Interspeech 2024
1-5 September 2024, Kos, Greece

A New Approach to Voice Authenticity

Nicolas M. Müller[1], *Piotr Kawa*[2], *Shen Hu*[3], *Matthias Neu*[4], *Jennifer Williams*[5], *Philip Sperl*[1], *Konstantin Böttinger*[1]

[1]Fraunhofer AISEC [2]Wrocław University of Science and Technology [3]Technical University of Munich [4]Bundesamt für Sicherheit in der Informationstechnik [5]University of Southampton

nicolas.mueller@aisec.fraunhofer.de

Abstract

Voice faking poses significant societal challenges. Currently, the prevailing assumption is that unaltered human speech can always be considered genuine, while fake speech usually comes from text-to-speech (TTS) synthesis. We argue that this type of binary distinction is oversimplified. For instance, altered playback speeds can maliciously deceive listeners, as in the 'Drunken Nancy Pelosi' incident. Similarly, editing of audio clips can be done ethically, e.g. for brevity or summarization in news reporting or podcasts, but editing can also create misleading narratives. In this paper, we propose a conceptual shift away from the longstanding binary paradigm of speech audio being either 'fake' or 'real'. Instead, we focus on pinpointing 'voice edits', which encompass traditional modifications like filters and cuts, as well as neural synthesis. We delineate six categories of voice edits and curate a new challenge dataset, for which we present baseline voice edit detection systems.
Index Terms: voice edits,voice anti-spoofing, deepfake

1. Introduction

The rapidly advancing field of machine learning has significantly enhanced the quality and computability of text-to-speech (TTS) synthesis technology, opening the door to a myriad of beneficial applications ranging from assistive technology for the hearing impaired to spoken audio books for children [1, 2, 3, 4]. However, this same forward progress also brings forth serious threats to our understanding and perception of authentic speech, including the creation of deepfakes, which facilitate misinformation, fake news, slander, fraud, AI-mediated pornography, and deceptive calls [5, 6, 7]. Speech synthesis technology becomes particularly perilous when used to trick automatic speaker verification and biometric identification systems, a practice known as 'spoofing' [8].

The anti-spoofing community has responded by establishing challenge datasets and spoofing detection algorithm benchmarks such as ASVspoof 2019 [9] and 2021 [10], pouring extensive efforts into differentiating between genuine ('bonafide') and counterfeit ('spoof') speech samples. While these datasets, algorithms, and evaluation metrics are adequate within the scope of anti-spoofing protection for speaker verification, they fall short in addressing broader societal challenges. Particularly, the binary idea that all TTS/VC-generated synthetic speech is inherently deceptive (and fake), while all other content is benign (and genuine), is an oversimplification. Despite this false equivalence, it is a viewpoint widely held both inside and outside of the speech technology community [9, 10, 11, 12].

The reality of assessing fake and genuine audio is complex and context-dependent. For instance, altering playback speeds - a technique maliciously used in scenarios like the 'Drunken Nancy Pelosi' [13] incident - can also serve benign purposes, such as in language learning tools or assistance for the hearing-impaired. Similarly, even the removal or reordering of words can fabricate misleading narratives [14] that use real human speech, yet these same editing techniques are valid in concise, legitimate formats such as news broadcasts or podcasts. Audio can also be edited through equalization to subtly alter the quality of a politician's voice by adjusting bass and treble levels, as well as overall pitch. Equalization adjustments can affect how sincere or confident a politician sounds. Even unintentional manipulation can degrade speech sound quality, thereby negatively influencing public perception [15], which has already been observed in political campaigns [16].

Not all synthetic speech warrants skepticism, and arguably if a human is controlling synthetic speech for genuine communication purposes, perhaps it cannot be defined as wholly synthetic or fake. Consider the speech synthesis used by the late physicist Stephen Hawking - his synthetic voice is both memorable and recognisable. Synthetic speech from the 'Google Parrotron' [1] project can help improve perceived speaking intelligibility for people who have difficulty producing speech due to hearing loss. Effectively, such voice conversion techniques, while they are generative and produce synthetic speech, have a benevolent and practical place in society. This is despite the fact that they share similar underlying machine learning techniques, and differ mostly in application. Lastly, in the field of politics, TTS synthesis has been used both ethically and unethically in automated calls. On the one hand, it has been used to impersonate US President Biden and encourage people to skip the 2024 primary election in the state of New Hampshire [17]. On the other hand, the US Democratic Party has used it to connect with new potential voters [18]. This underscores the necessity for a nuanced understanding of 'synthetic' and 'authentic' in audio content—a delineation that is not merely black and white but incorporates the myriad shades of intention and context.

