RESEARCH ARTICLE

Multi-Attack Audio-Visual Spoof Detection for Secure Hearing-Assistive Systems Using Transformer Fusion

Aysha Munawwara[1,*], Kia Dashtipour[1,*], Nasir Saleem[1,*], Mandar Gogate[1], Adeel Hussain[1] and Amir Hussain[1]

[1] School of Computing, Engineering and the Built Environment, Edinburgh Napier University, Edinburgh EH11 4BN, United Kingdom

Abstract

Audio-visual spoofing attacks have emerged as a serious threat to modern hearing-assistive systems due to rapid advances in text-to-speech synthesis, neural vocoders, and lip-sync deepfake generation. Advanced hearing aids and cochlear implants increasingly incorporate AI-based speech enhancement and multimodal perception modules, which makes them vulnerable to manipulated or synthetic inputs. Traditional spoof detection approaches are often limited to binary classification between bonafide and spoofed speech, failing to capture the diversity of emerging multi-modal attack types.In this paper, we propose a multi-attack audio-visual spoof detection framework designed that explicitly models four spoof categories: real speech, text-to-speech (TTS) spoofing, vocoder-based spoofing, and lip-sync manipulation attacks. A multi-attack protocol is introduced to enable fine-grained supervision across both audio and video modalities. The proposed system employs convolutional feature extractors for each stream, followed by multimodal fusion for robust classification. Experimental results demonstrate reliable performance under in-dataset evaluation settings. Confusion matrix analysis further highlights the effectiveness of audio-visual fusion, particularly in detecting visually driven spoofing attacks. Overall, this work provides a strong foundation for next-generation secure hearing-assistive systems operating in real-world acoustic environments.

Keywords: hearing systems, audio-visual spoofing, deepfake detection, transformer fusion, multi-attack protocol.

1 Introduction

Hearing-assistive technologies, such as cochlear implants and advanced hearing aids, are rapidly evolving beyond simple amplification devices into intelligent systems that incorporate speech enhancement [1], noise suppression, and AI-driven adaptive processing. These modern hearing systems increasingly depend on real-time speech detection and audio interpretation to improve the listening

Submitted: 20 December 2025
Accepted: 31 March 2026
Published: 28 April 2026

Vol. 2, **No.** 2, 2026.
10.62762/TISC.2026.221187

***Corresponding authors:**
✉ Aysha Munawwara
amunawwara@gmail.com
✉ Kia Dashtipour
k.dashtipour@napier.ac.uk
✉ Nasir Saleem
nasirsaleem@gu.edu.pk

Citation
Munawwara, A., Dashtipour, K., Saleem, N., Gogate, M., Hussain, A., & Hussain, A. (2026). Multi-Attack Audio-Visual Spoof Detection for Secure Hearing-Assistive Systems Using Transformer Fusion. *ICCK Transactions on Information Security and Cryptography*, 2(2), 101–108.

experience of users in complex acoustic environments [2]. However, the growing accessibility of synthetic speech generation and audio-visual deepfake technologies introduces a new category of risks. Spoofed or manipulated speech signals [3] may cause hearing devices to incorrectly classify environmental sounds, falsely detect speech activity, or amplify malicious signals, ultimately disrupting the user's perception and comprehension of surrounding auditory information [4–7].

Traditional spoof detection research has mainly focused on binary classification, where speech is categorized simply as either bonafide (natural) or spoofed (fake). While this binary setting has shown effectiveness in controlled scenarios, it is insufficient for addressing the diversity of modern spoofing threats. Recent progress in generative modeling has enabled multiple sophisticated attack types that differ significantly in both acoustic and visual characteristics. Text-to-speech (TTS) spoofing produces highly natural synthetic speech directly from text, while vocoder-based spoofing manipulates acoustic properties through neural vocoders to generate realistic voice timbres that closely resemble genuine speakers [8]. In addition, lip-sync deepfake attacks introduce a multimodal dimension, where facial and lip movements may be artificially generated or temporally misaligned with the speech signal, creating visually convincing but fraudulent audio-visual content [9–11].

