Journal of
Cybersecurity and Privacy

MDPI

Article

Partial Fake Speech Attacks in the Real World Using Deepfake Audio

Abdulazeez Alali and George Theodorakopoulos

School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, UK;
Correspondence: aliaa8@cardiff.ac.uk (A.A); theodorakopoulosg@cardiff.ac.uk (G.T.)

Abstract: Advances in deep learning have led to dramatic improvements in generative synthetic speech, eliminating robotic speech patterns to create speech that is indistinguishable from human voice. Although these advances are extremely useful in various applications, they also facilitate powerful attacks against both humans and machines. Recently, a new type of speech attack called partial fake (PF) speech has emerged. This paper studies how well humans and machines, including speaker recognition systems and existing fake-speech detection tools, can distinguish between human voice and computer-generated speech. Our study shows that both humans and machines can be easily deceived by PF speech, and the current defences against PF speech are insufficient. These findings emphasise the urgency of increasing awareness for humans and creating new automated defences against PF speech for machines.

Keywords: neural networks; speech synthesis; biometric security; Deepfake audio; Partial Fake Audio

Citation: Alali, A.; Theodorakopoulos, G. Partial Fake Speech Attacks in the Real World Using Deepfake Audio. *J. Cybersecur. Priv.* **2024**, *1*, 1–24. https://doi.org/

Academic Editor:

Received:
Revised:
Accepted:
Published:

1. Introduction

All one needs is a few spoken words to identify family, friends, and even famous figures based on their voices. This is because the human voice is highly distinctive [1]. Even if one is unfamiliar with the speaker, they can still determine some basic identity characteristics, such as their gender [2] and approximate age, based on their voice. In the future, our voices may become our personal passwords, which will allow us to access various applications and systems linked to our identity.

Recent advances in deep neural network (DNN) technology have raised concerns about the uniqueness of individual voices. It is now possible to clone voices in a way that makes them undetectable to human listeners. This can be achieved through both text-to-speech (TTS) techniques, which allow cloning a voice to read any text aloud [3–11] and voice conversion (VC) techniques, which convert the voice of one speaker to sound like another while maintaining the same speech content [3,12–14]. There are also commercial services available that offer synthesised speech generation and voice cloning [15,16].

The uniqueness of the human voice has led to its use in various applications, such as identity verification and credential authentication. Automatic speaker recognition (ASR) systems, which encompass automatic speaker identification (ASI) and automatic speaker verification (ASV), rely heavily on voice. ASV is integrated into everyday transactions such as call centre and e-banking systems at institutions like HSBC Bank [17], user authentication in mobile apps such as WeChat [18], payment authorisation systems like Alipay [19], and user verification in the Internet of Things (IoT) devices like Amazon Alexa in echo series devices [20]. Although the use of voice recognition speeds up transactions and simplifies the usage of these services, it also opens up a new avenue for attack by cloning one's voice using DNN techniques and embedding the cloned voice segment within real speech to create partial fake speech. Partial fake (PF) speech can bypass the verification and authentication processes, causing severe consequences for individuals and organisations [21,22].

Furthermore, PF speech can be used to perform various types of phishing attacks [23–25], which will be discussed in detail in Section 2.

DNN advances that generate synthetic speech pose a critical threat to both computer-based and human-based systems. Yet, until now, no study has measured the success of PF speech attacks, as all prior works have measured such attacks using entirely fake speech [26–28] which is in principle easier to detect.

We believe that it is important to investigate and quantify the extent to which PF speech attacks based on deep learning can succeed against two different entities: machines and humans. Can these attacks defeat existing speaker recognition systems or bypass voice verification systems in mobile apps? Moreover, can PF speech that imitates a specific human voice trick listeners into believing that it is real?

In this paper, we present the findings of our detailed analysis of the threat posed by deep-learning PF speech to both humans and machines. Building upon previous research [26], we conducted similar attacks against modern speaker recognition systems and we also performed questionnaire-based user studies. However, instead of using entirely fake speech, we opted for partially fake speech. This study is the first of its kind to experimentally analyse the impact of deep-learning-based PF speech attacks on both human listeners and machines. In addition to testing speaker recognition systems in both commercial and open-source models, mobile app verification, and user studies, we also evaluated the effectiveness of existing defences against PF speech attacks. All our experiments used publicly available deep-learning speech synthesis systems. Our results underline the need for new defences against deep learning-based PF speech attacks for both humans and machines.

Key Findings. We have discovered several important outcomes from our study:

- We conducted a series of experiments with over 80 speakers and found that using DNN speech synthesis tools to create PF speech can effectively deceive modern public and commercial speaker recognition systems with a success rate of 95% to 97%.
- A survey of 148 participants revealed that humans can only distinguish PF speech from real speech with very low accuracy: 16% for unknown voices and 17.5% for known voices.
- Our study, which involved 148 participants, found that humans can only identify partially fake speech within a completely authentic video with 24% accuracy when the speaker's face is close to the camera and 11% accuracy when the speaker's face is far from the camera.
- A detailed evaluation of three state-of-the-art defence algorithms reveals an inability to prevent and detect PF speech, highlighting the need for new defences.

2. Motivation

The emergence of deep-fake technology has made it more difficult to distinguish between real and fake audio. Deepfake audio is a new take on the impersonation tactics that have been utilised for a long time in social engineering and phishing attacks. Partial fake speech is a new challenge that needs to be addressed and detected. It can be performed using existing attacks that are based on entirely fake audio. Here are some of the motivations for PF voice fraud:

2.1. Financial Gain

The first motivation is the desire to steal money. Financial institutions have increased their efforts to combat fraud by using voiceprints as the primary method for user identification. However, the security of using voiceprints as a check for uniqueness is becoming increasingly vulnerable due to the rapid advances in deep voice-faking technology. A real-world example occurred when a BBC reporter established an HSBC voice-ID authenticated account, and his non-identical twin managed to deceive the system, gaining unauthorised access to his brother's funds [29]. Another method for implementing attacks based on fake voices is by impersonating CEO voices over the phone. Such an instance occurred when

an energy company based in the UK fell victim to a CEO fraud that used AI-generated deepfake audio, resulting in a loss of $243,000 [23]. In another similar case, criminals utilised AI voice cloning to deceive a bank in the United Arab Emirates and managed to illegally acquire a staggering sum of $35 million [24].

2.2. Political Leader and Public Figure Reputation Attacks

The use of deepfake voices is becoming a major concern for public figures, celebrities, and political leaders. Deepfake cloned voices can be manipulated to create fake statements, speeches, or endorsements. This can harm their reputations and spread misinformation. The consequences of such actions can be far-reaching, with a significant impact on political processes and public trust. Regarding the implication of deepfake audio and video on the 2024 United States presidential election, A.J. Nash, vice president of intelligence at the cybersecurity firm ZeroFox, stated, "We're not prepared for this, the big leap forward is the audio and video capabilities that have emerged. When you can do that on a large scale and distribute it on social platforms, well, it's going to have a major impact" [21]. Celebrities also face reputational damage when their voices are cloned and spread through social media platforms. Such an instance occurred when Taylor Swift and Kelly Clarkson's cloned voices were used in product advertisements, which led to disappointed fans [22].

2.3. Virtual Kidnapping

An attacker could acquire a recording of a child, possibly taken from a movie script, to artificially simulate crying, screaming, and deep distress. Subsequently, the attacker might use this deepfake voice as proof of having the targeted victim's child in their custody, using this "proof" to pressure the victim into meeting their demands and sending significant ransom amounts. CNN published a real-life case where an attacker called a mother using the cloned voice of her 15-year-old daughter. The attacker asked her to pay US $1 million as ransom for her daughter, but luckily her daughter called her before she transferred the money [25].