Contribution. Our paper introduces a structured approach and new way of thinking about challenges for handling voice-edited audio. We introduce a toolkit to generate data that reflects this paradigm shift. We also propose several baseline machine learning models capable of identifying and classifying audio modifications in the dataset, including the nature of the edit and its location in the time domain. Our evaluation confirms that models are effective at accurately detecting a variety of audio modifications and edits. Using this dataset as the basis, we call for a fundamental shift in voice authenticity research efforts, away from the simplistic real/fake classification and toward a more nuanced approach.

10.21437/Interspeech.2024-31

2. Related Work

The domain of audio modification detection encompasses various challenges, including the identification of double compression, codec recognition, and the detection of copy-move or splicing operations, which involve the deletion or insertion of audio segments [19, 20, 21]. Efforts in this area include the analysis of electric network frequency (ENF) signals [22], as well as investigations into microphone and acoustic characteristics [21]. Despite these endeavors, the literature remains sparse, with no existing work providing a detailed classification of vocal alterations or a current overview of neural network-based detection methodologies compared to what we present in our study. Further, the research community has not agreed on standard benchmarking metrics for comparing or evaluating performance of audio modification detection approaches.

Voice authenticity attacks are commonly categorized into three primary scenarios: physical access spoofing, logical access spoofing, and deepfakes. This taxonomy is reflected in the datasets developed by the research community. Physical access spoofing attacks target automatic speaker verification (ASV) systems with replayed speech, where original utterances are recorded and then played back under varied acoustic conditions [9, 10, 23]. Logical access spoofing attacks entail remote interference with ASV systems, such as bypassing biometric identification in a telephone call. Related attack techniques include the injection of spoofed audio, created using text-to-speech and voice conversion techniques, into the communication channel [10, 24, 25]. Deepfake technology aims to generate artificial speech that is purpose-built to deceive human listeners, focusing on disinformation spread via social media, using similar techniques as logical access spoofing but targeting humans instead of voice biometric systems [10, 12, 26, 27].

Anti-spoofing and deepfake detection algorithms primarily rely on deep neural networks, diverging into two main research branches based on the nature of the input data. The first branch utilizes audio waveforms as input, requiring no prior transformations [28, 29, 30], while the second focuses on transformed audio signals to highlight features indicative of spoofed material [31, 32]. Across both branches, self-supervised learning models and embeddings from other audio processing architectures have also been proposed [33, 34, 35]. Challenges in the detection of voice spoofs and deepfakes include ensuring the generalization of trained methods. This is currently addressed by incorporating a wide array of generation methods, codecs, and audio quality within the training datasets. Other focal points include partial spoofs, multi-modal deepfakes, synthesized singing voices detection, real-world deepfake utterances, and language diversity [11, 27, 36, 37, 38, 39, 40]. Despite the extensive taxonomy, existing classifications do not consider traditional attack vectors such as slicing and pitch alteration. There is no comprehensive database or benchmark addressing both neural and 'traditional' audio manipulations.

3. Voice Edit Categories

In Table 1, we identify 6 overarching categories of voice edits that contribute to voice authenticity. Each category is based on a different type of audio modification which can affect the perception of speech and voice as well as their perceived quality. From these high-level categories, we present 20 unique voice edits (encircled in numerals).

