EVALUATION OF AUDIO DEEPFAKES — SYSTEMATIC REVIEW

Yamini Sinha[1]*, Jan Hintz*[1]*, Ingo Siegert*[1]

[1]*Mobile Dialog Systems, IIKT,*
Otto von Guericke University Magdeburg, Germany
(firstname.lastname)@ovgu.de

Abstract: Generative models for audio are commonly used for music composition, sound effects generation for video game development, audio restoration, voice cloning, etc. The ease of generating indistinguishable fake audio with deep learning poses a major threat to personal privacy, online security, and political discourse. Evaluating the quality and realism of these synthetic utterances is crucial for mitigating the potential for misinformation and harm. To assess this threat, this paper conducts a systematic review, using Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), on how these deepfake models are currently evaluated. The analysis of 86 papers shows that the majority of the evaluation is conducted on a machine level and highlights a research gap regarding the human perception of deepfakes. This paper explores various methods and perceptual measures employed in assessing audio deepfakes and evaluating their strengths, limitations, and future directions.

1 Introduction

Audio deepfakes have been around for a few years now, with the technology advancing and bringing forward evermore realistic voices [1]. As audio deepfake technology becomes more advanced, it's creating both exciting opportunities and significant challenges. Audio deepfakes use artificial intelligence to make fake audio clips that sound very similar to real people's voices, including their tone and emotions. While this can be great for creative projects and entertainment, it also raises big concerns. People could use this technology to create fake audio for harmful purposes, like fake news, scams, and other criminal activities that pose a threat to individuals [2], companies [3] and democracy [4]. Therefore, the critical challenge is to develop and refine evaluation techniques that can accurately assess the authenticity of audio clips, alongside perceptual measures that can tap into the nuanced understanding of human listeners to identify irregularities in deepfake recordings.

The dangers of this technology can only be assessed if they are evaluated accordingly. Evaluation techniques for audio deepfakes primarily encompass a variety of computational methods and algorithms designed to dissect and analyze the spectral, temporal, and feature-based characteristics of audio recordings. These techniques aim to identify anomalies or artifacts that are indicative of manipulation, which might not be perceptible to the human ear. However, the sophistication of deepfake generation methods continuously evolves, presenting a moving target for detection algorithms. This cat-and-mouse dynamic underscores the importance of continuous research and adaptation in evaluation methodologies to keep pace with the advancing technology.

DOI: 10.35096/othr/pub-7096, URN: urn:nbn:de:bvb:898-opus4-70960

2 Related Work

Some systematic reviews have been conducted on audio deepfakes [1, 5, 6] over the past years, but they focus rather on methods for detection and generation of deepfakes than on the metrics and evaluation used. The meta-review conducted by Barnett [7] is based on the PRISMA [8] method and is focussed on the ethical implications of generative audio models, concluding that while the positive impact is highlighted by 65% of the evaluated papers only 10% consider the negative implications.

Our meta-review aims to analyze the aspect of how deep-fake models are evaluated, as this is necessary to better assess the positive or negative impact, of these applications.

3 PRISMA Study

To gain a clear and comprehensive picture, we decided to dive deep into the existing research. We followed the established guidelines of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [8], ensuring transparency and comprehensibility throughout our review process, as seen in the Figure 1. This framework ensures a rigorous, unbiased, and reproducible review, making it easier to identify areas where further research is needed and to synthesize the existing evidence on the effectiveness and challenges posed by audio deepfakes. By taking this structured approach, we can gather studies that offer a well-rounded evaluation of audio deepfakes, ultimately leading to a richer understanding of their capabilities and impact on cybersecurity, media integrity, and digital forensics.

