Navigating the Soundscape of Deception: A Comprehensive Survey on Audio Deepfake Generation, Detection, and Future Horizons

Other titles in Foundations and Trends® in Privacy and Security

Reverse Engineering of Deceptions on Machine- and Human-Centric Attacks
Yuguang Yao, Xiao Guo, Vishal Asnani, Yifan Gong, Jiancheng Liu, Xue Lin, Xiaoming Liu and Sijia Liu
ISBN: 978-1-63828-340-9

Identifying and Mitigating the Security Risks of Generative AI
Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, Kathleen Fisher, Tatsunori Hashimoto, Dan Hendrycks, Somesh Jha, Daniel Kang, Florian Kerschbaum, Eric Mitchell, John Mitchell, Zulfikar Ramzan, Khawaja Shams, Dawn Song, Ankur Taly and Diyi Yang
ISBN: 978-1-63828-312-6

Cybersecurity for Modern Smart Grid Against Emerging Threats
Daisuke Mashima, Yao Chen, Muhammad M. Roomi, Subhash Lakshminarayana and Deming Chen
ISBN: 978-1-63828-294-5

Decentralized Finance: Protocols, Risks, and Governance
Agostino Capponi, Garud Iyengar and Jay Sethuraman
ISBN: 978-1-63828-270-9

Navigating the Soundscape of Deception: A Comprehensive Survey on Audio Deepfake Generation, Detection, and Future Horizons

Taiba Majid Wani
Sapienza University of Rome
majid@diag.uniroma1.it

Syed Asif Ahmad Qadri
National Tsing Hua University
syedasif@m110.nthu.edu.tw

Farooq Ahmad Wani
Sapienza University of Rome
wani@diag.uniroma1.it

Irene Amerini
Sapienza University of Rome
amerini@diag.uniroma1.it

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Privacy and Security

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

T. M. Wani *et al.*. *Navigating the Soundscape of Deception: A Comprehensive Survey on Audio Deepfake Generation, Detection, and Future Horizons*. Foundations and Trends® in Privacy and Security, vol. 6, no. 3-4, pp. 153–345, 2024.

ISBN: 978-1-63828-493-2

Foundations and Trends® in Privacy and Security
Volume 6, Issue 3-4, 2024

Editorial Board

Editorial Scope

Foundations and Trends® in Privacy and Security publishes survey and tutorial articles in the following topics:

- Access control
- Accountability
- Anonymity
- Application security
- Artifical intelligence methods in security and privacy
- Authentication
- Big data analytics and privacy
- Cloud security
- Cyber-physical systems security and privacy
- Distributed systems security and privacy
- Embedded systems security and privacy
- Forensics
- Hardware security
- Human factors in security and privacy
- Information flow
- Intrusion detection
- Malware
- Metrics
- Mobile security and privacy
- Language-based security and privacy
- Network security
- Privacy-preserving systems
- Protocol security
- Security and privacy policies
- Security architectures
- System security
- Web security and privacy

Information for Librarians

Foundations and Trends® in Privacy and Security, 2024, Volume 6, 4 issues. ISSN paper version 2474-1558. ISSN online version 2474-1566. Also available as a combined paper and online subscription.

Contents

Navigating the Soundscape of Deception: A Comprehensive Survey on Audio Deepfake Generation, Detection, and Future Horizons

Taiba Majid Wani[1], Syed Asif Ahmad Qadri[2], Farooq Ahmad Wani[1] and Irene Amerini[1]

[1] *Sapienza University of Rome, Italy; majid@diag.uniroma1.it, wani@diag.uniroma1.it, amerini@diag.uniroma1.it*
[2] *National Tsing Hua University, Taiwan; syedasif@m110.nthu.edu.tw*

ABSTRACT

The rise of audio deepfakes presents a significant security threat that undermines trust in digital communications and media. These synthetic audio technologies can convincingly mimic a person's voice, enabling malicious activities like impersonation, fraud, and misinformation. Addressing this growing threat requires robust detection systems to ensure the authenticity of digital content.

In this survey, we provide a comprehensive analysis of the state-of-the-art techniques in audio deepfake generation and detection. We examine various methods used to generate audio deepfakes, including Text-to-Speech (TTS) and Voice Conversion (VC) technologies, and discuss their capabilities in producing highly realistic synthetic audio. On the

Taiba Majid Wani, Syed Asif Ahmad Qadri, Farooq Ahmad Wani and Irene Amerini (2024), "Navigating the Soundscape of Deception: A Comprehensive Survey on Audio Deepfake Generation, Detection, and Future Horizons", Foundations and Trends® in Privacy and Security: Vol. 6, No. 3-4, pp 153–345. DOI: 10.1561/3300000048.

detection front, we explore a wide range of approaches, encompassing traditional machine learning and deep learning models for feature extraction and classification. The importance of publicly available datasets for training and evaluating these models is emphasized, showcasing their role in advancing detection capabilities.

Additionally, the integration of audio and video deepfake detection systems is discussed, providing a comprehensive defense against sophisticated attacks. This survey critically assesses existing methods and datasets, highlighting challenges like the high realism of deepfakes, limited data diversity, and the need for models that generalize well. It aims to guide future research in enhancing detection and safeguarding digital media integrity.

1

Introduction

In recent years, social media platforms have revolutionized information dissemination, marking a significant departure from traditional communication channels. By breaking down geographical and cultural barriers, they have become essential for knowledge sharing, fostering global connections, and enabling in-depth discussions. Social media has significantly broadened our horizons, allowing for the exchange of diverse ideas, experiences, and perspectives that would otherwise remain localized (Ali *et al.*, 2023). The capacity of social media to unite individuals from varied backgrounds, amplify underrepresented voices, and initiate collaborative ventures has fundamentally altered our engagement with the world. However, this landscape is not without its challenges. The openness and immediacy that facilitate information flow also expose these platforms to misuse. A concerning aspect has emerged, characterized by the spread of harmful content intended to mislead and manipulate public perception (Chen *et al.*, 2023). In this interconnected era, the swift spread of misinformation and deceptive narratives poses notable risks to societal well-being. This juxtaposition

reflects the delicate balance between the beneficial and adverse impacts of digital platforms, highlighting the need for careful vigilance and responsible management in the digital area.

Central to this digital transformation is the advent of artificial intelligence (AI), particularly its subset, deep learning (DL), which mimics the complex processes of the human brain. DL has emerged as a transformative force across various sectors, enabling organizations to innovate their products and services significantly (Saxena *et al.*, 2023). An illustrative example is Instagram's use of DL to address cyberbullying, showcasing how sophisticated algorithms can foster safer digital communities by identifying and mitigating harmful interactions (Sachdeva, 2021; Yi and Zubiaga, 2023). Such applications of AI and DL not only demonstrate technology's potential to enhance digital spaces but also reflect a commitment to social responsibility. Moreover, DL's influence extends to the realm of communication, as seen in Gmail's smart replies and the development of AI-driven chatbots (Chen *et al.*, 2019). These innovations, characterized by their ability to offer personalized and context-aware interactions, represent the synergy between human intelligence and machine efficiency. They pave the way for a new era of human-machine communication, enhancing user experiences across industries through seamless, natural language-based interactions (Olujimi and Ade-Ibijola, 2023).

However, this digital utopia is counterbalanced by the insidious rise of deepfake technology, a phenomenon that tests the boundaries of media manipulation (Xiao *et al.*, 2023). The term "deepfake," derived from "deep learning" and "fake," encapsulates the essence of this AI-driven manipulation. Through the employment of deep neural networks, hyper-realistic yet entirely fabricated content, spanning images, audio, and videos, are generated (Yan, 2023). The use of multimedia content as evidence in the legal world has become increasingly common, but it presents a significant challenge due to the rise of sophisticated manipulation tools. The authenticity and integrity of audio-visual evidence must be rigorously verified to ensure its credibility in legal proceedings. However, the emergence of easily accessible manipulation tools,

such as FaceApp[1], Sound Forge[2], DeepFaceLab[3], Wombo[4], REFACE[5], Wav2Lip[6], Avatarify[7], and Deepart.io[8], have made it easier to create realistic fabricated data.

