and intonation are amended to be reflected in its characteristics. The spectrum modifications are performed by statistical parameters, frequency warping, and lastly unit selection algorithm[25]. The spectral modification techniques like Vector Quantization[26], GMM[27], Restricted Boltzmann Machines (RBM)[28] and Deep Belief Networks (DBN) (29) are explored in producing VC speech. The frequency warping techniques modify the source's frequency axis to the target's speaker. These modification techniques preserve the spectral content producing naturally sounding target speech[29,30]. Furthermore, the unit selection technique gave promising results, producing converted speech similar to the target speaker's voice.

Along with spectral parameters, the prosody modification would contribute to closer and natural synthetic speech. Pitch and duration are looked up when mentioning about speaker's prosody in case of voice conversion framework[31]. The threat to the ASV systems has increased over the period due to improvements in converted speech signals' quality.

2.1.4 Replay Speech

The pre-recorded speech fed into the ASV poses a potential risk to the system's configuration as may lead to giving unauthorized entry to the adversary. Such a spoofed speech could be procured at a given time without the consent of the victim. The speech samples may be concatenated or even clipped to obtain the desired utterance. Such attacks work well in the text-dependent ASV that has a fixed text phrase for getting access to the system. Spoofing attacks using replayed speech has now been a common practice due to the availability of affordable, good-quality recording instruments like mobile phones and laptops. Therefore , these attacks occur at the microphone level more often than the transmission level. The spectral similarity between natural and replay speech turn out to be quite close; thus it becomes rightful to conclude that the spectral features are susceptible to replay speech-based attacks[32]. From the point of view of objective score, the False Acceptance Ratio (FAR) has increased due to these attacks.

2.2 Detection of Spoofed Speech

The spoofed speech is a by-product of a human mimicry or machine's effort to mimic natural speech traits. The former is a low-risk attack hence not quite popular in the anti-spoof detection community while the latter involves synthetic speech produced by a TTS or VC framework or replayed through a recording device. There seems to be a need for detecting spoofed speech from the natural speech with the sole purpose of protecting unauthenticated access to crucial information. The developments in the VC field are more established than the TTS due to early work that began at the start of the 1990s. This implies more breaches were uncovered using VC speech than SS systems. The preliminary VC attack consisted of Harmonic and Noise Model (HNM) and HMM-based synthesis[33]. As opposed to VC, the SS gained momentum only after developments in HMM-based synthesis[34]. The study in developing countermeasures for attacks began with prior knowledge of the attack that yielded biased results yet gave a kick start in developing algorithms. The work initiated with f0-based contours along with time stability to make distinction between genuine and spoofed speech[35]. The algorithm had a setback for capturing generality as there was scarce variation in the number of speakers. The visual cues such as images from video were a fine choice for representing speech through Mean Pitch Stability (MPS) and MPSr (range) along with jitter in[36]. In another approach Cosine Normalization Phase Spectrum (CosPhase) along with Modified Group Delay Function (MGDF) were proposed. The MFCC based features ignore the phase-related information during feature extraction; hence resulted in an increased EER. Furthermore, Magnitude Modulation (MM) and Phase Modulation (PM) super-vectors improved EER respectively when fused with MGDF features[37,38]. These short-term features produce a lower EER and lead to another inference that the long and short-term features produce related yet reciprocal content when fusion is performed. Having said that, the short-term features produce artefacts due to framing which is potential scope for improvement for speech researchers. Table 1 summarises various features and associated attack types while Table 2 lists the past five years research in the spoof detection area with regards to features along with the detection results.

Table 1. Various types of Attacks and their associatedfeatures

Type of Attack	Practical-ity	TI-ASV	TD-ASV	Features
Imperson-ation	Low	Low	Low	None
Speech Synthesis	Medium-to-high	High	High	CFCCIF, LFCC, CQCC, ICQC, Deep features, Scattering Cepstral Coefficients fundamental frequency, Phase based features: GD, MGD, PSP
Voice Conversion	Medium-to-high	High	High	
Replay Speech	High	High	Low (random phrase) High (fixed phrase)	MFCC, LFCC, PLP, CFCCIF, CQCC IMFCC CNN features Others: RFCC, SCMC, SSFC, VESA-IFCC

Table 2. Feature Extractiontechniques used in spoof detection

Author	Feature Set	Classifier	Results
Wu et al. 2016	MFCC, CosPh, MGD, PP, SMS	GMM-UBM	10.05 (FAR), 10.57(FAR), 8.62(FAR), 22.36(FAR), 26.41(FAR)
	UMS	SVM	19.47(FAR)
	Fusion		7.69(FAR)
Kamble and Patil 2017	ESA-IFCC	GMM	6.79
	MFCC		9.15
	ESA-IFCC + MFCC		7.16
Paul et al. 2017	MFCC, IMFCC, LFCC, MFCC+CFCCIF, Sub-band, SFCC, MOBT, SOBT, ISFCC, IMOBT, ISOBT	GMM	1.99, 0.95, 0.92, 1.21, 1.33, 1.05, 1.85, 2.49, 0.86, 1.46, 1.69
Kavya S. 2018	Mean, Variance, Log Spectrum	Siamese NN and LCNN	6.40
Rahmeni et al. 2019	IAIF	SVM	0.8635
	IAIF	ELM	0.8407
	MFCC	SVM	0.9329
	MFCC	ELM	0.5135
Phapatanaburi et al. 2019	LPR-RP + RP+ CQCC	GMM	9.26
Kumar et al. 2019	CQCC + IMFCC + LFBE (x-vector)	DNN	6.14
Volkova 2019	FFT Spectrograms	LCNN	5.5
Halpern 2020	CQT	ResNet	2.63
Chintha 2020	CQCC	CRNN	4.02

The developments in known attacks are not in ample for real-time scenarios and open up doors for building algorithms that can detect spoofed speech irrespective of the attack type. One such study started with the development of the SAS dataset and ASV Spoof 2015 challenge[11,39]. Following which, the Local Binary Patterns (LBP)-DCT for computing the MGDF and CosPhase[40], Magnitude features such as LMS, RLMS, and phase features like GD, MGD, Instantaneous Frequency (IF), Baseband Phase Difference (BPD), Pitch Synchronous Phase (PSP)[41] are also found to perform well in spoof detection task. Apart from phase based features, the Linear Prediction Coefficients (LPC) and its residual (LPR) are also considered for detecting known attacks[42]. Some distinct feature sets like Cross-Teager Cepstral Co-efficient (TECC)[43], Energy Separation[44] and Time-frequency based LFCC[45] are amongst novel representation techniques.

The heterogeneity in feature sets has thrown researchers challenges and further training/testing of these features is supposed to be performed through appropriate speaker modelling techniques. The i-vectors approach is a breakthrough in the speaker verification scenario; hence it has been proposed for spoof detection scenario with filter-bank based features and Deep Neural Networks (DNN)[46,47]. Furthermore, a comparative study using the Mel Wavelet Packet Coefficients (MWPC) was conducted to investigate Support Vector Machines (SVM) and Deep Belief Networks (DBN). The SVM performed better than the DBN[48]. Various ensemble based approaches have also been proposed in the literature paving way for research in Deep Learning area[49–52].

3 Database for ASV

Primitively for the development of anti-spoofing and spoofing measures in ASV, we need to decide the dataset based on our goals and specifications. The following section describes various datasets used in ASV-system development. The corpora are labelled as licensed and open source for ease of understanding and requirements in this research, as summarised in Table 3 and Table 4 respectively.

Table 3. Licensed dataset for Automatic SpeakerVerification

Type of Attack	Dataset	#bonafide speech	#spoofed speech	Number of speakers	Language
Impersonation	YOHO	5520	-	138 (106M, 32F), 2 naive	English
Voice Mimicry	NIST	-	-	5, 1 professional	Finnish
	WSJ	157000	-	284	English
	WSJ	157000	-	284	English
VC	NIST-SRE 2006	1570/ 3978	20561/ 2782	504	Many
VC	NIST-SRE 2006	1570	20561	-	English
VC	NIST-SRE 2006	1570	20561	-	English
VC, SS, Synthetic spoof	NIST-SRE 2006	1344	12648	298	English
SS, Replay, VC	BioCPqD-PA	7941	114111	222 (124M, 98F)	Portuguese

Table 4. Open-source datasets for Automatic Speaker Verification

Type of Attack	Dataset	#bonafide speech	#spoofed speech	#speakers	Language
SS, VC	SAS	33431	309592	106(45M, 61F)	English
Replay	RSR 2015	133243	-	300(157M, 143F)	English
VC and replay	RSR 2015	133243	-	300(157M, 143F)	English
VC and SS	ASV2015	9404	Kn92000, Unk92000	46(20M,26F)	English
SS, VC, Replay	AV Spoof	5576	LA20060, PA43320	44(31M, 13F)	English
SS, Replay, VC	VoicePA	5576	129988	16	English
Replay	RedDots	2346	16067	62	English
Replay	ASV Spoof 2017	1298	12008	24	English
SS, Replay, VC	ASV Spoof 2019	71747	137457	20	English

3.1 Licensed / Proprietary Datasets

The YOHO dataset has large utterances with office space recordings but lacks variations pertaining to vocabulary[53]. The dataset offers more number speakers with 106 Male while 32 Female speaker voices. There were 24 utterances in 4 sessions each. The utterances are low pass filtered at 3.4kHz and up sampled to 8kHz comprising 5500 utterances in all. Contrarily, the WSJ is a multi-speaker dataset but not created for ASV, as was the case for YOHO. As the speaker variability and size is large, it may be entitled for producing new synthetic speech and then treating the original corpus as genuine speech samples. The work[36] used the corpus SI-284 of the WSJ (WSJ0 and WSJ1) for the synthetic speech generation which comprises 81 hours of recording for 284 speakers.

The NIST-SRE is a speaker recognition corpus developed by a joint collaboration of NIST and LDC. There are multiple speakers with conversational telephonic speech[54]. BioCPqD-PA is a proprietary database, that has 222 speakers recorded in the Portuguese language. The dataset is known to be versatile as a result of variations in recording environments. It comprises in all 27,253 samples with 7,941 evaluation samples while another condition in which 3,91,678 spoofed samples are present amongst which 1,14,111 are evaluation samples[55].

3.2 Open Source datasets

The SAS (Speaker Verification and Anti-spoofing) dataset is built from the VCTK corpus with 106 speakers sampled at 16kHz frequency. The corpus contains 22,831 natural speech as opposed to 2,03,592 spoofed speech. The VC and SS are the source of spoofed speech[39].

The RSR 2015 corpus is a text-dependent Speaker Recognition database with 151 hours and 30 minutes of speech recorded in English. There are nearly 300 speakers and 1,96,844 utterances segregated for development and evaluation motives[56,57].

