ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/353933140

# Automatic speaker verification systems and spoof detection techniques: review and analysis

**Article** *in* International Journal of Speech Technology · March 2022

DOI: 10.1007/s10772-021-09876-2

CITATIONS
13

READS
1,116

**2 authors:**

Aakshi Mittal
National Institute of Technology, Kurukshetra
**7** PUBLICATIONS **36** CITATIONS
SEE PROFILE

Mohit Dua
National Institute of Technology, Kurukshetra
**108** PUBLICATIONS **1,112** CITATIONS
SEE PROFILE

**Some of the authors of this publication are also working on these related projects:**



Image Encryption View project



Hindi QA System View project

All content following this page was uploaded by Mohit Dua on 19 September 2021.

The user has requested enhancement of the downloaded file.

International Journal of Speech Technology
https://doi.org/10.1007/s10772-021-09876-2

# Automatic speaker verification systems and spoof detection techniques: review and analysis

**Aakshi Mittal**[1] · **Mohit Dua**[1]

Received: 18 September 2020 / Accepted: 27 July 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

**Abstract**
Automatic speaker verification (ASV) systems are enhanced enough, that industry is attracted to use them practically in security systems. However, vulnerability of these systems to various direct and indirect access attacks weakens the power of ASV authentication mechanism. The increasing research in spoofing and anti-spoofing technologies is contributing to the enhancement of these systems. The objective of this paper is to review and analyze these important advancements proposed by different researchers and scientists. Various classical, autoregressive, cepstral, etc., and modern deep learning based feature extraction techniques that are chosen to design the frontend of these systems are discussed. Extracted features are learned and classified in the backend of an ASV system, which can be classical machine learning or deep learning models that are also the main focus of the presented review. Experimental studies use constantly modified datasets and evaluation measures to develop robust systems since emergence of practical work in this area. This paper analysis most of the contributing spoofed speech datasets and evaluation protocols. Speech synthesis (SS), voice conversion (VC), replay, mimicry and twins are the potential spoofing attacks to ASV systems. This work provides the knowledge of generation techniques of these attacks to empower the defence mechanism of ASV. This survey marks the start of a new era in ASV system development and highlights the start of a new generation ($G_4$) in SS attack development methods. With the increase in advancement of deep learning techniques, the paper makes best efforts to give the complete idea of ASV to new comers to this area and also, puts some light on some of the spoofing attacks that can be targeted during implementation of the future ASV systems.

**Keywords** ASV · Feature extraction · Spoofing attacks · Deep learning

## 1 Introduction

Authentication is one of the pillars of information assurance, which attaches a valid identification with the information. Classically and currently, word passwords have been enough for protecting the applications from unauthorized access. However, it is a tiresome and considerable time taking technique, due to keying the data. Various human physiological characteristics like retina, fingerprint, voice, etc. can be used to identify a person uniquely. Voice is the easiest and comfortable means of association with objects as compared to other traits. And also, it has more than one characteristic like the shape of the vocal tract, pitch, time-delay, etc. to differentiate individuals. Voice based authentication systems, i.e., automatic speaker verification (ASV) systems have become popular and convenient alternatives to existing security systems due to the technology advancements occurred in recent years. Unlike others, these systems offer no discomfort and health risks to the user as there is no direct contact with the machine. Studies have revealed that 90% of people are excited about using speech signal based biometrics instead of the classical ones (Beranek, 2013).

An ASV system processes the voice inserted through a microphone and either accepts or rejects the claimed identity. The task of speaker verification is to check whether the applied speech from a claimer is genuine or not. Frontend and backend are the two equally important parts of such systems for acquiring the desired functionality. As shown in Fig. 1, the input voice signal is processed at the front end of the ASV system and then, validity check and verification of the speaker (by comparing genuineness of his/her voice

✉ Mohit Dua
er.mohitdua@nitkkr.ac.in

Aakshi Mittal
aakshi8755@gmail.com

[1] Department of Computer Engineering, National Institute of Technology, Kurukshetra, Kurukshetra, India

Published online: 16 August 2021

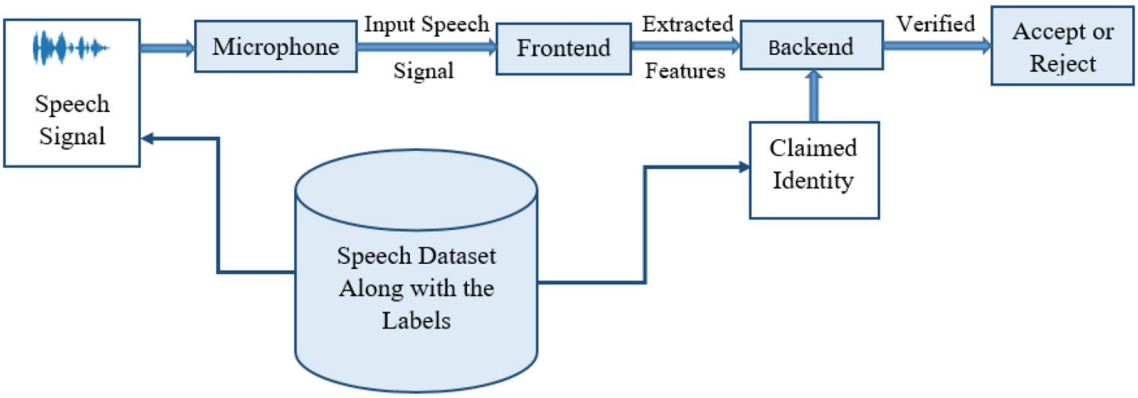



**Fig. 1** Components of ASV system

with the already existing legal user's speech in the database) are accomplished in the backend part of the system to accept or reject the claimed identity.

The frontend of the system extracts information of the uniqueness of the speaker and signal being genuine, which is intact in the input speech signal in the form of its characteristics. Characteristics of speech signal like phase, time-delay, frequency, sampling rate, pitch, magnitude, etc. varies signal to signal. These characteristics can be compared to differentiate signals produced by different sources. Features defining these characteristics can be categorized into three categories, i.e. short-term power spectrum features, short-term phase features, and features involving long term processing steps (Sahidullah et al., 2015). A wide range of feature extraction techniques that can capture the clues of speech manipulation are being used for designing the frontend of ASV systems. Feature extraction techniques that extract the features of the cepstrum domain have been dominating for speech and speaker related tasks since they can model the human auditory and human vocal tract properly (Dinkel et al., 2018; Paul et al., 2015). The classification model of the backend identifies the processable artifacts from the applied speech features. As the speaker verification lies under the class of classification problems, machine learning approaches are suitable to conclude the seen data. Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), etc. models dominated this area of verification over the years. These models are suitable to process speech related data and were being used consistently for different ASV related activities. However, these are not efficient for non-linear or near to non-linear spread of data. In the last few years, with the development of more advanced algorithms, the research community is shifted to deep learning models that can process a huge dataset with complex relationships. A good speech dataset is essential for providing the robustness to the system i.e. to make it efficient in highly varying acoustic conditions. Such speech datasets are recorded under different acoustic conditions and might be containing a large speaker and speech related variations. There are various dataset releasing communities that are cooperating with the release of acoustically and phonetically enrich datasets. Audios, speaker identity and meta data for audio and speaker are the key components of any speech corpora. Training the ASV system with a good corpus makes it gain enough experience to take the decision of acceptance or rejection for further real life data. This paper discusses and analyses various state of the art feature extraction techniques, classification models and datasets proposed by the researchers (Kumar & Aggarwal, 2020a, b).

Although these systems are being used effectively in different forms, for example, unlocking policies of smartphones, voice surveillance systems in finance, voice chat application, etc. but they are not completely reliable due to their vulnerability to various kinds of spoofing attacks. ASV systems can be attacked directly or indirectly at different stages, from the entry of the speech signal into the system to the verification of the claimer. Direct access attacks are inserted via the microphone or channel, whereas whole ASV arrangement is vulnerable to indirect access attacks at different stages. Speech Synthesis (SS)/Text-to-Speech (TTS), Voice Conversion (VC), replay, mimicry, and twins are the potential attacks to these systems. This paper shows a precise classification of these spoofing attacks based on the requirement of the access level of the system to accomplish them. INTERSPEECH (ASVspoof consortium, 2019; Lavrentyeva et al., 2017; Sahidullah et al., 2016; Zhizheng et al., 2017) has initiated sessions to enlighten these risks, as well as challenges in the design and implementation of ASV systems. The first session, ASVspoof 2013, held in Lyon, France, was aimed to spread awareness about the vulnerability of ASV systems against the spoofing attacks. Challenge of ASVspoof 2015 was to propose countermeasures fit to differentiate original speech and the speech tricked by techniques (TTS and VC). Replay attack was taken care

of by the ASVspoof 2017. Designing the countermeasure solutions to deal with the replay attack was the focus of this session. ASVspoof 2019 challenge was aimed to extend previous challenges. And it identified the need for new ASV centric evaluation measures to assess the countermeasures.

## 2 Related surveys

Several surveys have been carried out on ASV systems and their spoof detection techniques. The main focus of already existing survey works is spoofing attacks to the system. Work Wu et al., (2015a, 2015b, 2015c) provides a good classification of spoofing attacks along with the potential attack points of the ASV system. Even replay attack got attention separately for ASVspoof 2017 challenge in a survey done by Patil and Kamble (2018). A lot of research has been done in designing the spoof free countermeasures. Survey of Sahidullah et al. (2019) discusses types of different attacks, their spoofing procedure and countermeasures designed especially in their direction. However, all these discussions still have room for improvement as presenting various feature extraction techniques, different backend designing classification models, datasets used and various spoof attacks have may be explored in single discussion. For instance, work of Sahidullah et al. (2019) covers different feature extraction techniques applied for frontend design of countermeasures focusing on SS, VC, replay and mimicry attack types individually or in combination. Speech corpora and their protocols along with the evaluation metrics for countermeasures contribute equally in development of ASV system. However, these are being part of the recent studies partially. Kamble et al. (2020) covers some of the dedicated speech corpora and almost all the evaluation measures in this area. Motivated by all these discussions, the proposed survey in this paper tries to explores various important contributions involved in development chain of ASV systems. Table 1 provides the comparative view of the proposed survey with other recent surveys. Below listed points describe the main contributions of the proposed survey work:

i. Knowledge of speech signal processing is essential for frontend design of any kind of speech based system. This paper presents the detailed computation mechanism of traditional and modern speech feature extraction techniques applied for ASV frontend design.
ii. Machine learning techniques are suitable to adapt the classification clues from the huge dataset. This work provides the study of architectures of classical machine learning and deep learning networks adopted by ASV systems.
iii. This paper describes approximately all datasets applied in speaker verification systems with the best of our knowledge. Dataset enrich with the speaker, spoofing, etc. variations plays a remarkable roll in the development of these systems.
iv. Evaluation measures are also covered in this paper on which accuracy of the countermeasure is marked. Advancements in spoof generating techniques contributes in empowerment the defence mechanism of any security systems. Some attack types, added recently into this field, and new insights for generating all applied spoofing attacks are one of the major parts of discussion in this paper.
v. This survey figures out the revolutionary time periods in ASV systems. It analysis traditional and modern countermeasures, a combination of different frontend and backed techniques trained with different datasets, to check the status of accuracy of these systems.
vi. This survey finds out some attack types that are not being targeted by todays countermeasures. It discusses all the methodologies, techniques and concepts while keeping from new readers to researchers and developer of ASV system in mind. It highlights new noticeable facts and system requirements to researchers and developers for future work.

**Table 1** Comparison of several surveys on ASV systems

| Existing surveys | Feature extraction techniques | Classification models | Speech corpora | Evaluation measures | Spoofing attacks taxonomy |
|---|---|---|---|---|---|
| Wu et. al. in 2015 | Partial | No | No | Yes (t-DCF No) | Yes |
| Patil & Kamble in 2018 | No | No | Partial (AVSpoof & ASVspoof2017) | Partial (EER & t-DCF) | No (Only Replay) |
| Sahidullah et. al. in 2019 | Yes | Partial | Partial | No | Yes |
| Kamble et. al. in 2020 | No | No | Partial | Yes | Yes |
| Proposed survey | Yes | Yes | Yes | Yes | Yes |

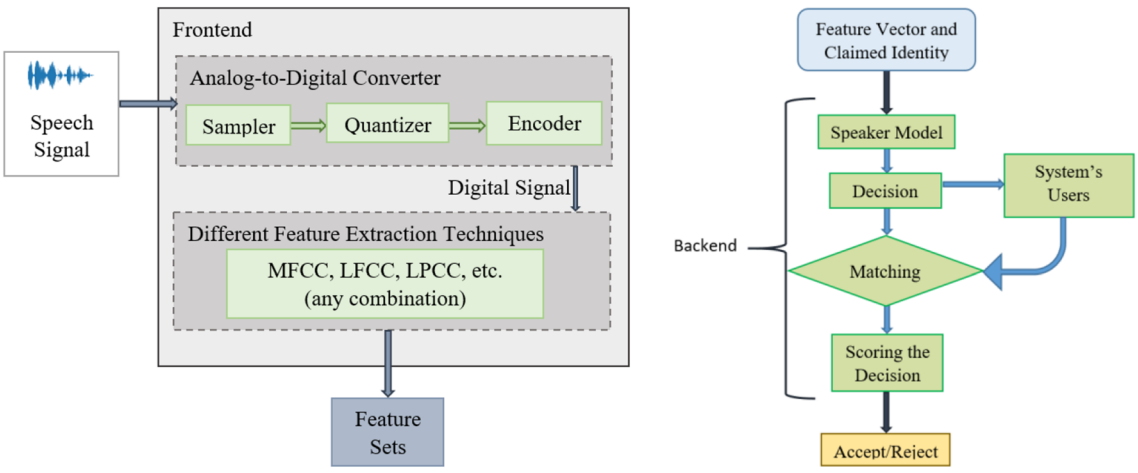
## 3 Automatic speaker verification (ASV) system

Automatic speaker verification systems are successfully contributing to the development of human behavioural and physiological characteristics based biometric surveillance and authentication systems (Koolwaaij & Boves, 1999; Singh et al., 2018). One can misunderstand the task of speaker verification with speaker identification. In the case of speaker identification, an unknown speaker is matched with the already existing pool of the known speakers of the database, and the closest matching speaker is declared the desired identity. Whereas for speaker verification, a claimed known speaker is accepted or rejected based on the genuineness of his/her voice. The verification model checks whether the applied speech is original (directly coming from the speaker) or generated by tricks (spoofed). The decision of acceptance or rejection is taken based on a threshold value**.** These systems can be classified into text-dependent ASV (TD-ASV) and text-independent ASV (TI-ASV) systems. TD-ASV system requires a registered text-password to be said correctly without bothering about the voice of the speaker whereas the TI-ASV system finds the identity of the user only by verifying his/her voice with freely accepting the speech content (Marinov, 2003; Reynolds & Rose, 1995). Former systems are suitable only for authentication purposes, whereas later ones are applicable in authentication as well as in biometric surveillance. Frontend and backend parts, shown in Fig. 2a–b, of the ASV system play equally significant roles in delivering the desired functionality of the system.

### 3.1 Approaches to design frontend of ASV system

Frontend extracts the features from the applied speech signal after converting the analog signal into digital by passing it to a sampler, quantizer, and then to an encoder. Continuous time speech signal is converted into discrete time by the sampler, which is then passed to a quantizer to get the finite values of amplitude. Then each quantized value is associated with a digital word by the encoder (Ochiai et al., 2014; Picone, 1993). Linear Predictive Coding (LPC), Mel Frequency Cepstrum Coefficients (MFCC), Linear Predictive Cepstrum Coefficients (PLCC), etc. feature extraction techniques are applied to the digital signal to achieve the required features (Chen et al., 2018). Extracted features are analyzed on the backend to verify the speaker's genuineness. Most of the classical feature extraction techniques involve the tasks of filtering, Linear Predictive Coding, and cepstrum calculation in any combination. First we highlight some important characteristics of these techniques and in the next subsection we discuss different feature extraction techniques that are also summarised in Table 2.