Table 1: *Voice Edit Categories*

Source Origin
Original ⓪: Unaltered human speech.
Synthesized ①: Speech generated using text-to-speech synthesis algorithms.
Converted ②: Speech generated by voice conversion.
Temporal Edits
Concatenation/Trimming ③: Inserting or removing portions of the audio.
Mixing ④: Merging two separate audio files together.
Modulation and Effects
Pitch Shifting ⑤, ⑥: Increase or decrease the pitch without altering the speed.
Time Stretching ⑦, ⑧: Slow down or speed up the pace without altering the pitch.
Encoding and Compression
Lossless Encodings: Formats like WAV, FLAC are considered the 'default' encoding.
Lossy Encodings ⑨, ⑩: Formats like MP3 or AAC.
Telephony Encodings ⑪, ⑫: Specific to voice communications, e.g., alaw, ulaw.
Frequency and Spectral Edits
Low/High Pass Filters ⑬, ⑭: Filters designed to allow only low or high frequency ranges to pass through.
Equalization ⑮: Adjusting the balance between frequency components.
Autotune ⑯: Correcting pitch in vocal performances.
Spatial and Environmental Edits
Room Impulse Response ⑰: The modification of audio to simulate different room acoustics.
Reverb ⑱: Adding reflections.
Overlay ⑲: Mix with background noise.
Noise Cancellation ⑳: Removing background noises.

4. Data

Human Speech Baseline: We introduce a framework designed for the efficient, real-time generation of the those 20 voice edits. To this end, we use the M-AILABS Speech Dataset [41] of original human speech. This dataset comprises eight languages: English, French, German, Italian, Polish, Russian, Spanish, and Ukrainian. It is based on LibriVox and Project Gutenberg, encompassing almost 1000 hours of audio across 493,900 baseline human utterances ⓪, and sampled at 16kHz.

TTS and VC Speech Creation: In order to generate corresponding text-to-speech data ①, we use the MLAAD dataset [40], which also uses the M-AILABS dataset as a reference point, and re-synthesizes the audio files using 52 different text-to-speech algorithms. Since MLAAD does not feature any voice-conversion ②, we create the corresponding data ourselves. To this end, we use FreeVC [42] and Phoneme Hallucinator [43] to create 1000 converted audio samples for each language of the M-AILABS dataset and each Voice Conversion method. The target speaker for each source speaker is a randomly chosen speaker of the same language.

On-the-Fly Modifications: The remaining modifications

spanning types ③ to ⑳ offer a computational efficiency significantly surpassing that of neural TTS systems. This enables their real-time execution during both training and testing phases; thus serving as on-the-fly data augmentations. These alterations are generally straightforward; for instance, the 'Concatenation/Trimming' operation ③ entails choosing two arbitrary points within audio files to concatenate together. Similarly, other modifications such as altering speech or pitch are readily achievable through efficient utilization of the *librosa*, *ffmpeg* or *sox* libraries. All modifications necessitate the specification of hyperparameters. For instance, adjustments like increasing pitch ⑤ or decreasing pitch ⑥ demand a specification of the semi-tone interval for the shift. We randomly determine this value within a pre-established spectrum, in this case, between 1 and 12 semi-tones. The modification is then applied and the file labelled accordingly. A similar approach applies to other voice modifications, where suitable hyperparameters are not fixed, but chosen from a range of suitable values. Thus, this process results in a dynamically created set of audio files.

5. Experiments

Given such a data framework comprising 20 types of voice edits, we evaluate the detection performance using several neural architectures from related work, which we describe in more detail in this section. We apply the following procedure: the M-AILABS is split into train and test utterances (90% and 10%, respectively), and each audio-file is transformed on-the-fly and used to train or evaluate the models. We use batch sizes of 16, the Adam optimizer with a learning rate of 10^{-3}, and train for 500 epochs. We employ early stopping when there is no increase in training accuracy of at least 0.5% within 5 epochs. We use four machine learning models common in the field of audio deepfake detection and anti-spoofing, discussed below.

5.1. Model Descriptions

Light Convolutional Neural Networks (LCNN) [31] employ a mel-spectrogram or CQT frontend, where only the magnitude of the time-frequency spectrogram is analyzed. With only $195,367$ trainable parameters, this approach can be fast and reliable, but incurs some information loss because the phase information is discarded.

ComplexNet [32] efficiently processes CQT-features [44] extracted directly from audio waveforms without discarding the phase data. This method ensures no loss of information during feature extraction. However, it requires the use of complex-valued neural layers, necessitating 2.6 million trainable parameters.