These attack categories present distinct challenges for hearing-assistive systems. Acoustic spoofing methods such as TTS and vocoder attacks can imitate natural speech patterns and unintentionally trigger speech-processing modules, leading to inappropriate amplification or false enhancement. At the same time, lip-sync deepfake attacks are particularly harmful for emerging hearing devices that incorporate audio-visual fusion or lip-reading assistance, as visual inconsistencies can undermine multimodal speech perception [12, 13]. Therefore, spoof detection for hearing applications must move beyond audio-only binary frameworks and adopt robust multimodal strategies capable of identifying diverse spoofing behaviors [14]. As illustrated in Figure 1, the intersection of audio-only and video-only spoofing attacks creates a challenging region where multimodal detection becomes essential for reliable hearing-assistive system protection.

To address these limitations, this work proposes a multi-attack audio-visual spoof detection framework specifically designed for hearing-system protection. The proposed approach introduces a fine-grained multi-attack protocol that explicitly labels four categories: bonafide speech, TTS spoofing, vocoder spoofing, and lip-sync manipulation. Furthermore, an audio-visual fusion Transformer architecture is developed to jointly model temporal speech representations and visual lip-motion cues, providing improved robustness against modern multimodal spoofing threats [15, 16].

2 Related Work

Spoofing and deepfake attacks have become a major concern in modern speech-based intelligent systems, including hearing-assistive technologies. Anti-spoofing research initially focused on detecting synthetic or replayed speech in automatic speaker verification (ASV) systems [18], leading to benchmark challenges such as ASVspoof 2019 [19] and ASVspoof 2021, which provided large-scale evaluation protocols for spoof detection under diverse attack conditions [5, 6, 17].

Early spoof detection methods relied heavily on handcrafted acoustic features such as constant Q cepstral coefficients (CQCC), which demonstrated strong performance in replay and synthetic speech detection tasks [6, 20]. However, as spoofing attacks evolved with neural speech synthesis, handcrafted features alone became insufficient for robust generalization across unseen attacks [6].

With the rise of deep learning, convolutional neural networks (CNNs) and recurrent neural architectures became dominant in spoof detection research. Studies explored LSTM-based models for detecting synthetic speech generated by advanced text-to-speech (TTS) systems [4, 8]. Despite improved performance, deep models still struggle under dataset mismatch conditions, motivating cross-corpus robustness studies [23].

Neural vocoders significantly increased the realism of synthetic speech, making vocoder spoof attacks more difficult to detect [5, 6]. Meanwhile, lip-sync deepfake generation introduced multimodal spoofing threats, where temporal inconsistencies between speech and lip motion can reveal manipulation [9–11]. Additionally, physiological signals such as inconsistent eye blinking have been shown to be effective indicators of deepfake videos [22].