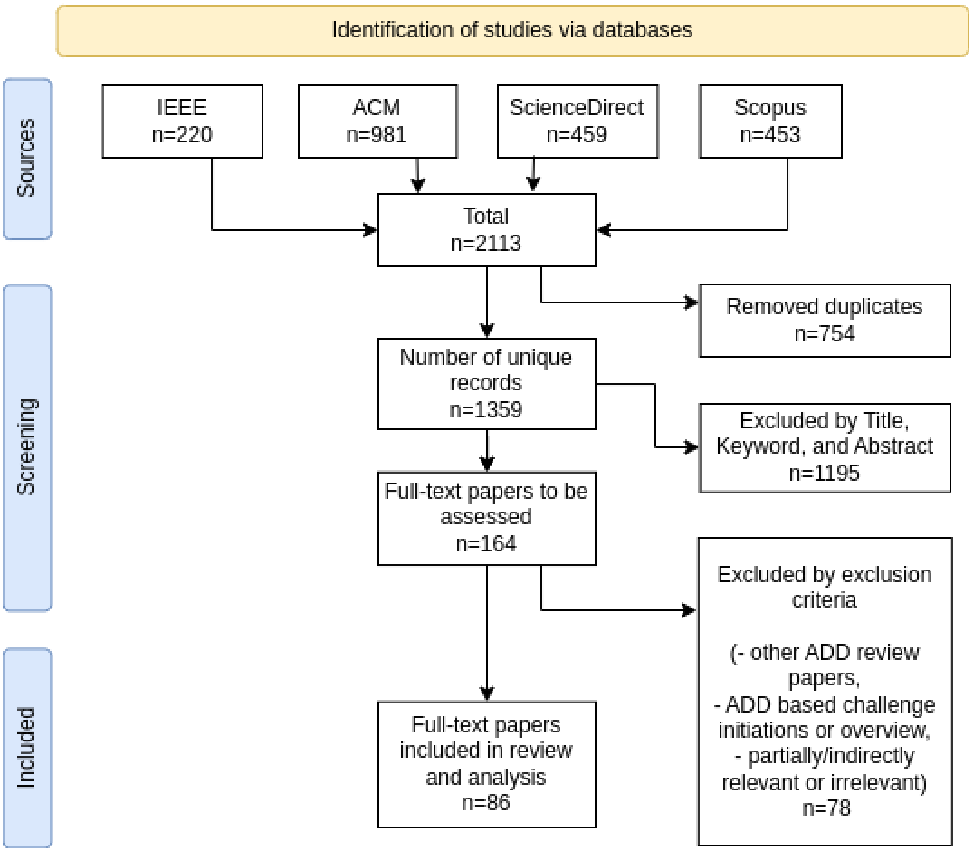
Identification of studies via databases
Sources
Screening
Included
IEEE n=220
ACM n=981
ScienceDirect n=459
Scopus n=453
Total n=2113
Removed duplicates n=754
Number of unique records n=1359
Excluded by Title, Keyword, and Abstract n=1195
Full-text papers to be assessed n=164
Excluded by exclusion criteria (- other ADD review papers, - ADD based challenge initiations or overview, - partially/indirectly relevant or irrelevant) n=78
Full-text papers included in review and analysis n=86

Figure 1 – PRISMA flow diagram specifying databases screened and reviewed in this study

3.1 Information Sources

To ensure a comprehensive review of the existing literature, we conducted a systematic search across multiple peer-reviewed academic databases, including IEEE Xplore, Scopus, ScienceDirect, and ACM Digital Library. We employed the search terms "deepfake AND (audio OR speech)" and yielded an initial pool of 2,113 papers. Following standard protocol, we removed duplicate entries, ultimately resulting in 1,359 unique publications for further screening. To refine this selection, we meticulously reviewed titles, keywords, and abstracts, adhering to established exclusion criteria, in section 3.2.

3.2 Exclusion Criteria

In this study, we analyzed research papers within the field of audio deepfake detection. We intentionally omitted other research works, such as extended abstracts and book chapters. The primary criterion for selecting papers was that their main focus must be firmly rooted in the evaluation of audio deepfake detection models. In essence, the papers needed to either propose methods for detecting audio deepfakes or offer perceptual evaluations to facilitate a critical discussion on audio deepfake detection models.

Our research approach involved a deliberate exclusion of certain types of studies to sharpen our focus. First, we disregarded studies that delved into the societal impact and implications of deepfakes. This strategic omission allowed us to direct our attention towards the technical aspects of deepfake technology rather than its broader societal ramifications. Additionally, we chose not to include studies centered on watermarking as a countermeasure to deepfakes. While we recognize the potential of watermarking in increasing the detectability of deepfakes by automated systems, this technique's applicability is limited to the generation phase of the spoofed audio.

