Genuine and Spoof Speech Signal Classification

Subjects: Engineering, Electrical & Electronic

Contributor: Hiren Mewada , Jawad F. Al-Asad , Faris A. Almalki , Adil H. Khan , Nouf Abdullah Almujally , Samir El-Nakla , Qamar Naith

Voice-controlled devices are in demand due to their hands-free controls. However, using voice-controlled devices in sensitive scenarios like smartphone applications and financial transactions requires protection against fraudulent attacks referred to as “speech spoofing”. The algorithms used in spoof attacks are practically unknown; hence, further analysis and development of spoof-detection models for improving spoof classification are required. A study of the spoofed-speech spectrum suggests that high-frequency features are able to discriminate genuine speech from spoofed speech well. Typically, linear or triangular filter banks are used to obtain high-frequency features. However, a Gaussian filter can extract more global information than a triangular filter. In addition, mel frequency cepstral coefficients (MFCCs) features are preferable among other speech features because of their lower covariance.

anti-spoofing convolutional neural network genuine speech detection voice conversion

1. Introduction

Automation plays an essential role due to more responsive and efficient operations and tighter fraud-detection compliance. Automation saves time, effort, and money while decreasing manual errors and focusing on our primary goals. Automatic speaker authentication is a system that uses samples of human audio signals to recognize people. Entry controls to limited locations, access to confidentiality, and banking applications, including cash transfers, credit card authorizations, voice banking, and other transactions, can all benefit from speaker verification. With the increasing popularity of smartphones and voice-controlled intelligent devices, all of which contain a microphone, speaker authentication technology is expected to become even more prevalent in the future.[1]

However, this technology’s vulnerability to manipulation of the voice using presentation attacks, also known as voice spoofing, poses a challenge. Various spoofs, such as speech synthesis (SS), voice conversion (VC), replay speech, and imitation, can be used to spoof automated voice-detection systems [1]. These possible attacks in speaker-based automation systems have been intensively examined in Reference [2], for example, microphone-based voice generation, feature extraction, and classifier- or decision-level attacks. In a replay attack, the perpetrator tries for physical access by playing a previously recorded speech that sounds like a registered speaker’s speech. The system is particularly vulnerable to replay attacks, as voices can easily be recorded in person or through a telephone conversation and then replayed to manipulate the system. Since replay attacks do not need a lot of training or equipment, these attacks are the most common and likely to happen. The ASVspoof

2017 dataset addresses the issue of replay spoofing detection. Previous works have extracted features that reflect the acoustic level difference between genuine and spoof speech for replay speech detection.

Mel frequency cepstral coefficients (MFCCs), linear prediction coefficients (LPCs), linear prediction cepstral coefficients (LPCCs), line spectral frequencies (LSFs), discrete wavelet transform (DWT) [3][4], and perceptual linear prediction (PLP) are speech feature extractions commonly used in speaker recognition as well as speaker spoofing identification [5]. A wavelet transform was used to obtain spectral features, and these features were integrated with convolution neural network (CNN)'s spatial features in Reference [6] for ECG classification. In Reference [7], the authors analyzed a 6–8 kHz high-frequency subband using CQCC features to investigate re-recording distortion. To record the distortions caused by the playback device, Singh et al. [8] derived the MFCC from the residual signal. A low-frequency frame-wise normalization in the (constant Q transform) CQT domain was suggested in Reference [9] to capture the playback speech artifacts. Deep feature-utilizing neural networks have also been studied for the recognition of playback speech in addition to these manually created elements. For instance, Siamese-embedded spectrogram and group delay were employed to teach deep features to the CNN [10]. However, feature extraction is highly dependent on DNN training, and it might be difficult to generalize it to ASV tasks that are performed outside of their intended domain.