2.4. Breaking Customer Trust

Attackers can use fake voices to impersonate customers or trusted figures within organisations. This can lead to various types of attacks, such as fraudulent transactions, social engineering, deceptive communication, false representation, and phishing. It is important to be aware of the risks associated with these types of attacks and to take appropriate measures to prevent them.

2.5. Deepfake Audio in the Court

During a child custody battle in the UK, a "deepfake" audio recording was presented as evidence to discredit a Dubai resident. The lawyer representing the resident, Byron James, stated that a heavily edited recording of his client was used in court to support the opposing side's argument in a family dispute [30].

3. Background

3.1. User Identification Based on Voice

Speaker Voice Recognition by Humans. Humans consistently and accurately identify individuals by their voices, particularly when there is a high degree of familiarity with the person, such as close acquaintances or public figures. Often, a brief non-linguistic cue, like a laugh, is sufficient for us to recognise a familiar person [31]. Although human speaker identification is not perfect, it is highly accurate and has inspired the creation of speaker recognition systems for security purposes.

Automatic Speaker Recognition (ASR). Speech signals reveal different speaker characteristics, such as their origin, identity, gender, and emotion. This unique feature of speech allows speaker profiling through speech-based techniques, which can be used in different areas like forensics and recommendation systems. Speaker recognition is a widely

researched subject with two primary objectives: automatic speaker identification, which involves determining the speaker's identity, and automatic speaker verification, which involves confirming the claimed identity [32].

Automatic Speaker Identification (ASI). Automatic speaker identification is a technology that uses algorithms and computational techniques to automatically identify the speaker using an audio sample [33]. ASI systems analyse various voice features, such as pitch, tone, rhythm, and spectral characteristics, to create a unique user voiceprint or template.

Automatic Speaker Verification (ASV). Automatic speaker verification is a technology that uses computational methods to automatically verify whether a speaker's claimed identity matches their actual identity based on their voice characteristics. ASV is often used in applications where secure access or authentication is required.

3.2. Automatic Speaker Recognition Challenges

Speech Pathology. Medical issues, such as dysphonia, dysarthria, and cleft lip and palate, can lead to breaks or interruptions in speech. As stated by [34], pathological speech is more vulnerable to privacy breaches compared to healthy speech when applied to ASV systems.

Background Noise. The accuracy of speaker recognition systems is notably affected by background noise [35]. This is because clean environmental recordings are utilised during training, while speakers often speak in noisy conditions during testing.

Children's Voices. Recognising and accurately identifying children's speech can be more difficult than identifying adults' speech. This is because children's vocal tracts and language skills are still developing, which causes their speech to be less clear, less stable, and more variable compared to adults. Additionally, children may not have the same level of control over their speech as adults do. This makes it challenging for recognition systems to identify both the speech and the speaker accurately [36].

3.3. Speech Synthesis Generation

Speech synthesis is a process that involves the artificial production of human speech using computer algorithms and synthetic voices. Speech synthesis encompasses two main categories: text-to-speech (TTS) and voice conversion (VC). Both TTS and VC systems have been significantly improved by deep neural networks (DNNs), which enable more natural-sounding voice synthesis. TTS systems convert any given text into spoken words, mimicking the voice of a specified target speaker [7,11,37–40]. Traditional TTS systems used concatenative synthesis, which involved stitching together short pre-recorded speech segments to form longer sentences [41]. By contrast, VC is a category of speech synthesis focused on modifying the characteristics of a source speaker's voice to make it sound like a target speaker's voice [12–14,42].

Zero-shot TTS and VC. Recently, researchers have dedicated significant attention to zero-shot speech synthesis generation [3–6,10,43–45]. This technique allows for the adoption of a new voice with just one utterance or a few seconds of target voice speech without requiring additional training. The main advantage of this technique is that it enables the adaptation of a new voice without the need for retraining [46].

Several methods may be used to carry out synthetic speech generation. There are two main methods for creating synthetic speech using DNN: commercial cloud services and open-source tools. In order to use a commercial cloud service to train a model for a specific voice, the person's consent must be obtained. If consent is not obtained, the user can utilise the voices provided by the cloud service to generate TTS audio. Commercial services offer full voice training [47], zero-shot TTS service [16], and zero-shot TTS and VC service [15]. On the other hand, numerous DNN speech synthesis open-source tools are publicly accessible for both TTS [9,48,49] and VC [50–52] including support for zero-shot tools.

3.4. Partial Fake (PF) Speech Generation

A new type of synthetic speech called PF has appeared in recent years. This synthetic speech includes both real and fake segments within the audio file. The fake segments could be a word or a few words that are generated using the DNN methods used to generate TTS or VC.

Zero-shot generation techniques simplify the generation of convincing PF speech because they require just a few seconds of target speech recording, which is not required to be part of the training data. The fake segment shown in Figure 1 can be generated using TTS or VC methods.

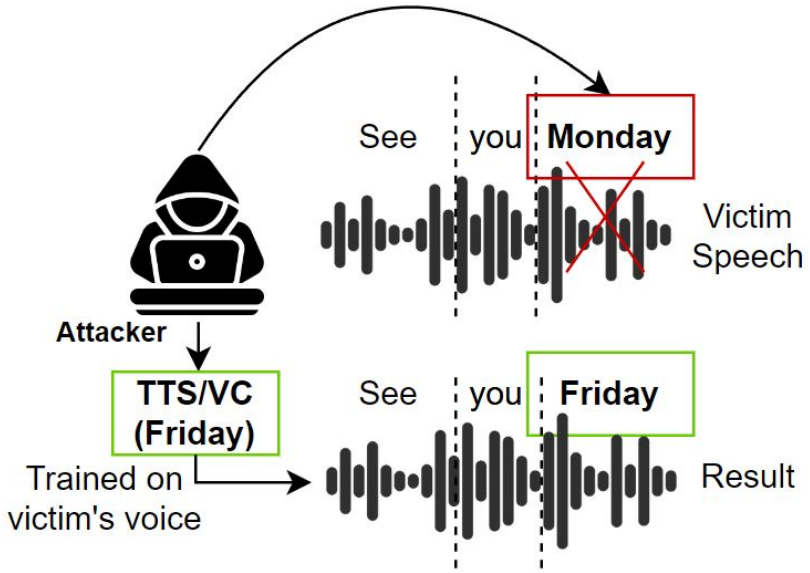

Figure 1. Example of PF generation.

3.5. Partial Fake Speech Attacks

The attacker can execute various attacks against different systems using PF speech. This includes attacks on ASV cloud services, open-source tools, mobile applications, and any system that depends on voice for identification or authentication purposes. To the best of our knowledge, no study has yet used PF speech audio to carry out such attacks. Therefore, our primary focus will be on the attack potential of PF speech.

3.6. Partial Fake Speech Detection

Until the Audio Deepfake Detection Challenge (ADD) [53,54] started, all research was focused on distinguishing between entirely fake and real audio. The ADD challenge motivated many researchers to work on PF detection methods [29,55–57]. We found that none of the partial fake detection tools are available to the public, so we were not able to reproduce their results or examine their detection efficiency on other PF datasets or entirely fake datasets.

4. Threat Model

4.1. Attacker Model

The aim of attacker ***A*** is to steal or represent victim ***B's*** identity. To do so, ***A*** must first obtain a set of real recording samples ***RB*** from ***B***. If ***A*** is a personal acquaintance of ***B***, these speech clips may also be acquired from private media sources. The attacker needs only 15–30 s to clone someone's voice using zero-shot speech synthesis generation method described earlier in Section 3.3. Next, ***A*** inputs ***RB*** into DNN state-of-the-art speech synthesizers, which produce a synthetic voice for ***B***, called ***RF***. The content of ***RF*** is chosen

by ***A***. Finally, the attacker will replace a word or a few words in the original clip of ***B*** with chosen words from fake speech ***RF***, resulting in a partial fake audio clip ***RP***.