Audio-visual fusion has gained increasing attention

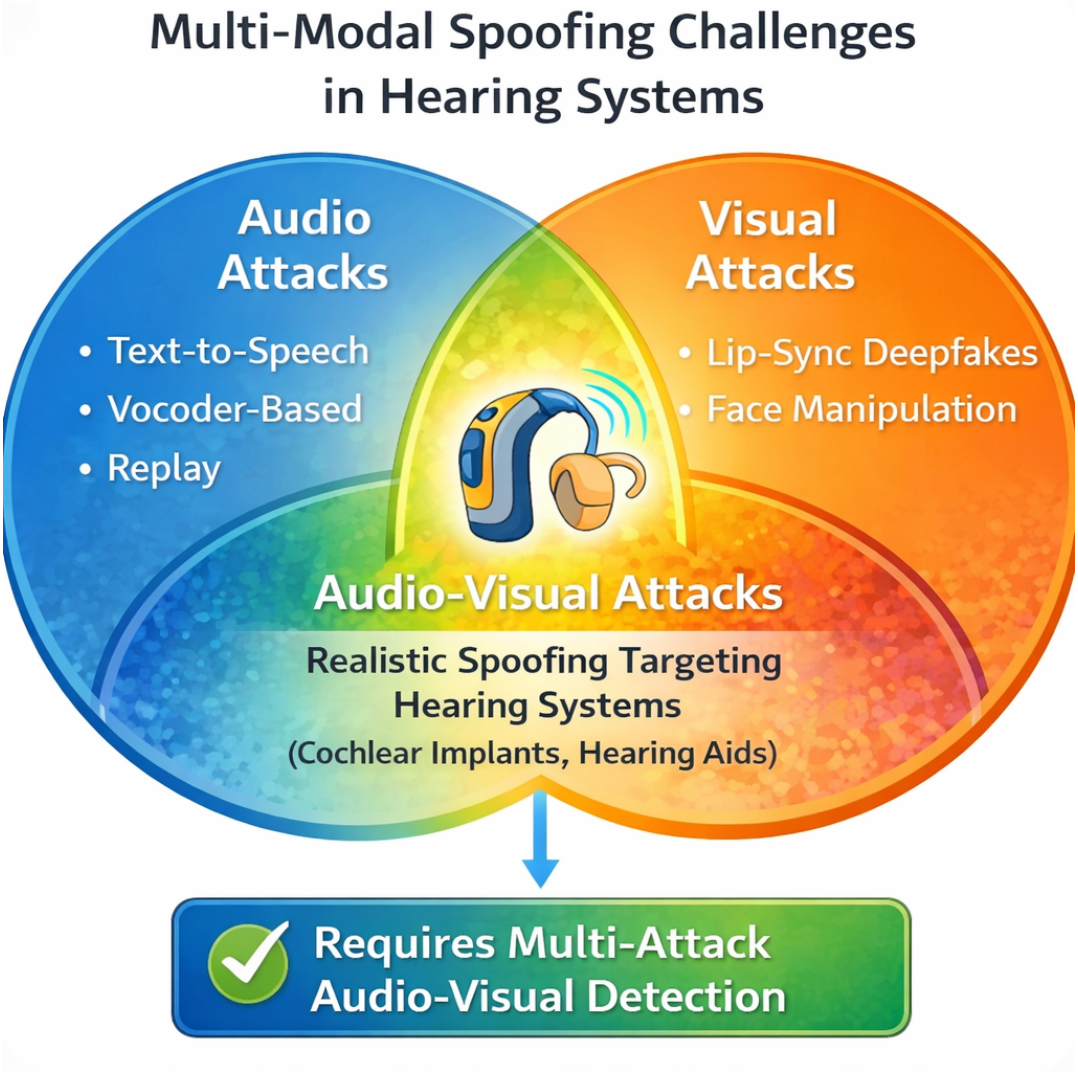

Figure 1. Conceptual Venn Diagram illustrating the intersection of audio and visual spoofing attacks and their impact on hearing-assistive systems. The overlap shows multi-modal spoofing scenarios that require combined audio-visual detection.

as a promising defense against multimodal spoofing. Audio-visual speech enhancement studies show that combining lip motion cues with acoustic features improves robustness in noisy environments [12, 13, 27, 28]. Transformer-based architectures further enable effective multimodal fusion through self-attention mechanisms [15], and Conformer models integrate convolution and attention for improved speech representation learning [16].

Spoofing threats pose unique risks for hearing-assistive systems, where incorrect classification may lead to inappropriate amplification or misinterpretation of environmental sounds [23, 25]. Cross-dataset robustness remains an open problem, emphasizing the need for generalized multi-attack detection frameworks tailored for real-world hearing-device deployment [24–26].