- *Filtering* Filter-Bank based feature extraction can imitate the way of human auditory. Artificial Cochlea-based (Patel & Patil, 2015; Patil & Kamble, 2018) and Fourier-Transform-based filters are being used for frequency analysis of speech signals as standards. The applied filter outputs a short-time energy signal. A short-time energy signal obtained by Fourier-Transform based filter undergoes logarithm, etc. functions to deliver an acoustic feature vector (Dua et al., 2018a, b; Ochiai et al., 2014).




**(a)** Frontend **(b)** Backend

**Fig. 2** Frontend and backend of ASV system

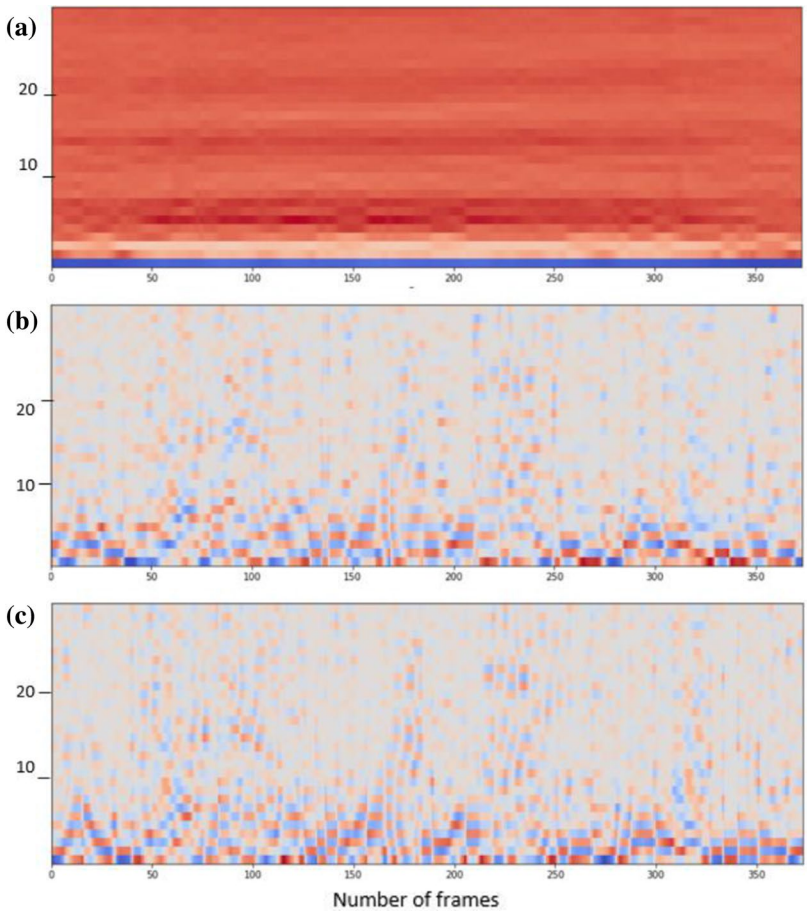



**Fig. 3** **a** 30 Static CQCC features, **b** 30 First order CQCC features, **c** 30 Second order CQCC Features, y-axis are the number of features

**Table 2** Summary of all classical feature extraction techniques

| Feature extraction | Filter type | Filter shape | Models | Coefficient type | Frequency region |
|---|---|---|---|---|---|
| MFCC | Mel | Triangular, Gaussian | Human Auditory | Cepstral | Low |
| IMFCCs | Inverse-Mel | Triangular | Human Auditory | Cepatral | High |
| LFCC | Linear Scale | Triangular | Vocal Tract | Cepstral | High |
| CQCC | Constant Q Transform | Linear | Human Auditory | Cepstral | High and Low Time–Frequency Analysis OR High, Low |
| LPCC | Linear Predictive Coding | Linear | Vocal Tract | Cepstral | Low and Medium |
| PLP | Linear Predictive Coding (Perceptually motivated) | Linear | Human Auditory | Spectral | – |
| PNCC | Gammatone | – | Human Auditory | Cepstral | – |
| APGDF | Group Delay Function | – | Phase Property of Speech Signal | Phase Coefficients | Models Phase |
| MODGDF | Group Delay Function | – | Phase Property of Speech Signal | Phase-magnitude | Models Phase |
| SCFC | – | – | Formant | Frequency magnitude | Average of Frequency |

- *Linear Predictive Coding (LPC)* The state of sample $s(t)$ of a speech signal at some point of time $t$ can be predicted on the basis of its previous state at $t$-$1$ because the current state of a time varying signal depends on the former state (Ochiai et al., 2014). This concept is formularized (Eq. 1) for feature extraction under Linear Predic-

tive Coding (LPC), also called autoregressive modeling. If the order of modeling is $o$ and noise in prediction is $n(t)$ then sample $s(t)$ can be modeled as:

$$s(t) = \sum_{i=1}^{o} \beta_i s_{t-1} + n(t) \tag{1}$$

where $\beta_i$ denote the coefficients of prediction found to minimize the Mean-Square-Error (MSE) for the speech segment's window starting and ending at $p_1$ and $p_2$, respectively. MSE is defined as

$$P_n = \sum_{t=p_1}^{p_2} \{n(t)\}^2 \tag{2}$$

- *Cepstrum* Cepstrum is another approach for modeling the linguistic class (vocal tract information) of a speech signal. Speech signal, a time varying signal, can be obtained by the convolution of vocal tract filter and excitation signal. To obtain its cepstrum, a signal undergoes some operators, including Fourier transformation, logarithmic operator, and then to inverse Fourier transformation. By applying the Fourier transformation, speech signal becomes the product of vocal tract filter and excitation spectrums. This power spectrum is a function of frequency. Log operator converts these power spectrums into log spectrums and returns their summation. Now Inverse Fourier transform is applied on these summed up log power spectrums to get the cepstrum of the initial speech signal.
- *Static and Dynamic Features* Static features capture general information from the speech, whereas dynamic features can capture contextual variations also from the speech. They can model speaker specific information more precisely. Dynamic features, delta, and delta-delta ($\Delta$, $\Delta\Delta$), are the first order derivative and second order derivative of static features, respectively. Dynamic features are proved to be better for speaker verification systems. Figure 2a–c show the spectrographic view of these features extracted for an utterance of the ASVspoof 2019 dataset. Firstly, Statics, first order, and second order 30 CQCC features are extracted for 400 frames in MatLab than these features are plotted by functions of python's (Cheuk et al., 2019; Glover et al., 2011) library librosa.

#### 3.1.1 Mel-frequency cepstral coefficients (MFCCs)

Cepstral analysis based Mel-frequency Cepstral Coefficients (MFCCs) are the most common feature coefficients used in spoof detection. In this approach, pitch comparisons based mel-scale unit is used for frequency representation. Mel scale is perceptually motivated by the human auditory system. For extraction of coefficients by MFCC technique Fast Fourier Transform (FFT) (Prithvi & Kumar, 2016) or (Todisco et al., 2018) Discrete Fourier Transform (DFT) is applied on the input speech signal, which results in an audio spectrum. Then, applied triangular, gaussian, etc. filter bank changes the scale of spectrum to mel-cale (Chakroborty & Saha, 2009). The logarithm is calculated for the spectrum before applying the Discrete Cosine Transform (DCT), which results in MFCCs (Cai et al., 2019). Generally static, first order and second order derivatives of first 12 to 14 coefficients are able to perform well for ASV systems (Balamurali et al., 2019; Dua et al., 2018a, b). Figure 3 shows a general approach of MFCC extraction. The mathematical process of MFCC extraction is as:

$$D_{DFT}(i) = DFT(f) \tag{3}$$

$$MFB(j) = \sum_{i=1}^{L} \left|D_{DFT}(i)\right|^2 W(i) \tag{4}$$

$$MFCC(r) = \sum_{j=1}^{J} log\left[MFB(j)\right] cos\left\{\frac{r(j-0.5)\pi}{J}\right\} \tag{5}$$

where $f$ is the audio frame, $MFB(j)$ is the mel scaled frequency spectrum calculated by $i^{th}$ mel filter bank $W(i)$ having $j$ number of filter banks, $L$ represents the DFT indices in total and $r$ MFCC features are carried out by $MFCC(r)$.

#### 3.1.2 Inverse mel-frequency cepstral coefficients (IMFCCs)

MFCCs are concerned about the low frequency regions of the spectrum, whereas high frequency regions are taken care of by the Inverse Mel-frequency Cepstral Coefficients (IMFCCs) (Mohammadi & Mohammadi, 2017). These features are retrieved by applying a similar process to MFCC except the use of inverted mel-scale filters (Cai et al., 2019; Saranya & Murthy, 2018) that introduce the inversion of frequency

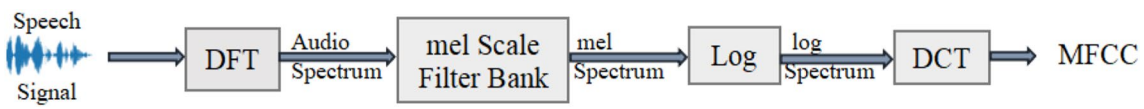



**Fig. 4** MFCC feature extraction

domain from low to high to offer complementary information (Pritam et al., 2018). Figure 4 shows the steps for IMFCC feature extraction technique.

#### 3.1.3 Linear frequency cepstral coefficients (LFCC)

Linear scale based Linear Frequency Cepstral Coefficients (LFCC) features can model all frequency regions equally (Mohammadi & Mohammadi, 2017) and proved to be better than MFCCs for ASV systems (Pritam et al., 2018). These features are extracted by applying a triangular filter bank that provides a constant resolution of the spectrum. LFCCs are helpful in speech recognition as well as in speaker identification as they can model the length of the vocal tract in higher frequency regions (Sahidullah et al., 2015; Pritam et al., 2018). Figure 5 depicts the complete process of LFCC extraction.

#### 3.1.4 Constant Q cepsral coefficients (CQCC)

In the case of CQCCs, frequency bins are geometrically spaced, whereas in Fourier-based transforms, frequency bins are regularly spaced, which leads to the variable Q factor. To generate CQCCs perceptually motivated Constant Q Transform (CQT) is applied to the speech signal, which ensures the constant Q factor. At lower frequencies, higher determination of frequency, and at higher frequencies higher determination of time, are the speech distinguishing properties offered by the CQT (Balamurali et al., 2019; Saranya & Murthy, 2018). These features are especially being applied for LA and PA spoof detection in ASV systems, and are achieving better performance than instantaneous frequency cosine coefficient (IFCC), MFCC, LFCC, ICMC, etc. features (Jelil et al., 2017; Todisco et al., 2018). The extraction of these features from speech starts with the application of CQT that converts the time domain into the frequency domain. Then the power of the spectrum is followed by a logarithm operation. Before applying the DCT on the log power spectrum, uniform re-sampling is done to convert geometrically spaced CQT bins to linearly spaced bins that make spectrum compatible to DCT operation (Brown, 1991; Brown & Puckette, 1992; Todisco et al., 2017). This whole process can be summarized by the Eqs. 6 and 7.

$$C_{CQT}(l) = CQT(s(t)) \tag{6}$$

$$CQCC(r) = \sum_{l=1}^{L} log\left|C_{CQT}(l)\right|^2 cos\left\{\frac{r(j-0.5)\pi}{L}\right\} \tag{7}$$

where $C_{CQT}(l)$ is the CQT of speech sample $s(t)$, $l$ is used for indexing into total $L$ number of linearly spaced bins to extract and $CQCC(r)$ denotes the total $r$ number of extracted coefficients. Figure 6 depicts this process clearly.

Yang et al. (2018) have proposed Extended Constant Q Cepstral Coefficients (eCQCC) that are the concatenation of DCT of log scale and log linear power spectrums, which is preceded by the same procedure as CQCCs. eCQCCs are performing better than the baseline CQCCs of the ASVspoof2015 and ASVspoof2017 datasets.

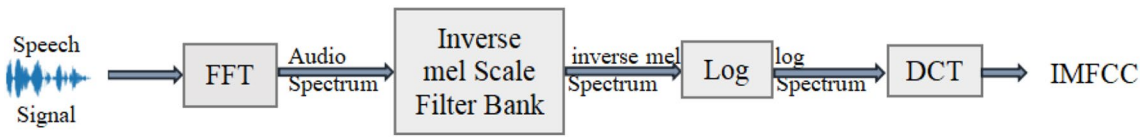



**Fig. 5** IMFCC feature extraction

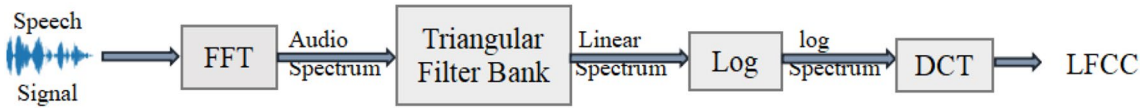



**Fig. 6** LFCC feature extraction

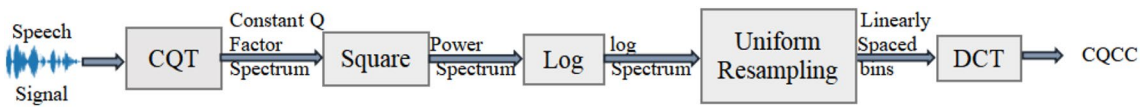



**Fig. 7** CQCC feature extraction

#### 3.1.5 Linear predictive cepstrum coefficients (LPCC)

Linear Predictive Cepstrum Coefficients (LPCC) extract the characteristics of a speaker's voice by using one of the oldest techniques called Linear Predictive Coding (LPC) on a speech frame. Application of LPC gives the LPC coefficients, which are converted into LPCC using the below given autoregressive (recursive) function (Balamurali et al., 2019; Prithvi et al., 2016). Speech sample $s(t)$ is the input to LPC filter having $l$ linear predictive coefficients $[\beta_0, \beta_1,\ldots, \beta_{l-1}]$ with error signal $n(t)$. These linear predictive coefficients are converted into $r$ LPCC $[C_0, C_1,\ldots, C_{r-1}]$ as:

$$C_j = \begin{cases} \ln\left(P_n\right) & if\ j = 0 \\ -\beta_j + \frac{1}{j}\sum_{k=1}^{j-1}\left\{-(j-k)\beta_k C_{j-k}\right\} & if\ 1 \le j \le l \\ \frac{1}{j}\sum_{k=1}^{l}\left\{\frac{-(j-k)}{j}\beta_k C_{j-k}\right\} & if\ l < j < r \end{cases} \tag{8}$$

where $j$ is for indexing into $l$ linear predictive coefficients (or linear predictive filter banks), $P_n$ is the power of error signal. Figure 7 shows this process clearly.

#### 3.1.6 Perceptual linear prediction (PLP)

Perceptual Linear Prediction (PLP) is one of the popular short term spectral features (Wu et al., 2015a, b, c). These are perceptually motivated linear predictive coding based features, which models the characteristics of the human auditory system (Dua et al., 2018a, b; Zouhir & Ouni, 2014). PLP features are computed by an estimated spectrum that is expected by windowed periodogram through DFT. Therefore, these features have a high variance. Alam et al. (2013) have proposed a multi-windowing technique for spectrum estimation to get a reduced variance spectrum that performs better than baseline PLP features for speaker verification task with i-vector classifier.

#### 3.1.7 Power normalized cepstrum coefficients (PNCC)

Power Normalization Cepstrum Coefficients (PNCC) features can perform better than other features, like MFCC, for noisy (with 0 dB to 15 dB Signal to Noise Ratio (SNR)) speeches (Al-Kaltakchi et al., 2016). These features are the human auditory system based, and they try to simulate its process too. The process of PNCC extraction starts with the application of gammatone filter bank for frequency analysis of spectrum, which is preceded by a Short-term Fourier Transform (STFT) operation. For noise reduction and addition of robust reverberation, this frequency spectrum is passed to a sequence of operations that are nonlinear and time varying in nature. After this medium time processing, the mean power normalization stage is invoked that lowers the effect of changing amplitude values. This stage processes the input by a power-low nonlinearity with the exponent value of 1/15 (Kim & Stern, 2016). This operation simulates the biological auditory system with its best efforts. Finally, the application of DCT outputs the PNCC features, as shown in Fig. 8. These features are performing better in the combination of other features like MFCC, LFCC, IMFCC, etc. for speaker identification (Al-Kaltakchi et al., 2016; Mohammadi & Mohammadi, 2017) but the computational complexity of these features is higher than others (MFCC, PLP, etc.) (Kim & Stern, 2016).