Deepfake technology encompasses various categories, including face-swap, lip-synching, puppet-master, face synthesis and attribute manipulation, and audio-only deepfakes. Face-swap deepfakes involve replacing a person's face with another person's face, often targeting famous individuals in scenarios they never appeared in (Walczyna and Piotrowski, 2023). Lip-synching-based deepfakes manipulate a target person's lip movements to sync with a specific audio recording, making it appear as though they are saying something they did not (Kumar *et al.*, 2017). Puppet-master deepfakes mimic a target person's expressions, including eye movement and facial expressions, to create a video that animates the impersonator's desires (Pantelić and Gavrovska, 2022). Face synthesis and attribute manipulation involve generating photo-realistic face images and editing facial attributes, often used for spreading disinformation on social media. These technologies have witnessed a rising presence in society, with discernible repercussions across various dimensions. This is exemplified by instances such as manipulated videos altering the public perception of figures like Nancy Pelosi (Funke, 2020) or misleading political campaign content featuring Joe Biden (Kessler, 2020). Deepfakes have also found a place in entertainment and creative applications, with programs like Spangler and Murphy (Spangler, 2020) and Huang's face-swapping tool (Murphy and Huang, 2019).

Audio deepfakes are centred on the generation of a target speaker's voice, employing advanced deep learning methodologies to convincingly replicate speech patterns and vocal characteristics. This technology offers the potential to make individuals appear as if they are uttering statements they have never actually articulated. Two prevalent approaches in audio deepfake creation are text-to-speech synthesis (TTS)

[1] www.faceapp.com
[2] www.magix.com/us/music-editing/sound-forge/
[3] https://www.deepfakevfx.com/downloads/deepfacelab/
[4] www.wombo.ai/
[5] reface.ai
[6] github.com/Rudrabha/Wav2Lip
[7] https://avatarify.ai/
[8] https://creativitywith.ai/deepartio/

and voice conversion (VC) (Masood *et al.*, 2023). In TTS-based deepfakes, the system generates natural-sounding voice waveforms based on provided text input, effectively mimicking the target speaker's voice (Taylor, 2009). For instance, a malicious actor could use TTS to fabricate an audio clip of a political figure endorsing a policy they never supported. On the other hand, VC techniques transform the speech signal of a source speaker to make it appear as though it was spoken by the target speaker while preserving linguistic content. An example of VC-based deepfake might involve altering a recorded statement from one individual to make it sound like it was said by a different person, potentially leading to significant misinformation and misattribution (Mohammadi and Kain, 2017). As audio deepfake technology advances, it poses substantial challenges to the integrity of voice-based communication and authentication; the instance of a CEO falling victim to a deepfake voice impersonation scam demonstrates the tangible financial risks associated with this technology (Stupp, 2019; Levine, 2020).

The emergence of deepfake technology has raised both challenges and opportunities, offering significant potential for diversifying into various commercial ventures and fostering innovation (Johnson and Diakopoulos, 2021). Deepfakes possess the capacity to reshape and advance business models, particularly as consumers increasingly engage in virtual environments (Kietzmann *et al.*, 2020). Companies like Meta, formerly Facebook, are actively investing substantial amounts, such as $10 billion in 2021 alone, in developing a virtual reality world known as the Metaverse, featuring deepfake-generated objects, signalling novel prospects alongside new challenges (Mateo, 2023). This dualistic nature of deepfake technology motivates exploration of both its risks and opportunities, a facet that remains relatively uncharted in the current business literature.

Furthermore, deepfakes on the internet and social media platforms have become integral to personal and professional interactions, providing easy-to-use avenues for real-time discussions, ideological expression, and information sharing (Karnouskos, 2020). This rapid dissemination, combined with the increasing integration of digital technologies into society, is poised to have far-reaching consequences in the marketplace. However, due to the intricate and emergent nature of deepfake technol-

ogy, the current comprehension of its implications remains fragmented, limited, and in its early stages of development (Dwivedi *et al.*, 2021). In light of the multifaceted implications of deepfake technology, it's essential to note that, thus far, the predominant focus of research and development efforts has been directed towards video deepfakes, given their visually compelling nature and potential for misuse (Chesney and Citron, 2018). A substantial body of work has emerged, encompassing everything from detection methodologies to ethical and legal considerations, all geared towards mitigating the negative impacts of video-based deepfakes. However, it's important to acknowledge that, while video deepfakes have been a primary focus, audio deepfakes have received comparatively limited attention in both research and public discourse (Somoray and Miller, 2023).

The availability of deepfake databases and generation algorithms has democratised the creation of convincing deepfake content, leading to an exponential increase in their dissemination across online platforms, amplified by the rapid reach and sharing capabilities of social media. This surge in deepfake-related issues is mirrored in the growing body of scientific literature, which delves into the technological aspects of deepfake generation and detection and explores the ethical, social, and legal dimensions. While there are existing reviews in specific subfields, such as deepfake creation and detection (Heidari *et al.*, 2024; Masood *et al.*, 2023; Dagar and Vishwakarma, 2022), (Mirsky and Lee (2021) and Abu-Ein *et al.* (2022)), legal considerations (Akpuokwe *et al.*, 2024; Kaddoura and Al Husseiny, 2023; Silva, 2021; Perot and Mostert, 2020), forensics (Kingra *et al.*, 2023; Amerini *et al.*, 2021; Verdoliva, 2020), social spam (Qazi *et al.*, 2024; Aljabri *et al.*, 2023; Rao *et al.*, 2021; Yurtseven *et al.*, 2021) and social impact (Wazid *et al.*, 2024; Al-Khazraji *et al.*, 2023; Hancock and Bailenson, 2021; Gamage *et al.*, 2021), none comprehensively encompass the entire spectrum of deepfake research areas. This gap presents an opportunity for researchers seeking to contribute to this rapidly evolving field, which spans diverse disciplines and continuously adapts to emerging trends and funding opportunities. Despite its relative novelty, deepfake research holds immense potential for interdisciplinary collaboration and innovation, with the potential to shape the future of digital content and its societal implications.

Figure 1.1 shows the trend in the number of research publications related to *Deepfakes* from 2015 to 2024, based on data from Dimensions.ai.[9] The graph indicates a significant increase in publications on both *Video Deepfakes* and *Audio Deepfakes* over this period. Notably, publications on Video deepfakes consistently exceed those on Audio Deepfakes, suggesting a stronger research focus on visual deception in deepfake technology. This trend reflects the broader impact and challenges associated with video deepfakes, particularly in areas like multimedia manipulation, digital forensics, and societal impacts. Although Audio deepfakes have garnered attention, the volume of research remains lower, highlighting a need for more focused studies on the implications and challenges of audio deepfakes. The objective of this survey is to address this gap by exploring the unique aspects of audio deepfakes within the broader context of deepfake technology, which has predominantly focused on visual deception.

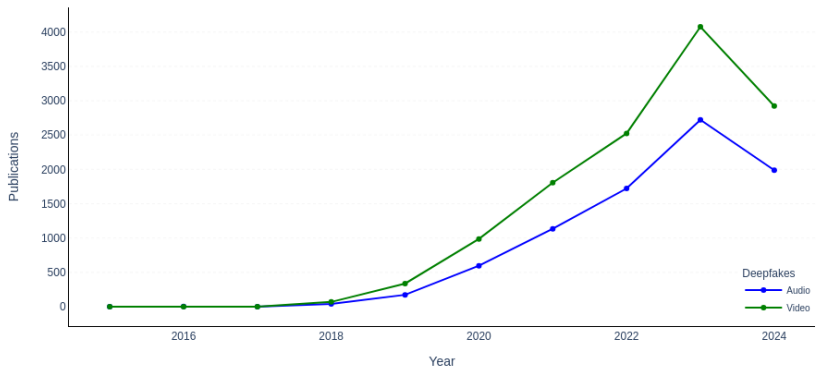

Figure 1.1: Trends in research publications on 'Deepfakes' from 2015 to 2024. The graph illustrates a higher number of publications on Video Deepfakes compared to Audio Deepfakes throughout the period, reflecting a predominant research focus on visual deception technologies.

[9] app.dimensions.ai/discover/publication

In this survey, we set the stage by providing an overview of the significance of audio deepfakes, their potential threats to security and trust in digital communications, and the necessity for robust detection systems. Section 2 traces the development of deepfake technology from its inception to its current state. We discuss the advancements in generative models and the increasing sophistication of deepfake creation techniques over time. We discuss the specific methods used to generate audio deepfakes in Section 3. This includes an exploration of Text-to-Speech (TTS) and Voice Conversion (VC) technologies, and the tools that enable these processes. Section 4 discusses audio deepfake detection systems. We examine the various techniques employed to extract meaningful features from audio data for the purpose of deepfake detection. This includes both traditional and deep learning approaches. Also, discuss the different classifiers used in detecting audio deepfakes. We cover a range of methods from machine learning to deep learning classifiers, highlighting their strengths and weaknesses in various detection scenarios. Section 5 provides an extensive review of the datasets commonly used for training and evaluating deepfake detection models. Section 6 provides the comparative analysis of various strategies applied in audio deepfake detection. Section 7 provides a detailed overview of the evaluation metrics considered throughout the cited works. Section 8 identifies and elaborates on the significant challenges in the field. Finally, in Section 9, we summarize the key findings of the survey and draw the conclusions.