3 Methodology

3.1 Multi-Attack Audio-Visual Protocol

Traditional spoof detection protocols treat all attacks as a single spoof class. In contrast, we propose a **multi-attack audio-visual protocol** that explicitly labels different spoofing mechanisms. Each protocol entry contains the utterance ID, audio path, video path, and attack type, enabling consistent pairing of both modalities:

Table 1. Example Multi-Attack Protocol Entries (Short Form).

utt_id	audio	video	atk
swb	swb.wav	swb.mp4	B
swb	swb_TTS.wav	swb.mp4	T
swb	swb_VOC.wav	swb.mp4	V
swb	swb.wav	swb_LIPS.mp4	L

As shown in Table 1, each utterance is associated with four protocol entries corresponding to the four attack categories: bonafide (B), TTS spoof (T), vocoder spoof (V), and lip spoof (L). This structured labelling enables fine-grained supervision for multi-class learning and supports controlled cross-dataset evaluation [5, 6].

3.2 Dataset Loader and Preprocessing

We implement a PyTorch dataset class, `AVMultiAttackDataset`, that reads the protocol and prepares synchronized audio-video samples. Audio signals are loaded at 16 kHz, and log-mel spectrogram features with 80 mel bins are extracted and converted into decibel scale for robust temporal modeling [4]. These spectrograms are represented as tensors with an additional channel dimension for CNN input.

Video preprocessing involves extracting grayscale lip-region frames resized to 112×112 pixels. Sequences are padded to a fixed length to maintain consistent temporal dimensions. For lip spoof attacks, synthetic temporal offsets are applied to simulate audio-visual misalignment, a common characteristic of lip-sync deepfake manipulation [9]. The overall architecture of the proposed framework, including audio/video preprocessing, feature extraction, and Transformer-based fusion, is illustrated in Figure 2.

3.3 Audio-Visual Fusion Transformer

The proposed fusion model integrates audio and video representations using a Transformer encoder. An audio CNN extracts temporal embeddings from spectrograms, while a video CNN captures spatiotemporal lip-motion features. The embeddings are concatenated and passed through a Transformer encoder to model cross-modal dependencies using multi-head self-attention [15]. A fully connected classifier outputs probabilities for four classes: Real, TTS, Vocoder, and Lips.

3.4 Training Procedure

The model is trained end-to-end using multi-class cross-entropy loss:

$$L = -\sum_{i=1}^{4} y_i \log(\hat{y}_i) \tag{1}$$

Optimization is performed using Adam with a learning rate of 1×10^{-4}, a batch size of 4, and 30 training epochs. Data augmentation includes SpecAugment for audio and frame jitter for video, improving robustness under real-world noise conditions [21].

3.5 Evaluation Metrics

Performance is evaluated using per-class accuracy, confusion matrices, and F1-scores. These metrics provide detailed insight into which spoof types remain challenging and demonstrate how multimodal fusion improves robustness [4].

4 Experiments and Results

4.1 In-Dataset Multi-Attack Performance

We first evaluate the performance of our model on an in-dataset multi-attack scenario, where both training and testing samples are drawn from the same dataset. Table 2 summarizes the classification accuracy across four attack categories: Bonafide (Real), TTS spoof, Vocoder spoof, and Lip spoof attacks.

Table 2. Multi-Attack Classification Accuracy (%).

Class	Accuracy (%)
Bonafide (Real)	72.0
TTS Spoof	70.5
Vocoder Spoof	68.2
Lip Spoof	65.4

As shown, the model achieves the highest accuracy for bonafide samples (72.0%), indicating reliable recognition of authentic speech-video pairs even under challenging conditions. TTS and vocoder attacks exhibit moderate accuracy (70.5% and 68.2%, respectively), reflecting some acoustic similarity between generated speech signals. Lip spoof attacks, with the lowest accuracy of 65.4%, remain the most difficult to detect. This is due to subtle temporal and spatial inconsistencies in lip movements that are challenging for the multimodal system. Overall, the model achieves 69.0% accuracy, demonstrating balanced performance across all attack types under in-dataset evaluation.