The speech signal is processed in a ten-millisecond time frame without overlapping for future extraction from the speech. The speech signal is cleaved into two zones: the silent and speech zones. An area with low energy and an excessive zero-crossing rate is considered a silent zone, and an area with high energy is regarded as a speech zone. Huang and Pun [11] experimented with the same person's genuine and spoofed speech signals using a replay attack. **Figure 1** shows the actual and replayed speech signal, and it is observed that there is a difference in the silent segment shown in the red box. Thus, the use of a silent zone with a high-frequency region can discriminate the spoofed speech easily. A precise recording system is required for a replay attack. The background noise of a recording device is easily noticeable in a silent zone due to its low energy relative to a highly energized speech zone. However, finding a silent zone accurately is tricky. Therefore, the endpoint method of finding a zero-crossing rate and energy can be used to approximate the silent zone [12]. By adjusting the threshold of the zero-crossing rate detection and short-term energy, a speech and silent zone can be judged systematically.

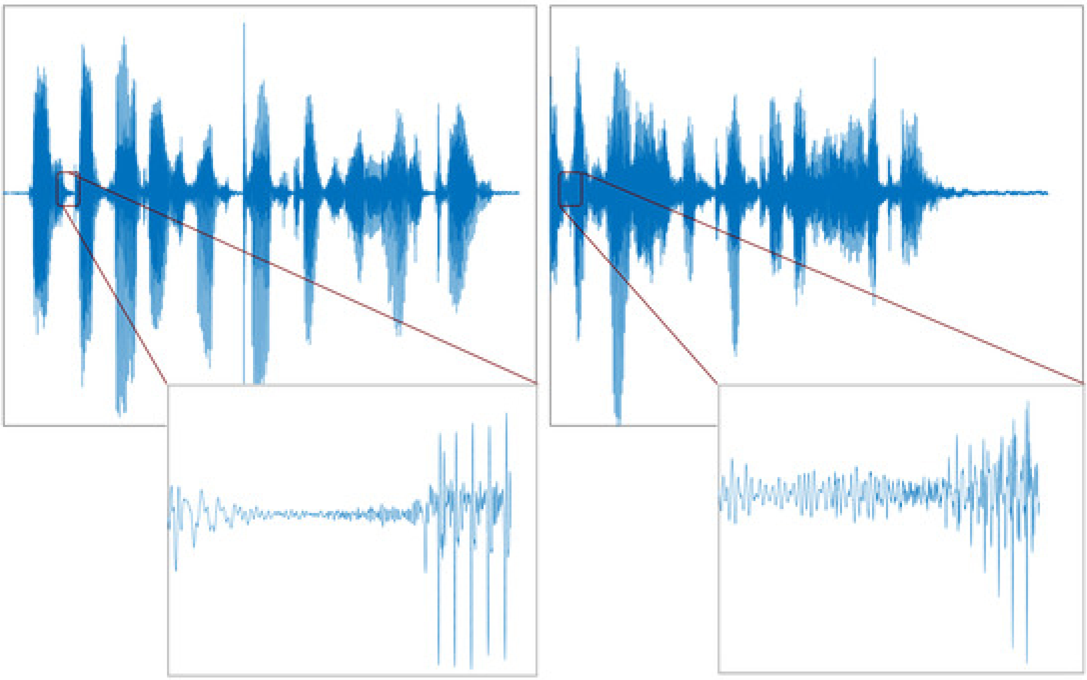

Figure 1. Time-domain speech signal of the same person's (**upper**) genuine waveform and (**lower**) spoofed waveform.

Although the MFCC and CQCC are considered reliable features, the classifier's performance can be significantly improved by combining them with complementary features, which can be done at the feature or score level [13]. Pitch, residual phase, and dialectical features are a few examples of complementary features. These complementary features, i.e., high pitch and corresponding phase, can easily be obtained at high frequencies [14].

