ETHIOPIAN INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

A REVIEW ON STATE-OF-THE-ART AUTOMATIC SPEAKER VERIFICATION SYSTEM FROM SPOOFING AND ANTI-SPOOFING PERSPECTIVE

Nosirov Asadbek Ulugbek o'g'li

Kimyo International University in Tashkent

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 availabilityopen 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

eISSN: 2349-5715 pISSN: 2349-5707
Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

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 realtime 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

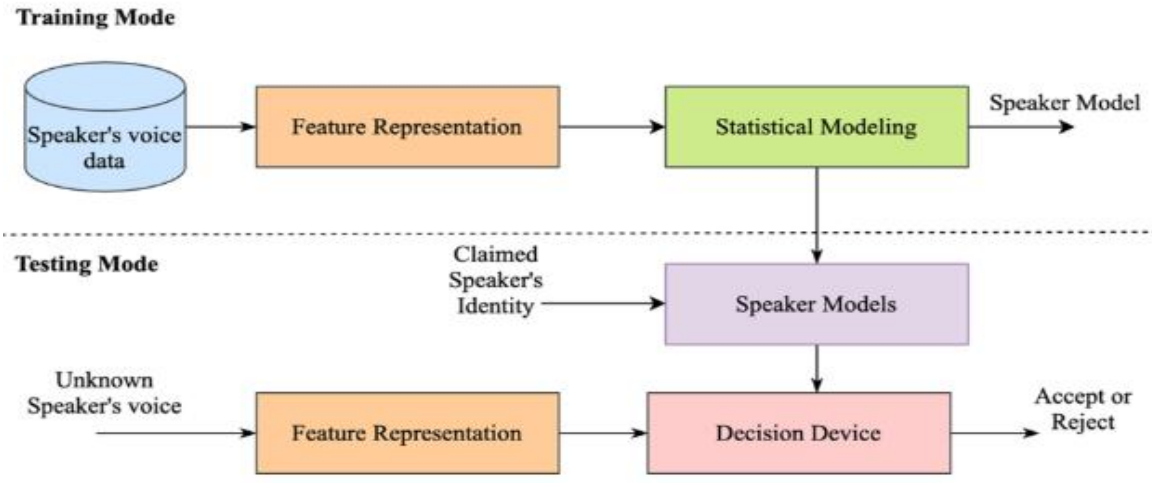

Fig 1. A block schematic of Automatic Speaker Verification framework

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

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

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 stateof-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,

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

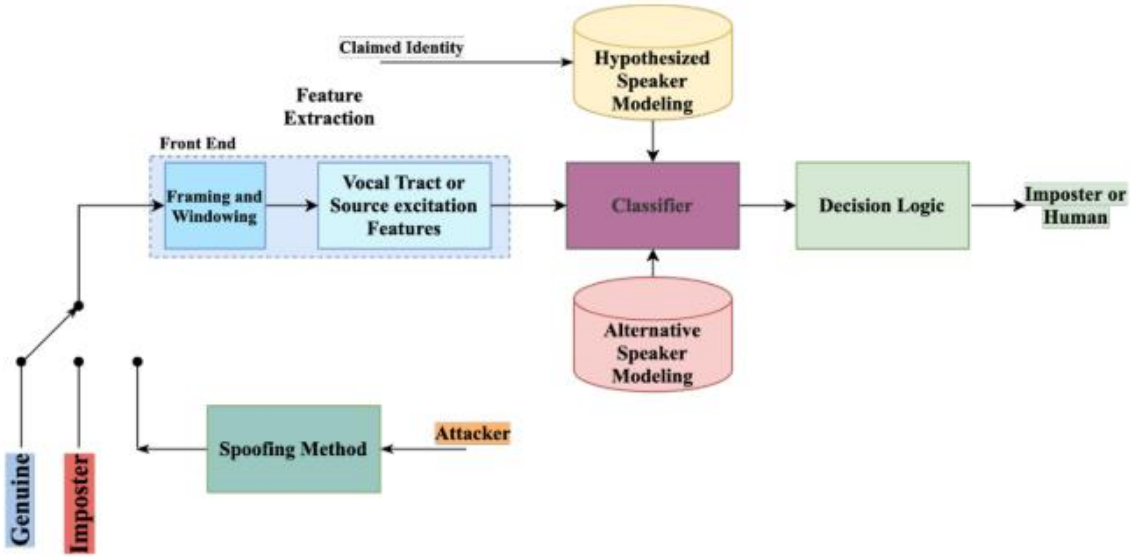

Fig 2. Automatic Speaker Verification with spoofing attack

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

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

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

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

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 lowrisk 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 phaserelated 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

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

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		6.79
	MFCC	GMM	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

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

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

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

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 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

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

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.

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

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