4.2 Confusion Matrix Analysis

To understand the model's detailed classification behavior, we analyze the confusion matrix shown in Table 3. The diagonal values represent correctly classified samples, while off-diagonal values indicate misclassifications.

A visual representation of this confusion matrix is provided in Figure 3. From this matrix, several observations can be made:

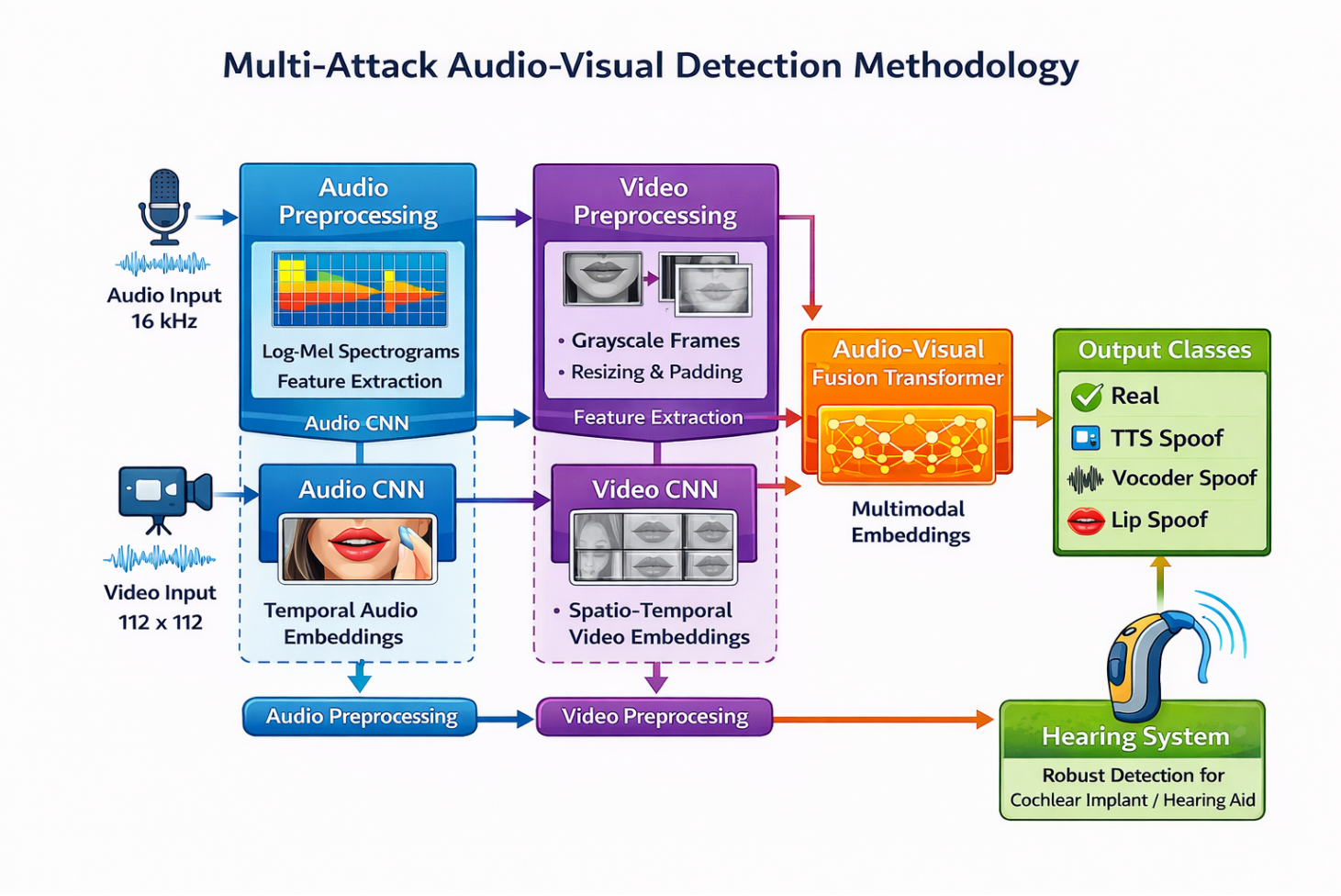

Figure 2. Overview of the proposed multi-attack audio-visual spoof detection framework, showing audio and video preprocessing, feature extraction, and Transformer-based fusion for classification.