#### 3.1.8 All-pole group delay function (APGDF)

The human ear cannot recognize the phase factor of sound. Also, phase based features are less used as they are computationally complex to extract, whereas magnitude-based features like MFCC are getting more attention as these can be perceived by the human auditory system (Rajan et al., 2013). However, the phase is a crucial characteristic of a speech signal, and it can be used to distinguish utterances generated by different sources. It can be achieved from a speech signal by group-delay methodology with the help of all pole model. Phase based speaker discrimination is introduced in ASV recently (Pal et al., 2018; Sahidullah et al., 2015). A group-delay function given by Eq. 9 plays a major role in the extraction of these features (Pal et al., 2018). The application of this function offers bogus high amplitude projections that are reduced by all pole models of the audio signal. APGDF function $A(z)$ can be calculated as:

$$A(z) = \frac{F_r(z)XF_r(z) + F_i(z)XF_i(z)}{|F(z)|^2} \tag{9}$$

$$F(z) = FFT(s(t)) \tag{10}$$

$$XF(z) = FFT(X(t)) \tag{11}$$

$$X(t) = ts(t) \tag{12}$$

where $F(z)$ is the FFT of speech sample $s(t)$, $XF(z)$ is the FFT of $X(t)$ and, $i$ and $r$ in subscripts of $F(z)$ and $XF(z)$ represents

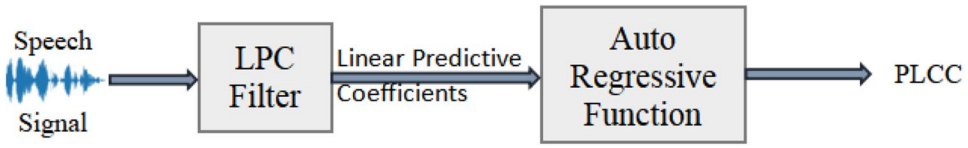



**Fig. 8** LPCC feature extraction

their imaginary and real parts. Modified Group Delay Function (MODGDF) is also a feature extraction technique that models the phase of speech signal (Hegde et al., 2004; Wu et al., 2016).

#### 3.1.9 Sub-band centroid frequency coefficients (SCFC)

Sub-band Centroid Frequency Coefficients (SCFC) are formant based features, which have been investigated as an alternative of cepstral features for speech recognition tasks (Dua et al., 2017; Paliwal, 1998). Further, these features offer complementary information of sub-bands that is not captured by cepstral features. Balamurali et al. (2019) have taken SCFC features in their set of features, including CQCC, IMFCC, LPCC, etc. for replay attack detection. In the extraction process of these features, firstly, k sub-bands are marked with their initial and ending frequency edges on the spectrum of a speech signal. Then weighted average frequency of each sub-band taken that is called sub-band centroid frequency. Sub-band centroid frequency of nth sub-band is given by:

$$C_n = \frac{\sum_{j=l_n}^{u_n} j\left|Z[j]X_n[j]\right|}{\sum_{j=l_n}^{u_n}\left|Z[j]X_n[j]\right|} \tag{13}$$

where $Z[j]$ is the spectrum of speech frame corresponding to signal $s(t)$. $n$ number of sub-bands are marked on $Z[j]$ starting at $l_n$ lower frequency edge and ending at $u_n$ upper frequency edge and frequency sampled response of the filter is given by $X_n[j]$.

#### 3.1.10 Deep learning based feature extraction techniques

As analysed, use of deep learning for feature extraction tasks emerged from 2011 (Chen et al., 2015). Deep learning is providing a new era to feature extraction task. Usually, CNNs and RNNs are used in the context of computer vision completing tasks like object detection and image recognition. But CNN's give indications that if an audio signal is represented appropriately, it can be made suitable to audios too (Shuvaev et al., 2017). In this case, hidden layers of deep learning networks are extracted as feature vectors of speech data. d-vectors, j-vectors, x-vectors, etc. come under these types of techniques. In the case of d-vectors, a Deep Neural Network (DNN) is used for feature extraction where the output layer of the network is ignored, and the values of the activation functions at the last hidden layers are taken as feature vectors. j-vectors are the extension of d-vectors (Kinnunen & Li, 2010). Another advanced technique x-vectors uses the Time Delay Neural Network (TDNN) embedding architecture and extracted features outperforms the classical i-vectors (Chen & Salman, 2011).

### 3.2 Approaches to design backend of ASV system

Backend, as described by Fig. 9b, of an ASV system is comprised of a classification model that takes features of the speech signal and claimed identities as input. During training classification model finds out discriminating patterns associated with bonafide and spoofed utterances in these applied features and learns out the characteristics of different classes well. Then the trained model takes a decision of acceptance or rejection for the claimed person of an utterance of testing data by matching it with the speech characteristics of the system's user. Various classical machine learning and deep learning approaches used to design the backend are discussed below.

#### 3.2.1 Classical machine learning approaches

Some classical machine learning approaches are generative in nature and some are classifiers. These approaches are suitable for spoof detection in applied dataset from the initial research in ASV systems. We are discussing some of the mostly used classical machine learning approaches in backend design of these systems below.

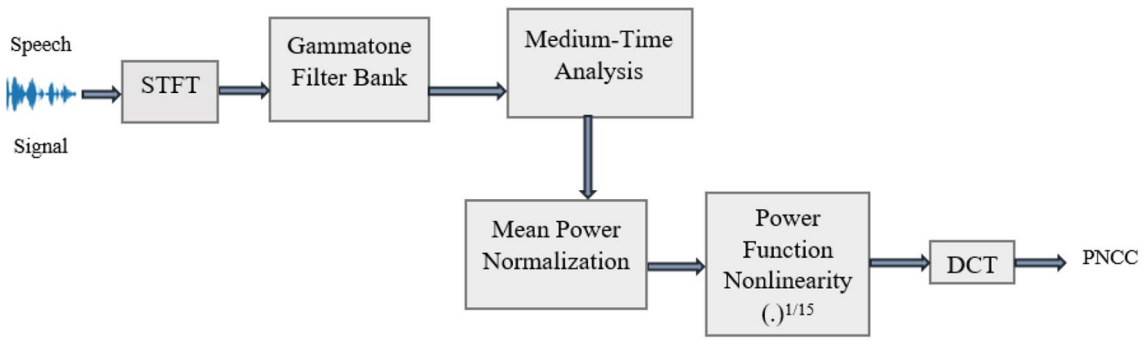



**Fig. 9** PNCC feature extraction

**3.2.1.1 GMM based models** Gaussian Mixture Model (GMM) has been a de-facto standard for developing the ASV system since the emergence of the idea of ASV. The use of the GMM in audio spoofing is inspired by the observation that general speaker-dependent shapes are found in Gaussian components. And it can model the difference between genuine and synthetic speeches effectively (Chettri et al., 2020; Pal et al., 2018). A special property of GMM is that it can plot even curves for randomly spread densities. Classical, GMM represents the distribution of the speaker's features by poison distribution and elliptic shape. For implementing the model, firstly, its parameters are randomly initialized; then, the Expectation Maximization (EM) algorithm is used to optimize the parameters by taking the maximum likelihood estimate (Suthokumar et al., 2017). GMM with Universal Background Model (UBM) is mainly being used in speech related tasks. In this case, an additional GMM, i.e., a UBM, is trained by a huge development dataset. Then Maximum a posteriori (MAP) estimation is used to obtain the speaker/task specific models (Al-Kaltakchi et al., 2016).

**3.2.1.2 SVM based models** The support vector machine has discriminative properties. It generates a hyperplane on 2-dimensional space, which classifies the data in two classes where each class lies on different sides of the plane. SVM separates the classes with a maximum margin between them. It does regularization to avoid the misclassification of example. SVM is being used successfully for speaker verification tasks as well as for spoof detection (De Leon et al., 2012; Godoy et al., 2015). One class SVM is also being used to identify the abnormal data by running only for genuine speeches (Hanilçi et al., 2015). Radial Bias Function (RBF) kernel based SVM can detect unknown spoofing attacks mixed in the speech data well (Godoy et al., 2015).

**3.2.1.3 HMM based models** Hidden Markov Model (HMM) is a well-known technique for speech recognition, speaker identification, and speaker verification tasks (De Leon et al., 2012; Dua, et al., 2019a, b; Varchol et al., 2008). These models are widely used for designing TD-ASV systems. Rich mathematical framework and robust architecture of the model are the two strengths of HMM (Gong & Yang, 2020; Rose & Juang, 1996). HMM has two parts first one is the sequence of states or Markov chain, and the second one is the collection of output distribution. Former part of the model characterizes the information from the speech signal and the later part converts the speech sequences of Markov chain into the observations to hide it from the observer (Dua et al., 2012a, 2012b).

**3.2.1.4 K-means algorithm** As Speaker Verification is a classification problem, so unsupervised clustering approach is suitable for it. K-means clustering algorithm can find out distinct clusters in the input vectors of speech features. This clustering algorithm firstly assigns random centroids for chosen k clusters and sets the vectors to different clusters by minimizing the distortion between it and the centroid. This iterative algorithm redistributes the vectors to achieve the minimum value of distortion within the cluster. This algorithm is applied to train the GMM model and can make excellent classification performance for speech data also.

### 3.2.2 Deep learning approaches

It has been perceived that deep learning is suitable for the audio spoofing community. Deep learning can process large datasets with complex distribution structure. These approaches can be used successfully with features vectors extracted by various feature extraction techniques as well as with raw speech signal. Various deep learning based architectures, individually or ensemble, are being used as backend of the ASV system. Application of raw waveform directly to the model is also famous as hidden layers of this network can build relevant features from the raw data. Some deep learning based models are discussed below.

**3.2.2.1 DNN based models** Deep neural networks are discriminative in nature. They are trained to capture the discrimination among the classes rather than enhancing the classification ability of their hidden layers. These networks are capable of representing the features as after training deep features can be drawn out from the veritable hidden layer. Training is vital for the preprocessing of data. The adjacent context of each frame is fed along with it, which enhances its performance. Common feature vectors have the same dimension; therefore, this approach is practical. Figure 10 shows a general architecture of the DNN network. A similar architecture is used in the work of Dinkel et al. (2018) with five fully connected blocks made of linear and batch normalization layers, each having 1024 parameters. These blocks are using the Relu activation function for non-linearity.

**3.2.2.2 CNN based models** Convolutional neural networks are suitable to find local and discernible patterns in the dataset, which helps to differentiate bonafide to spoofed speech. These networks require a very less preprocessing of data as they use kernels on convolutional layers. Pooling layers are the essential building block of the CNN. It reduces the spacial size of the dataset to make the computations easy and measure of parameters less (Mittal & Dua, 2021b). Hidden layers make the use of different activation functions to take the decision of firing a neuron. Relu is the most used activation function, and Softmax is used in output layers. When a raw speech signal is passed to a CNN, its hidden layers learn features of the signal well by adjustment of weights and flat-

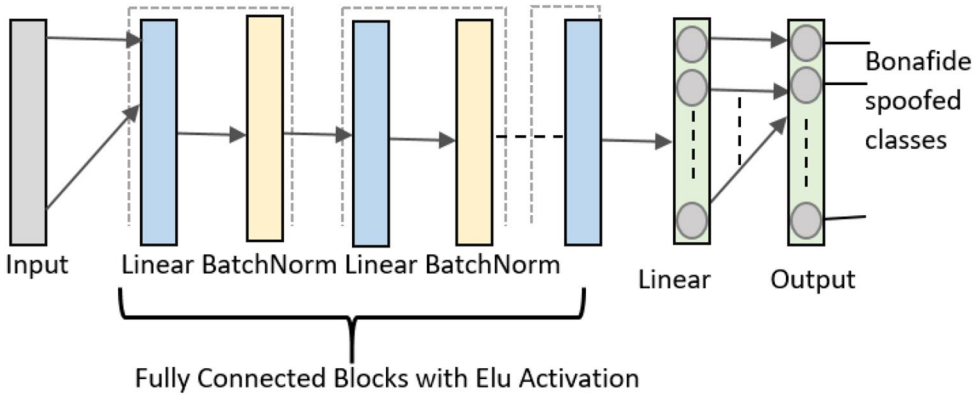



**Fig. 10** Architecture of CNN with Elu activation function for non-linearity

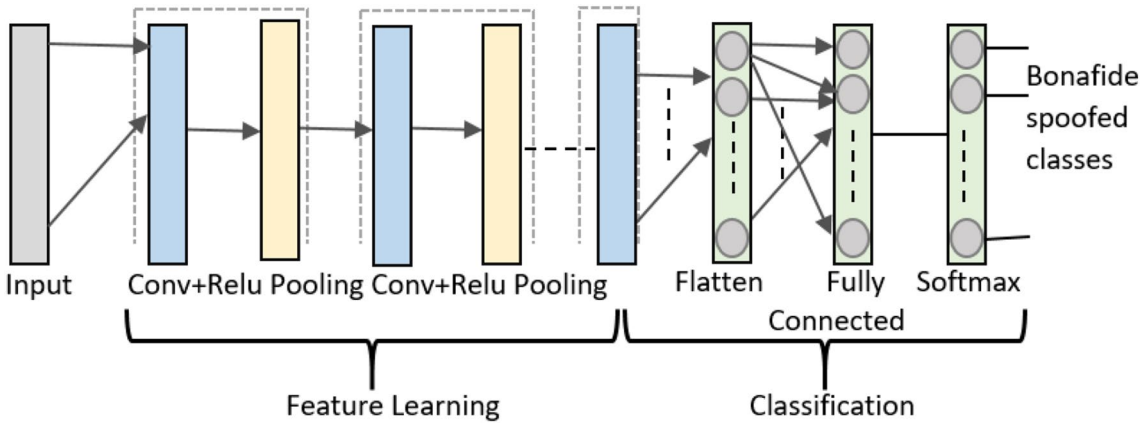



**Fig. 11** General architecture of CNN

ten, fully connected, etc. layers participate in a classification of the claimer, as shown by Fig. 11.

**3.2.2.3 RNN based networks** A recurrent neural network is capable of capturing the temporal history of the speech signal. RNN belongs to the sequence-based scoring group means it does not make any adjustment into the network during training and evaluation phases. RNN can remember a sequence of complex past events by a mechanism called cell mechanism. It calculates an output vector at time $t$ for every input vector $x_t$. It gives a single label either spoofed or bonafide to each sequence. As it provides a label for each sequence or each time step, so there is a need for reducing these output vectors into a single vector $v$. Dinkel et al. (2018) have proposed three approaches to doing so. These are: (1) If T is the last time step then output vector $o_t$ at that time is passed to $v$ ($v = o_t$). (2) Mean is taken over time by summing all output vectors $o_t$ up to time T then dividing by T and pass it to $v$. (3) Making $v$ equal to attention where attention is weighted average calculated by taking the mean of summation of weight multiplied with output vector $o_t$ over time T.

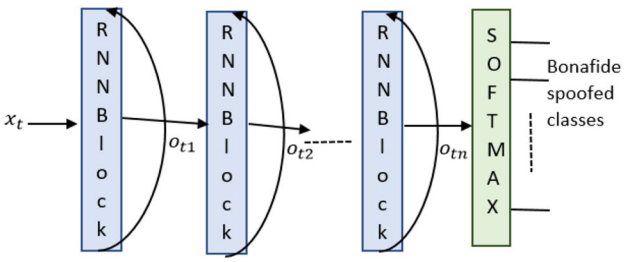



**Fig. 12** Architecture of RNN for ASV systems

Figure 12 shows a common RNN model, where $x_t$ is the input signal and $o_t$ denote the output of different RNN units at different time, Softmax is the output layer. Being a non-sequential model, it can achieve more information from time varying speech signal (Dua et al., 2019a, b; Sahidullah et al., 2019).