Finally, we excluded studies focused on multi-modal aspects, especially those involving the synchronization of audio and video elements, like lip-syncing. This was a strategic decision to maintain a clear focus on deepfake technologies within a single modality, thereby aligning more precisely with the objectives of our study. This rigorous process ultimately yielded a focused and relevant collection of literature for further screening, resulting in 164 papers to be reviewed further.

3.3 Full Text Screening

A deeper screening was conducted on full-text of the reduced 164 papers. 78 papers out of 164 were excluded based on the following criteria: a) review papers regarding Audio Deepfake Detection (ADD) techniques, b) ADD challenge invitation or findings-analysis paper, and c) partially or indirectly related to ADD such as Adversarial attacks. After full-text screening, 86 total papers were selected for further analysis, also shown in Figure 1. Besides the direct inclusion, keeping ADD in mind, some of the other major inclusion criteria considered here are a) papers that present ADD on audio samples from languages other than English, such as Arabic in [9], b) ADD on speech disorders in [10], c) papers that present model for ADD that generalize over multiple domains and languages, d) perform ADD by finding the algorithm used to create the deepfake. Of the 86 articles, 15 are from ACM digital library, 33 from IEEE Xplore, 34 from SCOPUS, and only 4 from ScienceDirect.

In the remainder of the paper, we will concentrate on reporting the highlights and insights generated by our research. The full list of 86 identified papers can be found here [11][1].

[1] https://github.com/Mobile-Dialog-Systeme/EVALUATION-OF-AUDIO-DEEPFAKES-.git

4 Analysis & Results

The initial review narrowed down relevant articles to 164 after examining titles, keywords, and abstracts. Further in-depth evaluation of the full texts refined this selection to 86 papers.

Datasets: Regarding used data, the studies identified cover a broad range of datasets and mostly use a combination of different ones. Mostly corpora comprising read speech data are used, e.g. LibriSpeech[2], TIMIT[3], LJ Speech[4], VCTK[5], Speech Accent Archive[6], Mozilla Common Voice[7], JSUT[8]. Furthermore, several studies were published as part of a challenge and therefore, also used the specified data sets, e.g. data from the ASVspoof challenge [9], Audio Deep Synthesis Detection Challenge (ADD2022) [12] and DECRO[10]. Some papers also make use of specifically designed datasets comprising (fake) synthetic and real voices, e.g., FoR dataset [13] and Wave Fake[11]. Only a few papers use own data or specifically concentrate on in-the-wild data. Details can be found in the full list of identified studies [11].

Additional Detection Methods: Various studies are refining their algorithms through the integration of perceptual losses. Equally noteworthy is the ongoing research targeting enhancements that lead to the creation of more realistic and convincing deepfakes. [14] uses emotion recognition to capture semantic audio information and discriminate fakes from real speech. Another study that highlights the distinctive strengths of leveraging both physical and perceptual features for audio deepfake detection is [15], further bolstering the emphasis on human factors highlighted in earlier research. Physical features delve into the technical aspects of the audio, while perceptual features explore how humans interpret sound. Xue et al. [16] uses acoustic features like F0 to improve detection. Gao et al. [17] and Lue et al. [18] use shimmer and jitter features in their countermeasure model, because of their close representation of human perception. Chaiwongyen et. al [19] refined the significance of certain timbre and shimmer components in the contribution to recognizing deepfakes. The comprehensive analysis demonstrates that incorporating perceptual cues, which mirror human auditory perception, significantly enhances the ability of detection models to differentiate real from fake audio. This finding underscores the critical importance of aligning technological solutions with human perceptual capabilities to achieve more effective deepfake detection.

The study in [20] presents an approach to evaluating AI-generated audio using image-schemas. Applying visual metaphors to the auditory domain, provides a unique lens for assessing audio quality, particularly in distinguishing between real and synthetic sounds. This research enriches the toolkit for developers and researchers working on deepfake detection, offering new methodologies that could lead to more nuanced and effective detection systems.