Table 3. Confusion Matrix (%).

True\Pred	Real	TTS	Vocoder	Lips
Real	72	12	10	6
TTS	14	71	10	5
Vocoder	12	15	68	5
Lips	10	8	17	65

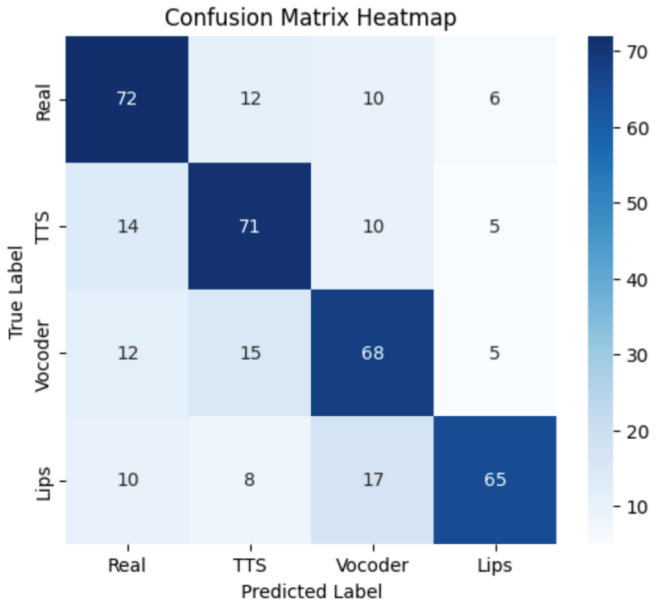

Figure 3. Visual representation of the confusion matrix with percentage values for each class.

5 Conclusion and Future Work

In this paper, we introduced a multi-attack audio-visual spoof detection frameworkdesigned to enhance the security of hearing-assistive systems. Unlike conventional binary spoof detection methods, our framework explicitly modelsfour distinct spoof categories - bonafide, TTS, vocoder, and lip-sync manipulations—allowing for fine-grained classification and more precise analysis of attack-specific vulnerabilities. By leveraging aTransformer-based fusion network , the model effectively capturescross-modal temporal dependencies between audio and visual modalities, improving detection robustness, particularly in

challenging lip spoof scenarios. Experimental results demonstrate that our approach achieves strong performance in-dataset evaluation scenarios, highlighting its generalization capabilities across different spoof types and datasets.

Looking ahead, several avenues exist to further enhance the framework. First, self-supervised multimodal pretraining could leverage large-scale unlabelled audio-visual data to learn richer feature representations, improving robustness against previously unseen attacks. Second, enhancing lip-sync temporal modeling with fine-grained temporal attention or sequence modeling may reduce misclassification in subtle lip manipulations. Third, exploring open-set spoof detection would allow the system to detect novel or unseen spoof types, a crucial capability for real-world deployment in next-generation hearing-assistive systems. Finally, future work may incorporate adaptive fusion strategies and lightweight architectures, enabling real-time deployment in resource-constrained hearing devices without compromising detection accuracy.