RawNet2 [28], with 17.6 million trainable parameters, is one of the most popular models in voice anti-spoofing, because it directly processes audio waveforms using sinc-layers, bypassing the need for a dedicated spectral frontend. Despite its notable success on ASVspoof, RawNet2 faces challenges in generalizing to out-of-domain data [11].

SSL W2V2 [34] capitalizes on unsupervised pre-training via wav-to-vec-2 representations [45], demonstrating significant potential in cross-domain anti-spoofing scenarios [40]. However, with 317 million parameters, its computational demands are high, and it requires at least 4.05 seconds of input audio.

5.2. Time Resolution vs. Model Accuracy

It is crucial to not only determine the type of manipulation an audio file has undergone, but also pinpointing the moment of alteration. For instance, in a public debate recording, an adversary might alter the voice of specific individuals to discredit specific words or phrases to selectively change meaning. Thus, in voice edit detection, two primary objectives emerge: (1) achieving high detection accuracy and (2) maintaining fine time resolution. These objectives often conflict because analyzing a larger audio sample enhances the model's ability to identify voice edits but simultaneously reduces the temporal resolution.

Moreover, some detection models such as SSL W2V2 necessitate a minimum audio input length, inherently limiting their capacity for fine time resolution. Extending duration through audio repetition would itself be considered a form of voice manipulation by concatenation ③. To assess this trade-off, we examine all models - excluding SSL W2V2 - at three levels of time resolution: fine (0.35s), medium (1.2s), and coarse (4.05s). This approach involves segmenting the original or modified audio into subsections of these specified lengths, and then supplying the subsections to the neural architecture. We choose 0.35s, because it is the lowest limit allowed by the LCNN model, 4.05s as the minimum required by SSL W2V2, and 1.2s as the intermediate value, calculated based on a geometric progression.

6. Results

We assess the performance of the trained models with respect to two metrics: overall test accuracy and individual F1 scores for each voice edit. The experiments are repeated twice, and the aggregated results, including mean and standard deviation, are presented in Table 2. Our evaluation extends beyond mere detection performance; we also examine performance relative to input length, highlighting the best performance for each category and input length. We use color coding for clarity: yellow for fine, red for medium, and purple for coarse time resolution.

Our observations are as follows. First, as the resolution becomes finer, the detection performance decreases: an overall test accuracy of 80.8% is observed at coarse time resolution, 73.1% at medium resolution, and 55.0% at fine resolution, as shown in Table 2. This is expected, since finer time resolution means less data for the model to base its decision on. Second, we identify a strong trend in detection performance across resolutions: the SSL W2V2 model excels at coarse resolution, while the LCNN and ComplexNet models perform best at medium and fine resolutions, respectively. RawNet2 underperforms across the board. Third, not all voice edits pose the same level of detection difficulty. The 'Concatenation/Trimming' edit ③ is detectable only by the SSL W2V2 model, suggesting that pre-training experience is crucial in this case. Many edits are detectable even at fine resolution, whereas some, like 'Alaw Encoding' ⑪ and 'Equalization' ⑮, are consistently challenging due to minimal audio changes. Nevertheless, models like SSL W2V2 demonstrate satisfactory performance when coarse time resolution is adequate.

7. Conclusion

Voice edits have multiple real-world use-cases, which we expand on here. As observed in the real world, changes in playback speed, equalization, insertion or cutting of audio, as well as TTS and VC recordings can be used to **spread misinformation and facilitate fraud** [15, 16, 46], making the detection thereof a paramount challenge –① to ⑧, and ⑮. Determining the authenticity of audio evidence is critical for audio evidence in **legal court cases** to discern between genuine and synthetic

Table 2: *Voice edit detection performance (F1 and accuracy) for each input duration. Color coding denotes optimal performance per input length: yellow for the best results at 0.35 seconds, red for 1.2 seconds, and purple for 4.05 seconds. Results are presented as mean ± standard deviation.*