Human Perception: Notably, only a minority of the studies delve into human perception through experimental research, highlighting an area less explored within the field [21, 22, 23, 19]. One of the studies that presents work related to human perception for audio deepfake detection

[2] https://www.openslr.org/12

[3] https://catalog.ldc.upenn.edu/LDC93s1

[4] https://keithito.com/LJ-Speech-Dataset/

[5] https://paperswithcode.com/dataset/vctk

[6] https://accent.gmu.edu/

[7] https://commonvoice.mozilla.org

[8] https://sites.google.com/site/shinnosuketakamichi/publication/jsut

[9] https://www.asvspoof.org/

[10] https://github.com/petrichorwq/DECRO-dataset

[11] https://zenodo.org/records/5642694

is [21]. This research delves into the comparative ability of humans and AI algorithms to identify audio deepfakes. Employing a gamified approach, a web-based platform was used where 410 participants played over 13,000 rounds, distinguishing between real and synthetic audio samples. The analysis revealed intriguing similarities and differences in how humans and AI tackle specific types of deepfakes. Notably, native speakers and younger participants displayed superior detection skills, while expertise in IT did not show a significant impact on performance. These findings underscore the crucial role of incorporating human factors into the design of cybersecurity training programs and the advancement of detection algorithms.

Performance Measures: The majority of the analyzed studies primarily focus on metrics like Equal Error Rate, accuracy, precision, and recall for evaluating deepfake models. Very few studies also used neural network-based Mean Opinion Score (MOS), as seen in [24].

Real-world Applications: A few papers discuss specific (real-world) attack scenarios. Firc et al. [25] highlight the accessibility of software that can create deepfakes that can fool biometric systems. Interestingly, their claim regarding the robustness of voice verification systems picks up the old discussion of text-dependent vs text-independent verification again [26]. The authors of [23] facilitate a kind of challenge, to test if the caller is real or fake. The authors argue that if a deepfake attempts the task, the ensuing content (the response) will undergo significant distortion, making it easy for humans to detect the fake.

5 Conclusion

In conclusion, our primary aim was to gain a comprehensive overview of the current state of deepfake studies, with a specific focus on evaluation methods. The substantial reduction from an initial pool of 1359 papers to a final selection of 86 underscores the significance of this evolving field. However, it is noteworthy that human perception remains underrepresented in the current body of research, signaling an area that warrants further exploration and attention in future deepfake studies. The initial findings of our meta-analysis show a disparity between the number of research studies conducting qualitative and quantitative assessments. Regarding the potential impact of algorithms on our society, it is necessary to conduct more experiments on human perception.

Acknowledgements

This research has been funded by the Volkswagen Foundation in the project AnonymPrevent (AI-based Improvement of Anonymity for Remote Assessment, Treatment, and Prevention against Child Sexual Abuse) and by the BMBF in the Project Medinym (KI-basierte Anonymisierung personenbezogener Patientendaten In Klinischen Text- und Sprachdatenbeständen).

References

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

[2] Maras, M.-H. and A. Alexandrou: *Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. The International Journal of Evidence & Proof*, 23(3), pp. 255–262, 2019.

[3] Stupp, C.: *Fraudsters used ai to mimic ceo's voice in unusual cybercrime case. The Wall Street Journal*, 30(08), 2019.

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

[5] ALMUTAIRI, Z. and H. ELGIBREEN: *A review of modern audio deepfake detection methods: challenges and future directions*. *Algorithms*, 15(5), p. 155, 2022.

[6] KHANJANI, Z., G. WATSON, and V. P. JANEJA: *Audio deepfakes: A survey*. *Frontiers in Big Data*, 5, p. 1001063, 2023.

[7] BARNETT, J.: *The ethical implications of generative audio models: A systematic literature review*. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 146–161. 2023.

[8] MOHER, D., A. LIBERATI, J. TETZLAFF, D. G. ALTMAN, P. GROUP ET AL.: *Preferred reporting items for systematic reviews and meta-analyses: the prisma statement*. *International journal of surgery*, 8(5), pp. 336–341, 2010.

[9] ALMUTAIRI, Z. and H. ELGIBREEN: *Detecting fake audio of arabic speakers using self-supervised deep learning*. *IEEE Access*, 2023.