References

Abu-Ein, A. A., O. M. Al-Hazaimeh, A. M. Dawood, and A. I. Swidan. (2022). "Analysis of the current state of deepfake techniques-creation and detection methods". *Indonesian Journal of Electrical Engineering and Computer Science*. 28(3): 1659–1667.

Adiga, A., M. Magimai, and C. S. Seelamantula. (2013). "Gammatone wavelet cepstral coefficients for robust speech recognition". In: *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)*. IEEE. 1–4.

Akpuokwe, C. U., A. O. Adeniyi, and S. S. Bakare. (2024). "Legal challenges of artificial intelligence and robotics: a comprehensive review". *Computer Science & IT Research Journal*. 5(3): 544–561.

Alam, J. and P. Kenny. (2017). "Spoofing detection employing infinite impulse response—constant Q transform-based feature representations". In: *2017 25Th european signal processing conference (EUSIPCO)*. IEEE. 101–105.

AlBadawy, E. A., S. Lyu, and H. Farid. (2019). "Detecting AI-Synthesized Speech Using Bispectral Analysis." In: *CVPR workshops*. 104–109.

Ali, I., M. Balta, and T. Papadopoulos. (2023). "Social media platforms and social enterprise: Bibliometric analysis and systematic review". *International Journal of Information Management*. 69: 102510.

Ali, Z., I. Elamvazuthi, M. Alsulaiman, and G. Muhammad. (2016). "Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model". *Journal of voice*. 30(6): 757–e7.

Aljabri, M., R. Zagrouba, A. Shaahid, F. Alnasser, A. Saleh, and D. M. Alomari. (2023). "Machine learning-based social media bot detection: a comprehensive literature review". *Social Network Analysis and Mining*. 13(1): 20.

Almutairi, Z. and H. Elgibreen. (2022). "A review of modern audio deepfake detection methods: challenges and future directions". *Algorithms*. 15(5): 155.

Almutairi, Z. M. and H. Elgibreen. (2023). "Detecting fake audio of Arabic speakers using self-supervised deep learning". *IEEE Access*. 11: 72134–72147.

Alzantot, M., Z. Wang, and M. B. Srivastava. (2019). "Deep residual neural networks for audio spoofing detection". *arXiv preprint arXiv:1907.00501*.

Amerini, I., A. Anagnostopoulos, L. Maiano, L. R. Celsi, *et al.* (2021). "Deep learning for multimedia forensics". *Foundations and Trends® in Computer Graphics and Vision*. 12(4): 309–457.

Aravind, P., U. Nechiyil, N. Paramparambath, *et al.* (2020). "Audio spoofing verification using deep convolutional neural networks by transfer learning". *arXiv preprint arXiv:2008.03464*.

Arif, T., A. Javed, M. Alhameed, F. Jeribi, and A. Tahir. (2021). "Voice Spoofing Countermeasure for Logical Access Attacks Detection". *IEEE Access*. 9: 162857–162868. DOI: 10.1109/ACCESS.2021.3133134.

Arık, S. Ö., M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, *et al.* (2017). "Deep voice: Real-time neural text-to-speech". In: *International conference on machine learning*. PMLR. 195–204.

Attorresi, L., D. Salvi, C. Borrelli, P. Bestagini, and S. Tubaro. (2022). "Combining automatic speaker verification and prosody analysis for synthetic speech detection". *arXiv preprint arXiv:2210.17222*.

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

Ballesteros, D. M., Y. Rodriguez, and D. Renza. (2020). "A dataset of histograms of original and fake voice recordings (H-Voice)". *Data in brief*. 29.

Barnett, J. (2023). "The ethical implications of generative audio models: A systematic literature review". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 146–161.

Barrault, L., Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, *et al.* (2023). "SeamlessM4T-Massively Multilingual & Multimodal Machine Translation". *arXiv preprint arXiv:2308.11596*.

Barrington, S., M. Bohacek, and H. Farid. (2024). "DeepSpeak Dataset v1. 0". *arXiv preprint arXiv:2408.05366*.

Bartusiak, E. R. and E. J. Delp. (2022). "Frequency domain-based detection of generated audio". *arXiv preprint arXiv:2205.01806*.

Baumann, R., K. M. Malik, A. Javed, A. Ball, B. Kujawa, and H. Malik. (2021). "Voice spoofing detection corpus for single and multi-order audio replays". *Computer Speech & Language*. 65: 101132.

Bhat, H. R., T. A. Lone, and Z. M. Paul. (2017). "Cortana-intelligent personal digital assistant: A review". *International Journal of Advanced Research in Computer Science*. 8(7): 55–57.

Bhatia, K., A. Agrawal, P. Singh, and A. K. Singh. (2022). "Detection of AI Synthesized Hindi Speech". *arXiv preprint arXiv:2203.03706*.

Bhatt, S., A. Jain, and A. Dev. (2021). "Feature extraction techniques with analysis of confusing words for speech recognition in the Hindi language". *Wireless Personal Communications*. 118: 3303–3333.

Bińkowski, M., J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan. (2019). "High fidelity speech synthesis with adversarial networks". *arXiv preprint arXiv:1909.11646*.

Blue, L., K. Warren, H. Abdullah, C. Gibson, L. Vargas, J. O'Dell, K. Butler, and P. Traynor. (2022). "Who Are You (I Really Wanna Know)? Detecting Audio {DeepFakes} Through Vocal Tract Reconstruction". In: *31st USENIX Security Symposium (USENIX Security 22)*. 2691–2708.

Bogach, N., E. Boitsova, S. Chernonog, A. Lamtev, M. Lesnichaya, I. Lezhenin, A. Novopashenny, R. Svechnikov, D. Tsikach, K. Vasiliev, *et al.* (2021). "Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching". *Electronics*. 10(3): 235.

Borrelli, C., P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro. (2021). "Synthetic speech detection through short-term and long-term prediction traces". *EURASIP Journal on Information Security*. 2021(1): 1–14.

Boyd, J., M. Fahim, and O. Olukoya. (2023). "Voice spoofing detection for multiclass attack classification using deep learning". *Machine Learning with Applications*. 14: 100503.

Cáceres, J., R. Font, T. Grau, J. Molina, and B. V. SL. (2021). "The Biometric Vox system for the ASVspoof 2021 challenge". In: *Proc. ASVspoof2021 Workshop*.

Camacho, S., D. M. Ballesteros, and D. Renza. (2021). "Fake speech recognition using deep learning". In: *Applied Computer Sciences in Engineering: 8th Workshop on Engineering Applications, WEA 2021, Medellín, Colombia, October 6–8, 2021, Proceedings 8*. Springer. 38–48.

Cao, D., Z. Zhang, and J. Zhang. (2024). "NeuralVC: Any-to-Any Voice Conversion Using Neural Networks Decoder For Real-Time Voice Conversion". *IEEE Signal Processing Letters*.

Capes, T., P. Coles, A. Conkie, L. Golipour, A. Hadjitarkhani, Q. Hu, N. Huddleston, M. Hunt, J. Li, M. Neeracher, *et al.* (2017). "Siri on-device deep learning-guided unit selection text-to-speech system." In: *Interspeech*. 4011–4015.

Casanova, E., J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti. (2022). "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone". In: *International Conference on Machine Learning*. PMLR. 2709–2720.

Channing, G., J. Sock, R. Clark, P. Torr, and C. S. de Witt. (2024). "Toward Robust Real-World Audio Deepfake Detection: Closing the Explainability Gap". *arXiv preprint arXiv:2410.07436*.

Chen, M. X., B. N. Lee, G. Bansal, Y. Cao, S. Zhang, J. Lu, J. Tsay, Y. Wang, A. M. Dai, Z. Chen, *et al.* (2019). "Gmail smart compose: Real-time assisted writing". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2287–2295.

Chen, M., X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu. (2021). "Adaspeech: Adaptive text to speech for custom voice". *arXiv preprint arXiv:2103.00993*.

Chen, S., S. Liu, L. Zhou, Y. Liu, X. Tan, J. Li, S. Zhao, Y. Qian, and F. Wei. (2024). "VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers". *arXiv preprint arXiv:2406.05370*.

Chen, S., L. Xiao, and A. Kumar. (2023). "Spread of misinformation on social media: What contributes to it and how to combat it". *Computers in Human Behavior*. 141: 107643.

Chen, T., A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury. (2020). "Generalization of Audio Deepfake Detection." In: *Odyssey*. 132–137.

Cheng, X., M. Xu, and T. F. Zheng. (2019). "Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019". In: *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 540–545.

Chesney, R. and D. Citron. (2018). "Deep fakes: A looming crisis for national security, democracy and privacy". *The Lawfare Blog*.