Furthermore, the very initial dataset built for the sole purpose of boasting anti-spoofing development started with ASV Spoof 2015. The dataset has 1,93,404 spoofed samples generated using VC, and SS(known attacks, LA) while 9,404 genuine samples. Additionally, the corpus has unknown attack-based spoofed samples for developing attack independent algorithms[28]. The AV Spoof corpus was developed as a major part of the BTAS 2016 Challenge[58]. The dataset includes various presentation attacks in particular VC and SS-based; with 20,060 LA spoofed samples and 43,320 PA spoofed samples. There are 5,578 natural speaker samples.

The Voice Presentation Attack corpus has emerged from AV Spoof corpus only for genuine samples. VoicePA dataset contains replay speech that is recorded using a laptop where speech is replayed using internal and external speakers. Moreover, there are replay speech samples present from iPhone and Samsung phone devices (internal speakers). There is broad range of spoofed utterances including 3,91,678 samples from which 1,14,111 samples are fixed for evaluation. Contrarily, the natural speech samples are 27,253 from which 7,941 samples are again reserved for assessment purposes[59].

The RedDots is a text-dependent replay speech corpus[60] that contains native and non-native 62 English speakers with small phrases recorded on multiple devices. The corpus contains 16,067 replayed spoofed samples while 2,346 natural samples. Since the dataset is designed by considering crowdsourcing during replaying and recording, it was entitled to be included in the ASV Spoof 2017 challenge.

After successfully conducting ASV Spoof 2015 challenge, the organizers refined and launched a new dataset ASV Spoof 2017 that was adapted from the RedDots replay dataset. There are 24 speakers, 1,298 genuine, and 12,008 spoofed samples[61] . The ReMASC (Realistic replay attack Microphone Array Speech Corpus) is a replay speech dataset that has been designed considering the voice-controlled device. There are 45,472 spoofed and 9,240 natural speech samples with 55 speakers. The recording areas include outdoors, inside the home, and in vehicles as well[62].

The ASV Spoof 2019 corpus is a third consecutive challenge for anti-spoofing measures development broadened to synthetic (VC, SS) and replay speech. There are 20 speakers, 71,747 LA, and 1,37,457 PA samples, with tandem-DCF introduced in addition to the other evaluation measures for the challenge (5,64).

4 Feature Representation for Spoof Detection

The voice signal when sampled and stored in digitized data form contains alot of application-independent content which may not be required for performing the dedicated task of ASV. Thus, the speech signal is represented using appropriate features through framing windowing and conversion. This conversion operation may be time, spectral, cepstral, or some form of filtering to reduce the redundant contents. Taking this into consideration, speech features may be categorized as low-level (or short-term), long-term (or prosodic), and high-level features. Additionally, there are deep features that obtained from using DNN as a feature extraction scheme. The low-level or short-term features are linked with a speaker's timbre. The aim here is to capture local information within a frame of 20-40ms. So the spectral representation parameters like MFCC, LPCC and Cochlear Model contributing to glottal parameters may be categorised as short-term features when extracted over the defined frame duration. These kind of features are synthesizable with simplicity and hence are more susceptible to spoofing attacks (65,66). The long-term or prosody-based parameters are linked to human-like auditory traits including pitch and duration of the speaker, intonation, and Constant Q-transform Cepstral Coefficients (CQCC) (45,52,67). The prosodic parameters are obtained from long segments of spoken speech like words and syllables representing speaking rate, style, and intonation levels. These features are less likely to channel distortions yet the training process requires a larger data size (68). Furthermore, the algorithms involving the extraction of pitch do not perform well in noisy scenarios. Likewise, the high-level parameters that are obtained from lexicons to characterize behaviour of speaker and lexical cues. Phonemes and ideolect are also high-level features. These unique parameters are preferred over short-term and long-term parameters due to the fact that they are less affected by noise and channel distortions. Yet, there seems to be a possibility for exploring more as researchers are hesitating to apply these high-end features in standalone ASV because they need a high-front-end like speech recognition framework[63,64].

The features may be segregated based on the duration of the segment as sub-segmental, segmental, and supra-segmental. The speaker utterance is framed using a frame duration of 3 ms to 5 ms in the sub-segmental parameters. The research clarifies that source excitation parameters are extracted through sub-segmental features[65,66]. The segmental features involve the same framing as in sub-segmental features except for the duration of the frame is changed and increased to 10 to 30 ms with shifting of frames with overlap to maintain continuity and avoid loss of information due to sharp edges of frame boundaries. The speech as a signal is generally non-stationary in nature and yet observed to be stationary for 10 ms to 30 ms duration, hence the segmental parameters' frame duration is justified. So the vocal tract parameters are obtained via segmental parameters. Lastly, the supra-segmental parameters are extracted with 100 ms to 300 ms frame size. These features symbolise the behavioural traits of the speaker like accent, word duration, speaking style, etc(65,67). The popular features used in ASV and countermeasures are described below:

4.1 Mel Frequency Cepstral Coefficients (MFCC)

The MFCC based perceptual features are preferred in most speech processing frameworks such as automatic speaker recognition(68,69). The frequencies are transformed into Mel scale through the standard procedure of short-term representation using framing, windowing, and spectral transformation as portrayed in Figure 3. The real cepstrum is processed through a triangular filter bank with T^{th} order. The triangular filtering is used to average out the centre frequency energies. The Mel scale is known to have linear spacing for lower frequencies while the logarithmic distribution for higher frequencies. The mel frequency f_{MEL} is computed as

$$f_{MEL} = 2595log\left(1 + fr/700\right) \tag{1}$$

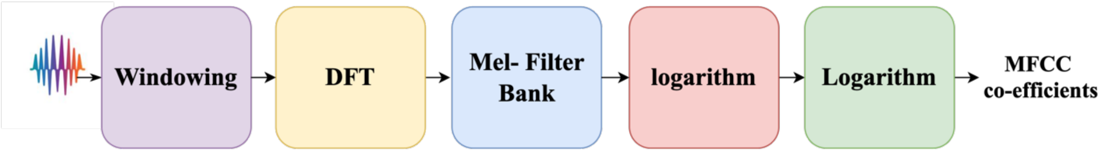

Fig 3. Feature Extraction block schematic for MFCC

Thus, the MFCC coefficient C_h is given as

$$C_h = \sum_{m=1}^{T}\left[\log X\left(m\right)\right]\cos\left[\pi h/T\left(m-0.5\right)\right] \tag{2}$$

Where h is the cepstral coefficient index. The MFCC coefficients along with their first and second derivatives are usually included in the feature set. Another form of representation is through an inverse mel filter bank for Inverse MFCC coefficients that represent high-frequency regions efficiently(51).

4.2 Mel-warped Overlap Block Transformation (MOBT)

Prior to speech processing, the MOBT parameters are efficient in capturing discriminative information provided by the formants . The filter-bank energies (MFLE) are segregated into overlapping and non-overlapping frames. Besides that, for computing the cepstral parameters, the filter-bank energies are block transformed, and the DCT of every block yields MOBT parameters as seen from Figure 4. Similar to IMFCC, inverse mel scale might be considered in place of Mel scale in MOBT to extract IMOBT parameters(70,71).

4.3 Speech-Signal-Based Frequency Cepstral Coefficient (SFCC)

For investigating the significant role of frequency content in the speech production model, warping of frequency is performed. A similar purpose is inculcated in SFCC features too(71). The input time-domain speech signal v(t) is firstly passed through STFT followed by power spectral density operation for every frame is computed which is given as

$$P\left(i,w\right) = 1/N \vee V\left(i,w\right)^2 \tag{3}$$

Here,N is the total samples for one window. When $P\left(i,w\right)$ is averaged over entire speech data, the ensemble energy $P\left(w\right)$ is calculated along with log function and lastly, distributed in a manner as below

$$Aj = \int_{w_j}^{w_{j+1}} \log \bar{P}(w)dw \text{ and } j = 1,\ldots\ldots,P \tag{4}$$

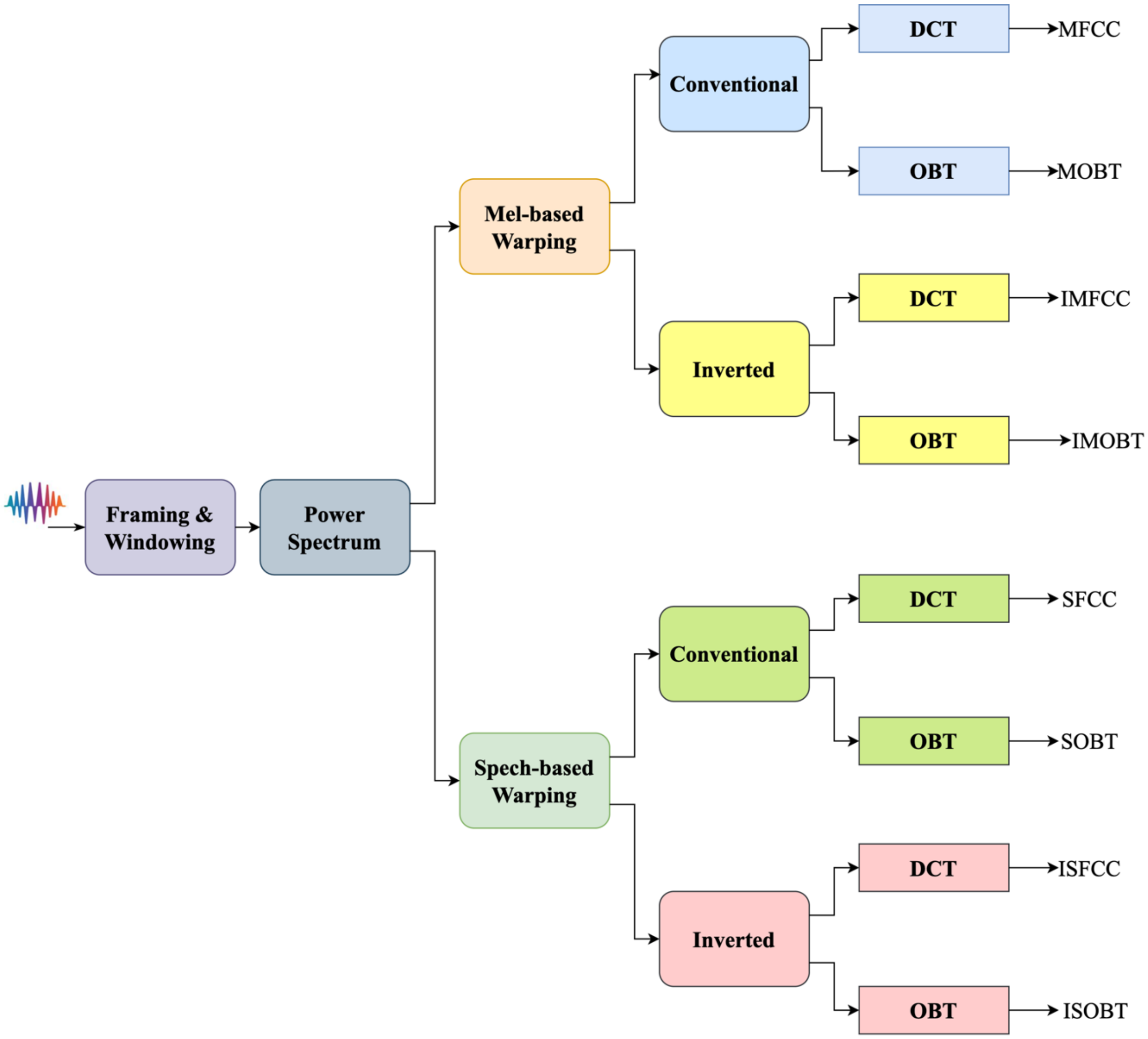

Fig 4. Feature Extraction block schematic for MFCC, MOBT, IMFCC, IMOBT, SFCC, SOBT, ISFCC, ISOBT techniques

$$A_j = A_{j+1}, \quad j = 1,\ldots\ldots,P-1 \tag{5}$$

Where, A_j is the j^{th} area interval, w_j is lower cut-off and, $w_{j+1}$is the higher cut-off frequencies. The P point speech-based frequency warping is given as

$$F\left[\left(w_j + w_{j+1}\right)/2\right] = j/P, \quad j = 1,\ldots\ldots,L \tag{6}$$

Where, $F\left(w\right)$ is a continuous function when P approaches infinity and lie within 1.