**3.2.2.4 LSTM based networks** Long short-term memory network is a special kind of RNN which can hold the information for an extended period. Basically, it is designed to

reduce the dependency problem for long periods. RNN is a chain of simple neural network modules. Unlike the RNN repeating module of LSTM has a different structure comprised of four neurons, each connected in a very special way. For implementation of LSTM, RNN blocks of Fig. 12 have LSTM units in them. LSTMs overcomes the gradient disappearance drawbacks of CNN and less prone to overfitting (Mittal & Dua, 2021a). A two LSTM layers network in addition to DenseNet can be applied well for replay attack detection when trained with MFCC features (Huang & Pun, 2019).

**3.2.2.5 Wave-U-Net based networks** Audios have a high temporal correlation in the long-range, so they need high-quality separation. Wave-U-Net computes and combines the feature map at contrasting time domains by repeatedly resampling them (Chettri et al., 2019).

**3.2.2.6 Ensemble of different models** Ensemble is a technique of combining different machine learning models dedicated to solve a same problem. Individual models do not perform well due to high bias or high variance, but they might have learnt different kind of facts from data (Kadyan et al., 2021a, b; Kumar & Aggarwal, 2020a, b, c). So to combine all the usefulness of single models into one, ensemble is done by stacking, boosting or bagging methods. Ensemble of ResNets, CNNs, etc., achieves better performance for speech data too (Fawaz et al., 2019).

## 4 Contributing datasets and evaluation measures

This section of the paper describes and analyses different datasets used in various state of the art ASV system and, discusses evaluation metrics in latter part.

### 4.1 ASV datasets

Dataset is also a crucial question for the development of ASV systems. For training, development, and testing of designed methodology, a suitable database is always required, especially for machine learning techniques. A database including all kinds of circumstances and rich with the protocols contributes highly to the reliability of countermeasure (Chen & Salman, 2011). Below discussed datasets are assisting remarkably in the development of threat free ASV systems. Table 3 gives the complete summary of these datasets, where M denotes Male, F indicates Female, T denotes Training, D stands for Development, E stands for Evaluation, I has been used for Imposter and, B and S are used for Bonafide and Spoofed, respectively.

#### 4.1.1 YOHO

The English Language based YOHO speaker verification dataset was developed to support the research for TD-ASV in 1989 under the contract of the US government. Total of 138 speakers (106 males and 32 females) took part in the collection of high quality speeches. Enrolment data was recorded in four sessions (each with 24 speeches) for each speaker, whereas verification data was recorded in ten sessions, each with four speeches. They all were said to pronounce two digit numbers from 21 to 97 in the combination of three, e.g., "Thirty-Six, Forty-Five, Eighty-Nine." All the utterances are created at 8000 Hz Sampling rate (Campbell, 1995).

#### 4.1.2 Wall Street Journal (WSJ)

DARPA initiated the creation of the Wall Street Journal (WSJ) dataset, especially to support speech recognition during 1991. WSJ0 and WSJ1 are the two accessible databases built under his program. The speech data of WSJ0 was recorded by two microphones. It has three sets having speeches recorded individually by them along with a mixture of them. Each set has all documents, transcripts, tests, etc. WSJ1 has 78,000 utterances in the training set, making 73 h of speech data and 8200 utterances in the testing set, making 8 h speech data (Paul & Baker, 1992).

#### 4.1.3 TIMIT

TIMIT dataset was recorded in 1993 at Texas Instruments, Inc. All sounds were recorded at 16 kHz sampling rate by 630 speakers. Each speaker was said to speak 10 American English sentences enrich of phonetics. This dataset is divided into the training and testing subsets. Both of the subsets have useful variations in phonemes, speaker characteristics, etc. (Garofalo, 1990).

#### 4.1.4 NIST

NIST Speaker Recognition Evaluation (SRE) dataset has speech files recorded in the American English language for 2,225 h via microphone and telephone system. SRE is an ongoing series in the development of text independent speaker recognition systems. This series was initiated in 1996 with the objects of improving the performance of speech based systems, to provide a common platform to the researchers in this field and to support the community in their idea of advancement in voice based technologies. NIST

**Table 3** Summary of ASV datasets used in development of ASV systems

| Dataset | Contributing community / authors | Language | Attack type | Number of speakers | Number of utterances | Remarks |
|---|---|---|---|---|---|---|
| YOHO | Campbell et al. in 1995 | English | Mimicry | 106 M + 32 F | Approximately 13,248 in T<br>Approximately 5520 in E | Audios with 8 kHz sampling rate<br>Developed to support TD-ASV systems |
| WSJ | Paul et al. in 1992 | English | SS and VC | – | WSJ 1<br>78,000 in T<br>8200 in E | Recorded at 16 kHz sampling rate<br>Source of speech data is microphone |
| TIMIT | Garofalo et al. in 1990 | American English | – | 630 | More than 6300 | Suitable for Acoustic–Phonetic studies<br>Basically designed for speech recognition but getting used in ASV too |
| NIST | Sturim et al. in 2016 | American English | – | – | – | Designed for speaker recognition tasks but being used in ASV too<br>Getting modified constantly under an ongoing series |
| SAS | Wu et al. in 2015 | English | SS and VC | 45 M + 61 F in D<br>45 M + 61 F in E | 222,600 (10,600 B + 106,000 I + 106,000 S) in D<br>430,015 (22,831 B + 203,592 I + 203,592 S) in E | Specially designed for speaker verification<br>Suitable for TI-ASV systems |
| RedDots | Lee et al. in 20; Kinnunen et al. in 2017 | 5 different languages | Replay | 500 to 5000 | – | Designed for text dependent speaker recognition<br>Recorded by mobile devices |
| AVspoof | ASVspoof | English | Replay, SS and VC | 31 M + 13 F | 148,955 by M<br>27,365 by F | Replayed trials are recorded by 10 different setups<br>Audios are down sampled to 16 kHz as actually recorded on 44.1 kHz |
| Vox Celeb | VoxCeleb | Indian, American, Finnish, etc | Mimicry | 1251 in Vox Celeb 1<br>6112 in vox Celeb 2<br>(for both 61% M + 39% F) | 100,000 in Vox Celeb 1<br>1,000,000 in Vox Celeb 2 | Enrich with speakers from 6 different nations |
| voicePA | Korshunov et al. in 2018 | – | SS and VC | 44 (10 M + 4 F in T<br>10 M + 4 F in D<br>11 M + 5 F in E) | – | Attacked data is recorded by 5 different devices |
| BioCPqD-PA | Korshunov et al. in 2018 | Portuguese | Presentation | 222 | 418,940(27,253 B + 391,687 S) | 27 utterances were recorded in each session |

**Table 3** (continued)

| Dataset | Contributing community / authors | Language | Attack type | Number of speakers | Number of utterances | Remarks |
|---|---|---|---|---|---|---|
| ASVspoof 2015 | Wu et al. in 2015 | English | SS and VC | 45 M + 61 F | 16,375 (3750 B + 12,625 S) for T<br>53,372 (3497 B + 49,875 S) for D<br>193,406 (9406 B + 184,000 S) for E | Released for first ASVspoof challenge<br>Evaluation data has five unknown attack types |
| ASVspoof 2017 | Delgado et al. in 2018 | English | Replay | 42 | 3016 (1508 B + 1508 S) for T<br>1710 (760 B + 950 S) for D<br>< 50,000 E | Removed errors of labeling, zero sequence artifacte and empty file, etc. in version 2.0 |
| ASVspoof 2019 | Wang et al. in 2019 | English | SS, VC and replay | 8 M + 12 F for T<br>4 M + 6 F for D<br>21 M + 27 F for E | LA<br>25,380 for T<br>24,844 for D<br>71,177 E<br>PA<br>54,000 for T<br>29,700 for D<br>134,730 for E | Data set is partitioned in LA and PA subsets<br>Includes protocols for t-DCF |

2019 SRE is the new step of this ongoing series (Sturim et al., 2016).

#### 4.1.5 Spoofing and antispoofing (SAS) corpus

Spoofing and Antispoofing (SAS) corpus contains a diverse range of attacks generated by nine speech synthesis (SS) and VC algorithms. This database includes two protocols, one for evaluating the ASV system and another for creating the spoofed utterances. Synthetic speeches are also the part of the corpus along with the natural speeches. Non-realistic silence has been removed from the utterances, which leads the dataset to be more realistic SS and VC spoofed corpus (Wu et al., 2015a, 2015b, 2015c).

#### 4.1.6 RedDots

RedDots dataset has more numbers of recording sessions with less number of English utterances in each. The goal while designing the dataset was to provide 52 sessions on per week basis (one year) to each speaker. These sessions were two minutes longer with 24 sentences (10 commons, 10 unique, 2 free choices, 2 free texts) in each. This dataset has large variations of inter-speaker and intra-speaker type (Lee et al., 2015). After this, the Replayed RedDots dataset is created by re-recording the utterances of the original corpus under different environmental conditions. Both of these databases support the development of replay attack free ASV systems as original RedDots provides genuine utterances, and Replayed RedDots provides their related spoofed data (Kinnunen et al., 2017).

#### 4.1.7 AVspoof

AVspoof dataset is designed to support ASV systems as well as anti-spoofing techniques. This dataset contains SS, VC, and replay attacks balanced ratio. Replay attacks are generated by various recording devices to include the variations in spoofs, SS attacks are generated by HMM techniques mostly, and VC attacks are generated by Festvox. 31 males and 13 females participated in the recording of these sessions under various environmental conditions with different recording devices. Speakers were said to read out sentences, phrases and speak out about any topic freely for 3 to 10 min (ASVspoof).

#### 4.1.8 Vox Celeb

It is a collection of audio visual data extracted from videos chosen from youTube. Data set has a good diversity in the nationality of speakers as there are speakers with Indian, American, Finnish, etc. accents. 61% of speakers are male, and 39% of speakers are female. Utterances have at least 3 s in length (VoxCeleb, 2019). VoxCeleb1 and VoxCeleb2 are the two versions of this dataset, each having audio files, face videos, meta data about speakers, etc. in its training and testing set. Table 3 shows a number of speakers and utterances in these versions. Their Finnish language based

sets are contributing to mimicry attack detection for ASV systems (Vestman et al., 2020).

#### 4.1.9 voicePA

voicePA dataset is created with the help of AVspoof datase. Its genuine data is the subset of genuine data of AVspoof dataset spoken by 44 speakers, each contributed in four recording sessions organized in different environments (ASVspoof, 2019). These sessions were recorded by high quality microphones of a laptop, Samsung S3, and iPhone 3GS. Spoofed data consists of 24 types of presentation attacks recorded by five different devices in three different environments. These spoofed utterances are based on genuine data. SS and VC spoofed audios taken from the original dataset were also replayed (Korshunov et al., 2018a, b). Table 3 shows the number of male and female speakers in different sets of dataset.

#### 4.1.10 BioCPqD-PA

This dataset was recorded by 222 participants under different environmental conditions in the Portuguese language. The dataset contains 27,253 genuine, and 391,687 presentation attacked data. Presentation attacked audios were recorded under 24 configurations made by 8 loudspeakers and 3 microphones in an isolated room, whereas genuine data was recorded by one laptop. Dataset is partitioned into training, development and evaluation sets recorded by different pairs of the microphone and loudspeakers. Every set has voices of all participating speakers (Korshunov et al., 2018a, b).

#### 4.1.11 ASVspoof 2015

ASVspoof 2015 dataset is derived from the Spoofing and Anti-spoofing (SAS) dataset. It has TTS and VC spoofed utterances along with the genuine utterances. 45 males and 61 females contributed in the creation of genuine speeches and spoofed speeches are generated by ten different $S_1$ to $S_{10}$ algorithms. Here $S_1$ to $S_5$ are known attacks and $S_6$ to $S_{10}$ are unknown attacks that are introduced in the evaluation set. The whole dataset is partitioned into training (3750 genuine, 12,625 spoofed), development (3497 genuine, 49,875 spoofed) and evaluation (9404 genuine, 184,000 spoofed) sets (Wu et al., 2015a, 2015b, 2015c).

#### 4.1.12 ASVspoof 2017

ASVspoof 2017 dataset is designed from RedDots corpus to focus replay attack in ASV. It contains the voices of 42 speakers recorded under 61 different combinations of recording devices, replay devices and environmental conditions. It is collected during 179 sessions. Version 2.0 of this dataset was released in 2018 after removing the errors of labeling, empty files, zero-sequence artifacts, etc. (Delgado et al., 2018).

#### 4.1.13 ASVspoof 2019

ASVspoof 2019 dataset is extracted from the VTCK corpus. It is divided into Logical Access (LA) and Physical Access (PA) subsets. LA has TTS and VC spoofed speeches, and PA has replay spoofed speeches. Both of these subsets are further partitioned into training, development and evaluation subsets. The training subset is generated by 8 males and 12 females, development subset by 4 males and 6 females, and evaluation subset by 21 males and 27 females. These subsets are recorded under the same recording conditions; only the sets of speakers are disjoint. Training and development subsets contain known attacks generated by the same algorithms, and evaluation subset has unknown attacks too made by different synthesizing algorithms. This dataset includes two evaluation measure protocols, namely Equal Error Rate (EER) and Tendum Detection Cost Function (t-DCF) (Wang et al., 2019).

### 4.2 Evaluation measures

Evaluation of ASV systems is done on the basis of a predefined threshold. All the evaluation metrics consider the false acceptance and false rejections of the system. They indicate to reduce the false acceptance rates or to balance the trade-off between them. This section highlights some of the threshold based evaluation metrics of ASV.

#### 4.2.1 Equal error rate (EER)

ASV system either accepts or rejects the claimed identity. There are four possibilities for a classification to be correct or incorrect. These are True Acceptance (TA), True Rejection (TR), False Acceptance (FA) and False Rejection (FR). TA and TR are the desirable possibilities, but FR and FA are harmful situations for the system. These possibilities are set on the basis of a predefined threshold $\tau$ (Todisco et al., 2017). In the case of FA, a spoofed utterance having score more than or equal to $\tau$ gets accepted, and in case of FR, a genuine utterance having score less than $\tau$ gets rejected. To measure the performance of ASV Equal Error Rate (EER) is used, which is the value where False Acceptance Rate (FAR) (Eq. 14) and False Rejection Rate (FRR) (Eq. 15) become equal.

$$FAR = \frac{Count(FA)}{Count(spoofed\ utterances)} \tag{14}$$

$$FRR = \frac{Count(FR)}{Count(genuine\ utterances)} \tag{15}$$

#### 4.2.2 Detection error tradeoff (DET) curve

Initial Researches used Detection Error Tradeoff (DET) curve for evaluation of ASV systems. It is suitable for binary classification problems. It plots curves for EER taking FAR on the x-axis and FRR on the y-axis. Reynolds et at. (2000) are showing the comparative representation of their models with DET curves.

#### 4.2.3 Half total equal error rate (HTER)

FAR and FRR are inversely proportional to each other, so they can be illustrated as a function of predefined threshold $\tau$ for a particular dataset DS. Equation 16 shows the calculation method of HTER.

$$HTER_{(DS,\tau)} = \frac{FRR_{(DS,\tau)} + FAR_{(DS,\tau)}}{2} \tag{16}$$

#### 4.2.4 Tendum detection cost function (t-DCF)

Tendum Detection cost function (t-DCF) is an ASV centric evaluation measure (Kinnunen et al., 2018) ASVspoof, 2019 challenge has provided ASV and countermeasure protocols explicitly (Todisco et al., 2019). t-DCF function for a system can be calculated by Eq. 17.

$$tDCF_{SYST}(\tau) = Val_{FA}^{ASV}.\alpha_{tar}.L_{FA}^{ASV}(\tau) + Val_{FR}^{ASV}.\alpha_{non-tar}.L_{FR}^{ASV}(\tau) \tag{17}$$

where $Val_{FA}^{ASV}$ and $Val_{FR}^{ASV}$ are the cost values of false acceptance of an non-target utterance and false rejection of an target utterance respectively. $L_{FA}^{ASV}(\tau)$ and $L_{FR}^{ASV}(\tau)$ are the values of FAR and FRR for ASV system, respectively, at threshold $\tau$. $\alpha_{tar}$ is the probability of utterance being target and $\alpha_{non-tar}$ the probability of utterance being non-target. These all computations are performed with the assumption that an ideal countermeasure has 0% FR and FA that makes EER 0% (Eq. 18).

$$L_{FA}^{CM} = L_{FR}^{CM} = 0 \tag{18}$$

$L_{FA}^{CM}$ is the value of FAR of countermeasure and $L_{FR}^{CM}$ is the value of FRR of countermeasure at threshold $\tau$.