2. Genuine and Spoof Speech Signal Classification

Enormous studies on genuine and spoof speech signal classification have been proposed in past years. Major classification algorithms have two stages: a design of features extraction algorithm from the speech signal and a classifier to discriminate these features for speech classification. Many feature sets have been proposed with statistical and deep learning-based classifiers. A few widely used feature sets are as follows: Mel frequency cepstrum coefficients (MFCCs); inverse MFCCs (IMFCCs) [15]; linear frequency cepstrum coefficients (LFCCs); constant Q cepstrum coefficients (CQCCs) [16]; log-power spectrum using discrete Fourier transform (DFT) [17]; Gammatonegram, group delay over the frame, referred to as GD-gram [18]; modified group delay; All-Pole Group Delay [19]; Cochlear Filter Cepstral Coefficient—Instantaneous Frequency [20]; cepstrum coefficients using single-frequency filtering [21][22]; Zero-Time Windowing (ZTW) [23]; Mel-frequency cepstrum using ZTW [24]; and polyphase IIR filters [25]. The human ear uses Fourier transform magnitude and neglects the phase information [26]. Therefore, the phase spectrum has yet to gain attention in classification.

Along with features, the classifier also plays an important role. Many machine learning models have been proposed, including the Gaussian mixture model (GMM), K-nearest neighborhood (KNN), the hidden Markov model

[27], support vector machine (SVM) [28], and convolution neural networks (CNNs). Multi-layer perceptron [29], deep CNN (DNN), and recurrent neural network (RNN) [30] are examples of widely used neural networks. The LSTM network is a type of RNN giving more memory power for an extended period, and it has been widely used in many applications. Ghosh et al. [31] used LSTM to remove the muscular artifacts from EEG signals. An energy-efficient speech recognition algorithm using LSTM was proposed in Reference [32]. This LSTM was implemented in CMOS, reducing energy requirements 2.19 times to the baseline model. The spikes' temporal dependencies were captured from the EEG signals using LSTM for the brain–computer interface, which can help to evaluate emotion recognition [33].

In 2015, the first challenge, "Automatic Speaker Verification Spoofing and Countermeasures" [34], provided the dataset of spoofed speech signals based on synthetic speech, voice conversion, and other unknown attacks. The base algorithm using CQCC features and GMM as a classifier was presented with 24.77% EER. In this challenge, CQCC-based features showed promising results with an Equal Error Rate (EER) of 0.255% in Reference [35]. However, this ASPSpoof 2015 dataset does not contain replay attacks. Therefore, the dataset was revised, and the new dataset of ASVSpoof 2017 [2] was published, focusing on replay attacks. Again, using CQCC features and GMM as a classifier, the base algorithm secured a 24.77% EER, where GMM was trained using training and development datasets.

Xue et al. [36] presented a fusion approach using facial and speech features using convolution neural networks. The results were tested on ASVSpoof 2019 datasets, achieving a 9% EER rate. In Reference [37], the authors observed that the block-based approach missed the instantaneous spectral features. Therefore, single-frequency filtering was proposed, presenting high spectral and temporal resolution. Their model performed well, with a 0.05% EER on BTAS test data. A similar approach was presented in Reference [38], where instantaneous frequency and instantaneous energies were obtained using Hilbert transform, and genuine speech was differentiated from spoofed speech using empirical mode-decomposition features. They integrated these features with CQCC and group delay to improve performance. Their work also focused on replay attacks only. The voice quality features were combined with CQCC features to identify the replay attacks in speech signals in Reference [39]. Their work is limited to binary classification with replay attacks only. Chaudhari et al. [40] discussed three features, including LPC, CQCC, and MFCC, with GMM classifiers. They showed that combining MFCC and CQCC features enhanced the performance with a 10.18% EER. Glottal flow and an acoustic-based total of 106 features obtained from the speech signals were used in SVM and XGBoost classifier in Reference [41]. The XGBoost outperformed the SVM, resulting in a 98.8% classification accuracy. However, this model used extensive feature sets in the classification. Compatibility testing among a large number of devices is also challenging. Naith [42] conducted a test for Android and IoS devices. A total of 42 speakers participated in the creation of 219 datasets, a good and sufficient participation number for such empirical studies.

The integration of the well-established speaker modeling model "i-vector space" and the synthesis-channel subspace model was proposed with two-stage probabilistic linear discriminant analysis [43]. However, they tested the model with two voice-conversion attacks only. A capsule network is modified by replacing the ReLU with a leaky ReLU layer and a modified routing algorithm for better attention to the speech artifacts [44]. They focused on text-to-speech-based attacks in spoofing. The authors in Reference [45] extracted features using two partitioned datasets in logical and physical access. Later, they assembled the features by normalizing them and trained the CNN model by evaluating the loss function.