Chintha, A., B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, and R. Ptucha. (2020). "Recurrent convolutional structures for audio spoof and video deepfake detection". *IEEE Journal of Selected Topics in Signal Processing*. 14(5): 1024–1037.

Choi, H.-Y., S.-H. Lee, and S.-W. Lee. (2024). "Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 16. 17862–17870.

Christensen, M. and A. Jakobsson. (2022). *Multi-pitch estimation*. Springer Nature.

Conner, K. and H. Farid. (2015). "Photo tampering throughout history". *Four and Six. Accessed*. 9.

Conti, E., D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, A. Sarti, M. C. Stamm, and S. Tubaro. (2022). "Deepfake speech detection through emotion recognition: a semantic approach". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 8962–8966.

Dagar, D. and D. K. Vishwakarma. (2022). "A literature review and perspectives in deepfakes: generation, detection, and applications". *International journal of multimedia information retrieval*. 11(3): 219–289.

Das, R. K., J. Yang, and H. Li. (2019). "Long range acoustic and deep features perspective on ASVspoof 2019". In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 1018–1025.

Das, R. K., J. Yang, and H. Li. (2020). "Assessing the scope of generalized countermeasures for anti-spoofing". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6589–6593.

De Leon, P. L., B. Stewart, and J. Yamagishi. (2012). "Synthetic Speech Discrimination using Pitch Pattern Statistics Derived from Image Analysis." In: *Interspeech*. 370–373.

Delgado, H., M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi. (2018). "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements". In: *Odyssey 2018-The Speaker and Language Recognition Workshop*.

Dhar, S., N. D. Jana, and S. Das. (2023). "GLGAN-VC: A guided loss-based generative adversarial network for many-to-many voice conversion". *IEEE Transactions on Neural Networks and Learning Systems*.

Dixit, A., N. Kaur, and S. Kingra. (2023). "Review of audio deepfake detection techniques: Issues and prospects". *Expert Systems*: e13322.

Doan, T. P., K. Hong, and S. Jung. (2023a). "GAN Discriminator based Audio Deepfake Detection". In: *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*. 29–32.

Doan, T.-P., L. Nguyen-Vu, S. Jung, and K. Hong. (2023b). "BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 1–5.

Donahue, C., J. McAuley, and M. Puckette. (2018). "Adversarial audio synthesis". *arXiv preprint arXiv:1802.04208*.

Driemel, A., A. Krivošija, and C. Sohler. (2016). "Clustering time series under the Fréchet distance". In: *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 766–785.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Vol. 3. Springer Science & Business Media.

Dwivedi, Y. K., L. Hughes, E. Ismagilova, G. Aarts, C. Coombs, T. Crick, Y. Duan, R. Dwivedi, J. Edwards, A. Eirug, *et al.* (2021). "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy". *International Journal of Information Management*. 57: 101994.

Ebden, P. and R. Sproat. (2015). "The Kestrel TTS text normalization system". *Natural Language Engineering*. 21(3): 333–353.

Elias, I., H. Zen, J. Shen, Y. Zhang, Y. Jia, R. Skerry-Ryan, and Y. Wu. (2021a). "Parallel Tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling". *arXiv preprint arXiv:2103.14574*.

Elias, I., H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu. (2021b). "Parallel tacotron: Non-autoregressive and controllable tts". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 5709–5713.

Firc, A., K. Malinka, and P. Hanáček. (2023). "Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors". *Heliyon*.

Forbes. (2019). *A Voice Deepfake Was Used To Scam A CEO Out Of $243,000*. URL: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.

Frank, J. and L. Schönherr. (2021). "Wavefake: A data set to facilitate audio deepfake detection". *arXiv preprint arXiv:2111.02813*.

Funke, D. (2020). "Nancy Pelosi doesn't drink so why do false claims about her being drunk keep going viral?" URL: https://www.politifact.com/article/2020/aug/03/why-false-claims-about-nancy-pelosi-being-drunk-ke/.

Gamage, D., K. Sasahara, and J. Chen. (2021). "The Emergence of Deepfakes and its Societal Implications: A Systematic Review." *TTO*: 28–39.

Ge, W., J. Patino, M. Todisco, and N. Evans. (2021). "Raw differentiable architecture search for speech deepfake and spoofing detection". *arXiv preprint arXiv:2107.12212*.

Gibiansky, A., S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou. (2017). "Deep voice 2: Multi-speaker neural text-to-speech". *Advances in neural information processing systems*. 30.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. (2014). "Generative adversarial nets". *Advances in neural information processing systems*. 27.

Govindu, A., P. Kale, A. Hullur, A. Gurav, and P. Godse. (2023). "Deepfake audio detection and justification with Explainable Artificial Intelligence (XAI)".

Guo, H., F. K. Soong, L. He, and L. Xie. (2019). "Exploiting syntactic features in a parsed tree to improve end-to-end TTS". *arXiv preprint arXiv:1904.04764*.

Guo, H., S. Zhang, F. K. Soong, L. He, and L. Xie. (2021). "Conversational end-to-end tts for voice agents". In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 403–409.

Hamza, A., A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol. (2022). "Deepfake audio detection via MFCC features using machine learning". *IEEE Access*. 10: 134018–134028.

Han, C., P. Mitra, and S. M. Billah. (2024). "Uncovering Human Traits in Determining Real and Spoofed Audio: Insights from Blind and Sighted Individuals". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.

Hancock, J. T. and J. N. Bailenson. (2021). "The Social Impact of Deepfakes". *Cyberpsychology, Behavior, and Social Networking*. 24(3): 149–152.

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

Hathaliya, J. J., S. Tanwar, and P. Sharma. (2022). "Adversarial learning techniques for security and privacy preservation: A comprehensive review". *Security and Privacy*. 5(3): e209.

Heidari, A., N. Jafari Navimipour, H. Dag, and M. Unal. (2024). "Deepfake detection using deep learning methods: A systematic and comprehensive review". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 14(2): e1520.

Hsu, C.-C., H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang. (2017). "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks". *arXiv preprint arXiv:1704.00849*.

Huang, L. and J. Zhao. (2021). "Audio replay spoofing attack detection using deep learning feature and long-short-term memory recurrent neural network". In: *AIIPCC 2021; The Second International Conference on Artificial Intelligence, Information Processing and Cloud Computing*. VDE. 1–5.

Huang, W.-C., T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda. (2019). "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining". *arXiv preprint arXiv:1912.06813*.

Hunt, A. J. and A. W. Black. (1996). "Unit selection in a concatenative speech synthesis system using a large speech database". In: *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*. Vol. 1. IEEE. 373–376.

Hussain, S., P. Neekhara, B. Dolhansky, J. Bitton, C. C. Ferrer, J. McAuley, and F. Koushanfar. (2022). "Exposing vulnerabilities of deepfake detection systems with robust attacks". *Digital Threats: Research and Practice (DTRAP)*. 3(3): 1–23.

Ilyas, H., A. Javed, and K. M. Malik. (2023). "AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection". *Applied Soft Computing*. 136: 110124.

Iqbal, F., A. Abbasi, A. R. Javed, Z. Jalil, and J. Al-Karaki. (2022). "Deepfake Audio Detection Via Feature Engineering And Machine Learning." In: *CIKM Workshops*.

Ito, K. and L. Johnson. (2017). "The LJ Speech Dataset". URL: https://keithito.com/LJ-Speech-Dataset/.

Jang, W., D. Lim, and J. Yoon. (2020). "Universal melgan: A robust neural vocoder for high-fidelity waveform generation in multiple domains". *arXiv preprint arXiv:2011.09631*.

Jiang, Z., H. Zhu, L. Peng, W. Ding, and Y. Ren. (2020). "Self-Supervised Spoofing Audio Detection Scheme." In: *INTERSPEECH*. 4223–4227.

Johnson, D. G. and N. Diakopoulos. (2021). "What to do about deepfakes". *Communications of the ACM*. 64(3): 33–35.

Jung, J., S. Lee, J. Kang, and Y. Na. (2024). "WWW: Where, Which and Whatever Enhancing Interpretability in Multimodal Deepfake Detection". *arXiv preprint arXiv:2408.02954*.

Kaddoura, S. and F. Al Husseiny. (2023). "The rising trend of Metaverse in education: Challenges, opportunities, and ethical considerations". *PeerJ Computer Science*. 9: e1252.

Kalpokas, I. and J. Kalpokiene. (2022). "From GANs to deepfakes: getting the characteristics right". In: *Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation*. Springer. 29–39.

Kameoka, H., T. Kaneko, K. Tanaka, and N. Hojo. (2018). "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 266–273.