Input Length	0.35s			1.2s			4.05s			
Model Name	ComplexNet	LCNN	RawNet2	ComplexNet	LCNN	RawNet2	ComplexNet	LCNN	RawNet2	SSL W2V2
Epoch Time (min)	4.0±0.8	4.3±0.1	2.5±0.1	5.6±0.1	5.0±0.3	2.9±0.5	15.9±0.1	7.6±2.0	5.5±0.5	34.0±0.3
Test Accuracy	51.3±1.1	55.0±2.2	10.9±1.5	72.1±1.8	70.2±1.3	51.8±0.9	75.4±0.0	69.0±6.5	52.7±0.4	80.8±5.4
F1 Scores										
Original Voice ⓪	2.2±1.7	7.4±10.4	0.0±0.0	16.6±13.1	4.4±3.1	3.9±5.6	36.8±3.2	22.2±22.2	16.5±3.1	48.7±9.9
Text To Speech ①	35.5±25.2	67.8±3.5	0.0±0.0	77.6±3.0	78.0±7.4	54.4±4.8	90.9±1.3	72.8±22.6	58.0±8.8	97.8±1.0
Voice Cloning ⓪	75.3±11.8	81.3±12.4	46.9±3.2	95.9±0.8	94.6±4.0	81.6±14.1	98.4±0.5	89.9±9.8	74.9±27.9	99.8±0.2
Insert Or Cut Audio ③	0.0±0.0	0.0±0.0	0.0±0.0	16.1±8.8	15.9±3.2	8.0±4.4	12.2±8.1	5.7±8.1	1.1±0.1	92.8±2.4
Mixing ④	22.8±3.6	25.9±2.3	0.0±0.0	78.6±1.8	62.0±6.0	40.1±12.4	92.5±2.2	83.4±1.7	56.6±25.1	97.8±1.1
Pitch Up ⑤	81.1±3.0	78.0±1.7	0.0±0.0	94.3±1.2	89.7±1.5	74.0±3.5	96.7±0.7	86.6±13.4	81.9±1.7	98.0±1.0
Pitch Down ⑥	70.6±7.0	90.7±1.2	0.0±0.0	94.8±3.8	97.8±1.1	77.3±6.3	97.9±1.9	99.3±0.6	75.7±16.2	98.3±1.4
Speed Slower ⑦	70.9±3.0	34.5±18.0	0.0±0.0	75.8±4.1	29.4±26.5	40.1±1.8	46.1±32.5	53.9±6.5	48.7±16.6	83.3±3.3
Speed Faster ⑧	73.5±5.7	62.1±6.8	0.0±0.0	79.1±2.8	64.8±2.7	56.4±4.0	50.1±21.9	43.1±26.8	32.5±13.1	83.5±0.9
Mp3 Compression ⑨	54.9±1.9	86.4±1.1	1.3±1.8	84.4±5.4	96.8±1.4	15.2±4.8	90.3±4.5	98.1±2.4	25.3±6.9	79.7±19.5
Aac Compression ⑩	91.2±2.9	92.8±3.9	2.7±3.8	91.6±1.3	84.8±1.7	83.8±3.6	85.4±2.1	54.2±16.2	6.9±0.9	81.9±1.5
Alaw Encoding ⑪	15.9±22.5	41.5±1.1	0.0±0.0	12.1±17.1	40.2±17.0	5.7±2.1	45.0±20.4	35.4±19.5	2.8±4.0	32.6±41.8
Ulaw Encoding ⑫	14.4±13.2	24.7±2.2	0.0±0.0	47.0±12.0	44.4±18.0	0.4±0.6	24.6±33.2	53.8±12.3	0.0±0.0	12.4±17.5
High Pass Filter ⑭	88.6±0.6	48.3±3.7	14.0±3.6	93.6±0.5	91.6±1.9	92.5±3.4	96.1±0.6	91.6±2.1	92.6±1.9	96.4±0.6
Low Pass Filter ⑬	22.4±5.5	59.9±0.6	0.0±0.0	58.6±12.7	77.6±1.0	16.7±5.7	69.3±4.4	70.0±21.5	13.2±14.6	50.8±44.8
Equalization ⑮	18.2±6.5	17.0±9.8	5.2±7.4	17.9±25.3	43.9±3.3	7.1±4.2	19.4±27.5	7.4±10.5	0.0±0.0	0.0±0.0
Auto Tune ⑯	42.4±19.5	45.8±1.5	0.0±0.0	78.9±3.6	77.6±2.1	47.2±1.9	95.1±1.5	80.1±8.1	73.8±8.4	98.3±0.5
Room Impulse ⑰	91.7±0.5	84.9±2.4	42.6±1.1	97.9±0.9	90.4±3.0	92.6±1.5	98.7±0.4	84.1±8.4	88.0±2.5	99.4±0.1
Reverb ⑱	72.2±4.3	45.7±14.8	0.0±0.0	96.8±1.4	86.3±1.5	87.3±4.6	99.3±0.3	87.0±4.7	84.1±8.2	99.9±0.2
Overlay Background ⑲	80.5±0.9	75.6±6.8	0.0±0.0	96.4±0.6	91.8±0.1	86.3±9.3	98.3±0.0	95.5±0.2	85.6±8.0	99.5±0.4
Noise Reduce ⑳	39.8±17.8	54.0±8.2	0.0±0.0	82.4±4.4	62.5±1.5	70.5±1.2	99.7±0.0	99.4±0.9	92.1±0.3	99.3±0.3