4.2. Summary of Experiments

We performed a measurement study to examine the risks associated with publicly accessible DNN-based speech synthesis systems on machines and humans through the following activities:

- Empirical experiments to see if PF speech attacks can fool Speaker Recognition systems.
- User studies to explore whether humans can differentiate between PF speech and real speech, even when the PF speech is embedded within an entirely legitimate video.
- Empirical experiments to evaluate the effectiveness of existing defences against PF speech attacks.

Below, we provide details on the speaker datasets used in our experiments, as well as the DNN synthesis and Speaker Recognition systems.

4.3. DNN-Based Speech Synthesis

In our attack we used "zero-shot" systems that require less than one minute of target data for voice cloning. Our focus is on peer-reviewed papers that have accessible code implementations and pre-trained models. Despite testing several TTS and VC systems, including [58–61] we noticed that many did not effectively generalise to unseen speakers (individuals who are not in the training dataset). Generalisation is critical for low-resource attackers, as it provides flexibility in target selection. Ultimately, to quickly generate PF speech in a simple manner, we selected four systems that showed superior performance on unseen speakers: Tortoise-TTS [9], a text-to-speech system, FreeVC [49], an any-to-any and cross-lingual VC, DiffVC [78] a one-shot many-to-many voice conversion, and ppgVC [50] an any-to-many VC.

Tortoise-TTS: Tortoise TTS is an advanced zero-shot text-to-speech model that requires just a few (3–10) seconds of recording. It excels at providing high-quality voice cloning and proficient narration for large volumes of speech content, like books or articles. With its precise control over voice synthesis, Tortoise is suitable for a wide range of applications, including virtual assistants and audiobook creation. The Tortoise-TTS system consists of a pipeline with five individually trained neural networks: an autoregressive decoder, a contrastive language-voice pretraining (CLVP) model, a contrastive voice-voice pretraining (CVVP) model, a diffusion decoder, and a vocoder [9].

FreeVC: The training strategy is an end-to-end approach that utilizes a combination of a Conditional Variational Autoencoder (CVAE) enhanced with Generative Adversarial Network (GAN) training. The model architecture features a bottleneck extractor and a normalizing flow, which work together to refine the process of content information extraction. This refinement improves the clarity and separation of speaker-independent features [49].

DiffVC: DiffVC is a one-shot, many-to-many voice conversion model. It utilizes WavLM to extract self-supervised learning (SSL) features from waveforms and incorporates a bottleneck extractor to capture content information from these features. Additionally, the model employs spectrogram-resizing (SR) based data augmentation, which modifies speaker information without altering the content information. This approach enhances the model's ability to disentangle content and speaker characteristics [78].

ppgVC: This VC allows for any-to-many voice conversion, meaning it can take a speech sample from any speaker and transform it into multiple target voices, rather than being restricted to a one-to-one conversion. It combines a sequence-to-sequence phoneme recognizer (Seq2seqPR) with a multi-speaker duration-informed attention network (DurIAN) to facilitate synthesis. Additionally, this approach has been expanded to support any-to-any voice conversion. [50].

4.4. Speaker Recognition (SR) Systems

To investigate the potential threat of PF attacks on machines, we selected three state-of-the-art SR systems: an open-source tool, a cloud service, and a mobile app.

Resemblyzer: Resemblyser is an open-source tool that provides three major tasks, namely, speech detection, speech segmentation, and embedding extraction [62]. It is a side project of [63] that can extract a detailed representation of a voice by using a deep learning model. It creates a summary vector of 256 values that capture the unique characteristics of the voice when given an audio file of speech. This tool analyses and compares the similarities between potential fake speech and real speech to determine whether it is real. Resemblyzer speech detection requires a minimum of 30 s of recording speech for training.

AWS Cloud: The Amazon Cloud's voice ID service [64] is designed to authenticate customers during phone interactions. To enroll in the system, an approximate 30-s sample of the caller's speech is required to create a voice print. During subsequent calls, the caller's voice is compared to the registered template, and the voice ID system takes around 10 s to authenticate and confirm a match.

Amazon Alexa: Alexa is a virtual assistant that can perform various tasks and provide information in response to user voice commands using natural language processing and voice recognition. It is mainly available on Echo devices and as an app on mobile devices. One of Alexa's features is Voice ID, which allows users to create a voice profile by reading six sentences aloud. This feature adds different lanes for specific users without the need for account switching. Figure 2 illustrates various attacks performed in this study.

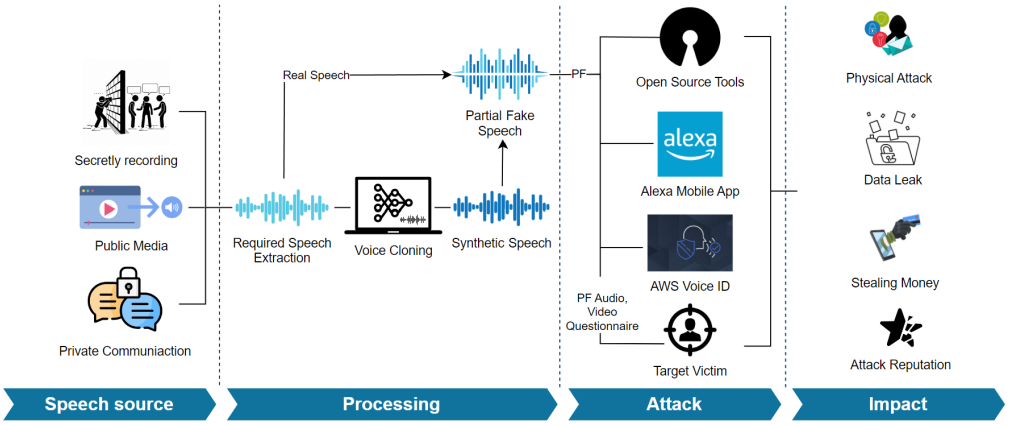

Figure 2. The various attack scenarios performed in this study.

4.5. Speaker/Speech Sources

We carried out our attacks using four different sources. The first dataset RFP [65] is a partial fake audio dataset (it also includes entirely real and entirely fake speech), while the second [66], third [67], and fourth [68] datasets are public and entirely real recording datasets. The last dataset is a custom dataset that we created specifically for the experiments conducted in this study.

- RFP [65] dataset comprises three types of audio: real, fake, and partially fake. It includes a total of 127,862 utterances spoken by 354 speakers, of which 184 are male and 170 are female. Eleven different methods were used to produce the fake voices, including seven methods for generating TTS and four methods for VC. This dataset is one of the few publicly available that includes PF audio.
- The UK and Irish English Dialect speech dataset [66] consists of 18,779 spoken utterances by 118 individuals. We opted to use this dataset as it includes participants from six different regions in the UK and Ireland who spoke various English dialects, namely Irish, Midlands, Northern, Scottish, Southern, and Welsh English.
- VCTK dataset [67] consists of speech data from 110 English speakers with age range 18-38 and varied accents. Each speaker reads around 400 sentences picked from a

newspaper, the rainbow passage, and an elicitation paragraph used for the speech accent archive.
- YouTube-8M [68] is a video dataset containing over 7 million videos, each labelled with one of 4716 classes using an annotation system. Our use case involves creating PF audio within a real video, which led us to use YouTube-8M.
- We created a custom dataset for our experiments. First, we collected Alexa commands spoken by 15 native male and female English speakers and used the recordings to perform Alexa attacks. Second, we created PF audio with different fake segment weights to perform AWS cloud and Resemblyzer attacks. Lastly, we created PF audio within an entirely real video and used it in the questionnaire.

4.6. Ethics

All our user study protocols were approved by the school's ethical team and were designed to protect the privacy and well-being of our participants. Our study involved 148 volunteers. Prior to their participation, we provided all the participants with detailed information about the scope, goals, and steps of the experiment. The participants followed all the steps without installing anything on their devices. Moreover, the participants did not incur any costs during the attacks. We only kept audio recordings of the Alexa attack, which was anonymised and stored on secure servers.