In Reference [46], cepstral features were obtained using single-frequency filtering. GMM and deep learning classifier models were compared. Later, a score-fusion approach was employed to improve the performance of the model by 17.82% EER in the evaluation dataset. Zhang et al. [30] employed a CNN and recurrent neural network (RNN) simultaneously. They trained this network using perceptual minimum variance distortionless response (PMVDR), teager energy operator-based critical auto-correlation envelope (TEO), and a spectrogram separately. They observed that spectrum-based features worked well with their network on ASVSpoof 2015 datasets, with an average EER of 0.36% compared with PMVDR and TEO, with EERs of 1.44% and 2.31%. Patil et al. [47] improved the potential of TEO using the signal mass in the front stage, and different classifiers, including GMM and light-CNN trained with 20 epochs, were tested in the second stage with ASVSpoof 2017 datasets. The GMM model performed well, with EERs of 5.55% and 10.75% on the development and evaluation datasets, respectively. In Reference [48], a group delay concatenated over the consecutive frames of the speech signal was used as a feature in the ResNET18 classifier. It showed a remarkable improvement, with zero EER on the development and evaluation datasets ASVSpoof 2017. However, the authors tested the model on a subset of the dataset, and the model's validation for different types of attacks was not presented in the paper. Various extensions of ResNET using the Squeeze Excitation Network, including SENET34, SENET50, Mean-Std ResNET, and Dialted ResNET, proposed using CQCC features sets by Lai et al. [49]. The EER rate was reduced to 0.59 for the physical access dataset and to 6.70 for the logical access dataset of ASVSPoof 2019. They observed that further meta-data analysis and refinement in the algorithm is required.

Analysis of the deep RNN network was presented by Scardapane et al. [50]. They evaluated four architectures with MFCC features, log-filter bank features, and a concatenation of these two feature sets using ASVSpoof 2015 datasets. They observed that three LSTM layers trained with MFCC features gave better EERs than a log-filter bank. In contrast, a network combining three dense layers and three LSTM layers with MFCC features performed well, with 2.91% EER. Mittal and Dua [51] presented a hybrid deep CNN using static and dynamic CQCC features sets. Hybrid CNN combined the CNN-LSTM model with a time-distributed wrapper integrated into the LSTM network. This hybrid approach achieved a 0.029% EER on the evaluation dataset with high computation power. A standard time-delayed CNN (TD-CNN) was modified with a statistical pooling operation instead of max pooling, and angular softmax was used in the architecture in Reference [1]. The training of the TD-CNN model using third- and fourth-order moments achieved a 3.05% EER.

Dinkel et al. [52] tried to remove the crucial feature extraction step. First, they used the row form of speech frames as an input to the LSTM model to obtain features in the form of likelihood, and later, CNN was used for classification. However, no validation for unknown attacks was presented. Mittal and Dua [53] converted the CQCC features in 3D-tensor into 2D space, and a 2D-CNN was used for classification. A 3D tensor was obtained by reshaping the 30 statics and first- and second-order CQCC features. An RNN network was trained with cross-entropy and KL divergence loss for audio spoof classification in Reference [54]. Three variants of RNN were proposed in Reference [55]. MFCC, CQCC, and log-magnitude STFT features were used in the RNN, and they obtained a 25% improvement compared with the base model of GMM.

A light-CNN has been proposed by Wu et al. [56] with feature genuinization. In the first phase, features obtained from genuine speech were used to train the genuinization transformer. In the second phase, this transformer was converted to enhance the genuine and spoof features' separation. This transformer was integrated with light-CNN and validated using the ASVspoof 2019 dataset with an EER rate of 4.07%. Li et al. [57] presented a high-frequency feature-based deep CNN model. They extracted long-term variable Q transform (L-VQT) features, and the light-DenseNET model was trained using these features. They validated the model using the ASVSpoof 2019 dataset with various CNN classifiers, including a 0.352% and 3.768% EER on the development and evaluation datasets, respectively.