or altered voices–①, ②, ③, and ①. The **security of voice biometric systems** must be protected to differentiate between real and synthetic voices and to ensure secure access–⓪, ① and ②. Unauthorized use or alterations of **copyrighted or pirated audio** content must be identified, including compression to evade detection–⑨ to ⑫, ④, and ⑮. Voice edit detection can help assess the authenticity of audio evidence in **insurance** claims, such as verifying the sound environment of an accident scene–⑰ to ⑳. Similarly, it can help to ensure that live or recorded performances in the **entertainment industry** are free from unauthorized enhancements or effects–⑮ and ⑯.

In this paper, we presented a new way of thinking about the conventional binary classification of voice recordings solely into 'fake' and 'benign' categories. In particular, a hyper-focus solely on speech synthesis oversimplifies the complexity of how voice edits can impact authenticity. A set of low-level edits can be utilized to disseminate misinformation, undermine political figures, and perpetrate fraud. Identifying, detecting, and locating such voice edits is therefore a critical new problem in our field. We introduce a detailed taxonomy of voice edits, compile a dataset framework corresponding to these categories, and assess the efficacy of existing anti-spoofing and deepfake detection models. Most importantly, we argue for a paradigm shift in research priorities – moving beyond the mere identification of AI-generated modifications to a more comprehensive examination that spans the full array of vocal edits.

8. Acknowledgement

This work has been (partially) funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy; and the Department of Artificial Intelligence, Wrocław University of Science and Technology.

9. References

[1] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation," in *Proc. Interspeech 2019*, 2019, pp. 4115–4119.

[2] "Respeecher synthesized a younger luke skywalker's voice for disney+'s the mandalorian," https://www.respeecher.com/case-studies/respeecher-synthesized-younger-luke-skywalkers-voice-disneys-mandalorian, (Accessed on 02/07/2024).

[3] Wired, "Audiobooks synthetic voices," https://www.wired.com/story/audiobooks-synthetic-voices/, (Accessed on 07/02/2024).

[4] Apple, "Siri product website," https://www.apple.com/siri/, 2024, (Accessed on 02/07/2024).

[5] Forbes, "Fraudsters cloned company director's voice in $35 million bank heist, police find," https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions, 2024, (Accessed on 02/05/2024).

[6] Avast, "Voice fraud scams company out of 243,000 dollars," https://blog.avast.com/deepfake-voice-fraud-causes-243k-scam, 2024, (Accessed on 02/07/2024).

[7] ——, "Ai-generated voice firm clamps down after 4chan makes celebrity voices for abuse," https://www.vice.com/en/article/dy7mww/ai-voice-firm-4chan-celebrity-voices-emma-watson-joe-rogan-elevenlabs, 2024, (Accessed on 02/07/2024).

[8] Vice, "How i broke into a bank account with an ai-generated voice," https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice, 2024, (Accessed on 07/02/2024).

[9] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

[10] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.

[11] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does Audio Deepfake Detection Generalize?" in *Proc. Interspeech 2022*, 2022, pp. 2783–2787.

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

[13] "Fact check: "drunk" nancy pelosi video is manipulated — reuters," https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated-idUSKCN24Z2BI, (Accessed on 08/23/2023).