The frequency warping helps convert spectral to the cepstral domain to get a triangular filter. The ISFCC is product of inverse warping operation as against SFCC(71). Furthermore, the SOBT parameters are produced as a combined effort of MOBT and SFCC while the inverse would yield ISOBT(71) as seen in Figure 4.

4.4 Linear Prediction Cepstral Coefficients (LPCC)

$$s\left(k\right) = a_1 s\left(k-1\right) + a_2 s\left(k-2\right) + \ldots. + a_j s\left(k-j\right) \tag{7}$$

The LPC is simple to compute in order to have low computations[72]. The speech production model is represented using Auto-regressive Moving Average (ARMA). The LPC model employs all-pole filters through the prediction of the k^{th} speech sample using linear combination of previous j samples.

$$e(k) = s(k) - \hat{s}(k) = s(k) - \sum_{m=1}^{j} a_m s(k-m) \tag{8}$$

Where $a_1, a_2 \ldots, a_j$ are the LPC parameters for every individual frame. Thus, error during prediction error is computed as

Where $s(k)$ is the original speech sample and $\widehat{s}(k)$ is the predicted sample. Further, the square of error is calculated to obtain unique coefficients,

$$Ek = \sum_p \left[s_k(k) - \sum_{m=1}^{j} a_m s_k(p-m) \right]^2 \tag{9}$$

The p stands for total samples in one analysis frame. To computer LPCC features, the squared error is differentiated wrt to LPC coefficients or filter weights as shown in Figure 5.

$$\delta E_m / \delta a_m = 0 \tag{10}$$

Therefore, the cepstral coefficients are

$$C_0 = log_e(j) \tag{11}$$

$$C_p = a_p + \sum m = 1^{P-1} m/p\, C_m a_{p-m} \ldots for 1 < p < j \tag{12}$$

$$C_p = \sum_{m=p-j}^{p-1} m/p C_m a_{p-m} \tag{13}$$

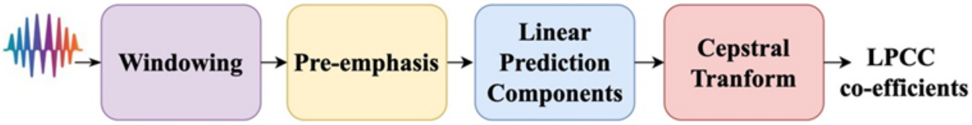

Fig 5. Feature extraction using LPCC technique

4.5 Cochlear filter cepstral coefficients (CFCC)

The conventional feature extraction techniques involve pre-processing such as windowing, framing, and low pass filtering. Contrastingly, human speech production does not operate on these pre-processing principles. Moreover, the framing and windowing operations surely introduce artefacts and discontinuities in the processed speech in contrast to the actual raw speech model. Hence, distinction of natural speech over a synthetic speech that is pre-processed becomes easier and will have the effects of distortions seen in their spectral response. One such feature is CFCC[73] that builds on the base of auditory cepstral coefficients and the block schematic is depicted in Figure 6.

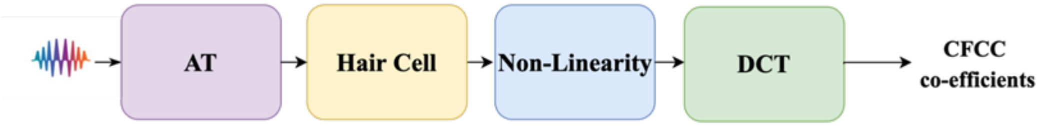

Fig 6. Feature extraction using CFCC technique

So, the speech $x(t)$ is processed by Auditory Transform (AT) producing traveling waves $W(a,b)$inside the Basilar Membrane given as

$$W(m,n) = x(t) * \psi_{m,n}(t)\, where \psi_{m,n}(t) = 1/\sqrt{m}\psi_{m,n}((t-n)/m] \tag{14}$$

The $x(t)$and $\psi(t)$ belong to Hilbert space. The factor m is scaling parameter while n is the time shift parameter, with energy that remains the same for all values of m and n given as

$$\int_{-\infty}^{\infty} \vee \psi_{m,n}(t)^{2} dt = \int_{-\infty}^{\infty} \vee \psi(t)^{2} dt \tag{15}$$

The cochlear filter is presented as

$$\psi_{m,n}(t) = 1/\sqrt{m}\left((t-m)/n^{\alpha} e^{(-2\pi(f_L\beta)\{(t-m)/n\}]} cos\,(2\pi(f_L)(t-m)/n] + \theta\right] u(t-n) \tag{16}$$

The α and β define the shape and width of the cochlear filter respectively. The value for θ must satisfy the admissibility property of mother wavelet $\psi(t)$,

$$\int_{-\infty}^{\infty} \psi(t)\,dt = 0 \rightarrow \psi(w)_{w=0} = 0 \tag{17}$$

Specifically, there exists a natural number C_ψ such that $C_\psi = \int_0^\infty \vee \psi(w) \vee wdw < \infty$ implying that wavelet $\psi(t)$is a Bandpass Cochlear filter with the lowest frequency $f_L$and center frequency f_C.

$$m = f_L/f_C \tag{18}$$

In specific sub-band filter, k^{th} sub-band filter, the value of m should be available in advance for a specific center frequency of cochlear filter at $k \in (1,28]$. After Cochlear filters the frequencies, the hair cell behaves like transducers to promote the vibration of BM. The hair cell vibrates in a positive direction only, hence -

$$h(m,n) = (W(m,n))^2 \forall W(m,n) \tag{19}$$

The output from the hair cell gets transformed into nerve spike density representation, which is given as,

$$s(a,b) = 1/g \sum_{n=q}^{q+g-1} h(a,n) \quad q = 1, Q, 2Q, \ldots \forall a,b \tag{20}$$

Where window length is g and window shift duration is Q. From the above function, the output obtained is further passed through the cube root and followed by the DCT function.

4.6 Cochlear Filter Cepstral Coefficients with Instantaneous Frequency (CFCCIF)

The CFCC-IF features were first applied in ASV through ASV Spoof 2015 Challenge[74]. The CFCC along with Instantaneous Frequency (IF) together builds up the CFCC-IF coefficients. The CFCC features are based on wavelet transform that utilizes the AT, Hair cell, and Nerve spike density computation. The product of Nerve spike density and IF is differentiated followed by non-linear log operation Int the end, DCT is applied to de-correlate the parameters producing CFCC-IF parameters as shown in Figure 7.

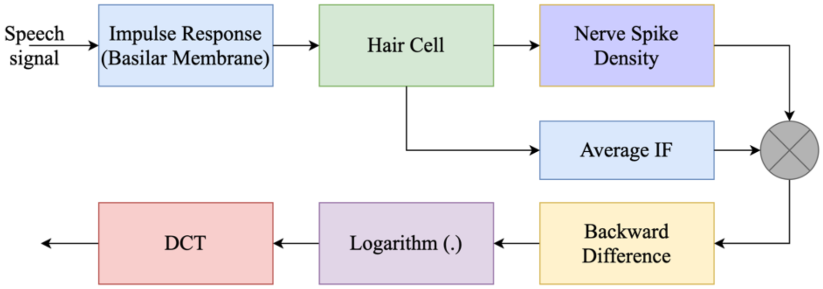

Fig 7. Feature extraction using CFCC-IF technique

4.7 Constant-Q Cepstral Coefficients (CQCC)

The CQCC was successfully utilized in anti-spoofing by[75]. Like STFT, the CQCC is known to produce conjointly time-frequency variations. The important highlight of CQCC is high-frequency resolution at low frequencies and high-time based resolution for high frequencies. The spectral response obtained because of Constant QT is then processed through a non-linear logarithmic scale and then linearized by the Constant QT scale. Yet again, DCT is applied for producing the CQCC parameters as portrayed in Figure 8.

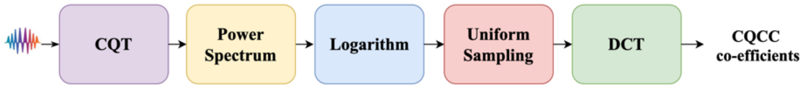

Fig 8. Feature extraction using CQCC technique

4.8 Magnitude based features

The time-domain speech signal is difficult to process and visually gives no clue of frequency contents. To do so, the STFT representation of speech yields magnitude and not to mention the phase contents too. On the whole, the spectral contents help process the data better. Thus, the magnitude-based features hold quite some weightage while detecting spoofed speech. The STFT of speech utterance is given as

$$Z(t,w) = S(t,w) \vee e^{j\theta(t,w)} \tag{21}$$

Here, $S(t,w) \vee$ signifies magnitude-related content while $\theta(t,w)$ holds the phase contents. The Log-Magnitude Spectrum (LMS) and Residual LMS (R-LMS) are worthy to detect spoofing attacks. LMS parameters are derived by the simple process of computing logarithmic of magnitude spectrum obtained because of STFT which is given as

$$M(w) = log \vee Z(t,w) \vee \tag{22}$$

Therefore, it may be confirmed that the LMS features hold crucial magnitude contents such as formants, pitch, and specifically the harmonics present in the vowel spectrum. Also, the logarithmic operation limits the dynamics of the speech spectrum[76].