Kameoka, H., K. Tanaka, and T. Kaneko. (2021). "Fasts2s-vc: Streaming non-autoregressive sequence-to-sequence voice conversion". *arXiv preprint arXiv:2104.06900*.

Kaneko, T. and H. Kameoka. (2018). "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks". In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE. 2100–2104.

Kaneko, T., H. Kameoka, K. Tanaka, and N. Hojo. (2019). "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6820–6824.

Karnouskos, S. (2020). "Artificial intelligence in digital media: The era of deepfakes". *IEEE Transactions on Technology and Society*. 1(3): 138–147.

Katamneni, V. S. and A. Rattani. (2023). "MIS-AVioDD: Modality invariant and specific representation for audio-visual deepfake detection". *arXiv preprint arXiv:2310.02234*.

Kawa, P., M. Plata, and P. Syga. (2022). "Specrnet: Towards faster and more accessible audio deepfake detection". In: *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE. 792–799.

Kenyon, J. (2020). "Clarifying the Probable Cause Standard in the Internet Age for Crimes Involving Child Pornography". *Cath. UL Rev.* 69: 633.

Kessler, G. (2020). "Trump Campaign Ad Manipulates Three Images to Put Biden in a'Basement'".

Khan, A., K. M. Malik, J. Ryan, and M. Saravanan. (2022). "Voice Spoofing Countermeasures: Taxonomy, State-of-the-art, experimental analysis of generalizability, open challenges, and the way forward". *arXiv preprint arXiv:2210.00417*.

Khanjani, Z., G. Watson, and V. P. Janeja. (2023). "Audio deepfakes: A survey". *Frontiers in Big Data*. 5: 1001063.

Al-Khazraji, S. H., H. H. Saleh, A. I. KHALID, and I. A. MISHKHAL. (2023). "Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications". *The Eurasia Proceedings of Science Technology Engineering and Mathematics*. 23: 429–441.

Khochare, J., C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi. (2021). "A deep learning framework for audio deepfake detection". *Arabian Journal for Science and Engineering*: 1–12.

Kietzmann, J., L. W. Lee, I. P. McCarthy, and T. C. Kietzmann. (2020). "Deepfakes: Trick or treat?" *Business Horizons*. 63(2): 135–146.

Kiliç, B. and M. E. Kahraman. (2023). "Current Usage Areas of Deepfake Applications with Artificial Intelligence Technology". *İletişim ve Toplum Araştırmaları Dergisi*. 3(2): 301–332.

Kim, J., J. Kong, and J. Son. (2021). "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech". In: *International Conference on Machine Learning*. PMLR. 5530–5540.

Kingma, D. P. and P. Dhariwal. (2018). "Glow: Generative flow with invertible 1x1 convolutions". *Advances in neural information processing systems*. 31.

Kingra, S., N. Aggarwal, and N. Kaur. (2023). "Emergence of deepfakes and video tampering detection approaches: A survey". *Multimedia Tools and Applications*. 82(7): 10165–10209.

Kinnunen, T., M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee. (2017a). "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection".

Kinnunen, T., M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, *et al.* (2017b). "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research". In: *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 5395–5399.

Klein, N., T. Chen, H. Tak, R. Casal, and E. Khoury. (2024). "Source tracing of audio deepfake systems". *arXiv preprint arXiv:2407.08016*.

Kong, J., J. Kim, and J. Bae. (2020). "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis". *Advances in Neural Information Processing Systems*. 33: 17022–17033.

Korshunov, P. and S. Marcel. (2018). "Deepfakes: a new threat to face recognition? assessment and detection". *arXiv preprint arXiv:1812.08685*.

Korshunov, P., S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. S. Mello, R. V. Violato, F. O. Simoes, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi, *et al.* (2016). "Overview of BTAS 2016 speaker anti-spoofing competition". In: *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*. IEEE. 1–6.

Kumar, K., R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville. (2019). "Melgan: Generative adversarial networks for conditional waveform synthesis". *Advances in neural information processing systems*. 32.

Kumar, R., J. Sotelo, K. Kumar, A. De Brebisson, and Y. Bengio. (2017). "Obamanet: Photo-realistic lip-sync from text". *arXiv preprint arXiv:1801.01442*.

Kumari, K. A., S. Ahamad, T. Patil, K. Sardana, E. Muniyandy, and D. Pilli. (2024). "Neural Network Pruning Techniques for Efficient Model Compression". *International Journal of Intelligent Systems and Applications in Engineering*. 12(15s): 565–575.

Kwon, P., J. You, G. Nam, S. Park, and G. Chae. (2021). "Kodf: A large-scale korean deepfake detection dataset". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10744–10753.

Lai, C.-I., N. Chen, J. Villalba, and N. Dehak. (2019). "ASSERT: Anti-spoofing with squeeze-excitation and residual networks". *arXiv preprint arXiv:1904.01120*.

Łajszczak, M., G. Cámbara, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski, *et al.* (2024). "BASE TTS: Lessons from building a billion-parameter text-to-speech model on 100K hours of data". *arXiv preprint arXiv:2402.08093*.

Lataifeh, M. and A. Elnagar. (2020). "Ar-DAD: Arabic diversified audio dataset". *Data in Brief*. 33: 106503.

Lavrentyeva, G., S. Novoselov, and K. Simonchik. (2017). "Anti-spoofing methods for automatic speaker verification system". In: *Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised Selected Papers 5*. Springer. 172–184.

Lei, Z., Y. Yang, C. Liu, and J. Ye. (2020). "Siamese Convolutional Neural Network Using Gaussian Probability Feature for Spoofing Speech Detection." In: *INTERSPEECH*. 1116–1120.

Levine, A. J. (2020). "Dollars, Deception, and Deepfakes: An Analysis of Deepfakes and Synthetic Media Fraud". *PhD thesis*. Utica College.

Li, L., T. Lu, X. Ma, M. Yuan, and D. Wan. (2023a). "Voice Deepfake Detection Using the Self-Supervised Pre-Training Model HuBERT". *Applied Sciences*. 13(14): 8488.

Li, M., Y. Ahmadiadli, and X.-P. Zhang. (2022). "A Comparative Study on Physical and Perceptual Features for Deepfake Audio Detection". In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 35–41.

Li, X., N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng. (2021). "Replay and synthetic speech detection with res2net architecture". In: *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 6354–6358.

Li, Y. A., C. Han, and N. Mesgarani. (2023b). "Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models". In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 920–927.

Liew, T. W., S.-M. Tan, W. M. Pang, M. T. I. Khan, and S. N. Kew. (2023). "I am Alexa, your virtual tutor!: The effects of Amazon Alexa's text-to-speech voice enthusiasm in a multimedia learning environment". *Education and information technologies*. 28(2): 1455–1489.

Lim, S.-Y., D.-K. Chae, and S.-C. Lee. (2022). "Detecting deepfake voice using explainable deep learning techniques". *Applied Sciences*. 12(8): 3926.

Liu, R., B. Sisman, F. Bao, G. Gao, and H. Li. (2020). "Modeling prosodic phrasing with multi-task learning in tacotron-based TTS". *IEEE Signal Processing Letters*. 27: 1470–1474.

Liu, R., J. Zhang, and G. Gao. (2024). "Multi-space channel representation learning for mono-to-binaural conversion based audio deepfake detection". *Information Fusion*. 105: 102257.

Liu, T., D. Yan, R. Wang, N. Yan, and G. Chen. (2021a). "Identification of fake stereo audio using SVM and CNN". *Information*. 12(7): 263.

Liu, X., F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. (2021b). "Self-supervised learning: Generative or contrastive". *IEEE transactions on knowledge and data engineering*. 35(1): 857–876.

Liu, X., X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, *et al.* (2023a). "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Liu, Z., Y. Guo, and K. Yu. (2023b). "Diffvoice: Text-to-speech with latent diffusion". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 1–5.

Luo, A., E. Li, Y. Liu, X. Kang, and Z. J. Wang. (2021). "A capsule network based approach for detection of audio spoofing attacks". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6359–6363.

Lv, Z., S. Zhang, K. Tang, and P. Hu. (2022). "Fake audio detection based on unsupervised pretraining models". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 9231–9235.

Ma, H., J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang. (2021). "Continual learning for fake audio detection". *arXiv preprint arXiv:2104.07286*.

Ma, H., J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, L. Xu, and R. Fu. (2022). "FAD: A Chinese dataset for fake audio detection". *arXiv preprint arXiv:2207.12308*.

Ma, Q., J. Zhong, Y. Yang, W. Liu, Y. Gao, and W. Ng. (2023). "A Lightweight and Efficient Model for Audio Anti-Spoofing". In: *Proceedings of the 5th ACM International Conference on Multimedia in Asia*. 1–7.