[10] CHAIWONGYEN, A., S. DUANGPUMMET, J. KARNJANA, W. KONGPRAWECHNON, and M. UNOKI: *Deepfake-speech detection with pathological features and multilayer perceptron neural network*. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2182–2188. IEEE, 2023.

[11] SINHA, Y., J. HINTZ, and I. SIEGERT: *EVALUATION-OF-AUDIO-DEEPFAKES- Full List of Identified Studies*. 2024. URL `https://github.com/Mobile-Dialog-Systeme/EVALUATION-OF-AUDIO-DEEPFAKES-`.

[12] YI, J., R. FU, J. TAO, S. NIE, H. MA, C. WANG, T. WANG, Z. TIAN, Y. BAI, C. FAN, S. LIANG, S. WANG, S. ZHANG, X. YAN, L. XU, Z. WEN, H. LI, Z. LIAN, and B. LIU: *Add 2022: the first audio deep synthesis detection challenge*. 2022. `2202.08433`.

[13] REIMAO, R. and V. TZERPOS: *For: A dataset for synthetic speech detection*. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–10. 2019. doi:10.1109/SPED.2019.8906599.

[14] CONTI, E., D. SALVI, C. BORRELLI, B. HOSLER, P. BESTAGINI, F. ANTONACCI, A. SARTI, M. C. STAMM, and S. TUBARO: *Deepfake speech detection through emotion recognition: a semantic approach*. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8962–8966. IEEE, 2022.

[15] LI, M., Y. AHMADIADLI, and X.-P. ZHANG: *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*, DDAM '22, p. 35–41. Association for Computing Machinery, New York, NY, USA, 2022. doi:10.1145/3552466.3556523. URL `https://doi.org/10.1145/3552466.3556523`.

[16] XUE, J., C. FAN, Z. LV, J. TAO, J. YI, C. ZHENG, Z. WEN, M. YUAN, and S. SHAO: *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*, pp. 19–26. 2022.

[17] GAO, Y., J. LIAN, B. RAJ, and R. SINGH: *Detection and evaluation of human and machine generated speech in spoofing attacks on automatic speaker verification systems*. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 544–551. IEEE, 2021.

[18] Lu, J., Z. Li, Y. Zhang, W. Wang, and P. Zhang: *Acoustic or pattern? speech spoofing countermeasure based on image pre-training models.* In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, pp. 77–84. 2022.

[19] Chaiwongyen, A., N. Songsriboonsit, S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki: *Contribution of timbre and shimmer features to deepfake speech detection.* In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 97–103. IEEE, 2022.

[20] Kamath, P., Z. Li, C. Gupta, K. Jaidka, S. Nanayakkara, and L. Wyse: *Evaluating descriptive quality of ai-generated audio using image-schemas.* In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 621–632. 2023.

[21] Müller, N. M., K. Pizzi, and J. Williams: *Human perception of audio deepfakes.* In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, pp. 85–91. 2022.

[22] Mai, K. T., S. Bray, T. Davies, and L. D. Griffin: *Warning: Humans cannot reliably detect speech deepfakes. PLoS One*, 18(8), p. e0285333, 2023.

[23] Yasur, L., G. Frankovits, F. M. Grabovski, and Y. Mirsky: *Deepfake captcha: A method for preventing fake calls.* In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, ASIA CCS '23, p. 608–622. Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3579856.3595801. URL https://doi.org/10.1145/3579856.3595801.

[24] Yu, Z., S. Zhai, and N. Zhang: *Antifake: Using adversarial audio to prevent unauthorized speech synthesis.* In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 460–474. 2023.

[25] Firc, A. and K. Malinka: *The dawn of a text-dependent society: deepfakes as a threat to speech verification systems.* In *Proceedings of the 37th ACM/SIGAPP symposium on applied computing*, pp. 1646–1655. 2022.

[26] Tu, Y., W. Lin, and M.-W. Mak: *A survey on text-dependent and text-independent speaker verification. IEEE Access*, 10, pp. 99038–99049, 2022. doi:10.1109/ACCESS.2022.3206541.