Furthermore, the R-LMS features are well established in speech recognition frameworks but still are not much explored in an anti-spoofing environment. Moreover, the synthetic speech from VC or TTS algorithms represents formants quite well. So, using formants to differentiate between speakers is tough. Furthermore, the LPC technique is popular for representing formants well enough, but the residual part has no presence of formants. The R-LMS parameters are obtained using the LMS algorithm on the LP-residual (LPR) signal[76].

4.9 Phase based features

The STFT representation of speech produces magnitude-related and phase-related spectrums. The phase contents from the speech are perceptually indistinguishable. Still when differentiating between speakers, the smallest of parameter counts; hence phase too is a potential choice for spoof detection. So, the phase-associated features including Group Delay (GD), Modified GD (MGD), Instantaneous Frequency (IF), Pitch Synchronous Phase (PSP), and Baseband Phase-Difference (BPD), are used by researchers for anti-spoofing[77,78]. Additionally, the phase spectrum is considered unstable making pattern matching tedious due to phase warping. So, phase spectrums are modified further to benefit from them in anti-spoofing scenario. The GD function of phase spectrum is obtained by computing derivation wrt frequency and is given below

$$G(w) = princ(\theta(w) - \theta(w-1)) \tag{23}$$

Where, the princ(.) functions maps the phase spectrum to $(-\pi,\pi)$. Despite its abilities to extract pitch and formants efficiently from speech, the standard GD function lacks in grasping short-time spectral contents as zeros are existing in z-plane which are nearer to the unit circle. Hence, the MGD function was introduced to overcome the shortcoming of the GD function[77]. The MGD parameters are computed as

$$D(w) = T(w)/T(w) \vee T(w)^{\alpha} \tag{24}$$

$$T\left(w\right) = S_{real}\left(w\right) C_{real}\left(w\right) + S_{imag}\left(w\right) C_{imag}(w)/M\left(w\right)^{2\gamma} \tag{25}$$

Here, the α and γ are meant to fine-tune the function, M(w) is the smoothened Z(w), while Z(w) is the complex speech spectrum. The derivation of phase spectrum wrt frequency yields GD parameters while the derivative wrt to time axis produces a different parameter called Instantaneous Frequency (IF)[(79)]. The IF can be computed as

$$F\left(t,w\right) = princ\left(\theta\left(t,w\right) - \theta\left(t,w-1\right)\right) \tag{26}$$

So, it may be inferred that GD and IF provide complimentary contents which may be useful for spoof detection. Furthermore, BPD parameters are more steady time derivative phase parameters that are also computed to support spoof detection studies[(70)]. The BPD parameters are given as

$$B\left(w\right) = princ\left(F\left(t,w\right) - \Omega_t P\right) \tag{27}$$

Here, P is frame shift expressed using total samples, Ω_t is frequency which is constant and is equal to $2\pi p/L$, FFT length is L. Additionally, the PSP is another choice when computing phase parameters. The speech signal consists of periodic and non-periodic signals. The periodic part is usually computed from fixed frame size while with regards to PSP parameters, it is extracted using pitch instances. The Glottal Closure instants (GCI) are important for deciding the start and end of the pitch period[(77)]. The algorithm begins with one pitch period preset and keeps updating from consecutive pitch periods. Another phase parameter is Cosine Normalization Function or more commonly addressed as Cosine Normalized Phase (CosPhase)[(21)]. Below are the steps for computing CosPhase parameters

1. Unwarp the phase spectrum
2. Compute cosine function to the spectrum obtained in step (i). This normalizes the function to -1 and +1.
3. Lastly, apply DCT after normalizing the function from step (ii). Choose initial eighteen parameters with their $\triangle$ and $\triangle^2$.

4.10 Miscellaneous features

Besides the categorical parameters which have a certain rigid way of classification like phase, magnitude, human speech production, or perceptual model; there exist other features that might not fit in the given categories but are specifically developed for spoof detection. So, indirectly they are handcrafted for dedicated tasks and may be fused to benefit from their combination rather than individual shortcomings. These parameters are Perceptual Linear Prediction (PLP), Rectangular Filter Cepstral-Co-efficients (RFCCs), Spectral Centroid Magnitude-Co-efficients (SCMC), Sub-band Spectral Flux Co-efficient (SSFC), and Variable length Teager energy operator energy separation algorithm- instantaneous frequency cosine coefficients (VESA-IFCC) (86,87).

4.11 Critical Evaluation of Feature Representation in ASV

The features are the key to any speech application because the manner in which raw speech frames are represented affects the performance of that application. Therefore, knowing the salient qualities of a feature set helps in choosing the right feature. The MFCC based features have established popularity amongst the entire speech and audio community because of their ability to represent human response accurately. Despite that, these features do not consider phase when extracting parameters from the speech[(80)]. Also, the synthetic speech production algorithms (TTS or VC) usually ignore the phase as it is imperceivable. This led to research on phase-based features in addition to magnitude features such as LMS, Residual-LMS, GD, MGD, IFD, BPD, and PSP. Furthermore, the VC speech that uses mel-frequency warping highlights the lower frequency regions as against the high-frequency regions. While all this time, the high-frequency components held vital speaker traits that contribute to differentiate synthetic speech. Thus, the long-term features are found to be more effective than short-term features[(81)], not to forget the CQCC and CFCCIF.

The sub-band-based not limited to LFCC and ESA-IFCC, gained importance since the artefacts in synthetic speech are spread across various sub-bands. The temporal features such as IF and magnitude envelope capture these artefacts. Additionally, the wavelet filter banks are also explored to represent scalograms. The conventional LP features are also explored lately as they represent spectral peaks more accurately than valleys[(82)].

The prosodic features represent the accent and speaker ques, yet they are easy to reproduce and hence susceptible to attacks. Also, there is a demand for a larger data size for training to extract prosody from speech.

Furthermore, the pitch extraction techniques do not perform up to the mark in noisy environments[11]. On the contrary, the high-level features are more reliable since they are less sensitive when exposed to variations in channel and noise. This is opposite to prosodic and spectral parameters (70,90).

5 Machine Learning techniques

The Machine learning algorithms either govern the pattern classification or learning of features. After feature representation, the statistical models train using significant features to further prepare for testing. The test sample may belong to a known/ unknown attack or is simply a genuine speaker. The task is complex but made solvable using efficient machine learning schemes for speaker modeling, and also decision-making tasks.

The speaker models may be subdivided into generative, discriminative, and mixed/fused approaches. The generative models comprise of Gaussian Mixture Models (GMM)[83,84], i-vectors[36] Vector Quantization (VQ)[21], and Hidden Markov Models (HMM) (93) while discriminative models include the Support Vector Machines (SVM) (94), deep learning, and neural networks[85,86]. The increasing developments in utilizing DNN in the speaker verification scenario are owing to accurate results and of course, their ability to discriminate between speakers[46]. Consider an unknown test utterance T that is claiming to be speaker A, then building a hypothesis for determining the class of the utterance.

$$H_A : T\ is\ uttered\ by\ A \tag{28}$$

$$\overline{H}_A : T\ is\ not\ uttered\ by\ A \tag{29}$$

The efficiency of speaker verification depends greatly on the model building; thus, the appropriate model choice will further improve results. The below sub-sections describe commonly used modeling algorithms that lead to a reduction in EER implying better anti-spoofing techniques[51].

5.1 Vector Quantization (VQ) technique

The VQ-based codebook mapping technique is suitable for text-dependent scenarios. When the training phase begins, the codebook is built through clustering techniques[87]. The clustering algorithm averages out the temporal information present in codebook. Moreover, there is no requirement for temporal alignment. The input vector is compared against every codeword from the codebook. The code word that has least distance is selected as a matched pair[37]. One such approximation technique is the nearest neighbor's algorithm which performs better than Dynamic Time Warping (DTW) and VQ[87]. As opposed to VQ technique, the nearest neighbor algorithm considers temporal content. The input frame is compared to past frames forming a distance-based inter-frame matrix. The nearest neighbor is one with the lowest distance amongst input and past frame. The match score produced is the average of distance for input frames. The match scores together give the log-likelihood ratio approximation.

5.2 Gaussian Mixture Models (GMM)

The GMM models are popular for generalization and assume that the nature of input data is Gaussian[70]. The individual gaussian has an associated mean, standard deviation, and feature vectors are multiple gaussians called mixture of gaussians. The GMM is represented using output probability function for X feature vector,

$$p\left(X/\lambda_k\right) = \sum_{m=1}^{G} w_m d_m \tag{30}$$

Where λ_k is a weighted sum of G components for k$^{\text{th}}$ speaker, $w_m$is the mixture weight with $\sum_{m=1}^{G} w_m = 1$ while the d_m is density for individual components, and a K-variate GMM gaussian function is given as

$$d_m = 1/\left[(2\pi)^{K/2}\left|\sum_m\right|^{1/2}\right] e^{-1/2(X-\mu_m)'\Sigma m^{-1}(X-\mu_m)} \tag{31}$$

Where $\mu_m$is a mean vector of the dimension (K) while $\sum_m$ signifies covariance matrix, with dimension $(K \times K)$. The GMM for a speaker k, λ_k is given as

$$\lambda_k = \mu_m, \sum_m, w_m \wedge m = 1,2,...,G \tag{32}$$

The criteria for selecting total mixtures are dependent on the kind of language, such as English language which has 45 phones. Thus, total mixtures selected must be higher than the total-phones so 64 is an appropriate measure. The mean, deviation, and mixture weights are obtained through parameter estimation including Expectation-Maximization (EM), Maximum Likelihood (ML), and Maximum A-Posteriori criteria (MAP)[3]. Out of these, the ML algorithm is quite common. Suppose for a given utterance$T = t_1, t_2, \ldots, t_N$, target model λ_{target} and imposter model $\lambda_{imposter}$, the ML ratio is given as

$$\frac{Pr\,(T \in target)}{Pr\,(T \in imposter)} = \frac{Pr\,(\lambda_{target} \vee T)}{Pr\,(\lambda_{imposter} \vee T)} \tag{33}$$

Thus, in logarithmic scale, Bayes Rule is given as,

$$\Lambda\,(T) = logp\,(T \vee \lambda_{target}) - logp\,(T \vee \lambda_{imposter}) \tag{34}$$

The prior probabilities are ignored as they are constant. The likelihood ratio is compared to a threshold φ .

$$\Lambda\,(T) \geq \varphi \rightarrow Acceptedastarget \tag{35}$$

$$\Lambda\,(T) < \varphi \rightarrow Rejectedastarget \tag{36}$$

The likelihood ratio is a score obtained by a fair comparison of the target model to the imposter model. The value of the threshold is updated regularly depending on these scores to steer clear of any false positives and negatives. The likelihood of sample belonging to the target is given as

$$logp\,(T \vee \lambda_{target}) = \frac{1}{N}\sum logp\,(t_n \overset{N}{\underset{n=1}{\vee}} \lambda_{target}) \tag{37}$$