The literature reveals that CQCC features and a lateral variant of the CQCC improved the spoofed-speech classification error rate with a statistical or machine learning model to a certain extent compared with other features. High-frequency features with CNN were more prominent in identifying speech with unknown attacks. In CNN, DenseNET, light-CNN, and recurrent neural networks, including RNN, LSTM, and BiLSTM networks, have mainly been used in spoof classification.

References

1. Wu, Z.; Evans, N.; Kinnunen, T.; Yamagishi, J.; Alegre, F.; Li, H. Spoofing and countermeasures for speaker verification: A survey. Speech Commun. 2015, 66, 130–153.
2. Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.A. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection; The International Speech Communication Association: Berlin, Germany, 2017.
3. Ghaderpour, E.; Pagiatakis, S.D.; Hassan, Q.K. A survey on change detection and time series analysis with applications. Appl. Sci. 2021, 11, 6141.
4. Mewada, H.K.; Patel, A.V.; Chaudhari, J.; Mahant, K.; Vala, A. Wavelet features embedded convolutional neural network for multiscale ear recognition. J. Electron. Imaging 2020, 29, 043029.
5. Alim, S.A.; Rashid, N.K.A. Some Commonly Used Speech Feature Extraction Algorithms; IntechOpen: London, UK, 2018.
6. Mewada, H. 2D-wavelet encoded deep CNN for image-based ECG classification. In Multimedia Tools and Applications; Springer: Berlin/Heidelberg, Germany, 2023; pp. 1–17.
7. Witkowski, M.; Kacprzak, S.; Zelasko, P.; Kowalczyk, K.; Galka, J. Audio Replay Attack Detection Using High-Frequency Features. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24

August 2017; pp. 27–31.

8. Singh, M.; Pati, D. Usefulness of linear prediction residual for replay attack detection. AEU-Int. J. Electron. Commun. 2019, 110, 152837.
9. Yang, J.; Das, R.K. Low frequency frame-wise normalization over constant-Q transform for playback speech detection. Digit. Signal Process. 2019, 89, 30–39.
10. Sriskandaraja, K.; Sethu, V.; Ambikairajah, E. Deep siamese architecture based replay detection for secure voice biometric. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 671–675.
11. Huang, L.; Pun, C.M. Audio Replay Spoof Attack Detection by Joint Segment-Based Linear Filter Bank Feature Extraction and Attention-Enhanced DenseNet-BiLSTM Network. IEEE/ACM Trans. Audio Speech Lang. Process. 2020, 28, 1813–1825.
12. Zaw, T.H.; War, N. The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection. In Proceedings of the 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 22–24 December 2017; pp. 1–5.
13. Singh, S.; Rajan, E. Vector quantization approach for speaker recognition using MFCC and inverted MFCC. Int. J. Comput. Appl. 2011, 17, 1–7.
14. Singh, S.; Rajan, D.E. A Vector Quantization approach Using MFCC for Speaker Recognition. In Proceedings of the International Conference Systemic, Cybernatics and Informatics ICSCI under the Aegis of Pentagram Research Centre Hyderabad, Hyderabad, India, 4–7 January 2007; pp. 786–790.
15. Chakroborty, S.; Saha, G. Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter. Int. J. Signal Process. 2009, 5, 11–19.
16. Jelil, S.; Das, R.K.; Prasanna, S.M.; Sinha, R. Spoof detection using source, instantaneous frequency and cepstral features. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 22–26.
17. Sahidullah, M.; Kinnunen, T.; Hanilçi, C. A comparison of features for synthetic speech detection. In Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
18. Loweimi, E.; Barker, J.; Hain, T. Statistical normalisation of phase-based feature representation for robust speech recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5310–5314.