Macharyas, J. P. (2015). "The malicious and forensic uses of Adobe software". *PhD thesis*. Utica College.

Maddocks, S. (2020). "'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political'deep fakes". *Porn Studies*. 7(4): 415–423.

Malik, H. (2019). "Securing voice-driven interfaces against fake (cloned) audio attacks". In: *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 512–517.

Martín-Doñas, J. M. and A. Álvarez. (2022). "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 9241–9245.

Martın-Donas, J. M., A. Alvarez, E. Rosello, A. M. Gomez, and A. M. Peinado. (2024). "Exploring Self-supervised Embeddings and Synthetic Data Augmentation for Robust Audio Deepfake Detection". In: *Proc. Interspeech*. Vol. 2024.

Masood, M., M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik. (2023). "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward". *Applied intelligence*. 53(4): 3974–4026.

Masuyama, Y., K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada. (2019). "Deep griffin–lim iteration". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 61–65.

Mateo, E. (2023). "A Deep Dive into Artificial Intelligence and Its Integration into Cybersecurity". *PhD thesis*. Utica University.

Mirsky, Y. and W. Lee. (2021). "The creation and detection of deepfakes: A survey". *ACM Computing Surveys (CSUR)*. 54(1): 1–41.

Mischie, S., L. Mâţiu-Iovan, and G. GăŠpăresc. (2018). "Implementation of google assistant on rasberry pi". In: *2018 International Symposium on Electronics and Telecommunications (ISETC)*. IEEE. 1–4.

Mittal, T., U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. (2020). "Emotions don't lie: An audio-visual deepfake detection method using affective cues". In: *Proceedings of the 28th ACM international conference on multimedia*. 2823–2832.

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

Mubarak, R., T. Alsboui, O. Alshaikh, I. Inuwa-Dute, S. Khan, and S. Parkinson. (2023). "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats". *IEEE Access*.

Muppalla, S., S. Jia, and S. Lyu. (2023). "Integrating audio-visual features for multimodal deepfake detection". *arXiv preprint arXiv:2310.03827*.

Murphy, C. and Z. Huang. (2019). "China's Red-Hot Face-Swapping App Provokes Privacy Concern".

Nercessian, S. (2020). "Zero-Shot Singing Voice Conversion." In: *ISMIR*. 70–76.

Nguyen, B. and F. Cardinaux. (2022). "Nvc-net: End-to-end adversarial voice conversion". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 7012–7016.

Nguyen, T. T., Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen. (2022). "Deep learning for deepfakes creation and detection: A survey". *Computer Vision and Image Understanding*. 223: 103525.

Novoselov, S., A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin. (2016). "STC anti-spoofing systems for the ASVspoof 2015 challenge". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 5475–5479.

Olujimi, P. A. and A. Ade-Ibijola. (2023). "NLP techniques for automating responses to customer queries: a systematic review". *Discover Artificial Intelligence*. 3(1): 20.

Oord, A., Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, *et al.* (2018). *Parallel wavenet: Fast high-fidelity speech synthesis*. PMLR.

Oord, A. v. d., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. (2016). "Wavenet: A generative model for raw audio". *arXiv preprint arXiv:1609.03499*.

Orrù, G., A. Galli, V. Gattulli, M. Gravina, M. Micheletto, S. Marrone, W. Nocerino, A. Procaccino, G. Terrone, D. Curtotti, *et al.* (2023). "Development of Technologies for the Detection of (Cyber) Bullying Actions: The BullyBuster Project". *Information*. 14(8): 430.

Ouyang, M., R. K. Das, J. Yang, and H. Li. (2021). "Capsule network based end-to-end system for detection of replay attacks". In: *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE. 1–5.

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

Pan, J., S. Nie, H. Zhang, S. He, K. Zhang, S. Liang, X. Zhang, and J. Tao. (2022). "Speaker recognition-assisted robust audio deepfake detection." In: *INTERSPEECH*. 4202–4206.

Pantelić, A. and A. Gavrovska. (2022). "From puppet-master creation to false detection".

Park, S.-W., D.-Y. Kim, and M.-C. Joe. (2020). "Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data". *arXiv preprint arXiv:2005.03295*.

Pascu, O., D. Oneata, H. Cucu, and N. M. Müller. (2024). "Easy, Interpretable, Effective: openSMILE for voice deepfake detection". *arXiv preprint arXiv:2408.15775*.

Patel, T. B. and H. A. Patil. (2015). "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech". In: *Sixteenth annual conference of the international speech communication association*.

Patel, T. B. and H. A. Patil. (2016). "Effectiveness of fundamental frequency (f 0) and strength of excitation (soe) for spoofed speech detection". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 5105–5109.

Patil, H. A., R. Acharya, A. T. Patil, and P. Gupta. (2022). "Non-Cepstral Uncertainty Vector for Replay Spoofed Speech Detection". In: *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE. 374–378.

Paul, D., M. Pal, and G. Saha. (2017). "Spectral features for synthetic speech detection". *IEEE journal of selected topics in signal processing*. 11(4): 605–617.

Pawelec, M. (2022). "Deepfakes and democracy (theory): how synthetic audio-visual media for disinformation and hate speech threaten core democratic functions". *Digital society*. 1(2): 19.

Perot, E. and F. Mostert. (2020). "Fake it till you make it: an examination of the US and English approaches to persona protection as applied to deepfakes on social media". *Journal of Intellectual Property Law & Practice*. 15(1): 32–39.

Phillips, H., S. Soffer, and E. Klang. (2022). "Oncological applications of deep learning generative adversarial networks". *JAMA oncology*. 8(5): 677–678.

Phukan, O. C., G. S. Kashyap, A. B. Buduru, and R. Sharma. (2024). "Heterogeneity over Homogeneity: Investigating Multilingual Speech Pre-Trained Models for Detecting Audio Deepfake". *arXiv preprint arXiv:2404.00809*.

Pianese, A., D. Cozzolino, G. Poggi, and L. Verdoliva. (2022). "Deepfake audio detection by speaker verification". In: *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE. 1–6.

Ping, W., K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller. (2017). "Deep voice 3: Scaling text-to-speech with convolutional sequence learning". *arXiv preprint arXiv:1710.07654*.

Prenger, R., R. Valle, and B. Catanzaro. (2019). "Waveglow: A flow-based generative network for speech synthesis". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 3617–3621.

Qais, A., A. Rastogi, A. Saxena, A. Rana, and D. Sinha. (2022). "Deepfake Audio Detection with Neural Networks Using Audio Features". In: *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*. IEEE. 1–6.

Qazi, A., N. Hasan, R. Mao, M. E. M. Abo, S. K. Dey, and G. Hardaker. (2024). "Machine Learning-Based Opinion Spam Detection: A Systematic Literature Review". *IEEE Access*.

Rabhi, M., S. Bakiras, and R. Di Pietro. (2024). "Audio-deepfake detection: Adversarial attacks and countermeasures". *Expert Systems with Applications*. 250: 123941.

Rao, S., A. K. Verma, and T. Bhatia. (2021). "A review on social spam detection: challenges, open issues, and future directions". *Expert Systems with Applications*. 186: 115742.

Raza, M. A. and K. M. Malik. (2023). "Multimodaltrace: Deepfake detection using audiovisual representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 993–1000.

Reimao, R. and V. Tzerpos. (2019). "For: A dataset for synthetic speech detection". In: *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE. 1–10.

Rekimoto, J. (2023). "WESPER: Zero-shot and realtime whisper to normal voice conversion for whisper-based speech interactions". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–12.

Ren, Y., W. Liu, D. Liu, and L. Wang. (2021). "Recalibrated bandpass filtering on temporal waveform for audio spoof detection". In: *2021 IEEE international conference on image processing (ICIP)*. IEEE. 3907–3911.

Ren, Y., C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. (2020). "Fastspeech 2: Fast and high-quality end-to-end text to speech". *arXiv preprint arXiv:2006.04558*.

Ren, Y., Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. (2019). "Fastspeech: Fast, robust and controllable text to speech". *Advances in neural information processing systems*. 32.

Rodríguez-Ortega, Y., D. M. Ballesteros, and D. Renza. (2020). "A machine learning model to detect fake voice". In: *International Conference on Applied Informatics*. Springer. 3–13.

Rupesh Kumar, S. and B. Bharathi. (2022). "Generative and discriminative modelling of linear energy sub-bands for spoof detection in speaker verification systems". *Circuits, Systems, and Signal Processing*. 41(7): 3811–3831.

Sachdeva, N. (2021). "Cyberbullying Detection on Social Media Using Deep Learning Models". *PhD thesis*. Delhi Technological University.