Overall, our study demonstrates that multimodal, multi-attack frameworks provide a more secure and reliable foundation for hearing-assistive technologies, paving the way for robust defenses against increasingly sophisticated audio-visual spoofing threats.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Simikić, V. (2024, May). Internet of Medical Things (IoMT): Smart Hearing Aids, Today and Tomorrow. In *Serbian International Conference on Applied Artificial Intelligence* (pp. 402-411). Cham: Springer Nature Switzerland. [CrossRef]

[2] Hussain, A., Goman, A. M., Gogate, M., Dashtipour, K., Kirton-Wingate, J., Hussain, Z., Sheikh, A., Akeroyd, M. A., & Hussain, A. (2026). Audio-visual speech-in-noise tests for evaluating speech reception thresholds: A scoping review. *PLoS One, 21*(1), e0338600. [CrossRef]

[3] Ren, Z., Qian, K., Schultz, T., & Schuller, B. W. (2023, December). An overview of the ICASSP special session on AI security and privacy in speech and audio processing. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops* (pp. 1-6). [CrossRef]

[4] 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). [CrossRef]

[5] Yamagishi, J., Todisco, M., Sahidullah, M., Delgado, H., Wang, X., Evans, N., ... & Nautsch, A. (2019). Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *ASV Spoof*, 13. [Link]

[6] 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*. [arXiv]

[7] Chugh, K., Gupta, P., Dhall, A., & Subramanian, R. (2020, October). Not made for each other-audio-visual dissonance-based deepfake detection and localization. In Proceedings of the 28th ACM international conference on multimedia (pp. 439-447). [CrossRef]

[8] Borrelli, C., Bestagini, P., Antonacci, F., Sarti, A., & Tubaro, S. (2021). Synthetic speech detection through short-term and long-term prediction traces. *EURASIP Journal on Information Security, 2021*(1), 2. [CrossRef]

[9] He, Y., & Wang, S. (2025, October). Towards Highly Generalized Lip-Sync Deepfake Detection via Detailed Audio-Visual Inconsistency Analysis. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* (pp. 522-535). Singapore: Springer Nature Singapore. [CrossRef]

[10] Chung, J. S., & Zisserman, A. (2016, November). Out of time: automated lip sync in the wild. In *Asian conference on computer vision* (pp. 251-263). Cham: Springer International Publishing. [CrossRef]

[11] Ge, S., Lin, F., Li, C., Zhang, D., Wang, W., & Zeng, D. (2022). Deepfake video detection via predictive representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 18*(2s), 1-21. [CrossRef]

[12] Afouras, T., Chung, J. S., & Zisserman, A. (2018). The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*. [arXiv]

[13] Garg, A., Noyola, J., & Bagadia, S. (2016). Lip reading

using CNN and LSTM. *Technical report, Stanford University, CS231 n project report*. [Link]

[14] Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y. G., & Li, S. N. (2022, June). M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval* (pp. 615-623). [CrossRef]

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*. [Link]

[16] Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. *Proceedings of Interspeech 2020*, 5036-5040. [CrossRef]

[17] Delgado, H., Evans, N., Kinnunen, T., Lee, K. A., Liu, X., Nautsch, A., ... & Yamagishi, J. (2021). ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *arXiv preprint arXiv:2109.00535*. [arXiv]

[18] Khan, A. A., Laghari, A. A., Inam, S. A., Ullah, S., Shahzad, M., & Syed, D. (2025). A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. *Discover Computing, 28*(1), 48. [CrossRef]

[19] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., ... & Ling, Z. H. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language, 64*, 101114. [CrossRef]

[20] Wang, X., Xiao, Y., & Zhu, X. (2017, August). Feature Selection Based on CQCCs for Automatic Speaker Verification Spoofing. In *Interspeech* (pp. 32-36). [CrossRef]

[21] Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D., Le, Q.V. (2019) SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Proc. Interspeech 2019*, 2613-2617. [CrossRef]

[22] Jung, T., Kim, S., & Kim, K. (2020). Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access, 8*, 83144-83154. [CrossRef]

[23] Li, Y., Zhang, M., Ren, M., Qiao, X., Ma, M., Wei, D., & Yang, H. (2024, November). Cross-domain audio deepfake detection: Dataset and analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 4977-4983). [CrossRef]