5. Partial Fake Speech Against Machines

Our first research question is "To what extent are machine-based speaker recognition (SR) systems susceptible to PF speech attacks?" Previous research has investigated this question using entirely fake speech; however, the effectiveness of real-world SR systems against DNN PF synthesis attacks remains uncertain. In this section, we address this question by assessing the resilience of three contemporary SR systems to DNN-based PF synthesis attacks. Our study includes a series of experiments, which are illustrated in Figure 2.

5.1. Partial Fake Attack Against Modern SR Systems

We conducted an evaluation to test the effectiveness of the DNN PF speech attack against modern SR systems. In a previous study [26], researchers assessed the Resemblyzer and Microsoft Azure cloud services' abilities to detect entirely fake audio. In our experiment, we launched a similar attack using PF speech against Resemblyzer and the AWS cloud voice ID service. In a 2021 paper [26] the researchers stated that Resemblyzer and Azure effectively detect entirely fake speech. The limitation of their test is that they only used one method of synthetic speech generation. To overcome this weakness in our study, we used seven different state-of-the-art and up-to-date synthetic speech generation methods, as listed in the following section.

5.2. PF Attack Against Resemblyzer

We conducted a PF speech attack on Resemblyzer, a widely used modern SR system, using the system's official open-source code [62].

Experiment Setup. Using attack success rate (ASR) as the metric, we evaluate the effectiveness of a DNN-based PF speech attack against Resemblyzer. We specifically analyse the impact of the following five factors on ASR:

- Synthetic speech generation methods: We evaluate TTS generation methods using AWS, Azure, Google Cloud, and the open-sourced tool Tortoise, and VC generation methods FreeVC, DiffVC, and ppgVC.
- The location of a fake signal within the audio file: We evaluate three options for the location of a fake signal within the audio file. All PF speeches are included in our customized dataset listed in Section 4.5. The three locations of PF speech are as follows: (1) Fake segment at the beginning of audio file, (2) Fake segment at the middle of

audio file, and (3) Fake segment at the end of audio file. Tables 1 and 2 list the attack results of the three fake segment locations within the audio file.
- The gender and dialect of the speaker which includes both male and female speakers, and speakers of the following dialects: Welsh, Northern, midland, and Scottish. Tables 3 and 4 include the results of each gender and dialect.
- The varying lengths of the fake segments within the audio file and their effect on the detection process.

Our experiment involves a total of 86 target speakers ***RB*** from three different speech sources. These sources include 30 random speakers chosen from the YouTube-8M dataset, seven random speakers from the VCTK dataset, and 49 speakers from the UK and Ireland English Dialect Speech dataset. To produce synthetic speech ***RF***, we used seven different synthetic speech generation methods, including TTS and VC. TTS was generated using AWS, Azure, Google Clouds, and the open-sourced tool Tortoise. VC was generated using FreeVC, DiffVC, and ppgVC. We combined real ***RB*** and synthetic speech ***RF*** to generate ***RP*** speech. To create ***RP*** files, we concatenate a specific segment of real audio with a specific segment of synthetic audio. PF files were generated in two segments (starting with fake audio, followed by real audio and vice versa) and in three segments where we placed the synthetic segment in the middle of the audio file.

Resemblyzer helps determine if two speech embeddings come from the same person. To set it up, we first enrolled the target speakers using their real speech samples. We then generated embeddings, which are unique numerical representations of their speech, and selected the threshold that minimises Resemblyzer's equal error rate (EER) on these speakers, using cosine similarity as the distance metric. During verification, the system compares the embedding of the evaluated utterance with the saved embedding in the database. If a synthesis attack occurs, success is achieved if the similarity between the attack and the enrolled embeddings exceeds the threshold. For training, we enrolled 10 real audio samples belonging to the same speaker, with each speaker's audio lasting over 30 s. Resemblyzer is well-known and widely used in research and verification applications.

Results. In the Resemblyzer experiments, a total of 21,760 ***RP*** audio files belonging to 86 speakers were used to perform the attack. The results show that the attack success rate of PF audio that starts with four seconds of real speech followed by two seconds of fake speech (whether generated using TTS or VC methods) is far higher than with other types of PF audio that starts with fake speech or embeds the fake segment in the middle of the speech. We also notice that gender and dialect slightly impact the attack success rate. However, some voices have a high impact on the attack success rate, although they used the same recording setup environment. Tables 1 and 2 show the averaged results of the experiment while Figure 3 visually compares the results of all TTS and VC generation methods.

Table 1. Attack success rate against Resemblyzer using AWS, Azure, Google Cloud, and Tortoise TTS generation methods. 2F refers to two seconds of fake segment, 4R is four seconds of real segment, and 2R is two seconds of real segment.

	PF (TTS-Based) – Fake Segment Location		
	2F, 4R	**4R, 2F**	**2R, 2F, 2R**
AWS (M)	0%	13.44%	5.5%
AWS (F)	0%	8.64%	0%
Azure (M)	0%	36.44%	9.28%
Google (M)	0%	1.75%	4.82%
Google (F)	0%	3.48%	0%
Tortoise (M)	74.95%	93.26%	86.61%
Tortoise (F)	3.48%	95.15%	74.95%

Table 2. Attack success rate against Resemblyzer using FreeVC, DiffVC, and ppgVC open-source VC tools.

	PF (VC-Based) – Fake Segment Location		
	2F, 4R	**4R, 2F**	**2R, 2F, 2R**
FreeVC (M)	6.39%	79.68%	60.73%
FreeVC (F)	1.62%	82.44%	49.83%
DiffVC (M)	0%	28%	9%
DiffVC (F)	0%	35.45%	5.45%
ppgVC (M)	0%	47.12%	9.95%

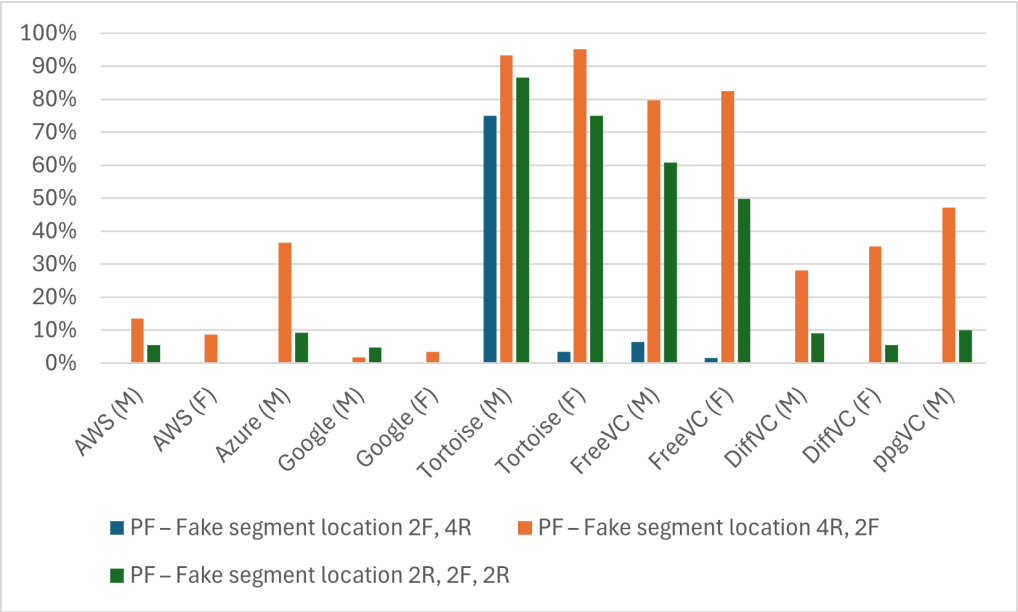

Figure 3. Attack success rate against Resemblyzer (TTS & VC).