5.3 GMM and Universal Background Models (UBM)

The ASV system acknowledges the test sample as known when the score obtained equals to or is greater than the set threshold. There is another model built which is the imposter model using GMM also called UBM. The UBM is known to represent any claimed speaker's identity efficiently. The data required for training a UBM is large which contributes to better parameter estimation. Subsequently, the number of components also increases (as against a single GMM like above 256). The motive of building a UBM is to reduce the speaker dependency i.e., speaker-independent features distributed speaker's data[3,84,88]. Along with imposter model training, the UBM overcomes the issue of training all the GMM in case of a new addition to the data. Only UBM is trained when new data is added and not the individual GMMs. Assume a UBM model with feature $F = F_1, F_2, \ldots., F_n$ wrt to a specific speaker. At i^{th}instant, with c components, to train this feature vector into the UBM, the probabilistic alignment is given as

$$Pr\,(c \vee F_i) = \frac{w_c p_c\,(F_i)}{\sum_{k=1}^{K} w_k p_k} \tag{38}$$

Here, $p_c\,(F_i)$ is probability density function for i^{th} feature vector with w_c is mixture weight. Thus mean, variance, and mixture weights are calculated as

$$\eta_c = \sum_{i=1}^{I} Pr\,(c \vee F_i)\, c^{th} weight \tag{39}$$

$$E_c\,(F) = \frac{1}{\eta_c}\sum_{i=1}^{I} Pr\,(c \vee F_i)\, F_i c^{th} mean \tag{40}$$

$$E_c\,\left(F^2\right) = \frac{1}{\eta_c}\sum_{i=1}^{I} Pr\,(c \vee F_i)\, F_i^2 c^{th} variance \tag{41}$$

The calculated mean, variance, and mixture weights are further used to update the previous values of UBM for the c^{th} component.

$$\widehat{w}_c = \left[\frac{\alpha_c w_{\eta_c}}{I} + \left(1 - \alpha_c^w\right) w_c\right] \gamma \tag{42}$$

$$\widehat{\mu}_c = \left[\alpha_c^m E_c\left(F\right) + \left(1 - \alpha_c^m\right) w_c\right] \mu_c \tag{43}$$

$$\sigma_c^2 = \alpha_c^v E_c\left(F^2\right) + \left(1 - \alpha_c^v\right)\left(\sigma_c^2 + \mu_c^2\right) - \widehat{\mu}_c^2 \tag{44}$$

Here, γ is scaling parameter for maintaining summation to unity, $\left(\alpha_c^w, \alpha_c^m, \alpha_c^v\right)$ are adaptation parameters that remove any mismatches between past and currently estimated parameters. The mean estimation involves setting weight and variance to zero.

5.4 Support Vector Machines (SVM)

The clubbing of generative and discriminative models is an interesting alliance in ASV research[89]. The GMM-UBM framework produces a speaker template that is fed to the SVM and SVM being a natural solution to ASV problem discriminates between speakers efficiently[89]. This framework implies that SVM is an appropriate choice for binary classification tasks.

The GMM-UBM framework assumes a diagonal covariance matrix and MAP is employed for training. The resultant adapted model is a stacked version comprising of mean with K dimensions called super vectors and K is the total mixtures. So, this GMM based super vector is a mapping function between speech utterances and the higher dimension matrix. Thus, these super vectors are treated as SVM features and the unknown sample is detected as

$$u\left(x\right) = \sum_{l=1}^{M} \alpha_l t_l G\left(x, v_l\right) + b \tag{45}$$

where t_l is the required output which might be equal to a positive one for acceptance and a negative one for imposter. The support vector is v_l, b is the learning constant, $\sum_{l=1}^{M} \alpha_l t_l$ is ideally zero and $\alpha_l > 0$. The kernel is given as

$$G\left(x, v_l\right) = m \tag{46}$$

Here, the mapping function is $m\left(.\right)$ that converts input features to super vectors, distance measure $u\left(x\right)$ separates hyperplane, and its polarity shows the category of the unknown sample. Regarding the x super vectors, the predicted label as zero, points to negative class while one signifies positive class.

5.5 Joint Factor Analysis (JFA)

The factors responsible for the efficiency of the GMM-UBM alliance are two-fold: the first being speaker variations and the second is session variability i.e. from training to testing[90]. These issues are solved by building models of individual speakers and channel distortions as is the case for JFA. The algorithm branches out GMM super vectors V, into individual speaker-dependent super vectors i and channel-dependent super vector, c

$$V = i + c \tag{47}$$

where, $i = j + Kf + Dg$ and $c = Nh$, K is a low-rank speaker variability matrix, D is a diagonal variability matrix that models residual variability that cannot be captured by the speaker, f and g are speaker factors and residuals respectively. The low-rank channel variability matrix is N and h is a channel factor vector. During the training stage, the GMM super vectors are obtained by JFA based training, and channel-dependent information is not considered. On the contrary in the testing stage, channel-dependent information is acquired from test utterance and the obtained super vector is ranked using linear dot product[87].

5.6 I-vectors

The JFA technique causes system degradation relating to performance due to loss of channel-dependent content being ignored during the training stage[91]. To get rid of this problem, i-vectors were introduced[4,16]. The i-vectors are known to use single variability space for GMM super vectors and are represented as

$$\mu = j + Bw \tag{48}$$

Where, j is the same super vector used in JFA, B is a lower-rank variability matrix for entire training data and w is the total variability factor. The cosine similarity score (CSS) gives the angular difference between i-vector w_{Test} and target i-vector, w_{Target} for classification objectives.

$$\text{score}\left(w_{\text{Target}}, w_{\text{Test}}\right) = \frac{< w_{\text{Target}}, w_{\text{Test}} >}{\left\|w_{\text{Target}}\right\| \left\|w_{\text{Test}}\right\|} \tag{49}$$

5.7 Hidden Markov Models (HMM)

The HMM comprise of a hidden stochastic process using an observation sequence[35]. The arcs and chain form a markov chain where arcs direct to the transitional probabilities that connect one state to another. The HMM differs from Markov chain with a slight variation of hidden state while state and transitional probabilities are already known in the Markov chain's case[92]. The conventional GMM is an obvious choice when building an ASV state-of-the-art model which does not take into account the temporal information present in the features. Along with temporal contents, linguistic information is also ignored when building a phone-based GMM[93]. For processing temporal information, HMM is considered. The HMM performs well in text-dependent scenarios while GMM takes a lead in text-independent tasks[3].

5.8 Multi-layer Perceptron (MLP)

The basic Feedforward Neural Network (FNN) is also termed a Multi-layer Perceptron (MLP) that uses back-propagation to train its weights. Usually, they perform binary classification when applied to ASV for clear distinction amongst known to unknown speakers. The construction of MLP is simple with nodes in every layer and multiple layers with interconnected nodes. The input to the nodes is utilized to calculate the weighted sum while the transfer function gives results of output nodes. The gradient descent algorithm is utilized to determine weights using back propagation. For an ASV, the MLP discriminates between the imposter and genuine speaker by computing a score from every frame of test utterance[94].

5.9 Deep Neural Networks (DNN) for ASV

The DNN is the new hope in not only the decision-making and speaker modeling but also as bottleneck features. In other words, for end-to-end ASV, DNNs act as features. The speaker representation is influenced by the speaker model, representation level, and loss function during training.

Furthermore, the DNN bottleneck features preliminarily acquire the speaker-specific information at frame level followed by utterance level. The DNN output i.e., DNN features are modified into i-vectors and at last, PLDA is employed to determine the verification score[95]. So, it is harmless to treat DNN alone as a feature representation technique or clubbing it with other existing features such as MFCC. The commendable performance by DNN is due to the reason that GMM does not predict the phonetic contents in the text-dependent scenarios[96]. More studies based on DNN can be accessed from[25,51,97].

5.10 Convolution Neural Network (CNN)

The conventional Feed-forward NN (FNN) has a similar layout as the CNNs. The components are identical like the weights, bias, non-linear conversion function; still, there seems to be a difference in local connectivity[51]. The FNN has connectivity between all the layers with the input nodes while the CNN comprises small filters that cover the entire input that gathers the summation of the result. This is the basis of convolution operation[70,98]. Thus, the layers in a CNN are a sandwich of convolution, max pooling, and fully connected layers. An extension to standard CNN layers, the addition of Max Feature Map (MFM) layer is used in Light CNN to boost in selecting local features. On the whole, variations in Convolution (Conv), MFM and Max pooling (Max) layer build various light CNN architectures such as AlexNet has 4 Conv, MFM and 4 Max, VGG has 5 Conv, 4 Network-In-Network (NIN)+ MFM, and 4 Max while Residual CNN has 5 Conv, Residual Block - 2 convolution, 2MFM with no batch normalization and 4 Max layers.

5.11 Recurrent Neural Network (RNN)

The RNN is a sequence-based NN that considers weight estimation over a timestamp[77]. When the RNN is unfolded a DNN is obtained that consists of layers with time step. The weight matrix W_x(where x can be input, hidden, and output) and biases b_y (where y is input and/or output) for input $r_1, r_2, ..., r_L$ has output given as

$$h_l = \phi_z \left(W_{input}.r_l + W_{hidden}.h_{l-1} + b_{input}\right) \tag{50}$$

$$y_l = \phi_o \left(W_{output}.h_l + b_{output}\right) \tag{51}$$

The conventional RNN efficiently captures temporal contents in a single direction only, hence the Bidirectional RNN[99] is proposed.

5.12 Critical Evaluation of Machine Learning algorithms in ASV

The GMM has been widely chosen on the account of their commendable performance in ASV task. Yet again at the same time require larger datasets for training with moderate to a high quality of data posing difficulty in noisy environments. The urge for a large dataset can be addressed through a diagonal covariance matrix that subsides the computational intricacy as well. Another concern is centered around unknown data, the GMM is unable to capture non-linearities due to its generative nature contributing to a low classification score. This concern may be addressed through data segmentation into training, development, and testing labels. While the GMM-UBM framework is preferred as against individual GMM overcoming unknown data problem. As a consequence, UBM is trained on a larger dataset gradual increment in mixture number makes them robust to unknown data.

Furthermore, the GMM-SVM alliance has advantages of generative and discriminative models leading to high accuracy scores. Nonetheless, the MLP needs larger data for an optimized performance and longer training time to reach that milestone. Above all, the DNN are found to outperform all networks with competence to adapt as features and learning unknown data which is contrasting to generative models. The CNN is used where variability is observed in time while RNN is preferred in case of temporal data[100].