## 5 Spoofing attacks to the ASV systems

Spoofing attacks are categorized into direct and indirect access attacks on the basis of access required to part of the system while conducting them. Direct attacks are introduced via microphone and transmission channel, whereas indirect attacks are injected during the speech processing, distribution of information internally, classification, and even just before the declaration of result after the verification of claimer (Wu et al., 2015a, 2015b, 2015c). Various known direct spoofing attacks are categorized into Logical Access (LA) and Physical Access (PA) attacks (Chettri et al., 2020). LA attacks, generated algorithmically, are comprised of voice Conversion (VC) and Text to speech (TTS) spoofing attacks. Their best algorithms produce much similar speech to the bonafide ones, and these synthetic speeches are injected directly into the system with no involvement of microphones. On the other hand, PA attacks are accomplished by transmitting impersonated speech physically or by playing the recorded speech back in front of the microphone. These spoofing attacks are replay, mimicry and twines attacks (Chettri et al., 2019). Potential of risk from direct spoofing attacks increases due to enhancement in speech synthesizing toolkits (Suthokumar et al., 2017) and (Vestman et al., 2020) availability of user data publicly via social media, audio/video sharing platforms, service provider websites, etc. Figure 15 shows the complete classification of all types of attacks to an ASV system.

### 5.1 Direct access attacks

Direct spoofing attacks are the most common and easily attainable threats to the ASV systems. These attacks can be performed without complete knowledge of the system (Wu et al., 2015a, 2015b, 2015c). LA attacks are performed at transmission level, and PA attacks are performed at the speech input level (via microphone). Standard datasets like AVspoof, ASVspoof 2015, ASVspoof 2019, voicePA, etc. are enriched with these attacks.

#### 5.1.1 Logical access (LA) attacks

Progress in the development of voice synthesis algorithms has promoted the Logical Access (LA) attacks. Some online platforms and open source software tools (Festival and Festvox) are available to generate these threats directly. Injection of these attacks takes place directly via the transmission channel. Accessing the channel becomes easy in case of (Reynolds & Rose, 1995) telephonic

communications applied in banking, e-commerce, etc. To perform the attack imposter, pretending as a legitimate user puts a synthesized speech utterance into the channel and gains access to the system. The most general LA attacks are discussed below.

**5.1.1.1 Voice conversion (VC) spoofing attack** Although Voice Conversion (VC) has been a useful technique for the development of personified speech driven systems (Lim & Kwan, 2011; Patil & Kamble, 2018) from nearly the last two decades, it is broadly being used as a threat to ASV systems (Mohammadi & Kain, 2017; Pellom & Hansen, 1999). VC attacks are conducted by applying an artificial speech directly to the system, which is generated by converting the imposter's voice into the target speaker's voice. Converted speech can be achieved by GMM based, bilinear based, codebook based, neural network based, etc. approaches; some are defined in (Helander & Gabbouj, 2012). During training, the process of voice conversion involves a transformation function that is applied on phone or frame aligned utterances of imposter and target speaker. This function converts the characteristics of the voice of imposter's utterance into that of the target speaker's voice's (Patil & Kamble, 2018). This threat affects the performance of ASV systems considerably.

**5.1.1.2 Text-to-speech (TTS) or speech synthesis (SS) spoofing attack** Speech Synthesis (SS) spoofing attack is generated by the Text to Speech (TTS) method. Synthetic speech is produced in two steps by this method. Firstly, input text is converted into the rhetoric that will be having elements like phonemes. In the second step generation of speech waveform from the rhetoric takes place. Speech waveform can be generated by different approaches (Patil & Kamble, 2018; Wu et al., 2015a, 2015b, 2015c). Indumathi & Chandra (Indumathi & Chandra, 2012) present three generations (G1, G2, G3) of speech synthesis on the basis of acoustic characteristics, usage of speech corpus, and statistical model in this order. Formant synthesis and articulatory speech synthesis belong to first generation (G1), where formant synthesis is the very first technique in the field of SS, which tries to model the transfer function of the human vocal tract. Formant synthesis produces low quality robot like sounding speech, but articulatory speech synthesis is able to produce far better natural like speech as it simulates the biological sound production system (Karpe & Vernekar, 2018). Corpus based second generation (G2) includes the concatenative speech and sinusoidal synthesis techniques. G2 got initiated in 1980 with the usage of small datasets in SS, then in 1990, large sized corpora were collected (Patil & Kamble, 2018). The year 2000 was the start of statistical model based third generation (G3) that contributed to SS by Hidden Markov Model (HMM) based and unit selection based techniques. HMM, synthesis produces very natural speech by using maximum likelihood criteria to model characteristics like fundamental frequency, etc. of speech (Indumathi & Chandra, 2012). Unit selection is another approach of G3, which makes use of a large variety of speech from the corpus to deliver a better quality voice (Karpe & Vernekar, 2018). From 2010 a succeeding generation G4 based on deep learning is introducing improvement in acoustic characteristics prediction and overcoming the limitations of traditional techniques (Ze et al., 2013). Since the start of G4 various Deep Neural Network (DNN), Recurrent Neural Network (RNN), etc. based techniques have been proposed that provides a drastic improvement in this field. Figure 13 shows the complete view of these generations. These emerging SS technologies are contributing to the implementation of TTS systems like text reader, digital personal assistant, etc.

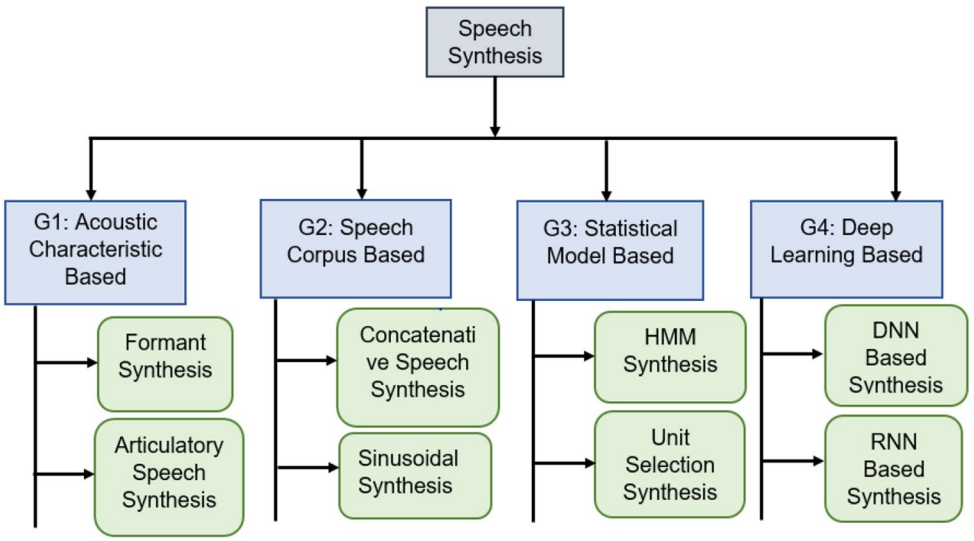



**Fig. 13** Generations of speech synthesis

#### 5.1.2 Physical access (PA) attacks

These attacks are carried out by presenting the spoofed utterances in front of the microphone of the ASV system. This is the easiest form of attack where imposter does not need to put extra efforts to generate the spoofed speech algorithmically to gain access to the system. The taxonomy of these attacks is presented below.

**5.1.2.1 Replay attack** Imposter needs a recording device that is easily available in today's era and good environmental conditions to conduct a replay attack on an ASV system (Wu et al., 2015a, 2015b, 2015c). To attack the system, he/she intentionally records the voice of a target registered user with the help of a recording device and plays it back in front of the input port of the system along with the insertion of the identity of the target user. A speech recorded in a noiseless environment is so much viable to attack the system successfully. But spoofed speech has distinctive clues like acoustic features, additive and convolutional distortions introduced by intermediate devices, etc. Mostly initial 400 ms are enough to classify an utterance spoofed or genuine (Chettri et al., 2018). ASVspoof 2017 challenge started taking care of replay attack by providing standard ASVspoof 2017 version 1.0 and version 2.0 datasets as previously there was very little data and research available for this attack type (Oo et al., 2019). With this dataset Lavrentyeva et al. (2017) show that deep learning based countermeasure identifies attacks better than the classical GMM approaches.

**5.1.2.2 Mimicry attack** The vulnerability of ASV systems for mimicry attacks was identified about fifty years ago (Vestman et al., 2020). After that, various experiments and analysis were negotiated to declare the fact officially. Lau et al. (2004) performed the experiment on a Gaussian Mixture Model based ASV countermeasure, trained by 138 speakers of the YOHO dataset (Lau et al., 2005), attacking it by two mimickers as imposters and they could verify that ASV systems can be attacked by mimicking the any valid user's voice. Other remarkable studies are presented by Hautamäki et al. (2013, 2014), where the authors are training a Cosine score based i-vector countermeasure and a Gaussian Mixture Model Universal Background Model (GMM-UBM) with Finnish language dataset. Because of mimicry attack, their experiment shows a significant increase in the False Acceptance Rate (FAR) for i-vector countermeasure. To conduct this attack, due to no involvement of technology, imposter needs zero efforts to cross the system's security barriers just by modifying his/her voice characteristics.

A Skillful attacker manipulates his/her prosodic features, lexical behaviour, etc. and produces the voice similar to a target user. In a very elegant approach (Fig. 14) to find out the attack, the professional finds out his/her matching speaker with the help of a general ASV system, practices on an experimental machine, and then attacks to a target ASV System. The same approach can be used to build the mimicry attack free ASV system with the help of a suitable dataset like VoxCeleb1, VoxCeleb2, etc. having voices of well-known people (Vestman et al., 2020).

**5.1.2.3 Twins attack** Identical twins are known to have similar voice features because they have similar or approximately similar shape and structure of the vocal tract. This makes twins attack a matching physiological characteristics kind of mimicry attack. But the spectrographic pattern of their speech samples shows the speaker specific variations of speech (Lindberg & Blomberg, 1999). Different sets of features like a set of MFCC and Target Energy Operator (TEO), a set of MFCC and Variable Length Target Energy Operator (VTEO), etc. are proved to be effective for different Indian languages in distinguishing the characteristics of

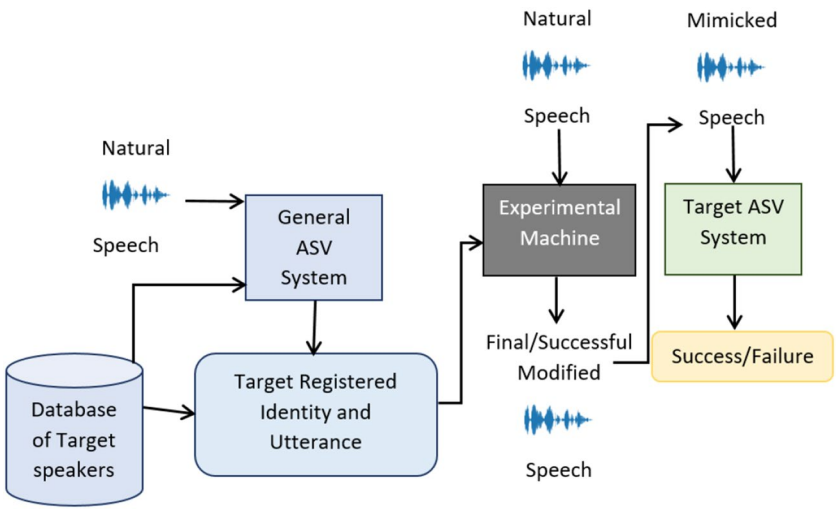



**Fig. 14** Mimicry attack scenario

identical twins or triplets (Patil & Parhi, 2009; Patil et al., 2017).

### 5.2 Indirect access attacks

The accomplishment of indirect access attacks needs the access to various parts of the ASV system. These parts include a feature extraction unit, the database of registered users, the verification model, and decision making unit, etc. (Wu et al., 2015a, 2015b, 2015c). Gaining access to these system components is a bit complicated and tricky. Therefore, indirect attacks are not so frequent. But even a single attack success hampers the privacy of users and confidentiality of data. So there is a need for identification of these attacks and prevention of system from them (Fig. 15).

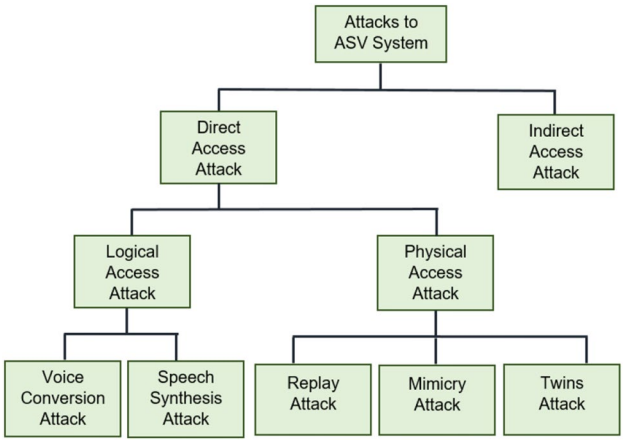



**Fig. 15** Types of attacks to the ASV system

## 6 ASV spoofing and countermeasures discussion

The work of designing spoof free countermeasures is consistently running since the arise of the speaker verification problem. Initially, classical machine learning was applied with different feature extraction techniques. But ASV countermeasures are enhanced in parallel with modification in feature extraction and classification technologies since always. A new era has begun in ASV by the introduction of efficient deep learning models in the machine learning field. We are providing a detailed discussion about the old and new era countermeasures below. Table 4 summarises the different aspects of countermeasures of these eras and Table 5 gives the all notations used in the paper.