Next, we discuss the results and the impact on the attack success rate of fake segment location, synthetic speech generation methods, gender, and target dialect in greater detail.

Table 3. Attack success rates against Resemblyzer based on different dialect speakers' real speech and using the Tortoise TTS generation method.

Speaker/Dialect	**Gender**	**ASR – (TTS: Tortoise)**		
		2F, 4R	**4R, 2F**	**2R, 2F, 2R**
Welsh speakers	Male	67.67%	97.69%	92.38%
Welsh speakers	Female	98.41%	88.94%	54.03%
Scottish speakers	Male	63.34%	92.36%	87.48%
Scottish speakers	Female	49.93%	91.27%	78.52%
Northern speaker	Male	95.26%	98.04%	69.62%
Northern speaker	Female	54.17%	93.62%	75.79%
Midlands speakers	Male	99.54%	90.02%	100%
Midlands speakers	Female	56.69%	95.48%	71.77%

Table 4. Attack success rates against Resemblyzer based on different dialect speakers' real speech and using FreeVC VC generation method.

Speaker/Dialect	Gender	ASR – (VC: FreeVC)		
		2F, 4R	4R, 2F	2R, 2F, 2R
Welsh speakers	Male	8.43%	93.80%	69.44%
Welsh speakers	Female	27.93%	88.94%	52.21%
Scottish speakers	Male	1.67%	90.63%	64.29%
Scottish speakers	Female	9.75%	84.16%	48.54%
Northern speaker	Male	19.93%	86.69%	42.51%
Northern speaker	Female	1.71%	89.97%	55.77%
Midlands speakers	Male	47.22%	99.18%	70.12%
Midlands speakers	Female	1.81%	64.30%	35.65%

5.2.1. Impact of Fake Segment Location

We conducted an experiment using the same content and speakers but with a different location for the fake segment in each crated file. Our aim was to determine whether the location of the fake segment would affect the detection process. We conducted the experiment three times and found that when the speech started with four seconds of real speech followed by two seconds of fake speech, the attack was most successful in both TTS and VC generation methods. This points to the possibility that the detection process does not examine the entire audio file. If it did, we would not have observed such a significant difference between 4R, 2F, and 2F, 4R.

5.2.2. Impact of Fake Segment Speech Content for the Same Speaker

We aimed to determine whether the speech content affects the detection process. To do so, we needed the final ***RP*** audio to have the same content spoken by the same person. To achieve this, we use Tortoise and Speech Recognition, which is a wrapper for various speech APIs, including the Google Web Speech API, which we used in our experiment to convert speech to text. Figure 4 in our report shows the process of generating PF using TTS for the same speaker and content.

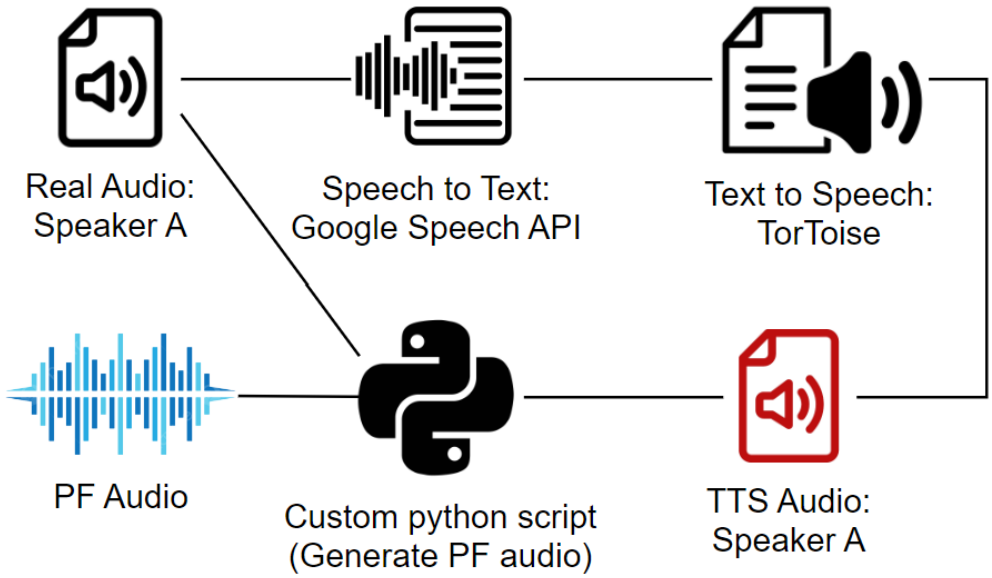

Figure 4. The flow of a custom Python script to create a PF speech using the same content as the original speech for the same speaker.

We used the FreeVC tool for voice conversion to create PF audio that includes both real and converted voices for the same speaker. We used the same file as the source and target,

resulting in a version of the same content but with a converted voice. We then entered both the real and VC audios into our custom Python script to generate the final PF audio that we used in the attack.

The PF attack that we created based on the same content and speaker in both TTS (Tortoise) and VC (FreeVC) was more successful than PF audio that did not belong to the same speaker or that had different content.

5.2.3. Impact of Synthetic Speech Generation Methods

It is important to investigate the different generation methods that can affect the attack outcome. The results in Tables 1 and 2 indicate that Resemblyzer can efficiently detect PF with an EER of 0% only against AWS, Azure, and Google from TTS generation methods. For VC, Resemblyzer detects DiffVC and ppgVC with 0% EER. For instance, Tortoise generation method has achieved up to 95.15% EER in TTS, while in voice conversion, a recorded EER of 82.44% has been observed in FreeVC generation method. The results show that Resemblyzer detection has a weakness against certain fake speech generation methods. Therefore, the attacker can choose Tortoise text-to-speech generation or FreeVC voice conversion to perform PF speech attack.

5.2.4. Impact of Gender

Based on the results presented in Tables 1 and 2, we noticed a marginal difference in the ASR outcomes for male and female PF speech against the Resemblyzer detection system, both in TTS and VC. The ASR accuracy for males ranged from 0% to 93.26% in TTS and from 0% to 79.68% in VC. For female TTS, the ASR ranged from 0% to 95.15% while in VC ranged from 0% to 82.44%.

5.2.5. Impact of Target Dialect

We noticed in Tables 3 and 4 that speakers from the Midlands had a higher ASR of 100% compared to other dialects. These tables display the average ASR of three speakers for each dialect. The PF speech was created using Tortoise for TTS and FreeVC for VC. We also found that the ASR was more successful for all dialects and genders when the fake segment at the end (4R, 2F) was used, except for TTS female Welsh speakers and Midland's male speakers.

5.2.6. Impact of the fake segment length

We created a subset of PF files, each with a length of 10 seconds. This subset includes 18 groups of files, each containing different lengths of fake segments. The files start with 1 second of fake segment followed by 9 seconds of real segment. We then gradually increased the length of the fake segment by one second while reducing the real segment accordingly until we reached a ratio of 9 seconds of fake segment and 1 second of real segment. Next, we repeated this process, starting with the real segment followed by the fake segments. We used two types of fake segments: TTS generated using Azure Cloud and VC files created with the DiffVC open-source tool. The results illustrated in 5 indicate that the attack is significantly more successful when the file contains 1 to 4 seconds of fake segments. This is an important objective for the PF attack, which focuses on replacing a word or a few words within the real speech.