19. Pal, M.; Paul, D.; Saha, G. Synthetic speech detection using fundamental frequency variation and spectral features. Comput. Speech Lang. 2018, 48, 31–50.
20. Patil, A.T.; Patil, H.A.; Khoria, K. Effectiveness of energy separation-based instantaneous frequency estimation for cochlear cepstral features for synthetic and voice-converted spoofed speech detection. Comput. Speech Lang. 2022, 72, 101301.
21. Kadiri, S.R.; Yegnanarayana, B. Analysis and Detection of Phonation Modes in Singing Voice using Excitation Source Features and Single Frequency Filtering Cepstral Coefficients (SFFCC). In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 441–445.
22. Kethireddy, R.; Kadiri, S.R.; Gangashetty, S.V. Deep neural architectures for dialect classification with single frequency filtering and zero-time windowing feature representations. J. Acoust. Soc. Am. 2022, 151, 1077–1092.
23. Kethireddy, R.; Kadiri, S.R.; Kesiraju, S.; Gangashetty, S.V. Zero-Time Windowing Cepstral Coefficients for Dialect Classification. In Proceedings of the The Speaker and Language Recognition Workshop (Odyssey), Tokyo, Japan, 2–5 November 2020; pp. 32–38.
24. Kadiri, S.R.; Alku, P. Mel-Frequency Cepstral Coefficients of Voice Source Waveforms for Classification of Phonation Types in Speech. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 2508–2512.
25. Mewada, H.K.; Chaudhari, J. Low computation digital down converter using polyphase IIR filter. Circuit World 2019, 45, 169–178.
26. Loweimi, E.; Ahadi, S.M.; Drugman, T. A new phase-based feature representation for robust speech recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7155–7159.
27. Dua, M.; Aggarwal, R.K.; Biswas, M. Discriminative training using noise robust integrated features and refined HMM modeling. J. Intell. Syst. 2020, 29, 327–344.
28. Rahmeni, R.; Aicha, A.B.; Ayed, Y.B. Speech spoofing detection using SVM and ELM technique with acoustic features. In Proceedings of the 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 2–5 September 2020; pp. 1–4.
29. Muckenhirn, H.; Korshunov, P.; Magimai-Doss, M.; Marcel, S. Long-term spectral statistics for voice presentation attack detection. IEEE/ACM Trans. Audio Speech Lang. Process. 2017, 25, 2098–2111.
30. Zhang, C.; Yu, C.; Hansen, J.H. An investigation of deep-learning frameworks for speaker verification antispoofing. IEEE J. Sel. Top. Signal Process. 2017, 11, 684–694.

31. Ghosh, R.; Phadikar, S.; Deb, N.; Sinha, N.; Das, P.; Ghaderpour, E. Automatic Eyeblink and Muscular Artifact Detection and Removal From EEG Signals Using k-Nearest Neighbor Classifier and Long Short-Term Memory Networks. IEEE Sens. J. 2023, 23, 5422–5436.

32. Jo, J.; Kung, J.; Lee, Y. Approximate LSTM computing for energy-efficient speech recognition. Electronics 2020, 9, 2004.

33. Gong, P.; Wang, P.; Zhou, Y.; Zhang, D. A Spiking Neural Network With Adaptive Graph Convolution and LSTM for EEG-Based Brain-Computer Interfaces. IEEE Trans. Neural Syst. Rehabil. Eng. 2023, 31, 1440–1450.

34. Wu, Z.; Kinnunen, T.; Evans, N.; Yamagishi, J.; Hanilçi, C.; Sahidullah, M.; Sizov, A. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.

35. Todisco, M.; Delgado, H.; Evans, N.W. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In Proceedings of the Odyssey, Bilbao, Spain, 21–24 June 2016; Volume 2016, pp. 283–290.

36. Xue, J.; Zhou, H.; Song, H.; Wu, B.; Shi, L. Cross-modal information fusion for voice spoofing detection. Speech Commun. 2023, 147, 41–50.

37. Alluri, K.R.; Achanta, S.; Kadiri, S.R.; Gangashetty, S.V.; Vuppala, A.K. Detection of Replay Attacks Using Single Frequency Filtering Cepstral Coefficients. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 2596–2600.

38. Bharath, K.; Kumar, M.R. Replay spoof detection for speaker verification system using magnitude-phase-instantaneous frequency and energy features. Multimed. Tools Appl. 2022, 81, 39343–39366.