### 6.1 ASV spoofing and countermeasures: old era

The most significant work for ASV systems can be traced from the latter decades of the twentieth century. Early ages of development used HMM, GMM, etc. These techniques contribute remarkably in designing the countermeasures, and they are compatible with almost all kinds of speech features. Villalba et al. (2015) are training a HMM with static and dynamic MFCC features, extracted from ATR speech data, for speaker verification task of a TD-ASV system that achieves 0% EER for human speakers and a HMM system is used for speech synthesis. The synthetic speech signals generated by the synthesis system are tested over the reference HMM model which, shows a remarkable value of false acceptance rate, i.e., 70% when they are using only a single synthetic statement from each user. Concatenation of isolated words, Re-synthesis, and diphone synthesis can provide the worst case scenario for the security of an ASV system. Among these techniques, word concatenation provides the most unbreakable utterances for HMM trained with LPCC features (Sztahó et al., 2019). Reynolds et al. (2000) state for TI-ASV systems that GMM has been the most successful model to find maximum likelihood. They are using GMM-UBM along with mel-scale cepstral coefficients to verify whether the speech is coming from the carbon button or electric based microphone handset. For carbon button microphone handset, they trained 2048 GMMs and for electric microphone handset 1024 gender dependent UBMs. Finally, they made a 2048 gender dependent UBM that is performing better than the others. HTK toolkit is a used for speech processing along with the HMM (Dua et al., 2012a, 2012b; Wong et al., 2001). MFCCs extracted by this method, along with the GMM, are used to check the vulnerability to mimicry attack (Lau et al., 2004). GMM with mixture of 32, 64 and 128 are performing best for three different sets of speeches made on the basis of duration (Shanmugapriya & Venkataramani, 2011). Speech synthesis (SS) and mimicry attacks were potential attacks to the system up to decades, but with the enhancement of technology, voice conversion (VC) also got added into the list of vulnerabilities to the ASV. This attack was included in ASVspoof 2015 dataset along with the SS. GMM of 128 mixtures trained with Cochlear Filter Cepstral Coefficients Instantaneous Frequency (CFCCIF) is used to take care of this attack (Kersta & Colangelo, 1970; Patel & Patil, 2015). SVM is used along with the neural networks and GMM (Godoy et al., 2015; Hanilçi et al., 2015). One-class SVM is used along with the DNN as a supportive model (Hanilçi et al., 2015). Similarly, k-means algorithm is playing the role of a supportive approach in ASV. Unvoiced speech frames have more information of spoof than the voiced speech. In the work of Wu et al., (2015a, 2015b, 2015c) an experiment has been carried out by separating low and high energy speech signal frames to detect the SS and VC attacks from ASVspoof 2015 and SAS and BTAS 2016

**Table 4** Summary of different aspects of old and new era countermeasures

| Author | Year | Dataset | Language | Attack types/speaker verification task | Feature extraction technique | Backend/classification model | Results |
|---|---|---|---|---|---|---|---|
| Masuko et al. | 1999 | ATR | Japanese | Speech Synthesis, TD-ASV | Mel-Cepstral Coefficients ($\Delta$, $\Delta\Delta$) | HMM | False Acceptance Rate (FAR) 0%, EER 70% |
| Lindberg J et al. | 1999 | Telephone Quality Speaker Database | English | Diphone Synthesis speech, re-synthesis speech | LPCC | HMM | EER 5.6% for re-synthesis one / Female And EER 0.6% For re-synthesis two / Male |
| Reynold et al. | 2000 | NIST SRE | English | Verification whether speech is from Carbon Button or Electret microphone handset, TI-ASV | mel-scale cepstral features | GMM-UBM | Shows good performance Detection Error Tradeoff (DET) curve |
| Wong & Russell | 2001 | YOHO and NOISEX-92, | English | Mimicry, TD-ASV | Static, $\Delta$, $\Delta^2$ of MFCC | Parallel Model Combination (PMC) with HMM | EER $<10\%$ at $+6$ dB SNR |
| Fenglei & Bingxi | 2002 | OGI | – | Verification Task, | MFCC | Speaker Clustering, SVM, GMM | EER 3.1% |
| Singh et al. | 2003 | KING, TIMIT | English | Speaker Verification | Mel Cepstrals | K-means Algorithm with GMM, Linde, Buzo, Gray (LBG) Algorithm with GMM | Expectation Maximization (EM) Accuracy 99.88% (LBG) 99.91% (K-means) |
| Ou & Ke | 2004 | Self-built database of 25 people | – | Speaker Verification | MFCC | GMM, Correlation Matrix (CM) | FRR GMM CK 21.7% 34.2% FAR GMM CK 1.75% 1.73% |
| Lau et al. | 2004 | YOHO | English | mimicry | MFCC | GMM | False Acceptance Error Rate (FAR) Imposter' FAR 0% Imitator1 FAR 5% Imitator2 FAR 20% Imitator3 FAR |
| Mezghani & O'Shaughnessy | 2005 | TIMIT | English | Speaker Verification | MFCC, Formants | Mahalanobis distance | EER 3.21% |
| Mariethoz & Bengio | 2006 | Pascal Couchepin, Swiss federal minister, Daniel Br´elaz, mayor of Lausanne, and Christian Constantin, head of the Sion Football Club | Swiss | mimicry | LFCC | GMM | n/a |
| Zetterholm | 2007 | well-known male Swedish voices, politicians and TV personalities | Swedish | mimicry | Mean Fundamental Frequency F0, Phoneme, Articulation Rate | Panel for Auditory Analysis | n/a |

**Table 4** (continued)

| Author | Year | Dataset | Language | Attack types/speaker verification task | Feature extraction technique | Backend/classification model | Results |
|---|---|---|---|---|---|---|---|
| Varchol et al. | 2008 | Speech data of 24 people TI-ASV | – | Impersonation | MFCC | GMM, UBM | EER 12.92% |
| Tadokoro et al. | 2009 | ATR SDB-I | – | Speaker Verification task | MFCC | HMM | EER 2.68% |
| Farrus et al. | 2010 | 5 Reputed Male Politicians | Spanish | mimicry | Prosodic Parameters | – | Identification Error Rate (IER) 5–22% |
| Munteanu & Toma | 2010 | Self-recorded 10 h speech data | Romanian | Speaker Recognition/ Verification | MFCC | HMM | $<1\%$ EER |
| Shanmugapriya & Venkataramani | 2011 | TIMIT | English | Speaker Verification | MFCC | Fuzzy Wavelet Neural (FWN) Model | 0.3297% EER |
| De Leon et al. | 2012 | WSJ | Engish | SS | MFCC, Relative Phase Shift (RPS) | GMM-UBM, SVM-GMM supervectors | For GMM-UBM 0.284%, For SVM-UBM Supervector 0.002% |
| Wu et al. | 2013 | WSJ0+WSJ1 | English | Speaker Adaptation, VC | Magnitude and Phase Spectrum of MFCC and MGDCC, Magnitude and Phase Modulation of MFCC and MGDCC, | HMM | For Fusion of Phase Modulaytion and Phase Features 0.89% |
| Hautamaki | 2013 | YLE | Finnish | mimicry | MFCC | GMM-UBM, i-vector Cosine | False Acceptance Rate-9–12% |
| Alam et al. | 2013 | NIST 2010 SRE | English | Speaker Verification | Multi-taper MFCC and PLP | GMM-UBM | For Multi- taper MFCC Reduction in 9.5% (Relative to baseline) For Multi-taper PLP Reduction in EER 5.0% (Relative to baseline) |
| Rajan et al. | 2013 | NIST 2010 SRE | English | Speaker Verification | Group Delay Function Generated from All Pole Method (SWLP) | GMM-UBM | For Male11.7% For Female 11.5% |
| Hautamaki et al. | 2014 | Voices of Presidents and Ministers (Sauli Niinisto, Martti Ahtisaari, Matti Vanhanen, Jouko Turkka), one businessman (Hjallis Harkimo) and one theatrical director (Jouko Turkka) | Finnish | mimicry | MFCC | GMM-UBM, i-vector | Table 2 EER% Material Test GMM-UBM i-vector Cosine Original audio Baseline 11.11 9.03 Mimicry attack 9.68 11.61 Enhanced audio Baseline 7.08 0.59 Mimicry attack 5.52 4.41 |
| Hanilci et al. | 2015 | ASVspoof 2015 | English | SS, VC | MFCC | GMM-ML | EER 3.01% |
| Patel & Patil | 2015 | ASVspoof 2015 | English | SS, VC | CFCCIF | GMM | EER 1.211% |

**Table 4** (continued)

| Author | Year | Dataset | Language | Attack types/speaker verification task | Feature extraction technique | Backend/classification model | Results |
|---|---|---|---|---|---|---|---|
| Paul et al. | 2015 | ASVspoof 2015 | English | SS, VC | Static, Δ, $\Delta^2$ of MFCC, MOBT, SFCC. SOBT, IMFCC, IMOBT, ISFCC, ISOBT | GMM-ML (Maximum likelyhood) | MOBT-<br>Static-3.212%,<br>Static + Δ $\Delta^{2-}$0.731%, Δ $\Delta^2$-0.098%<br>ISOBT-<br>Static-0.101%,<br>Static + Δ $\Delta^2$-0.004%, Δ $\Delta^2$-0.000% |
| Villalba et al. | 2015 | ASVspoof 2015 | English | SS, VC | Relative Phase Shift (RPS), Spectrum values | SVM + DNN (Fusion) | EER 0.002% |
| Janiki | 2015 | ASVspoof 2015 | English | SS,VC | LPC | Long Term Prediction (LTP) refered as Linear Predictive Analysis (LPA) | EER 11.616% |
| Godoy et al. | 2016 | ASVspoof 2015 | English | SS,VC | MFCC, MGDCC, DCT Coefficients | SVM, GMM,MLP | EER<br>SVM-6.719%<br>GMM-6.620%<br>MLP-6.527% |
| Al-Kaltakachi et al. | 2016 | TIMIT | English | Speaker Identification | MFCC and PNCC with Cepstral Mean and Varience (CMVN) and Feature wrapping (FW) | GMM-UBM | EER<br>SIA with Fusion 95.83% |
| Saranyaet al. | 2017 | NIST 2010 SRE | English | Speaker Verification | MFCC, MODGDF (Feature Switching) | GMM-UBM, i-vectors | Average Improvement over Feature Fusion and Baseline 34.5% (GMM-UBM) and 30.4% (i-vectors) |
| Jelil et al. | 2017 | ASVspoof 2017 | English | Replay | Instentaneous Frequency Cosine Coefficients, CQCC, MFCC | GMM Ensambal | EER<br>13.95% |
| Pal et al. | 2017 | ASVspoof 2015 | English | SS, VC | CQCC,APGDF, Fundamental Frequency Variation (FFV) | GMM | EER<br>0.05% |
| Suthokumar et al. | 2017 | ASVspoof 2015, BTAS 2016, Anti spoofing (SAS) | English | SS, VC | High and Low energy speech frames, IMFCC, LFCC, Δ, ΔΔ | GMM | EER<br>ASV 2015<br>ICMC–0.52%<br>LFCC–0.48%<br>BTAS<br>ICMC–0.10%<br>LFCC–0.08%<br>SAS<br>ICMC–0.55%<br>LFCC–0.40% |

**Table 4** (continued)

| Author | Year | Dataset | Language | Attack types/speaker verification task | Feature extraction technique | Backend/classification model | Results |
|---|---|---|---|---|---|---|---|
| Laverantyeva et al. | 2017 | ASVspoof 2017 | English | Replay | Log Power Magnitude + CQT, Log Power Magnitude + FFT | CNN + RNN | EER<br>6.73% |
| Scardapane et al. | 2017 | ASVspoof 2015 | English | SS, VC | MFCC | Deep RNN | EER<br>2.910% |
| Mohammadi & Mohammadi | 2017 | TIMIT, NOISEX92 | English | Noise | IMFCC, LFCC,MFCC,PNCC | GMM | EER<br>2.94% |
| Sriskandaraja et al. | 2018 | ASVspoof 2017 | English | Replay | Features generated by Light CNN (embeddings) | GMM | EER<br>6.4% |
| Shim et al. | 2018 | ASVspoof 2017 Version 1.0 | English | Replay | DNN Extracted Features (embeddings) | DNN | EER<br>9.56% |
| Korshunov et al. | 2018 | BioCPqD-PA | – | Presentation Attack/ Replay | MFCC | CNN | EER<br>7.01% |
| Zhao et al. | 2018 | ASVspoof 2015 | English | SS, VC | CQCC, SCC | GMM | EER<br>0.10% |
| Yang et al. | 2018 | ASVspoof 2015 and ASVspoof 2017 Version 2.0 | English | SS, VC, Replay | eCQCC | DNN | EER<br>With ASVspoof 2015–0.04%<br>With ASVspoof 2017–13.38% |
| Chettri et al. | 2018 | ASVspoof 2017 Version 2.0 | English | Replay | Log Power Magnitude + FFT | CNN + SLIME Algorithm | EER<br>10.6% |
| Dinke et al. | 2018 | ASVspoof 2015, BTAS 2016 | English | SS, VC | $CQCC_{8k}$-DD | CLDN (Convolutional LSTM Neural Network, Joint) | EER<br>With BTAS2016-0.171%<br>With ASVspoof 2015–4.56% |
| Balamurli et al. | 2019 | ASVspoof 2017 | English | Replay | CQCC,MFCC, LPCC, RFCC, CCC, SCFC, SCMC, IMFCC, LFCC, Spectrogram and Autoencoder reconstructed features | GMM-UBM<br>Fusion | EER<br>10.8 |
| Oo et al. | 2019 | ASVspoof 2017 | English | Replay | CQCC, MFCC, MGDCC, Gammatone Relative Phase (RP) Features, Mel-Scale-RP | GMM | EER<br>For CQCC + Gammatone Scale—RP 9.48% |

**Table 4** (continued)

| Author | Year | Dataset | Language | Attack types/speaker verification task | Feature extraction technique | Backend/classification model | Results |
|---|---|---|---|---|---|---|---|
| Chettri et al. | 2019 | ASVspoof 2019 | English | SS,VC, Replay | MFCC, IMFCC, SCMC, Long Term Average Spectrum, CQCC, etc | CNN + RNN + 1D CNN + Wave U Net + SVM + GMM(Fusion, For LA)<br>CNN + CRNN + Wave U Net + GMM with MFCC, IMFCC, SCMC + SVM with i-vector, long term average spectrum (Fusion for PA) | EER<br>For LA 2.64%<br>For PA 6.11%<br>tDCF<br>For LA 0.0755%<br>For PA 0.1492% |
| Haung et al. | 2019 | ASVspoof 2017 | English | Replay | CQCC, MFCC (Hybrid) | Dense Net LSTM | EER<br>8.84% |
| Cai et al. | 2019 | ASVspoof 2019 | English | Reply | CQCC, LFCC, IMFCC, STFT gram | DNN, ResNet | EER<br>DNN 0.66<br>tDCF<br>Fusion of ResNet with different features 0.0168 |
| Kumar et al. | 2020 | ASVSpoof 2019 | English | SS,VC,Replay | CQCC, LFCC, IMFCC, LFBC | Time-Delay Shallow Neural Network (TD-SNN) | EER<br>For SS and VC- 5.7%<br>For Replay 6.4% |

datasets. Static and dynamic features extracted by different feature extraction techniques like MFCC, Instantaneous Frequency Cosine Coefficients (IFCC), CQCC, etc. have been compared by applying on GMM based systems. CQCC features show the best performance when single system are compared, but an ensemble of all features performed the best for spoof detection task when trained with replay attacked data of ASVspoof 2017 dataset (Jelil et al., 2017). After ASVspoof 2017 challenge, research for replay attacks drastically increased, and a lot of reliable countermeasures are available now for this single attack type. This old era was entirely dominated by GMMs up to decades, but rise of deep learning attracted the research community towards neural network based models. And a new era started in the development of ASV systems. Since classical machine learning techniques or, more specifically, GMMs are easy to understand and suitable for speech involving systems, still a lot of research is going on involving these techniques.

### 6.2 ASV spoofing and countermeasures: new era

Deep learning based countermeasures are easy to implement and need less preprocessing of data. The noticeable time period when ASV research adopted deep learning can be marked from less than a decade ago. As the research is going on in deep learning from DNN to CNN than RNN, LSTM, autoencoders, etc. parallel adoption of technology is running by the ASV community. Speaker verification is a two class classification problem DNN can work well for this classification. Hanilci et al. (2015) are using DNN along with the SVM for ASVspoof 2015 dataset having SS and VC attacks. A fusion of these models is achieving less than 0.05% EER for nine spoofing attack types out of ten. This DNN model is using Softmax in the output layer.