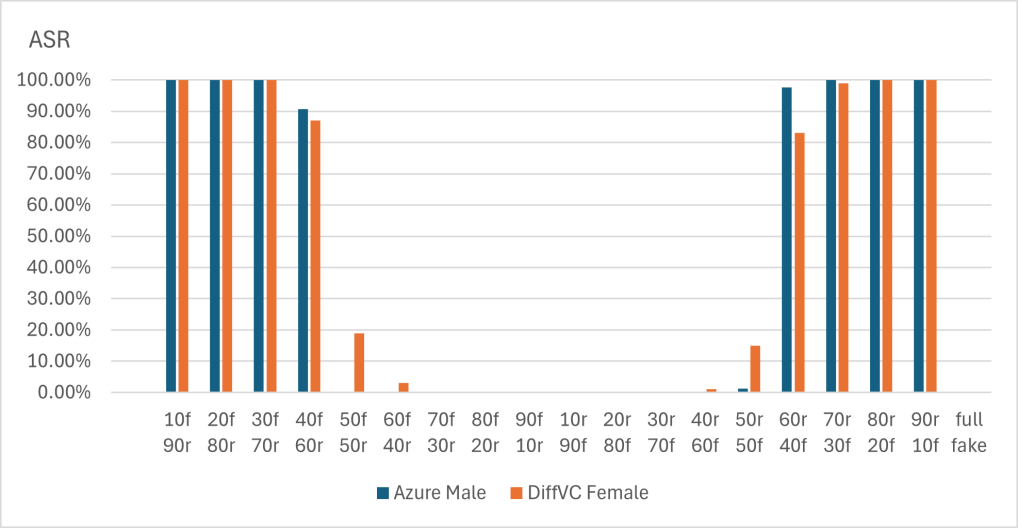

Figure 5. Attack success rate against Resemblyzer (TTS & VC) using 10 seconds PF files.

5.3. PF Attack Against AWS (Voice ID Service)

We carried out an experiment identical to the one in Section 5.2, but this time we used a PF speech attack against the AWS voice ID detection service. Unlike Resemblyzer, there was no need to set a threshold for this experiment. We enrolled all 86 target speakers from Section 5.2 into the AWS voice ID service and used their profiles for all tests. Tables 5 and 6 display the outcomes of the AWS voice ID attack using both TTS and VC while Figure 6 visually compares the results of all TTS and VC generation methods.

Table 5. Attack success rate against AWS voice ID service using AWS, Azure, Google Cloud and Tortoise TTS generation methods.

Generation Method	ASR - PF (TTS-Based) – Fake Segment Location		
	2F, 4R	**4R, 2F**	**2R, 2F, 2R**
AWS (M)	94.44%	94.44%	94.44%
AWS (F)	92.56%	92.56%	92.56%
Azure (M)	93.61%	93.61%	93.61%
Google (M)	97.6%	97.6%	97.6%
Google (F)	94.76%	94.76%	94.76%
Tortoise (M)	36.62%	36.62%	36.62%
Tortoise (F)	42.03%	42.03%	42.03%

Table 6. Attack success rate against AWS voice ID service using FreeVC, DiffVC, and ppgVC open-source VC tools.

Generation Method	ASR - PF (VC-Based) – Fake Segment Location		
	2F, 4R	**4R, 2F**	**2R, 2F, 2R**
FreeVC (M)	71.65%	71.65%	71.65%
FreeVC (F)	59.60%	59.60%	59.60%
DiffVC (M)	94.44%	94.44%	94.44%
DiffVC (F)	91.50%	91.50%	91.50%
ppgVC (M)	88.89%	88.89%	88.89%

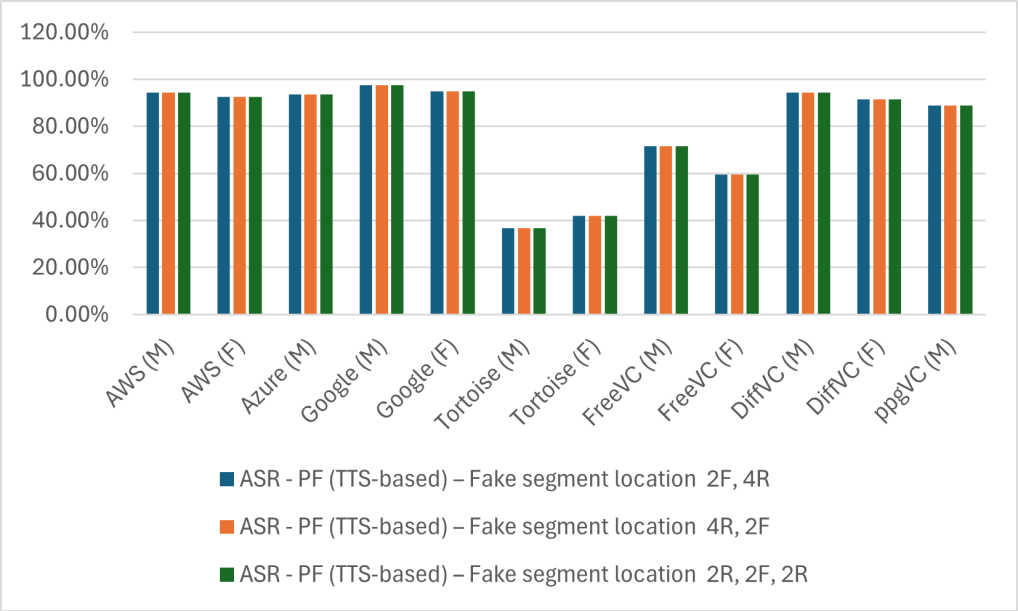

Figure 6. Attack success rate against AWS voice ID (TTS & VC).

5.3.1. Impact of Fake Segment Location

We used the same PF speeches that we used in Section 5.2 experiment to determine whether the location of the fake segment would affect the detection process. the location of the fake segment within the audio file does not have any impact on the ASR against AWS voice ID detection.

5.3.2. Impact of Synthetic Speech Generation Methods

We discovered from Table 5 results that TTS audio generated by AWS, Azure, and Google clouds was more successful in PF speech attack than Tortoise. Among male speakers, Google recorded the highest ASR of 97.6%, while Tortoise recorded the lowest ASR of 36.63%. Moving on to the VC attack listed in Table 6, the highest ASR of 94.44% was recorded by the DiffVC generation for male speakers, and FreeVC female speakers recorded the lowest ASR of 59.60%.

5.3.3. Impact of Gender

We have discovered that there is a slight difference in the accuracy of ASR between males and females who belong to the same TTS and VC generation methods. Tables 5 and 6 show that for males, the ASR accuracy ranged from 36.62% to 97.6% in TTS and from 71.65% to 94.44% in VC. For females in TTS, the ASR ranged from 42.03% to 94.76%, while in VC, it ranged from 59.60% to 91.50%.

5.3.4. Impact of Target Dialect

We found that Scottish male speakers have higher ASR than other speakers for both TTS and VC generation methods. In contrast, Welsh female speakers have the lowest ASR for both TTS and VC. Tables 7 and 8 present the average results for all dialect speakers based on both TTS and VC generation methods.

5.3.5. Impact of the fake segment length

We used the same subset of PF files that we used in section 5.2.6, we found that the length of the fake segment doesn't affect to the AWS Voice ID detection process as the Attack success rate using Azure TTS fake segments was 77.12% despite the fake segment length and that same for DiffVC fake segment with success rate of 91.25%.

Table 7. Attack success rates against AWS voice ID service based on different dialect speakers' real speech and using the Tortoise TTS generation method.

Speaker/Dialect	Gender	TTS: Tortoise		
		2F, 4R	4R, 2F	2R, 2F, 2R
Welsh speakers	Male	42.69%	42.69%	42.69%
Welsh speakers	Female	36.41%	36.41%	36.41%
Scottish speakers	Male	61.58%	61.58%	61.58%
Scottish speakers	Female	43.17%	43.17%	43.17%
Northern speaker	Male	44.62%	44.62%	44.62%
Northern speaker	Female	46.46%	46.46%	46.46%
Midlands speakers	Male	54.62%	54.62%	54.62%
Midlands speakers	Female	37.61%	37.61%	37.61%

Table 8. Attack success rates against AWS voice ID service based on different dialect speakers' real speech and using FreeVC VC generation methods.