6 Score Normalization

The unnecessary variations in the score are stabilized using score normalization techniques. The operation of normalization is equivalent to thresholding in speaker dependent scenarios. The normalization techniques found in literature are based on a unique assumption that the imposter's score to have gaussian distribution. Thus, the mean μ_G and standard deviation σ_G are accustomed normalize a given score $\aleph(Y)$

$$\hat{\aleph}(Y) = \frac{G(Y) - \mu_G}{\sigma_G} \tag{52}$$

The normalization process implemented by means of the target speaker's statistical information is Zero Normalization (ZNorm)[101,102]. The similarity score is measured through relational analysis of the target speaker's model with various set of imposters as in imposter similarity score. This similarity score is further used to compute μG and σG. The good thing about ZNorm is permitting offline parameter estimation. However, the Test Normalization (TNorm) utilizes the test sample to compute μG and σG[103]. During the testing stage, a set of imposters are involved in computing the imposter similarity score for specific test utterances. It is observed that the significant improvement due to TNorm is when low false positives are obtained. In contrast to the ZNorm, the TNorm has to be conducted in the online mode which is the testing stage.

The ZNorm variants like Handset Norm (HNorm) and Channel Norm (CNorm) may offer a reduction in the channel as well as microphone effects[104]. The HNorm and CNorm are conducted for every individual speaker model which has chances of being attacked through handset or channel. So during the testing stage, the prior knowledge of the effect (either handset or channel) leads to application to respective parameters for score normalization. The main drawback of ZNorm and TNorm is being informed about the imposter in advance which is technically impossible. To address this issue, the DNorm fits right in to predict pseudo-imposter information from the background model through appropriate algorithms like Monte-Carlo[105].

6.1 Limitations of Score Normalization technique

The ZNorm can estimate parameters offline while TNorm needs online estimation. The TNorm outperforms cohort normalization by employing variance for approximating the distribution of cohort population more efficiently. As this estimation is based on the same target speaker-test. Thus, acoustic mismatches are intervened. Though TNorm has a major setback of language dependency of the speaker[102].

7 Evaluation Metrics for ASV

The ASV system built with experimental features and classification needs to qualify closer to or even prove better than state-of-the-art techniques. This is feasible through standard evaluation measures used in ASV scenarios.

Let t be a verification trial linked to two speech samples $t = s1, s2$. In case the trial is supervised, there might be labels present for respective samples at $\{A1, A2\}$, which are indirectly controlled by samples belonging to the same or different speakers. Consider a test set T for the same speaker and different speakers with correct detection labelled as $A1$ or $A2$. Thus, based on these labels there are two potential errors reported in the case of Miss Detection (MD) also called False Negative or False rejection Ratio (FRR) and the second being False Positive (FP) or False Alarm (FA) or False Acceptance Ratio (FAR).

The probability for detecting error for a specific test set is given as

$$p(MD/T) = \frac{Numberofmissdetections N_{MD}}{Numberofsametrials \vee T_1 \vee = p(MD/T, d)} \tag{53}$$

$$p(FA/T) = \frac{Numberofmissdetections N_{FA}}{Numberofsametrials \vee T_2 \vee = p(FA/T, d)} \tag{54}$$

Where, d is the threshold that keeps track of both the errors giving the freedom to the user to choose a preferred operating point. The Detection Cost Function (DCF) was first presented in the NIST-SRE challenge that determines the overall cost for both errors[11]. The weighted aggregate of probabilities of FP and FN given as

$$C_{dcf} = C_{MD} \times p(MD/T) \times p(A_1) + C_{FA} \times p(FA/T) \times p(A_2) \tag{55}$$

Where, $p(A_2) = 1 - p(A_1)$. Here, C_{MD} is the relative cost of MD, C_{FA} is the relative cost of FA, $p(A_1)$and $p(A_2)$ are prior probabilities. The normalization of C_{dcf} by a-priori cost $C_{default}$ yields a more specific metric, C_{norm}. The $C_{default}$ is calculated by fixing all trials to the same speakers or different speakers whichever is less.

$$C_{default} = minC_{FN}\,p(A_1), C_{FP}\,p(A_2) \tag{56}$$

$$\therefore C_{norm} = C_{dcf}/C_{default} \tag{57}$$

Irrespective of the threshold, the cost is calculated from hard decisions and the threshold is chosen for the value of min DCF[15]. Thus, the min DCF is computed as

$$minC_{dcf} = min_h C_{FN}\,p(FN \vee T, d)\,p(A_1) + C_{FP}\,p(FP \vee T, d)\,p(A_2) \tag{58}$$

Furthermore, the Constellation plots are a better choice in case similar scores appear overlapping or closer on the DET graph[106]. The constellation plot utilizes chosen pairs of operating points in a 2D. Another popular metric, the EER gives a point of convergence where FP and FN are the same and ideally must be as low as possible[8]. Additionally, the visualization of plots sometimes makes decision-making easier with plots like Detection Error Tradeoff (DET). It is an option in place of ROC curves. The minimum DCF and EER are drawn on the DET curve [15, 117][15].

8 Recent Trends and Future Perspectives

Currently, there is an immediate demand to have transparent methods in order to evaluate spoofing attacks. This task isn't easy as it may seem, as a couple of parameters need to be considered while developing anti-spoofing techniques. Addressing only a single anti-spoofing method won't be enough as the atrocity of attack type and data acquiring environments also influence the performance. So a countermeasure that resumes its operation while the first one failed, might provide two-step authentication to ASV systems. With advancements in speaker recognition and verification, potential algorithms are being developed where the conventional features are no longer required and are replaced with deep features sometimes. The end-to-end DNN systems have gained momentum recently due to their low EER in contrast to the state-of-the-art GMM techniques. Thus there have been instances where fusion-based algorithms have been reported leading to 0% EER as well. To conclude, it is certainly time straining and difficult to bring out novelty in order to get low EER yet with proper literature review, the tasks level of complication may be reduced. Moreover, the developments in building stronger ASV system extends protection to our current speech biometric systems.

8.1 Acoustic conditions of the speaker

The impact of the acoustic environment such as background noise, reverberated speech, and effects due to windowing and overlapping speech frames influences outcome of the ASV framework. These conditions need to be incorporated prior to developing an ASV system for eluding FRR due to contaminated speech which may be from the natural environment. Hence, either speech filtering must be performed or a noise-resistant feature set and decision-making algorithms needs to be re-engineered to foster a better ASV system.

8.2 Traits of the speaker

Selecting a dataset does not involve listening to every speaker's voice but rather a significant amount of samples with variation in attack types is enough for building a generic anti-spoofing ASV framework. Yet, sometimes the machine learning system is unable to capture the speaker-specific information and performs badly by not reaching convergence. This happens when speaking rate, speaker's style, and accent are ruled out as possibilities that could hamper the ASV's performance. According to the author's knowledge, not much work has been carried in developing systems adapting according to speaker's traits.

8.3 Demand for unsupervised data and training

The requirement for unknown test data and hence the unsupervised training is both challenging and demanding. Since, the ASV Spoof 2017 challenge, test samples for unknown attacks have been considered. Hence, some form of unsupervised learning mechanism that helps capture the generality over the unlabelled data and equally detect the unknown attack needs to be explored. Ultimately, the sole aim of researchers must be to make quick amendments in the already trained model (if necessary) and also for the algorithms to continuously learn from their ideal role models that are "the humans".

8.4 Language independent anti-spoofing ASV

The demand for making the system capable is unending as we desire the machines to be exactly like us but simply without the physiological aspect. Similarly, the speaker's language should not be glitch anymore, in order to verify his identity. Imagine a scenario where a speaker needs access to his bank account (of course speech-based authentication) but has suspicions that an attacker is keeping a watch over him, so he might change his language to get access without the attacker being able to impersonate or record or take some action. Sounds like a three-level authentication scenario.

8.5 Lack of exploring human-based features and Universal anti-spoofing techniques

The studies conducted to date consider single features are several dissimilar features along with their fused versions. There is a need to develop features that are interdependent and influenced by the human speech production mechanism. So a chain of features that are highly similar to humans may act as mediator trait in feature engineering. Additionally while developing anti-spoofing systems, the focus should be on building algorithms that consider all attacks along with noisy real environments to obtain a universal measure. Diversity is surely expected to reach standardization and avoid biasing issues.

8.6 Lack of grading in database development

The databases available so far do not indulge in variation of environments like noisy places and different attacks in addition. For instance, ASV Spoof 2015 challenge only has a synthetic speech from TTS and VC frameworks while the ASV Spoof 2019 dataset has replay speech in addition to VC and TTS speech. Yet again, if datasets are recorded with clean and noiseless scenarios, there is a huge possibility the system is most likely to fail for unknown attacks with unknown noises in the test speech.

8.7 Short utterance testing and Text-dependent systems

The real-time instances involve smaller utterances and if the system has fixed the length of samples for authentication, it becomes vulnerable besides facing performance degradation. A user's voice must be unique and differentiable causing the fixed-length systems to fail. Furthermore, current ongoing researches are based on text-independent systems which are the right fit for surveillance applications. Conversely, the text-dependent ASVs are yet to reach an established stage with their usage in authentication scenarios.

9 Acknowledgement

This research was funded by Taylors University, Lakeside Campus, Selangor, Malaysia. The authors deeply grateful to the University and the School of Computer Science and Engineering for their encouragement and support.

10 Conclusion

This article gives a broad view of speech based spoofing attacks and anti-spoofing measures. Most recent and state-of-the-art feature representation and pattern matching techniques employed for building countermeasures are also discussed in this article along with their critical analysis. The score normalization techniques and evaluation metrics used to judge performance of the anti-spoofing system are mentioned as well. To support the anti-spoofing system, the dedicated spoofing datasets, their traits and limitations are also described in this article.

The studies relating to evaluation of ASV systems exposed to spoofing attacks have increased lately. But it has also been quite difficult and challenging to reconstruct the unbiased and unfeigned attack scenarios for building spoofing datasets. Moreover, the spoofing attack samples are generated in controlled environments and thus it is rare or impossible to assemble datasets with diverse characteristics. Additionally, in the real-attack conditions, the type of attack is certainly unknown and so there is a need to develop systems that work without any constraints. Hence, there are few open-ended queries in this decade-old anti-spoofing research, such as what are problems in the counter-measures developed so far?; what are the future directives that would contribute in improving the anti-spoofing frameworks?; and lastly, where to begin with in order to make a difference in this domain?

Before long, the most obvious part to begin with is evaluating the speech based attacks. This issue is not a candid task but in fact more demanding in terms of acquiring knowledge about newer irregularities involved while building the spoofing techniques. Furthermore, it may be rightful to say there is no such algorithm that can be considered as the best as not one but fusion of best algorithms may be capable to perform equally well for all spoofing attacks. Thus, the development of alternative counter-measures to the ones proposed by the researchers is the need of the hour. Last but not the least, developments of refined spoof detection schemes have led to deeper possibilities in terms of research as the need to prevent such attacks is even more crucial now. This opens up opportunities for fostering reliable spoof detection algorithms.