39. Woubie, A.; Bäckström, T. Voice Quality Features for Replay Attack Detection. In Proceedings of the 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 29 August–2 September 2022; pp. 384–388.

40. Chaudhari, A.; Shedge, D. Integration of CQCC and MFCC based Features for Replay Attack Detection. In Proceedings of the 2022 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 9–11 March 2022; pp. 1–5.

41. Rahmeni, R.; Aicha, A.B.; Ayed, Y.B. Voice spoofing detection based on acoustic and glottal flow features using conventional machine learning techniques. Multimed. Tools Appl. 2022, 81, 31443–31467.

42. Naith, Q. Thesis title: Crowdsourced Testing Approach For Mobile Compatibility Testing. Ph.D. Thesis, University of Sheffield, Sheffield, UK, 2021.

43. Sizov, A.; Khoury, E.; Kinnunen, T.; Wu, Z.; Marcel, S. Joint speaker verification and antispoofing in the i-vector space. IEEE Trans. Inf. Forensics Secur. 2015, 10, 821–832.

44. Luo, A.; Li, E.; Liu, Y.; Kang, X.; Wang, Z.J. A Capsule Network Based Approach for Detection of Audio Spoofing Attacks. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6359–6363.

45. Monteiro, J.; Alam, J.; Falk, T.H. An ensemble based approach for generalized detection of spoofing attacks to automatic speaker recognizers. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6599–6603.

46. Alluri, K.R.; Achanta, S.; Kadiri, S.R.; Gangashetty, S.V.; Vuppala, A.K. SFF Anti-Spoofer: IIIT-H Submission for Automatic Speaker Verification Spoofing and Countermeasures Challenge 2017. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 107–111.

47. Patil, A.T.; Acharya, R.; Patil, H.A.; Guido, R.C. Improving the potential of Enhanced Teager Energy Cepstral Coefficients (ETECC) for replay attack detection. Comput. Speech Lang. 2022, 72, 101281.

48. Tom, F.; Jain, M.; Dey, P. End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 681–685.

49. Lai, C.I.; Chen, N.; Villalba, J.; Dehak, N. ASSERT: Anti-spoofing with squeeze-excitation and residual networks. arXiv 2019, arXiv:1904.01120.

50. Scardapane, S.; Stoffl, L.; Röhrbein, F.; Uncini, A. On the use of deep recurrent neural networks for detecting audio spoofing attacks. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 3483–3490.

51. Mittal, A.; Dua, M. Static–dynamic features and hybrid deep learning models based spoof detection system for ASV. Complex Intell. Syst. 2022, 8, 1153–1166.

52. Dinkel, H.; Qian, Y.; Yu, K. Investigating raw wave deep neural networks for end-to-end speaker spoofing detection. IEEE/ACM Trans. Audio Speech Lang. Process. 2018, 26, 2002–2014.

53. Mittal, A.; Dua, M. Automatic speaker verification system using three dimensional static and contextual variation-based features with two dimensional convolutional neural network. Int. J. Swarm Intell. 2021, 6, 143–153.

54. Chintha, A.; Thai, B.; Sohrawardi, S.J.; Bhatt, K.; Hickerson, A.; Wright, M.; Ptucha, R. Recurrent convolutional structures for audio spoof and video deepfake detection. IEEE J. Sel. Top. Signal Process. 2020, 14, 1024–1037.

55. Alzantot, M.; Wang, Z.; Srivastava, M.B. Deep residual neural networks for audio spoofing detection. arXiv 2019, arXiv:1907.00501.

56. Wu, Z.; Das, R.K.; Yang, J.; Li, H. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. arXiv 2020, arXiv:2009.09637.

57. Li, J.; Wang, H.; He, P.; Abdullahi, S.M.; Li, B. Long-term variable Q transform: A novel time-frequency transform algorithm for synthetic speech detection. Digit. Signal Process. 2022, 120, 103256.

Retrieved from https://encyclopedia.pub/entry/history/show/114119