Speaker/Dialect	Gender	VC: FreeVC		
		2F, 4R	4R, 2F	2R, 2F, 2R
Welsh speakers	Male	58.17%	58.17%	58.17%
Welsh speakers	Female	42.69%	42.69%	42.69%
Scottish speakers	Male	64.32%	64.32%	64.32%
Scottish speakers	Female	45.97%	45.97%	45.97%
Northern speaker	Female	57.59%	57.59%	57.59%
Northern speaker	Male	44.62%	44.62%	44.62%
Midland speakers	Female	54.23%	54.23%	54.23%
Midland speakers	Male	55.43%	55.43%	55.43%

6. Amazon Alexa (Mobile App)

Finally, we carried out a partial fake speech attack on Alexa, a virtual assistant software that is widely used on Amazon Echo devices and as a mobile app on multiple mobile phone brands. Alexa depends on pre-defined voices to perform various sensitive and secure transactions, such as purchases, calling, and messaging [20]. It also handles other secure elements in smart homes, such as opening the main or garage door, switching off CCTV, and more. Therefore, it is crucial to determine how Alexa responds to partial fake speech.

Participants. We recruited 14 participants for our study through an online crowd-sourcing platform called Prolific (https://www.prolific.co/). All the participants self-identified as native English speakers and were currently residing either in the United Kingdom or the United States. Half of the participants identified themselves as female, while the other half identified as male. We only included participants who were 18 years of age or older. The survey was designed to take approximately 5 min to complete, and we compensated the participants with $1 for their time.

Experiment Setup. We evaluated the effectiveness of a DNN-based PF speech attack on Alexa using the attack success rate (ASR). To conduct this evaluation, we asked the participants to read and record six different sentences. Out of these six sentences, four were used for the enrolment of each participant's real voice ***RB*** to Alexa, while the remaining two sentences were used to clone their voice ***RF*** saying the same sentences and then create the PF speech ***RP*** to perform the attack and evaluate Alexa's responses. Then, we used ElevenLabs [15], an online commercial platform, to clone voices and create text-to-speech audio. After that, we replaced the original segments with the fake ones using Audacity. Lastly, we registered each speaker using the four recorded sentences in the Alexa (iPhone) mobile app version and performed the attack by re-playing the two PF sentences listed below:

- Alexa, can you recognise my voice?
- Alexa, who am I?

The first sentence contained ('my voice') as the fake segment, while we used the wake-up word ('Alexa') as the fake segment in the second sentence.

Results. We found that Amazon Alexa was able to identify all 14 participants when they used partial fake speech. The success rate of this attack was 100% since Alexa did not reject any of the partial fake speech. This result reveals a significant issue with Alexa's pre-defined voices, especially when it comes to secure and sensitive transactions.

7. Partial Fake Speech and Human Speech

In this section, the scenario that we will address is as follows: An attacker ***A*** aims to mimic the voice of a target person ***T***. The aim of the attacker is to impersonate ***T*** to a (human) victim ***V***, by creating partial fake speech ***Rp*** that ***V*** will believe is ***T's*** real voice. These attacks can be especially effective if ***V*** is not familiar with ***T's*** voice. For instance, ***A*** could employ partial fake speech to execute phishing attacks, which aim to deceive ***V*** into giving away money or leading them to a specific location where they are then physically attacked.

Attacker ***A*** can thus launch a partial fake speech attack against prominent individuals and companies. This attack can cause severe damage to ***V's*** reputation by spreading fabricated lies via social media. The consequences of such an attack can be negative and long-lasting. It can take significant effort to clarify the truth to the audience and minimise the damage caused by the attack.

7.1. Methodology and Key Findings

Our research aims to investigate the effect of partial fake speech on human listeners in two different scenarios. Firstly, we aim to understand how easily human listeners can be deceived by partial fake speech when it is presented as an audio-only stimulus. Secondly, we are interested in how they respond to partial fake speech when it is presented within a completely real video. To achieve these objectives, we have developed experimental protocols and procedures for the study, which have been carefully evaluated and approved by our local ethics committee.

User Study A (Online Survey: PF as audio-only). We conducted an experiment to determine whether humans can differentiate between real, fake, and, in particular, partially fake speech. We categorised the voices into two types: familiar voices (i.e., famous voices) and unfamiliar voices. Participants were asked to listen to 10 audio recordings and determine whether each recording contained real, fake, or partially fake speech. Four of the audio clips belonged to Donald Trump, whose voice is assumed to be familiar to the participants.

Findings. In this study, participants had difficulty identifying partially fake speech, regardless of familiarity with the speaker. As shown in Table 9, the PF clip accuracy of identifying familiar voice ranged from 11% and 24% while for unfamiliar voices ranged from 14% to 18%.

User Study B (Online Survey: PF within real video). For this survey, we asked the participants to watch six videos. All of the videos were real, but we manipulated the audio. As with User Study A, we gave the participants three options for each video: real, fake, and partially fake. Our goal in conducting this survey was to determine the participants' ability to identify partially fake audio in two different scenarios: when the speakers were in front of the camera and their lips were clearly visible, and when the speakers were far from the camera.

Findings. We found that in both cases, when the speaker was in front of the camera and when the speaker was far from the camera, participants were unable to identify partial fake audio. As presented in Table 10, the accuracy when the speaker's face is close to the camera ranged from 17% to 31%, while when the speaker's face is far from the camera the accuracy ranged from 11% to 23%.

Table 9. Average accuracy of participants' responses for 10 audio clips.

Audio Type	Familiar Voice?	Accuracy of Participant Responses		
		Real	Fake	Partial Fake
PF	No	34%	48%	18%
PF	No	12%	74%	14%
PF	Yes	73%	16%	11%
PF	Yes	63%	13%	24%
Fake	No	77%	17%	6%
Fake	yes	11%	85%	4%
Fake	No	36%	42%	22%
Fake	Yes	19%	54%	27%
Real	Yes	76%	12%	12%
Real	No	53%	23%	24%

Table 10. Participants' average responses to six video questions.

Embedded Audi	Face Distance of Camera	Participants Responses Accuracy		
		Real	Fake	Partial Fake
PF	far	85%	4%	11%
PF	far	63%	14%	23%
PF	front	77%	6%	17%
PF	front	52%	17%	31%
Real	front	69%	10%	21%
Fake	Front/far	60%	13%	27%

7.2. User Study A: Can Users Identify Partial Fake Speech (as Audio-Only)?

Our experiment aims to investigate whether human listeners can differentiate between real and partially fake audio for both familiar and unfamiliar speakers. To achieve this, we have designed the following survey:

Participants. We enrolled 148 participants through the online crowd-sourcing platform Prolific (https://www.prolific.co/). All participants identified themselves as native English speakers and were currently residing in either the United Kingdom or the United States. Half of the participants identified as female, while the other half identified as male. All the participants were aged 18 years or older. The survey was designed to take an average of six minutes to complete, and participants were compensated with $1 for their time.

Procedure. The participants were given an online survey that consisted of 10 audio and six video clips. The total duration of all audio and video recordings was approximately three minutes. The audio files contained speech that was real, fake, and partially fake. Six audio files were from unknown speakers, and the other four used Donald Trump's voice. The participants were required to listen to these audio files and watch the six videos before submitting their answers.

We used the Tortoise TTS zero-shot model to create partial fake speech segments. To generate the cloning voice of each speaker, we only used a six-second real speech recording. Additionally, we manually created the partial fake speech by replacing a word or two within a real speech. This manual editing of the audio was done to make the transition between real and fake segments smoother and to create a more challenging speech for participants to distinguish. The audio sources included in this survey are either custom recordings made for this survey or audio extracted from YouTube 8M videos.