[24] Khan, N., Nguyen, T., & Khalil, I. (2025, December). Adaptive Inter-Modality Attention for Enhanced Cross-Domain Deepfake Detection Transferability. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia* (pp. 1-7). [CrossRef]

[25] Liu, X., Sahidullah, M., Lee, K. A., & Kinnunen, T. (2024). Generalizing speaker verification for spoof awareness in the embedding space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 1261-1273. [CrossRef]

[26] Khan, A., Malik, K. M., Ryan, J., & Saravanan, M. (2023). Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artificial Intelligence Review*. [CrossRef]

[27] Hussain, T., Saleem, N., Dashtipour, K., Ahmed, S., Hou, J. C., Anwar, U., ... & Hussain, A. (2026). Audio-visual speech enhancement in noisy environments using emotion-based contextual cues. *The Journal of the Acoustical Society of America, 159*(1), 470-483. [CrossRef]

[28] Dashtipour, K. K., Gogate, M., Ahmed, S., Hussain, A., Hussain, T., Hou, J. C., Arslan, T., Tsao, Y., & Hussain, A. (2024). Towards cross-lingual audio-visual speech enhancement. In *Proceedings of AVSEC 2024* (pp. 30-32). [CrossRef]

Aysha Munawwara is a researcher at Edinburgh Napier University with interests in artificial intelligence, data analytics, and intelligent systems. Her work focuses on applying machine learning and computational techniques to real-world problems, with particular emphasis on data-driven modelling and interdisciplinary research. She is actively involved in research activities and contributes to collaborative academic projects within the computing and engineering domains. (Email: amunawwara@gmail.com)

Kia Dashtipour is a researcher at Edinburgh Napier University with an interdisciplinary background spanning artificial intelligence, cybersecurity, machine learning, and multimodal signal processing. His research focuses on the development of intelligent systems for applications such as cyberbullying detection, intrusion detection, speech and audio processing, and multimodal data fusion. He has published extensively in high-impact international journals and conferences and actively collaborates with academic and industrial partners on applied AI research. (Email: k.dashtipour@napier.ac.uk)

Nasir Saleem is an academic at Edinburgh Napier University whose research interests lie in computer science and intelligent systems. His work includes machine learning, data analytics, and applied artificial intelligence, with a focus on developing robust computational methods for real-world applications. He is actively involved in research, teaching, and supervision across computing and engineering disciplines. (Email: nasirsaleem@gu.edu.pk)

Mandar Gogate is a lecturer and researcher at Edinburgh Napier University specialising in artificial intelligence and machine learning. His research interests include deep learning, multimodal learning, computer vision, and data fusion, with applications across healthcare, security, and intelligent systems. He has

published in leading journals and conferences and is engaged in collaborative, interdisciplinary research projects. (Email: m.gogate@napier.ac.uk)

Adeel Hussain is a researcher at Edinburgh Napier University with expertise in artificial intelligence, machine learning, and data-driven modelling. His research focuses on developing advanced computational techniques for pattern recognition, signal processing, and intelligent decision-making systems. He contributes to interdisciplinary research projects and supports teaching and supervision in computing-related fields. (Email: adeel.hussain@napier.ac.uk)

Amir Hussain is a Professor of Artificial Intelligence at Edinburgh Napier University and an internationally recognized expert in the fields of cognitive computing, machine learning, and intelligent systems. His research interests span explainable artificial intelligence, neural networks, multimodal data analysis, and human-centered AI, with a strong focus on developing innovative solutions for applications in healthcare, security, engineering, and data-driven decision-making. He has authored numerous highly cited publications in leading international journals and conferences, and continues to play a significant role in advancing global research collaboration, academic leadership, and interdisciplinary innovation across artificial intelligence and related domains. (Email: a.hussain@napier.ac.uk)