Saha, S., M. Sahidullah, and S. Das. (2024). "Exploring Green AI for Audio Deepfake Detection". *arXiv preprint arXiv:2403.14290*.

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

Sailor, H. B., D. M. Agrawal, and H. A. Patil. (2017). "Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification." In: *InterSpeech*. Vol. 8. 9.

Salvi, D., B. Hosler, P. Bestagini, M. C. Stamm, and S. Tubaro. (2023a). "TIMIT-TTS: a Text-to-Speech Dataset for Multimodal Synthetic Media Detection". *IEEE Access*.

Salvi, D., H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, and S. Tubaro. (2023b). "A robust approach to multimodal deepfake detection". *Journal of Imaging*. 9(6): 122.

Sanderson, C. and B. C. Lovell. (2009). "Multi-region probabilistic histograms for robust and scalable identity inference". In: *Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3*. Springer. 199–208.

Saxena, P., V. Saxena, A. Pandey, U. Flato, and K. Shukla. (2023). *Multiple Aspects of Artificial Intelligence*. Book Saga Publications.

Scott, D. W. (2001). "Parametric statistical modeling by minimum integrated square error". *Technometrics*. 43(3): 274–285.

Shan, M. and T. Tsai. (2020). "A Cross-Verification Approach for Protecting World Leaders from Fake and Tampered Audio". *arXiv preprint arXiv:2010.12173*.

Sharma, G., K. Umapathy, and S. Krishnan. (2020). "Trends in audio signal feature extraction methods". *Applied Acoustics*. 158: 107020.

Shen, J., Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu. (2020). "Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling". *arXiv preprint arXiv:2010.04301*.

Shen, J., R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, *et al.* (2018). "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 4779–4783.

Shi, Y., H. Bu, X. Xu, S. Zhang, and M. Li. (2020). "Aishell-3: A multi-speaker mandarin tts corpus and the baselines". *arXiv preprint arXiv:2010.11567*.

Shofner, W. P. and G. Selas. (2002). "Pitch strength and Stevens's power law". *Perception & psychophysics*. 64(3): 437–450.

Silva, R. B. da. (2021). "Updating the authentication of digital evidence in the international criminal court". *International Criminal Law Review*. 22(5-6): 941–964.

Singh, A. K. and P. Singh. (2021). "Detection of ai-synthesized speech using cepstral & bispectral statistics". In: *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 412–417.

Sinha, S., S. Dey, and G. Saha. (2024). "Improving self-supervised learning model for audio spoofing detection with layer-conditioned embedding fusion". *Computer Speech & Language*. 86: 101599.

Sisman, B. and H. Li. (2018). "Wavelet Analysis of Speaker Dependent and Independent Prosody for Voice Conversion." In: *Interspeech*. 52–56.

Sisman, B., J. Yamagishi, S. King, and H. Li. (2020). "An overview of voice conversion and its challenges: From statistical modeling to deep learning". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 29: 132–157.

Šmídl, L., J. Švec, D. Tihelka, J. Matoušek, J. Romportl, and P. Ircing. (2019). "Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development". *Language Resources and Evaluation*. 53: 449–464.

Solak, I. (2019). "The M-AILABS speech dataset".

Somoray, K. and D. J. Miller. (2023). "Providing detection strategies to improve human detection of deepfakes: An experimental study". *Computers in Human Behavior*. 149: 107917.

Song, K., Y. Zhang, Y. Lei, J. Cong, H. Li, L. Xie, G. He, and J. Bai. (2023). "Dspgan: a gan-based universal vocoder for high-fidelity tts by time-frequency domain supervision from dsp". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 1–5.

Sotelo, J., S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio. (2017). *Char2wav: End-to-end speech synthesis*.

Spangler, T. (2020). "Snap Confirms Acquisition of Deepfakes Startup AI Factory, Used to Power'Cameos' Selfie Videos".

Stupp, C. (2019). "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case". *The Wall Street Journal*. 30(08).

Subramani, N. and D. Rao. (2020). "Learning efficient representations for fake speech detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 04. 5859–5866.

Suratkar, S., E. Johnson, K. Variyambat, M. Panchal, and F. Kazi. (2020). "Employing transfer-learning based CNN architectures to enhance the generalizability of deepfake detection". In: *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE. 1–9.

Swetha, P. and T. Swami. (2021). "AI Based Assistance for Visually Impaired People Using TTS (Text To Speech)". *International Journal of Innovative Research in Science and Technology*. 1(1): 8–14.

Tak, H., M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans. (2022). "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation". *arXiv preprint arXiv:2202.12233*.

Tanaka, K., H. Kameoka, T. Kaneko, and N. Hojo. (2019). "AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6805–6809.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

Team, S. (2017). "Deep Learning for Siri's Voice: On-device Deep Mixture Density Networks for Hybrid Unit Selection Synthesis. Retrieved May 10, 2020".

TechCrunch. (2016). *AdobesProjectVoCo*. URL: https://techcrunch.com/2016/11/03/adobes-project-voco-lets-you-edit-speech-as-easily-as-text/.

Tian, X., S. Du, X. Xiao, H. Xu, E. S. Chng, and H. Li. (2015). "Detecting synthetic speech using long term magnitude and phase information". In: *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*. IEEE. 611–615.

Tian, X., Z. Wu, X. Xiao, E. S. Chng, and H. Li. (2016). "Spoofing detection from a feature representation perspective". In: *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2119–2123.

Tirumala, S. S., S. R. Shahamiri, A. S. Garhwal, and R. Wang. (2017). "Speaker identification features extraction methods: A systematic".

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

Todisco, M., H. Delgado, and N. W. Evans. (2016). "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients." In: *Odyssey*. Vol. 2016. 283–290.

Ustubioglu, A., B. Ustubioglu, and G. Ulutas. (2023). "Mel spectrogram-based audio forgery detection using CNN". *Signal, Image and Video Processing*. 17(5): 2211–2219.

Valle, R., J. Li, R. Prenger, and B. Catanzaro. (2020). "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6189–6193.

Vaswani, A., S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, Ł. Kaiser, N. Kalchbrenner, N. Parmar, *et al.* (2018). "Tensor2tensor for neural machine translation". *arXiv preprint arXiv:1803.07416*.

Verdoliva, L. (2020). "Media forensics and deepfakes: an overview". *IEEE Journal of Selected Topics in Signal Processing*. 14(5): 910–932.

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

Walczyna, T. and Z. Piotrowski. (2023). "Overview of Voice Conversion Methods Based on Deep Learning". *Applied Sciences*. 13(5): 3100.

Wang, C., J. He, J. Yi, J. Tao, C. Y. Zhang, and X. Zhang. (2024). "Multi-Scale Permutation Entropy for Audio Deepfake Detection". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 1406–1410.

Wang, C., J. Yi, J. Tao, C. Zhang, S. Zhang, and X. Chen. (2023a). "Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features". *arXiv preprint arXiv:2305.13700*.

Wang, C., S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, *et al.* (2023b). "Neural codec language models are zero-shot text to speech synthesizers". *arXiv preprint arXiv:2301.02111*.

Wang, R., F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, and Y. Liu. (2020a). "Deepsonar: Towards effective and robust detection of ai-synthesized fake voices". In: *Proceedings of the 28th ACM international conference on multimedia*. 1207–1216.

Wang, X. and J. Yamagishi. (2021). "Investigating self-supervised front ends for speech spoofing countermeasures". *arXiv preprint arXiv:2111.07725*.

Wang, X., J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, *et al.* (2020b). "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech". *Computer Speech & Language*. 64: 101114.

Wang, X., B. Zeng, H. Suo, Y. Wan, and M. Li. (2023c). "Robust audio anti-spoofing countermeasure with joint training of front-end and back-end models". In: *Proc. INTERSPEECH*. Vol. 2023. 4004–4008.

Wang, Y., R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, *et al.* (2017). "Tacotron: Towards end-to-end speech synthesis". *arXiv preprint arXiv:1703.10135*.

Wang, Z., S. Cui, X. Kang, W. Sun, and Z. Li. (2020c). "Densely connected convolutional network for audio spoofing detection". In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 1352–1360.

Wani, T. M. and I. Amerini. (2023). "Deepfakes Audio Detection Leveraging Audio Spectrogram and Convolutional Neural Networks". In: *International Conference on Image Analysis and Processing*. Springer. 156–167.

Wani, T. M., R. Gulzar, and I. Amerini. (2024a). "ABC-CapsNet: Attention based Cascaded Capsule Network for Audio Deepfake Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2464–2472.

Wani, T. M., S. A. A. Qadri, D. Comminiello, and I. Amerini. (2024b). "Detecting Audio Deepfakes: Integrating CNN and BiLSTM with Multi-Feature Concatenation". In: *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*. 271–276.