References

1) Ferrer L, Mclaren M, Brümmer N. A speaker verification backend with robust performance across conditions. *Comput Speech Lang*. 2022;71:101258. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0885230821000656DOI:10.1016/j.csl.2021.101258. doi:10.1016/j.csl.2021.101258.
2) Jahangir R, Teh YW, Nweke HF, Mujtaba G, Al-Garadi MA, Ali I. Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Syst Appl*. 2021;171:114591. doi:10.1016/j.eswa.2021.114591.
3) Zeinali H, Sameti H, Burget L. HMM-Based Phrase-Independent i-Vector Extractor for Text-Dependent Speaker Verification. *IEEE/ACM Trans Audio, Speech*. 2017;25(7):1421–1456. Available from: http://ieeexplore.ieee.org/document/7902120/DOI:10.1109/TASLP.2017.2694708.
4) Mtibaa A, Petrovska-Delacrétaz D, Boudy J, Hamida AB. Privacy-preserving speaker verification system based on binary I-vectors. *IET Biometrics*. 2021;10(3):233–278. Available from: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/bme2.12013DOI:10.1049/bme2.12013.
5) Yamagishi J, Todisco M, Sahidullah M, Delgado H, Wang X, Evans N. Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. *ASV Spoof*. 2019. Available from: http://dx.doi.org/10.7488/ds/1994.
6) Nautsch A, Wang X, Evans N, Kinnunen TH, Vestman V, Todisco M. Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech. *IEEE Trans Biometrics, Behav Identity Sci*. 2019;3(2):252–265. Available from: 10.1109/TBIOM.2021.3059479.
7) Yan C, Long Y, X J, Xu W. The Catcher in the Field. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. ACM. 2019;p. 1215–1244. doi:10.1145/3319535.3354248.
8) Matsubara K, Okamoto T, Takashima R, Takiguchi T, Toda T, Shiga Y. High-Intelligibility Speech Synthesis for Dysarthric Speakers with LPCNet-Based TTS and CycleVAE-Based VC. In: IEEE International Conference on Acoustics, Speech and Signal Processing. Institute of Electrical and Electronics Engineers. ;p. 7058–7062. doi:10.1109/ICASSP39728.2021.9414136.
9) Mohammadi SH, Kain A. An overview of voice conversion systems. *Speech Commun*. 2017;88:65–82.
10) Marcel S, Nixon M, Fierrez J, Evans N. Handbook of biometric anti-spoofing. 2019.
11) Wu Z, Evans N, Kinnunen T, Yamagishi J, Alegre F, Li H. Spoofing and countermeasures for speaker verification: A survey. *Speech Commun*. 2015;66:130–53. doi:10.1016/j.specom.2014.10.005.

12) Wu Z, Yamagishi J, Kinnunen T, Hanilçi C, Sahidullah M, Sizov A. ASVspoof: The automatic speaker verification spoofing and countermeasures challenge. *J Sel Top Signal Process*. 2017;11(4):130–153. doi:10.1016/j.specom.2014.10.005.
13) Patil HA, Kamble MR. A Survey on Replay Attack Detection for Automatic Speaker Verification (ASV) System. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. doi:10.23919/APSIPA.2018.8659666.
14) Poddar A, Sahidullah M, Saha G. Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biometrics*. 2018;7(2). doi:10.1049/iet-bmt.2017.0065.
15) Todisco M, Wang X, Vestman V, Sahidullah M, Delgado H, Nautsch A. Future horizons in spoofed and fake audio detection. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2019*. 2019;p. 1008–1020. doi:10.21437/Interspeech.2019-2249.
16) Vestman V, Kinnunen T, Hautamäki RG, Sahidullah M. Voice Mimicry Attacks Assisted by Automatic Speaker Verification. *Comput Speech Lang*. 2020;59:36–54. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0885230818303863.
17) Farrús M. Voice disguise in automatic speaker recognition. *ACM Comput Surv*. 2018;51(4). doi:10.1145/3195832.
18) Kurihara K, Seiyama N, Kumano T, Fukaya T, Saito K, Suzuki S. AI News Anchor" With Deep Learning-Based Speech Synthesis. *SMPTE Motion Imaging J*. 2021;130:19–27. Available from: https://ieeexplore.ieee.org/document/9395678.
19) Daengsi T, Pornpongtechavanich P, Wuttidittachotti P. Comparison of TTS System Efficiency: A Pilot Study of Word Intelligibility between Siri and Google Translate with Thai Language. *International Conference on Artificial Intelligence and Computer Science Technology*. 2021;p. 196–205. doi:10.1109/ICAICST53116.2021.9497835.
20) Gujarathi P, Patil SR. Review on Unit Selection-Based Concatenation Approach in Text to Speech Synthesis System. 2021;p. 191–202. Available from: https://link.springer.com/chapter/10.1007/978-981-33-6691-6_22.
21) Sriskandaraja K. Spoofing countermeasures for secure and robust voice authentication system: Feature extraction and modelling. In: The University of New South Wales. 2018.
22) Chen F, Yang J, Zhao L. A Bilingual Speech Synthesis System of Standard Malay and Indonesian Based on HMM-DNN. *2020 International Conference on Asian Language Processing*. 2020;p. 181–187. doi:10.1109/IALP51396.2020.9310503.
23) Zangar I, Mnasri Z, Colotte V, Jouvet D. Duration modelling and evaluation for Arabic statistical parametric speech synthesis. *Multimed Tools Appl*. 2020;80(6):8331–8353. Available from: https://link.springer.com/article/10.1007/. doi:10.1007/S11042-020-09901-7.
24) Lorincz B, Stan A, Giurgiu M. Speaker verification-derived loss and data augmentation for DNN-based multispeaker speech synthesis. In: European Signal Processing Conference. Institute of Electrical and Electronics Engineers. 2021.
25) Xie FL, Li XH, Liu B, Zheng YB, Meng L, Lu L. An Improved Frame-Unit-Selection Based Voice Conversion System Without Parallel Training Data. *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2020;p. 7754–7762.
26) Wang D, Deng L, Yeung YT, Chen X, Liu X, Meng H. VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-shot Voice Conversion. 2021. Available from: https://arxiv.org/abs/2106.10132v1.
27) Shah NJ, Sreeraj R, Shah N, Patil HA. Novel Inter Mixture Weighted GMM Posteriorgram for DNN and GAN-based Voice Conversion. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. 2018;p. 1776–1781. Available from: https://ieeexplore.ieee.org/document/8659638/DOI:10.23919/APSIPA.2018.8659638.
28) Lee KS. Restricted Boltzmann Machine-Based Voice Conversion for Nonparallel Corpus. *IEEE Signal Process Lett*. 2017;24(8):1103–1110. doi:10.1109/LSP.2017.2713412.
29) Kannan S, Raju PR, Madhav R, Tripathi S. Voice Conversion Using Spectral Mapping and TD-PSOLA. *Lecture Notes in Eletrical Engineering Springer*. 2021;p. 193–205.
30) Sisman B, Yamagishi J, King S, Li H. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *Speech, and Language Processing*. 2021;p. 132–57.
31) Sisman B, Li H. Wavelet Analysis of Speaker Dependent and Independent Prosody for Voice Conversion. In: Interspeech 2018. 2018;p. 52–58.
32) Zhou Y, Liu Y. Replay Attack Anaysis Based on Acoustic Parameters of Overall Voice Quality. *6th International Conference on Intelligent Computing and Signal Processing*. 2021;p. 599–604. doi:10.1109/ICSP51882.2021.9408884.
33) Genoud D, Spoken GC. Speech pre-processing against intentional imposture in speaker recognition. *Fifth International Conference on Spoken Language Processing ISCA*. 1998.
34) Masuko T, Tokuda K, Kobayashi T, Imai S. Voice characteristics conversion for HMM-based speech synthesis system. *International Conference on Acoustics, Speech, and Signal Processing*. 1997;p. 1611–1615.
35) A O. Discrimination Method of Synthetic Speech Using Pitch Frequency against Synthetic Speech Falsification. *IEICE Transactions on Fundamentals of Electronics*. 2005;p. 280–286.
36) Leon PLD, Stewart B, Yamagishi J. Synthetic speech discrimination using pitch pattern statistics derived from image analysis. *International Speech Communication Association*. 2012. doi:10.1109/ICASSP.2012.6288895.
37) Kinnunen T, Wu Z, Lee KA, Sedlak F, Chng ES, Li H. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. *International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2012. doi:10.1109/ICASSP.2012.6288895.
38) Wu Z, Xiao X, Chng ES, Li H. Synthetic speech detection using temporal modulation feature. *International Conference on Acoustics, Speech and Signal Processing*. 2013.
39) Wu Z, Khodabakhsh A, Demiroglu C, Yamagishi J, Saito D, Toda T. SAS: A speaker verification spoofing database containing diverse attacks. *International Conference on Acoustics, Speech and Signal Processing*.
40) Liu Y, Tian Y, He L, Liu J, Johnson MT. Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing. In: International Speech Communication Association, INTERSPEECH. 2015.
41) Xiao X, Tian X, Du S, Xu H, Chng ES, Li H. Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge. In: International Speech Communication Association, INTERSPEECH. Dresden. 2015.
42) Alam MJ, Kenny P, Bhattacharya G, Stafylakis T. Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge. *International Speech Communication Association, INTERSPEECH Dresden*. 2015.
43) Acharya R, Kotta H, Patil AT, Patil HA. Cross-Teager Energy Cepstral Coefficients for Replay Spoof Detection on Voice Assistants. 2021. doi:10.1109/ICASSP39728.2021.9414847.
44) Prajapati GP, Kamble MR, Patil HA. Energy separation based features for replay spoof detection for voice assistant. *European Signal Processing Conference 2021*.