[14] "Fox news edits video of biden to make it seem he was being racially insensitive — fox news — the guardian," https://www.theguardian.com/media/2021/nov/13/fox-news-edits-biden-video-negro-leagues-satchel-paige, (Accessed on 02/06/2024).

[15] "The sound of a politician's voice impacts their success • earth.com," https://www.earth.com/news/politicians-voice-impacts-success/, (Accessed on 02/06/2024).

[16] "Tvp celowo zmienila glos trzaskowskiemu w sorkwitach? mozliwosci sa dwie — natemat.pl," https://natemat.pl/312589,tvp-celowo-zmienila-glos-trzaskowskiemu-w-sorkwitach-mozliwosci-sa-dwie, (Accessed on 02/06/2024).

[17] "Fake biden robocall tells voters to skip new hampshire primary election - bbc news," https://www.bbc.com/news/world-us-canada-68064247, (Accessed on 02/06/2024).

[18] "Meet ashley, the world's first ai-powered political campaign caller — reuters," https://www.reuters.com/technology/meet-ashley-worlds-first-ai-powered-political-campaign-caller-2023-12-12/?, (Accessed on 02/06/2024).

[19] Q. Liu, A. H. Sung, and M. Qiao, "Detection of double mp3 compression," *Cognitive Computation*, vol. 2, pp. 291–296, 2010.

[20] P. R. Bevinamarad and M. Shirldonkar, "Audio forgery detection techniques: Present and past review," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*. IEEE, 2020, pp. 613–618.

[21] M. Zakariah, M. K. Khan, and H. Malik, "Digital multimedia audio forensics: past, present and future," *Multimedia tools and applications*, vol. 77, pp. 1009–1040, 2018.

[22] P. A. A. Esquef, J. A. Apolinário, and L. W. Biscainho, "Edit detection in speech recordings via instantaneous electric network frequency variations," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2314–2326, 2014.

[23] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Proc. Interspeech 2017*, 2017, pp. 2–6.

[24] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2015, pp. 1–6.

[25] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech 2015*, 2015, pp. 2037–2041.

[26] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren *et al.*, "Add 2023: the second audio deepfake detection challenge," *IJCAI 2023 Workshop on Deepfake Audio Detection (DADA 2023)*, 2023.

[27] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, L. Xu, and R. Fu, "Fad: A chinese dataset for fake audio detection," *arXiv preprint arXiv:2207.12308*, 2022.

[28] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-End anti-spoofing with RawNet2," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6369–6373.

[29] W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw Differentiable Architecture Search for Speech Deepfake and Spoofing Detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 22–28.

[30] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

[31] X. Wang and J. Yamagishi, "A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection," in *Proc. Interspeech 2021*, 2021, pp. 4259–4263.

[32] N. M. Müller, P. Sperl, and K. Böttinger, "Complex-valued neural networks for voice anti-spoofing," in *Proc. INTERSPEECH 2023*, 2023, pp. 3814–3818.

[33] Z. Zhang, X. Yi, and X. Zhao, "Fake speech detection using residual network with transformer encoder," in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, 2021, pp. 13–22.

[34] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *The Speaker and Language Recognition Workshop*, 2022.

[35] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved DeepFake Detection Using Whisper Features," in *Proc. INTERSPEECH 2023*, 2023, pp. 4009–4013.

[36] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-Truth: A Partially Fake Audio Detection Dataset," in *Proc. Interspeech 2021*, 2021, pp. 1654–1658.

[37] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2023.

[38] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[39] Y. Zang, Y. Zhang, M. Heydari, and Z. Duan, "Singfake: Singing voice deepfake detection," *arXiv preprint arXiv:2309.07525*, 2023.

[40] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "Mlaad: The multi-language audio anti-spoofing dataset," *arXiv preprint arXiv:2401.09512*, 2024.

[41] T. M.-A. S. Dataset, "The m-ailabs speech dataset," https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/, 2023, accessed on 01/02/2024.

[42] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[43] S. Shan, Y. Li, A. Banerjee, and J. B. Oliva, "Phoneme hallucinator: One-shot voice conversion via set expansion," *arXiv preprint arXiv:2308.06382*, 2023.

[44] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[45] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[46] "A voice deepfake was used to scam a ceo out of $243,000," https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/, (Accessed on 02/06/2024).