Wazid, M., A. K. Mishra, N. Mohd, and A. K. Das. (2024). "A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges, and Societal Impact". *Cyber Security and Applications*. 2: 100040.

Wijethunga, R., D. Matheesha, A. Al Noman, K. De Silva, M. Tissera, and L. Rupasinghe. (2020). "Deepfake audio detection: a deep learning based solution for group conversations". In: *2020 2nd International Conference on Advancements in Computing (ICAC)*. Vol. 1. IEEE. 192–197.

Wong, R. Y., A. Chong, and R. C. Aspegren. (2023). "Privacy Legislation as Business Risks: How GDPR and CCPA are Represented in Technology Companies' Investment Risk Disclosures". *Proceedings of the ACM on Human-Computer Interaction*. 7(CSCW1): 1–26.

Wu, Y., X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin, and T.-Y. Liu. (2022). "Adaspeech 4: Adaptive text to speech in zero-shot scenarios". *arXiv preprint arXiv:2204.00436*.

Wu, Z., R. K. Das, J. Yang, and H. Li. (2020). "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks". *arXiv preprint arXiv:2009.09637*.

Wu, Z., P. L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z.-H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, *et al.* (2016a). "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., N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li. (2015a). "Spoofing and countermeasures for speaker verification: A survey". *speech communication*. 66: 130–153.

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

Wu, Z., O. Watts, and S. King. (2016b). "Merlin: An Open Source Neural Network Speech Synthesis System." In: *SSW*. 202–207.

Wu, Z., X. Xiao, E. S. Chng, and H. Li. (2013). "Synthetic speech detection using temporal modulation feature". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 7234–7238.

Xiao, S., Z. Zhang, J. Yang, J. Wen, and Y. Li. (2023). "Manipulation detection of key populations under information measurement". *Information Sciences*. 634: 1–13.

Xiao, X., X. Tian, S. Du, H. Xu, E. Chng, and H. Li. (2015). "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge." In: *Interspeech*. 2052–2056.

Xie, Y., H. Cheng, Y. Wang, and L. Ye. (2023a). "Domain Generalization Via Aggregation and Separation for Audio Deepfake Detection". *IEEE Transactions on Information Forensics and Security*.

Xie, Y., H. Cheng, Y. Wang, and L. Ye. (2023b). "Learning A Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection". In: *Proc. INTERSPEECH*. Vol. 2023. 2808–2812.

Xu, Y., P. Terhöst, M. Pedersen, and K. Raja. (2024). "Analyzing Fairness in Deepfake Detection With Massively Annotated Databases". *IEEE Transactions on Technology and Society*.

Xue, J., C. Fan, Z. Lv, J. Tao, J. Yi, C. Zheng, Z. Wen, M. Yuan, and S. Shao. (2022). "Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features". In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 19–26.

Xue, R., Y. Liu, L. He, X. Tan, L. Liu, E. Lin, and S. Zhao. (2023). "Foundationtts: Text-to-speech for asr customization with generative language model". *arXiv preprint arXiv:2303.02939*.

Yamagishi, J., X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, *et al.* (2021). "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection". *arXiv preprint arXiv:2109.00537*.

Yamamoto, R., E. Song, and J.-M. Kim. (2020). "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6199–6203.

Yan, R., C. Wen, S. Zhou, T. Guo, W. Zou, and X. Li. (2022). "Audio deepfake detection system with neural stitching for add 2022". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 9226–9230.

Yan, Y. (2023). "Deep Dive into Deepfakes—Safeguarding Our Digital Identity". *Brooklyn Journal of International Law*. 48(2): 767.

Yan, Y., X. Tan, B. Li, T. Qin, S. Zhao, Y. Shen, and T.-Y. Liu. (2021a). "Adaspeech 2: Adaptive text to speech with untranscribed data". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6613–6617.

Yan, Y., X. Tan, B. Li, G. Zhang, T. Qin, S. Zhao, Y. Shen, W.-Q. Zhang, and T.-Y. Liu. (2021b). "Adaspeech 3: Adaptive text to speech for spontaneous style". *arXiv preprint arXiv:2107.02530*.

Yang, J.-C., C. You, and Q. He. (2018). "Feature with Complementarity of Statistics and Principal Information for Spoofing Detection." In: *INTERSPEECH*. 651–655.

Yang, G., S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie. (2021). "Multi-band melgan: Faster waveform generation for high-quality text-to-speech". In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 492–498.

Yang, J. and R. K. Das. (2020). "Long-term high frequency features for synthetic speech detection". *Digital Signal Processing*. 97: 102622.

Yang, J., R. K. Das, and N. Zhou. (2019). "Extraction of octave spectra information for spoofing attack detection". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 27(12): 2373–2384.

Yang, J., J. Lee, Y. Kim, H. Cho, and I. Kim. (2020). "VocGAN: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network". *arXiv preprint arXiv:2007.15256*.

Yang, Y., Y. Kartynnik, Y. Li, J. Tang, X. Li, G. Sung, and M. Grundmann. (2024). "StreamVC: Real-Time Low-Latency Voice Conversion". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 11016–11020.

Yi, J., Y. Bai, J. Tao, Z. Tian, C. Wang, T. Wang, and R. Fu. (2021). "Half-truth: A partially fake audio detection dataset". *arXiv preprint arXiv:2104.03617*.

Yi, J., R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, *et al.* (2022). "Add 2022: the first audio deep synthesis detection challenge". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 9216–9220.

Yi, J., J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, *et al.* (2023). "ADD 2023: the Second Audio Deepfake Detection Challenge". *arXiv preprint arXiv:2305.13774*.

Yi, P. and A. Zubiaga. (2023). "Session-based cyberbullying detection in social media: A survey". *Online Social Networks and Media*. 36: 100250.

Yu, H., Z.-H. Tan, Z. Ma, R. Martin, and J. Guo. (2017a). "Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features". *IEEE transactions on neural networks and learning systems*. 29(10): 4633–4644.

Yu, H., Z.-H. Tan, Y. Zhang, Z. Ma, and J. Guo. (2017b). "DNN filter bank cepstral coefficients for spoofing detection". *Ieee Access*. 5: 4779–4787.

Yu, Y., X. Liu, R. Ni, S. Yang, Y. Zhao, and A. C. Kot. (2023). "Pvass-mdd: predictive visual-audio alignment self-supervision for multi-modal deepfake detection". *IEEE Transactions on Circuits and Systems for Video Technology*.

Yu, Z., Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao. (2022). "Deep learning for face anti-spoofing: A survey". *IEEE transactions on pattern analysis and machine intelligence*. 45(5): 5609–5631.

Yurtseven, İ., S. Bagriyanik, and S. Ayvaz. (2021). "A review of spam detection in social media". In: *2021 6th International Conference on Computer Science and Engineering (UBMK)*. IEEE. 383–388.

Zhang, J., G. Tu, S. Liu, and Z. Cai. (2023a). "Audio Anti-Spoofing Based on Audio Feature Fusion". *Algorithms*. 16(7): 317.

Zhang, M., Y. Zhou, L. Zhao, and H. Li. (2021a). "Transfer learning from speech synthesis to voice conversion with non-parallel training data". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 29: 1290–1302.

Zhang, X., J. Yi, J. Tao, C. Wang, and C. Y. Zhang. (2023b). "Do you remember? Overcoming catastrophic forgetting for fake audio detection". In: *International Conference on Machine Learning*. PMLR. 41819–41831.

Zhang, X., J. Yi, C. Wang, C. Y. Zhang, S. Zeng, and J. Tao. (2024a). "What to remember: Self-adaptive continual learning for audio deepfake detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 17. 19569–19577.

Zhang, Y., W. Lin, and J. Xu. (2024b). "Joint Audio-Visual Attention with Contrastive Learning for More General Deepfake Detection". *ACM Transactions on Multimedia Computing, Communications and Applications*. 20(5): 1–23.

Zhang, Y., J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi. (2022). "Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 7237–7241.

Zhang, Z., X. Yi, and X. Zhao. (2021b). "Fake speech detection using residual network with transformer encoder". In: *Proceedings of the 2021 ACM workshop on information hiding and multimedia security*. 13–22.

Zhang, Z., B. He, and Z. Zhang. (2020). "Gazev: Gan-based zero-shot voice conversion over non-parallel speech corpus". *arXiv preprint arXiv:2010.12788*.

Zhao, W., W. Wang, J. Chai, and J. Huang. (2021). "IVCGAN: An Improved GAN for Voice Conversion". In: *2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. Vol. 5. IEEE. 1035–1039.

Zhou, Y. and S.-N. Lim. (2021). "Joint audio-visual deepfake detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14800–14809.