45) Li Z, Wei J, Sun Q. Time-frequency Resolution Optimization Features on Spoof Detection. In: 2021 Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS). IEEE. 2021;p. 10–16.
46) Liu X, Sahidullah M, Kinnunen T. Learnable MFCCs for Speaker Verification. *2021 IEEE International Symposium on Circuits and Systems (ISCAS) IEEE; 2021*;p. 1–5.
47) Kumar MG, Kumar SR, Saranya MS, Bharathi B, Murthy HA. Spoof Detection Using Time-Delay Shallow Neural Network and Feature Switching. *Automatic Speech Recognition and Understanding Workshop, ASRU 2019*. 2019;p. 1011–1018.
48) Novoselov S, Kozlov A, Lavrentyeva G, Simonchik K, Shchemelinin V. STC anti-spoofing systems for the ASVspoof 2015 challenge. *International Conference on Acoustics, Speech and Signal Processing*. 2016.
49) Tak H, Jung J, Patino J, Kamble M, Todisco M, Evans N. End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection. *ASV Spoof 2021 Challenge 2021*.
50) Tak H, Patino J, Todisco M, Nautsch A, Evans N, Larcher A. End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection. 2021. Available from: http://arxiv.org/abs/2107.1271. doi:0.
51) Dua M, Jain C, Kumar S. LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems. *J Ambient Intell Humaniz Comput*. 2021;p. 1–16.
52) Kwak IY, Kwag S, Lee J, Huh JH, Lee CH, Jeon Y. Detecting voice spoofing attacks with residual network and max feature map. *Proceedings - International Conference on Pattern Recognition*. 2020;p. 4837–4881.
53) Panjwani S, Prakash A. Crowdsourcing Attacks on Biometric Systems. *Tenth Symposium On Usable Privacy and Security, SOUPS '14*. 2014.
54) Zhang C, Bahmaninezhad F, Ranjan S, Dubey H, Xia W, Hansen J. UTD-CRSS Systems for 2018 NIST Speaker Recognition Evaluation. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2019;p. 5776–80.
55) Voilato R, Biocpqd. . Available from: https://www.researchgate.net/publication/268811183_BioCPqD_uma_base_de_dados_biometricos_com_amostras_de_face_e_voz_de_individuos_brasileiros.
56) Jung JW, Heo HS, Shim YIH, Yu HJ, J H. A Complete End-to-End Speaker Verification System Using Deep Neural Networks: From Raw Signals to Verification Result. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2018;p. 5349–53.
57) Lapidot I, Delgado H, Todisco M, Evans N, Bonastre JF. Speech database and protocol validation using waveform entropy. *International Speech Communication Association*.
58) Paul D, Sahidullah M, Saha G. Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora. *International Conference on Acoustics, Speech and Signal Processing* .
59) Voicepa D, Ddp VD. 2018. Available from: https://www.idiap.ch/dataset/voicepa.
60) Kinnunen T, Sahidullah M, Falcone M, Costantini L, Hautamäki RG, Thomsen D. RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research. *International Conference on Acoustics, Speech and Signal Processing*.
61) Kinnunen T, Sahidullah M, Delgado H, Todisco M, Evans N, Yamagishi J. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. *INTERSPEECH ISCA: ISCA; 2017*;p. 2–6.
62) Delgado H, Todisco M, Sahidullah M, Evans N, Kinnunen T, Lee KA. Version 2.0: meta-data analysis and baseline enhancements. *Odyssey 2018 The Speaker and Language Recognition Workshop ISCA: ISCA*. 2017;p. 296–303.
63) Mary L. Significance of Prosody for Speaker, Language, Emotion, and Speech Recognition. *Springer Briefs Speech Technol*. 2019;p. 1–22.
64) Jia Y, Chen X, Yu J, Wang L, Xu Y, Liu S. Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network. *Complex Intell Syst*. 2021;7(4):1749–57.
65) Muckenhirn H, Magimai-Doss M, S M. Towards Directly Modeling Raw Speech Signal for Speaker Verification Using CNNS. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018.
66) Misra S, Laskar RH, Baruah U, Das TK, Saha P, Choudhury SP. Analysis and extraction of LP-residual for its application in speaker verification system under uncontrolled noisy environment. *Multimed Tools Appl*. 2017;76(1).
67) Ahuja P, Vyas JM. Forensic speaker profiling: the study of supra-segmental features of Gujarati dialects for text-independent speaker identification. *Aust J Forensic Sci*. 2018;50(2).
68) Sukvichai K, Utintu C, Muknumporn W. Automatic Speech Recognition for Thai Sentence based on MFCC and CNNs. *2nd International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics*.
69) Özcan Z, Kayıkçıoğlu T. Evaluating MFCC-based speaker identification systems with data envelopment analysis. *Expert Syst Appl*. 2021;168.
70) Yang J, Wang H, Das RK, Qian Y. Modified Magnitude-Phase Spectrum Information for Spoofing Detection. *IEEE/ACM Trans Audio Speech Lang Process*. 2021;29.
71) Paul D, Pal M, Saha G. Spectral features for synthetic speech detection. *IEEE J Sel Top Signal Process*. 2017;11(4).
72) Tak H, Patino J, Nautsch A, Evans N, Todisco M. An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification. *The Speaker and Language Recognition Workshop*.
73) Javed A, Malik KM, Irtaza A, Malik H. Towards protecting cyber-physical and IoT systems from single- and multi-order voice spoofing attacks. *Appl Acoust*. 2021;183.
74) Alanís AG, Peinado AM, Gonzalez JA, Gomez A. A Deep Identity Representation for Noise Robust Spoofing Detection. *Interspeech 2018 ISCA: ISCA*. 2018;p. 676–80.
75) Todisco M, Delgado H, Evans N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Comput Speech Lang*. 2017;45:516–535. doi:10.1016/J.CSL.2017.01.001.
76) Yu H, Tan ZH, Ma Z, Martin R, Guo J. Spoofing Detection in Automatic Speaker Verification Systems Using DNN Classifiers and Dynamic Acoustic Features. *IEEE Trans Neural Networks Learn Syst*. 2018;29(10):4633–4677.
77) Kuznetsov AY, Murtazin RA, Garipov IM, Fedorov EA, Kholodenina AV, Vorobeva AA. Methods of countering speech synthesis attacks on voice biometric systems in banking. *Sci Tech J Inf Technol Mech Opt*. 2021;21(1).
78) Hanilçi C. Data selection for i-vector based automatic speaker verification anti-spoofing. *Digit Signal Process*. 2018;72:171–180. doi:10.1016/J.DSP.2017.10.010.
79) Dutta K, Singh M, Pati D. Detection of replay signals using excitation source and shifted CQCC features. *Int J Speech Technol*. 2021.
80) Chadha AN, Zaveri MA, Sarvaiya JN. Optimal feature extraction and selection techniques for speech processing: A review. *International Conference on Communication and Signal Processing*. 2016.

81) Kamble MR, Patil HA. Effectiveness of Mel Scale-Based ESA-IFCC Features for Classification of Natural vs. Spoofed Speech. *Lecture Notes in Computer Science*. 2017. doi:10.1007/978-3-319-69900-4_39.

82) Janicki A. Increasing anti-spoofing protection in speaker verification using linear prediction. *Multimed Tools Appl*. 2017;76(6):9017–9032. doi:10.1007/s11042-016-3508-x.

83) Prajapati GP, Kamble MR, Patil HA. Energy separation based features for replay spoof detection for voice assistant. *European Signal Processing Conference European Signal Processing Conference, EUSIPCO; 2021*;p. 386–90.

84) Singh M, Pati D. Usefulness of linear prediction residual for replay attack detection. *AEU - Int J Electron Commun*. 2019.

85) Li C, Ma X, Jiang B, Li X, Zhang X, Liu X. Deep speaker: An end-to-end neural speaker embedding system. *arXiv*. 2017.

86) Faisal MY, Suyanto S. SpecAugment Impact on Automatic Speaker Verification System. . In: In: 2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2019. IEEE. 2019;p. 305–313. doi:10.1109/ISRITI48646.2019.9034603.

87) Slivova M, Voznak M, Tovarek J, Partila P. Detection of speaker liveness with CNN isolated word ASR for verification systems. *Multimed Tools Appl*. 2021.

88) Alam MJ, Kenny P, Gupta V. Tandem features for text-dependent speaker verification on the RedDots corpus. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2016.

89) Shahin I, Nassif AB, Nemmour N, Elnagar A, Alhudhaif A, Polat K. Novel hybrid DNN approaches for speaker verification in emotional and stressful talking environments. *Neural Comput Appl*. 2021.

90) Prasetio BH, Tamura H, Tanno K. Emotional variability analysis based I-vector for speaker verification in under-stress conditions. *Electron*. 2020;9(9).

91) Hemavathi R, Kumaraswamy R. Voice conversion spoofing detection by exploring artifacts estimates. *Multimed Tools Appl*. 2021.

92) Rosenberg A, Chin-Hui L, Soong F. Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification. In: Third International Conference on Spoken Language Processing. 1994;p. 1835–1843.

93) Liu Y, He L, Liu J, Johnson MT. Introducing phonetic information to speaker embedding for speaker verification. *EURASIP J Audio*. 2019;2019(1):19–19.

94) Gao Y, Lian J, Raj B, Singh R. Detection and Evaluation of human and machine generated speech in spoofing attacks on automatic speaker verification systems. *arXiv*. 2020. doi:10.1109/slt48900.2021.9383558.

95) Snyder D, Garcia-Romero D, Povey D, Khudanpur S. Deep Neural Network Embeddings for Text-Independent Speaker Verification. *IInternational Speech Communication Association, INTERSPEECH ISCA: ISCA; 2017*;p. 999–1003.

96) Nasr S, Quwaider M, Qureshi R. Text-independent Speaker Recognition using Deep Neural Networks. *2021 International Conference on Information Technology (ICIT) IEEE; 2021*;p. 517–538.

97) Liu L, Yang J. Study on Feature Complementarity of Statistics, Energy, and Principal Information for Spoofing Detection. *IEEE Access*. 2020;8.

98) Saranya S, Kumar R, Bharathi S, B. Deep Learning Approach: Detection of Replay Attack in ASV Systems. *Advances in Intelligent Systems and Computing*. 2020;p. 291–299.

99) Chintha A, Thai B, Sohrawardi SJ, Bhatt K, Hickerson A, Wright M. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. *IEEE J Sel Top Signal Process*. 2020;14(5).

100) Hannani A, El, Petrovska-Delacrétaz D, Fauve B, Mayoue A, Mason J, et al. Text-independent Speaker Verification BT - Guide to Biometric Reference Systems and Performance Evaluation. D PD, B D, G C, editors;London. Springer. 2009. Available from: https://doi.org/10.1007/978-1-.

101) Li P, Li G, Han J, Zhi T, Wang D. Channel Mismatch Speaker Verification Based on Deep Learning and PLDA. *In: Journal of Physics: Conference Series*. 2020;p. 12056–12056.

102) Cao W, Liang C, Cao S, Cao W, Liang C, Cao S. Speaker Verification Based on Log-Likelihood Score Normalization. *J Comput Commun*. 2020;8(11):80–87.

103) Rakhmanenko I, Kostyuchenko E, Choynzonov E, Balatskaya L, Shelupanov A. Score Normalization of X-Vector Speaker Verification System for Short-Duration Speaker Verification Challenge. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2020;p. 457–66.

104) Thienpondt J, Desplanques B, Demuynck K. Cross-Lingual Speaker Verification with Domain-Balanced Hard Prototype Mining and Language-Dependent Score Normalization. *Proceedings of the Annual Conference of the International Speech Communication Association*. 2020;p. 756–60.

105) Matějka P, Novotný O, Plchot O, Burget L, Sánchez MD, Cěrnocký JH. Analysis of score normalization in multilingual speaker recognition. *Proceedings of the Annual Conference of the International Speech Communication Association*.

106) Harper CA, Lyons L, Thornton MA, Larson EC. Enhanced Automatic Modulation Classification using Deep Convolutional Latent Space Pooling. *54th Asilomar Conference on Signals, Systems, and Computers*;p. 162–167.