Deep learning models are suitable for raw input audio signals also if the signal is presented in a proper way to the model, even they can walk through a music audio file (Lee et al., 2017; Morfi & Stowell, 2018). With the ASVspoof 2017 challenge replay attack got new insights. It promoted a huge study for this attack type. A CNN network with five convolutional, two fully-connected, five Softmax and four network-in-network layers is achieving good accuracy (Chettri et al., 2018). VoicePA and BioCPqD-PA datasets are used with a shallow CNN with 20 neurons in the convolutional layer and a deep CNN with three convolutional layers for Presentation Attack Detection (PAD) (Korshunov et al., 2018a, b). These networks are trained by MFCC features. The ensemble of ResNets, the ensemble of Fully Convolutional Neural Networks (FCN) and the ensemble of neural networks are proved to be most suitable for time series classification tasks (Fawaz et al., 2019). Chattri et al. (2019) are training CNN, Convolutional Recurrent Neural

**Table 5** Notations used in the paper

| Phrase | Annotation |
|---|---|
| Automatic Speaker Verification | ASV |
| Gaussian Mixture Model | GMM |
| Hidden Markov Model | HMM |
| Speech Synthesis | SS |
| Text-to-Speech | TTS |
| Voice Conversion | VC |
| text-dependent ASV | TD-ASV |
| text-independent ASV | TI-ASV |
| Linear Predictive Coding | LPC |
| Mel Frequency Cepstrum Coefficients | MFCC |
| Linear Predictive Cepstrum Coefficients | PLCC |
| Linear Predictive Coding | LPC |
| Mean-Square-Error (MSE) | MSE |
| delta | $\Delta$ |
| delta-delta | $\Delta\Delta$ |
| Mel-frequency Cepstral Coefficients | MFCCs |
| Fast Fourier Transform | FFT |
| Discrete Fourier Transform | DFT |
| Discrete Cosine Transform | DCT |
| Inverse Mel-frequency Cepstral Coefficients | IMFCCs |
| Linear Frequency Cepstral Coefficients | LFCC |
| Constant Q Cepsral Coefficients | CQCC |
| Constant Q Transform | CQT |
| instantaneous frequency cosine coefficient | IFCC |
| Extended Constant Q Cepstral Coefficients | eCQCC |
| Linear Predictive Cepstrum Coefficients | LPCC |
| Linear Predictive Coding | LPC |
| Perceptual Linear Predictive | PLP |
| Power Normalization Cepstrum Coefficients | PNCC |
| Signal to Noise Ratio | SNR |
| Short-term Fourier Transform | STFT |
| All-Pole Group Delay Function | APGDF |
| Modified Group Delay Function | MODGDF |
| Sub-band Centroid Frequency Coefficients | SCFC |
| Deep Neural Network | DNN |
| Time Delay Neural Network | TDNN |
| Gaussian Mixture Model | GMM |
| Expectation Maximization | EM |
| Universal Background Model | UBM |
| Maximum a posteriori | MAP |
| Radial Bias Function | RBF |
| Wall Street Journal | WSJ |
| Speaker Recognition Evaluation | SRE |
| Spoofing and Antispoofing | SAS |
| Logical Access | LA |
| Physical Access | PA |
| Equal Error Rate | EER |
| Tendum Detection Cost Function | t-DCF |
| True Acceptance | TA |
| True Rejection | TR |
| False Acceptance | FA |
| and False Rejection | FR |
| False Acceptance Rate | FAR |
| False Rejection Rate | FRR |
| Detection Error Tradeoff | DET |
| Half Total Equal Error Rate | HTER |
| Recurrent Neural Network | RNN |
| Target Energy Operator | TEO |
| Variable Length Target Energy Operator | VTEO |
| Cochlear Filter Cepstral Coefficients Instantaneous Frequency | CFCCIF |
| Instantaneous Frequency Cosine Coefficients | IFCC |
| Presentation Attack Detection | PAD |
| Fully Convolutional Neural Networks | FCN |
| Convolutional Recurrent Neural Network | CRNN |
| One Dimensional CNN | 1D-CNN |
| Parallel Model Combination | PMC |
| Linde, Buzo, Gray | LBG |
| Expectation Maximization | EM |
| Correlation Matrix | CM |
| Identification Error Rate | IER |
| Fuzzy Wavelet Neural | FWN |
| Relative Phase Shift | RPS |
| Relative Phase Shift | RPS |
| Long Term Prediction | LTP |
| Cepstral Mean and Varience | CMV |
| Feature wrapping | FW |
| Fundamental Frequency Variation | FFV |
| Time-Delay Shallow Neural Network | TD-SNN |

Network (CRNN), One Dimensional CNN (1D-CNN), and Wave-U-Net models along with the classical classification models with ASVspoof 2019 dataset. Time and frequency representation of speech signal is applied to the input layers of these networks. Models are trained with early stopping criteria with a binary cross-entropy loss function. Adam is the most used optimizer in ASV related as well as other purpose networks. Individual performance is measured by EER and t-DCF for each model, but an ensemble unit is performing the best for both the measures (Chettri et al., 2019). ASVspoof 2019 dataset has two partitions LA and PA, both focusing on SS, VC and replay attacks, respectively.

All the studies show deep learning is also suitable for time varying speech signals presented either directly or with the help of any feature extraction technique. The ensemble of different models achieves better performance than the individual models. Mimicry and Twins attacks are not targeted much with modern machine learning techniques. This deep

learning based new era is giving more prospects of enhancement in the situation of ASV systems.

## 7 Conclusion, discussion and future expectations

This paper focused to each part of ASV systems by analysing different research works in this field. This survey finds out that MFCCs are the most commonly used and reliable features, but CQCC features proved to be most efficient (achieving better performance). This survey notices that dynamic features can model speaker and speech specific information better than static features. Some of the prior databases or corpora were focusing only on speaker recognition tasks. Datasets provided by ASVspoof community in 2015, 2017, and 2019 are more enrich corpora with the direct access attacks than the classical datasets. Mimicry attack related research is focusing on different languages, but other attacks are conducted mostly in the English language. Although deep learning has put its step in ASV systems, but researchers are still attracted to GMM. In classical learning, GMMs are suitable for TI-ASV systems, whereas HMMs are suitable for TD-ASV systems as they model the temporal information of known text and not efficient for TI-ASV as compared to GMMs. Deep learning is being used to train with already extracted features from different feature extraction techniques as well as with raw speech waveforms. Use of deep learning has started a new era in these systems. Spoofing techniques are also getting improved in parallel along with the enhancement in ASV countermeasures. Deep learning has given new insights in SS or TTS spoofing attack also with the start of a new generation ($G_4$) in text to speech conversion. But the position of the ASV system is fine enough that industry is using these systems practically. This research provides sufficient information with the best efforts for development and, starting or continuing research in ASV systems to the beginners also. We are listing out some suggestions and facts to be considered for future work in ASV systems:

i. Till now, different countermeasures are suitable for different particular attack set and they are working well with different particular speech corpora. We seek the attention of the research community to design a single countermeasure working well with all aspects.
ii. Mimicry and Twins attacks are not being targeted much with modern machine learning techniques. Although, early research with GMM focused on mimicry thoroughly. So we advocate to focus on these attack types with deep learning in near future.
iii. Although, indirect access attacks require different kind of efforts to get encountered but they should also be in notice while installing the ASV system. Because even single threat can break the robustness of whole security mechanism.
iv. A perfect set of features that can model all kinds of variations of speech and a perfect combination of classification models to design a countermeasure is needed to be chosen.
v. Single standard dataset, including every kind of possible attacks, utterances spoken by males and females of each age group in different languages under different environmental conditions seems to be required. More specifically, recently designed datasets are not having twins and mimicry attacks.

This paper has reviewed various technologies and advancements proposed in this area, and it brings a huge knowledge base of this area at one place. For future work, one countermeasure that can cope with all kinds of spoofing attacks should be the next target of research in this area, and development of hybrid systems should be practiced to involve the benefits of different technologies.

## Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References


Aggarwal, R. K., & Kumar, A. (2020). Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modeling.

Alam, M. J., Kinnunen, T., Kenny, P., Ouellet, P., & O'Shaughnessy, D. (2013). Multitaper MFCC and PLP features for speaker verification using i-vectors. *Speech Communication, 55*(2), 237–251.

Al-Kaltakchi, M. T., Woo, W. L., Dlay, S. S., & Chambers, J. A. (2016, March). Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification. In *4th international conference on biometrics and forensics* (IWBF) (pp. 1–6). IEEE.

ASVspoof consortium. (2019). ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan*. http://www.asvspoof.org/.

ASVspoof. (2019): https://www.idiap.ch/dataset/avspoof

Balamurali, B. T., Lin, K. E., Lui, S., Chen, J. M., & Herremans, D. (2019). Toward robust audio spoofing detection: A detailed comparison of traditional and learned features. *IEEE Access, 7*, 84229–84241.

Beranek, B. (2013). Voice biometrics: Success stories, success factors and what's next. *Biometric Technology Today, 2013*(7), 9–11.

Brown, J. C. (1991). Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America, 89*(1), 425–434.

Brown, J. C., & Puckette, M. S. (1992). An efficient algorithm for the calculation of a constant Q transform. *The Journal of the Acoustical Society of America, 92*(5), 2698–2701.

Cai, W., Wu, H., Cai, D., & Li, M. (2019). The dku replay detection system for the asvspoof 2019 challenge: On data augmentation, feature representation, classification, and fusion. http://arxiv.org/abs/arXiv:1907.02663

Campbell, J. P. (1995, May). Testing with the YOHO CD-ROM voice verification corpus. In *1995 international conference on acoustics, speech, and signal processing* (vol. 1, pp. 341–344). IEEE.

Chakroborty, S., & Saha, G. (2009). Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter. *International Journal of Signal Processing, 5*(1), 11–19.

Chen, K., & Salman, A. (2011). Learning speaker-specific characteristics with a deep neural architecture. *IEEE Transactions on Neural Networks, 22*(11), 1744–1756.

Chen, N., Qian, Y., & Yu, K. (2015). Multi-task learning for text-dependent speaker verification. Sixteenth annual conference of the international speech communication association.

Chen, Z., Zhang, W., Xie, Z., Xu, X., & Chen, D. (2018, April). Recurrent neural networks for automatic replay spoofing attack detection. In *2018 IEEE international conference on acoustics, speech and signal processing* (ICASSP) (pp. 2052–2056). IEEE.

Chettri, B., Kinnunen, T., & Benetos, E. (2020). Deep generative variational autoencoding for replay spoof detection in automatic speaker verification. *Computer Speech & Language, 101092*.

Chettri, B., Mishra, S., Sturm, B. L., & Benetos, E. (2018, December). Analysing the predictions of a CNN-based replay spoofing detection system. In *2018 IEEE spoken language technology workshop* (SLT) (pp. 92–97). IEEE.

Chettri, B., Stoller, D., Morfi, V., Ramírez, M. A. M., Benetos, E., & Sturm, B. L. (2019). Ensemble models for spoofing detection in automatic speaker verification. arXiv preprint arXiv:1904.04589.

Cheuk, K. W., Anderson, H., Agres, K., & Herremans, D. (2019). nnAudio: An on-the-fly GPU audio to spectrogram conversion toolbox using 1D convolution neural networks. https://abs/arXiv:1912.12055.

De Leon, P. L., Pucher, M., Yamagishi, J., Hernaez, I., & Saratxaga, I. (2012). Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(8), 2280–2290.

Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K., & Yamagishi, J. (2018, June). ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements.

Dinkel, H., Qian, Y., & Yu, K. (2018). Investigating raw wave deep neural networks for end-to-end speaker spoofing detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26*(11), 2002–2014.

Dua, M., Aggarwal, R. K., & Biswas, M. (2017). Discriminative training using heterogeneous feature vector for Hindi automatic speech recognition system. In *International conference on computer and applications* (ICCA) (pp. 158–162).

Dua, M., Aggarwal, R. K., & Biswas, M. (2018a). Discriminative training using noise robust integrated features and refined HMM modeling. *Journal of Intelligent Systems, 29*(1), 327–344.

Dua, M., Aggarwal, R. K., & Biswas, M. (2018b). Performance evaluation of Hindi speech recognition system using optimized filterbanks. *International Journal, Engineering Science and Technology, 1*(3), 389–398.

Dua, M., Aggarwal, R. K., & Biswas, M. (2019a). Discriminatively trained continuous Hindi speech recognition system using interpolated recurrent neural network language modeling. *Neural Computing and Applications, 31*(10), 6747–6755.

Dua, M., Aggarwal, R. K., & Biswas, M. (2019b). GFCC based discriminatively trained noise robust continuous ASR system for Hindi language. *Journal of Ambient Intelligence and Humanized Computing, 10*(6), 2301–2314.

Dua, M., Aggarwal, R. K., Kadyan, V., & Dua, S. (2012a). Punjabi automatic speech recognition using HTK. *International Journal of Computer Science Issues (IJCSI), 9*(4), 359.

Dua, M., R. K. Aggarwal, Kadyan, V., Dua, S., (2012). Punjabi speech to text system for connected words, 206–209.

Dua, M., Jain, C., & Kumar, S. (2021). LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-02960-0

Farrus, M., Wagner, M., Erro, D., & Hernando, F. J. (2010). Automatic speaker recognition as a measurement of voice imitation and conversion. *International Journal of Speech, Language and the Law, 1*(17), 119–142.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019, July). Deep neural network ensembles for time series classification. In *International joint conference on neural networks* (IJCNN) (pp. 1–6). IEEE.

Fenglei, H., & Bingxi, W. (2002, August). Text-independent speaker verification using speaker clustering and support vector machines. In *International conference on signal processing* (Vol. 1, pp. 456–459). IEEE.

Garofalo, J. S., Lamel, L. F., & Fisher, W. M. (1990). The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NIST.

Glover, J. C., Lazzarini, V., & Timoney, J. (2011). Python for audio signal processing.

Godoy, A., Sim˜oes, F., Stuchi, J. A., Angeloni, M. d. A., Uliani, M., & Violato, R. (2015). Using deep learning for detecting spoofing attacks on speech signals. arXiv preprint arXiv:1508.01746.

Gong, Y., & Yang, J., (2020). Detecting replay attacks using multi-channel audio: a neural network-based method, arXiv:2003.08225v1 [cs.SD].

Hanilçi, C., Kinnunen, T., Sahidullah, M., & Sizov, A. (2015). Classifiers for synthetic speech detection: A comparison.

Hautamäki, R. G., Kinnunen, T., Hautamäki, V., & Laukkanen, A. M. (2014). Comparison of human listeners and speaker verification systems using voice mimicry data. TARGET, 4000, 5000.

Hautamäki, R. G., Kinnunen, T., Hautamäki, V., Leino, T., & Laukkanen, A. M. (2013). I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *Interspeech* (pp. 930–934).

Hegde, R. M., Murthy, H. A., & Rao, G. R. (2004, May). Application of the modified group delay function to speaker identification and discrimination. In *IEEE international conference on acoustics, speech, and signal processing* (Vol. 1, p. I-517). IEEE.

Helander, E., & Gabbouj, M. (2012). Jani Nurminen1, Hanna Silén2, Victor Popa2. Speech Enhancement, Modeling And Recognition–Algorithms And Applications, 69.

Huang, L., & Pun, C. M. (2019, May). Audio replay spoof attack detection using segment-based hybrid feature and DenseNet-LSTM network. In *IEEE international conference on acoustics, speech and signal processing* (ICASSP) (pp. 2567–2571). IEEE.

Indumathi, A., & Chandra, E. (2012). Survey on speech synthesis. *Signal Processing: An International Journal (SPIJ), 6*(5), 140.

Janicki, A. (2015). Spoofing countermeasure based on analysis of linear prediction error. In *Proc. Interspeech*.

Jelil, S., Das, R. K., Prasanna, S. M., & Sinha, R. (2017, August). Spoof detection using source, instantaneous frequency and cepstral features. In *Interspeech* (pp. 22–26).

Kadyan, V., Dua, M., & Dhiman, P. (2021a). Enhancing accuracy of long contextual dependencies for Punjabi speech recognition system using deep LSTM. *International Journal of Speech Technology,* 1–11.

Kadyan, V., Shanawazuddin, S., & Singh, A. (2021b). Developing children's speech recognition system for low resource Punjabi language. *Applied Acoustics, 178*, 108002.

Kamble, M. R., Sailor, H. B., Patil, H. A., & Li, H. (2020). Advances in anti-spoofing: From the perspective of ASVspoof challenges. *APSIPA Transactions on Signal and Information Processing*. https://doi.org/10.1017/ATSIP.2019.21

Karpe, R., & Vernekar, N. (2018). A survey: On text to speech synthesis. *International Journal for Research in Applied Science and Engineering Technology, 6*, 351–355.

Kersta, L., & Colangelo, J. (1970). Spectrographic speech patterns of identical twins. *The Journal of the Acoustical Society of America, 47*(1), 58–59.

Kim, C., & Stern, R. M. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(7), 1315–1329.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication, 52*(1), 12–40.

Kinnunen, T., Lee, K. A., Delgado, H., Evans, N., Todisco, M., Sahidullah, M., & Reynolds, D. A. (2018). t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. arXiv preprint arXiv:1804.09618.

Kinnunen, T., Sahidullah, M., Falcone, M., Costantini, L., Hautamäki, R. G., Thomsen, D., & Evans, N. (2017, March). Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research. In *IEEE international conference on acoustics, speech and signal processing* (ICASSP) (pp. 5395–5399). IEEE.

Koolwaaij, J. W., & Boves, L. W. J. (1999). On the use of automatic speaker verification systems in forensic casework.

Korshunov, P., Gonçalves, A. R., Violato, R. P., Simões, F. O., & Marcel, S. (2018, January). On the use of convolutional neural networks for speech presentation attack detection. In *2018 IEEE 4th international conference on identity, security, and behavior analysis* (ISBA) (pp. 1–8). IEEE.

Korshunov, P., Gonçalves, A. R., Violato, R. P., Simões, F. O., & Marcel, S. (2018, January). On the use of convolutional neural networks for speech presentation attack detection. In *4th international conference on identity, security, and behavior analysis* (ISBA) (pp. 1–8). IEEE.

Kumar, A., & Aggarwal, R. K. (2020a). A hybrid CNN-LiGRU acoustic modeling using raw waveform sincnet for Hindi ASR. *Computer Science, 2*, 89. https://doi.org/10.7494/csci.2020.21.4.3748

Kumar, A., & Aggarwal, R. K. (2020b). Hindi speech recognition using time delay neural network acoustic modeling with i-vector adaptation. *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-020-09757-0

Kumar, A., & Aggarwal, R. K. (2020d). A time delay neural network acoustic modeling for hindi speech recognition. In *Advances in data and information sciences* (pp. 425–432). Singapore: Springer.

Kumar, A., & Aggarwal, R. K. (2021). An exploration of semi-supervised and language-adversarial transfer learning using hybrid acoustic model for hindi speech recognition. *Journal of Reliable Intelligent Environments,* 1–16.

Kumar, M. G., Kumar, S. R., Saranya, M. S., Bharathi, B., & Murthy, H. A. (2019, December). Spoof detection using time-delay shallow neural network and feature switching. In *Automatic speech recognition and understanding workshop* (ASRU) (pp. 1011–1017). IEEE.

Lau, Y. W., Tran, D., & Wagner, M. (2005). Testing voice mimicry with the yoho speaker verification corpus. In *International conference on knowledge-based and intelligent information and engineering systems* (pp. 15–21). Springer.

Lau, Y. W., Wagner, M., & Tran, D. (2004, October). Vulnerability of speaker verification to voice mimicking. In *International symposium on intelligent multimedia, video and speech processing* (pp. 145–148). IEEE.

Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., & Shchemelinin, V. (2017, August). Audio replay attack detection with deep learning frameworks. In *Interspeech* (pp. 82–86).

Lee, J., Park, J., Kim, K. L., & Nam, J. (2017). Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. arXiv preprint arXiv:1703.01789.

Lee, K. A., Larcher, A., Wang, G., Kenny, P., Brümmer, N., Leeuwen, D. V & Li, H. (2015). The RedDots data collection for speaker recognition. In *Sixteenth annual conference of the international speech communication association.*

Lim, R., & Kwan, E. (2011, August). Voice conversion application (VOCAL). In *International conference on uncertainty reasoning and knowledge engineering* (Vol. 1, pp. 259–262). IEEE.

Lindberg, J., & Blomberg, M. (1999). Vulnerability in speaker verification-a study of technical impostor techniques. In *Sixth European conference on speech communication and technology.*

Mariéthoz, J., & Bengio, S. (2005). Can a professional imitator fool a GMM-based speaker verification system? (No. REP_WORK). IDIAP.

Marinov, S. (2003). Text dependent and text independent speaker verification systems. Technology and applications. Overview article.

Masuko, T., Hitotsumatsu, T., Tokuda, K., & Kobayashi, T. (1999). On the security of HMM-based speaker verification systems against imposture using synthetic speech. In *Sixth European conference on speech communication and technology*.

Mezghani, A., & O'Shaughnessy, D. (2005, May). Speaker verification using a new representation based on a combination of MFCC and formants. In *Canadian conference on electrical and computer engineering* (pp. 1461–1464). IEEE.

Mittal A., Dua M. (2021a). Constant Q Cepstral Coefficients and Long Short-Term Memory Model-Based Automatic Speaker Verification System. Proceedings of International Conference on Intelligent Computing, Information and Control Systems. Advances in Intelligent Systems and Computing, 1272, 895–904.

Mittal A., Dua M. (2021b). Automatic speaker verification system using three dimensional static and contextual variation-based features with two dimensional convolutional neural network. International Journal of Swarm Intelligence.

Mohammadi, M., & Mohammadi, H. R. S. (2017, May). Robust features fusion for text independent speaker verification enhancement in noisy environments. Iranian Conference on Electrical Engineering (ICEE), 1863–1868. IEEE.

Mohammadi, S. H., & Kain, A. (2017). An overview of voice conversion systems. *Speech Communication, 88*, 65–82.

Morfi, V., & Stowell, D. (2018). Deep learning for audio event detection and tagging on low-resource datasets. *Applied Sciences, 8*(8), 1397.

Munteanu, D. P., & Toma, S. A. (2010, June). Automatic speaker verification experiments using HMM. In 2010 8th International Conference on Communications, 107–110. IEEE.

Ochiai, T., Matsuda, S., Lu, X., Hori, C., & Katagiri, S. (2014, May). Speaker adaptive training using deep neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6349–6353. IEEE.

Oo, Z., Wang, L., Phapatanaburi, K., Liu, M., Nakagawa, S., Iwahashi, M., & Dang, J. (2019). Replay attack detection with auditory filter-based relative phase features. *EURASIP Journal on Audio, Speech, and Music Processing, 2019*(1), 8.

Ou, G., & Ke, D. (2004, December). Text-independent speaker verification based on relation of MFCC components. International Symposium on Chinese Spoken Language Processing, 57–60. IEEE.

Pal, M., Paul, D., & Saha, G. (2018). Synthetic speech detection using fundamental frequency variation and spectral features. *Computer Speech & Language, 48*, 31–50.

Paliwal, K. K. (1998, May). Spectral subband centroid features for speech recognition. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), 2, 617–620. IEEE.

Patel, T. B., & Patil, H. A. (2015). Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. Sixteenth Annual Conference of the International Speech Communication Association.

Patil, H. A., & Kamble, M. R. (2018, November). A survey on replay attack detection for automatic speaker verification (ASV) system. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1047–1053. IEEE.

Patil, H. A., & Parhi, K. K. (2009, December). Variable length Teager energy based mel cepstral features for identification of twins. In: *International conference on pattern recognition and machine intelligence* (pp. 525–530). Berlin: Springer.

Patil, H. A., Kamble, M. R., Patel, T. B., & Soni, M. H. (2017, August). Novel variable length Teager energy separation based instantaneous frequency features for replay detection. In *INTERSPEECH* (pp. 12–16).

Paul, D. B., & Baker, J. M. (1992, February). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the workshop on speech and natural language* (pp. 357–362). Association for Computational Linguistics.

Paul, D., Pal, M., & Saha, G. (2015, December). Novel speech features for improved detection of spoofing attacks. In *Annual IEEE India conference* (INDICON) (pp. 1–6). IEEE.

Pellom, B. L., & Hansen, J. H. (1999, March). An experimental study of speaker verification sensitivity to computer voice-altered imposters. In *International conference on acoustics, speech, and signal processing. proceedings*. ICASSP99 (Cat. No. 99CH36258) (Vol. 2, pp. 837–840). IEEE.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE, 81*(9), 1215–1247.

Pritam, L. S., Jainar, S. J., & Nagaraja, B. G. (2018). A comparison of features for multilingual speaker identification—A review and some experimental results. *International Journal of Recent Technology and Engineering* (IJRTE), *7* (4S2).

Prithvi, P., & Kumar, T. K. (2016). Comparative analysis of MFCC, LFCC, RASTA-PLP. *International Journal of Scientific Engineering and Research, 4*(5), 1–4.

Rajan, P., Kinnunen, T., Hanilci, C., Pohjalainen, J., & Alku, P. (2013, August). Using group delay functions from all-pole models for speaker recognition. In *Interspeech* (pp. 2489–2493).

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing, 3*(1), 72–83.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing, 10*(1–3), 19–41.

Rose, R. C., & Juang, B. H. (1996). Hidden Markov models for speech and signal recognition. *Electroencephalography and Clinical Neurophysiology. Supplement, 45*, 137–152.

Sahidullah, M., Delgado, H., Todisco, M., Kinnunen, T., Evans, N., Yamagishi, J., & Lee, K. A. (2019). Introduction to voice presentation attack detection and recent advances. *Handbook of biometric anti-spoofing* (pp. 321–361). Springer.

Sahidullah, M., Delgado, H., Todisco, M., Yu, H., Kinnunen, T., Evans, N., & Tan, Z. H. (2016). Integrated spoofing countermeasures and automatic speaker verification: An evaluation on ASVspoof 2015.

Sahidullah, M., Kinnunen, T., & Hanilçi, C. (2015). A comparison of features for synthetic speech detection.

Saranya, M. S., & Murthy, H. A. (2018). Decision-level feature switching as a paradigm for replay attack detection. In *Interspeech* (pp. 686–690).

Saranya, M. S., Padmanabhan, R., & Murthy, H. A. (2017). Feature-switching: Dynamic feature selection for anti-vector based speaker verification system. *Speech Communication, 93*, 53–62.

Scardapane, S., Stoffl, L., Röhrbein, F., & Uncini, A. (2017, May). On the use of deep recurrent neural networks for detecting audio spoofing attacks. In *International joint conference on neural networks* (IJCNN) (pp. 3483–3490). IEEE.

Shanmugapriya, P., & Venkataramani, Y. (2011, February). Implementation of speaker verification system using fuzzy wavelet network. In *International conference on communications and signal processing* (pp. 460–464). IEEE.

Shim, H. J., Jung, J. W., Heo, H. S., Yoon, S. H., & Yu, H. J. (2018, November). Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes. In *Conference on technologies and applications of artificial intelligence* (TAAI) (pp. 172–176). IEEE.

Shuvaev, S., Giaffar, H., & Koulakov, A. A. (2017). Representations of sound in deep learning of audio features from music. arXiv preprint arXiv:1712.02898.

Singh, G., Panda, A., Bhattacharyya, S., & Srikanthan, T. (2003, April). Vector quantization techniques for GMM based speaker verification. In *IEEE international conference on acoustics, speech, and signal processing, proceedings* (ICASSP'03) (Vol. 2(65)). IEEE.

Singh, N., Agrawal, A., & Khan, R. A. (2018). Voice biometric: A technology for voice based authentication. *Advanced Science, Engineering and Medicine, 10*(7–8), 754–759.

Sriskandaraja, K., Sethu, V., & Ambikairajah, E. (2018). Deep siamese architecture based replay detection for secure voice biometric. In *Interspeech* (pp. 671–675).

Sturim, D. E., Torres-Carrasquillo, P. A., & Campbell, J. P. (2016). Corpora for the evaluation of robust speaker recognition systems. In *Interspeech* (pp. 2776–2780).

Suthokumar, G., Sriskandaraja, K., Sethu, V., Wijenayake, C., & Ambikairajah, E. (2017). Independent modelling of high and low energy speech frames for spoofing detection. In *Interspeech* (pp. 2606–2610).

Sztahó, D., Szaszák, G., & Beke, A. (2019). Deep learning methods in speaker recognition: a review. arXiv preprint arXiv:1911.06615.

Tadokoro, N., Kosaka, T., Kato, M., & Kohda, M. (2009, August). Improvement of speaker vector-based speaker verification. In *Fifth international conference on information assurance and security* (Vol. 1, pp. 721–724). IEEE.

Todisco, M., Delgado, H., & Evans, N. (2017). Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language, 45*, 516–535.

Todisco, M., Delgado, H., & Evans, N. W. (2016, September). Articulation rate filtering of CQCC features for automatic speaker verification. In *Interspeech* (pp. 3628–3632).

Todisco, M., Delgado, H., Lee, K., Sahidullah, M., Evans, N., Kinnunen, T., & Yamagishi, J. (2018, September). Integrated presentation attack detection and automatic speaker verification: Common features and Gaussian back-end fusion.

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., & Lee, K. A. (2019). Asvspoof 2019: Future horizons in spoofed and fake audio detection. arXiv preprint arXiv:1904.05441.

Varchol, P., Levicky, D., & Juhar, J. (2008, April). Optimalization of GMM for text independent speaker verification system. In *18th International Conference Radioelektronika* (pp. 1–4). IEEE.

Vestman, V., Kinnunen, T., Hautamäki, R. G., & Sahidullah, M. (2020). Voice mimicry attacks assisted by automatic speaker verification. *Computer Speech & Language, 59*, 36–54.

Villalba, J., Miguel, A., Ortega, A., & Lleida, E. (2015). Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge. In *Sixteenth annual conference of the international speech communication association*.

VosxsCselesb. (2019). http://www.robots.ox.ac.uk/~vgg/data/vosxsceslseb/

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., & Juvela, L. (2019). ASVspoof 2019: A large-scale public database of synthetic, converted and replayed speech. arXiv, arXiv-1911.

Wong, L. P., & Russell, M. (2001, May). Text-dependent speaker verification under noisy conditions using parallel model combination. In *IEEE international conference on acoustics, speech, and signal processing. Proceedings* (Cat. No. 01CH37221) (Vol. 1, pp. 457–460). IEEE.

Wu, Z., De Leon, P. L., Demiroglu, C., Khodabakhsh, A., King, S., Ling, Z. H., & Yamagishi, J. (2016). Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance. *IEEE/ACM Transactions on Audio, Speech and Language Processing, 24*(4), 768–783.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., & Li, H. (2015a). Spoofing and countermeasures for speaker verification: A survey. *Speech Communication, 66*, 130–153.

Wu, Z., Khodabakhsh, A., Demiroglu, C., Yamagishi, J., Saito, D., Toda, T., & King, S. (2015, April). SAS: A speaker verification spoofing database containing diverse attacks. In *2015 IEEE international conference on acoustics, speech and signal processing* (ICASSP) (pp. 4440–4444). IEEE.

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., & Sizov, A. (2015). ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In *Sixteenth annual conference of the international speech communication association*.

Wu, Z., Xiao, X., Chng, E. S., & Li, H. (2013, May). Synthetic speech detection using temporal modulation feature. In *IEEE international conference on acoustics, speech and signal processing* (pp. 7234–7238). IEEE.

Yang, J., Das, R. K., & Li, H. (2018, November). Extended constant-Q cepstral coefficients for detection of spoofing attacks. In *Asia-Pacific signal and information processing association annual summit and conference* (APSIPA ASC) (pp. 1024–1029). IEEE.

Ze, H., Senior, A., & Schuster, M. (2013, May). Statistical parametric speech synthesis using deep neural networks. In *IEEE international conference on acoustics, speech and signal processing* (pp. 7962–7966). IEEE.

Zetterholm, E. (2007). Detection of speaker characteristics using voice imitation. In *Speaker classification II, ser. lecture notes in computer science* (pp. 192–205).

Zhao, Y., Togneri, R., & Sreeram, V. (2018, January). Spoofing detection using adaptive weighting framework and clustering analysis. In *Interspeech* (pp. 626–630).

Zhizheng, W., Junichi, Y., Tomi, K., Cemal, H., Mohammed, S., Aleksandr, S., & Hector, D. (2017). ASVspoof: The automatic speaker verification spoofing and countermeasures challenge.

Zouhir, Y., & Ouni, K. (2014). A bio-inspired feature extraction for robust speech recognition. *Springerplus, 3*(1), 651.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



View publication stats