Result. We aim to provide answers to the questions listed below:

(1) How accurately can participants distinguish the partial fake speeches? As we can see in Figure 7 and from the detailed results presented in Table 9, it is extremely difficult for humans to identify partially fake speech from both familiar and unfamiliar voices. The results show that most of participants cannot identify PF speech correctly: more than 75% of the participants were fooled by partially fake speech.

(2) Does listening to multiple samples from familiar speakers increase the ability to distinguish partial fake speech? We have included four audio clips of Donald Trump's voice in this study. These clips fall into four categories: entirely real, entirely fake, and two examples of partial fake (PF) speech. Our survey results highlight the difficulty humans face in identifying PF audio. As no previous studies have been conducted on partial fake audio, we compared our results to a user study [26] that focused on entirely fake audio. The results of that study showed that participants identified entirely fake voices 50% of the time for unfamiliar speakers and 80% of the time for familiar speakers. In our study, participants identified entirely fake voices for unfamiliar speakers 29.5% of the time, and for familiar speakers 69.5% of the time. Our study reveals that partial fake speech is a more successful attack than entirely fake audio against humans, with an average accuracy rate of 17.5% for identifying unfamiliar PF speech and 16% for identifying familiar PF speech.

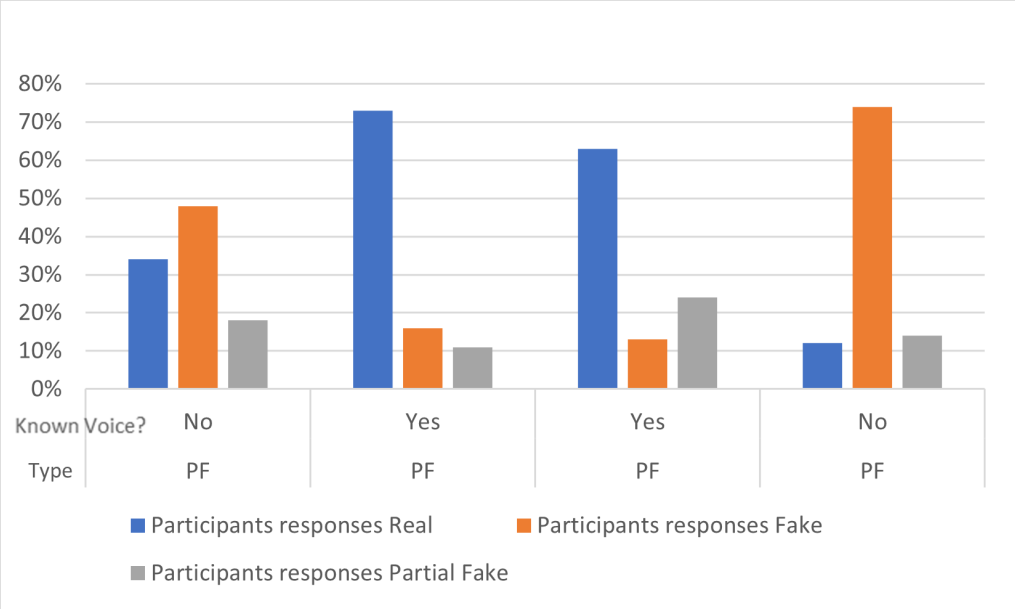

Figure 7. Average response accuracy to 4 PF speech questions.

(3) Does participant gender affect the ability to identifying PF speech? Analysing responses to partial fake questions based on male and female participants, we found that there is no advantage for one gender over another, even for familiar and unfamiliar voices.

7.3. User Study B: Can Users Identify Partial Fake Speech (Within Entirely Real Video)

The objective of the second user study is to determine how humans respond to a new type of deepfake media by embedding partial fake audio within entirely real video. To the best of our knowledge, this is the first time we have created partial fake audio within entirely real video and used it in user studies, as previous works focused on fake video [69–71,76].

Participants. The same participants involved in Section 7.2 answered the questions presented in this section.

Procedure. We have generated six videos sourced from the YouTube-8m dataset. We created and embedded partially faked audio segments into entirely real videos. To create the audio, we used the Tortoise zero-shot model for TTS segments and FreeVC for VC speech. We started by extracting the audio from the videos, then we generated a cloning voice of the actual speaker's voice that we had extracted from the video. Finally, we replaced a word or a few words of the speech with the cloning voice speech to process the audio.

As part of our experiment, we embedded partially fake audio into two types of real videos. The first type involved the speaker being in front of the camera, making it possible

for the viewers to see the speaker's face and lips while they were speaking. The second type involved the speaker being farther away from the camera, making it impossible for the viewers to see the speaker's lips while the fake audio segments were playing. We believe that it will be easier to identify partial fake speech when the camera is in front of the face, which is why we only created one partial fake segment within a video that was recorded with a camera at a distance. The videos feature three male and three female speakers.

Result. We aim to provide answers to the questions listed below:

(1) How accurately can participants distinguish the partial fake speech within an entirely real video? Similar to the experiment presented in Section 7.2, participants found it challenging to identify when audio was partially faked in a real video. Participants were only able to accurately identify partially fake audio between 11% and 31%. Table 10 presents the detailed results of the experiment. In contrast, participants were able to correctly identify entirely real audio and video 69% of the time. However, they were deceived when entirely faked audio was used in an entirely real video. Correct responses to entirely fake speech were as low as 13%. We used FreeVC voice conversion to generate the fake speech, resulting in a real video with a converted voice from a different speaker, while the content of the speech remained the same.

(2) Does seeing the face and lips during a fake segment make a difference in partial fake identification? Table 10 presents the results of the study, which found no significant difference in identifying PF speech within a real video whether the camera was positioned in front of the speaker's face or not. The study showed that when the camera was far from the speaker's face, the accuracy average of identifying PF was only 17%, while the accuracy average improved to 24% when the camera was placed in front of the speaker's face.

(3) Does participant gender affect responses? We found that female participants outperformed male by 19% when the camera in front of speaker face while male outperformed female by 14% when the camera is far from speaker face.

8. Assessing State-of-the-Art Defensive Measures

After considering the success of a speech attack against humans and speech recognition systems, we address the question of how we could efficiently detect such attacks. While searching for open-source defences against PF speech, we came across only one publicly available tool called PartialSpoof [72]. To test the effectiveness of this method against new PF speech data, we evaluated PartialSpoof using the RFP dataset. Additionally, we selected two open-source tools, SSL_Anti-spoofing [73] and M2S-ADD [74], which were created to defend against entirely fake speech and achieved state-of-the-art results. Our aim was to investigate how these tools could detect speech-spoofing attacks. The results of the EER of the three chosen detection models after being trained and evaluated using the RFP dataset are listed in Table 11.

Table 11. EER for three detection models: SSL_Anti-spoofing, M2S-ADD, and PartialSpoof.

Detection Methods	**SSL_Anti-Spoofing**		**M2S-ADD**		**PartialSpoof**	
Dataset	ASVspoof 2021	RFP	ASVspoof 2019	RFP	PS	RFP
EER	0.82%	50.16%	1.34%	18.69%	0.49%	3.70%

8.1. PartialSpoof: Detection of Short Fake Speech Segments

The PartialSpoof Database and Countermeasures paper [72] introduces the PartialSpoof database, which consists of short fake speech segments inserted into authentic utterances. To create these synthetic speech segments, the authors spliced synthetic speech into real recordings. The team developed countermeasures that leverage various feature extraction techniques and machine learning algorithms to differentiate between the real and fake speech segments. The proposed approach aims to detect short, fake speech segments that may be present in real recordings.

The researchers conducted a study that considered the possibility of an attacker embedding a fake segment using various audio segment lengths.

Procedure. The research assumes that fixed audio lengths should be generated for each segment within the audio file to test a dataset using the proposed CM. However, this assumption is an uncommon construction method for creating fake and partially fake audio since the detection tool should detect any fake segment regardless of the segment lengths. To begin with, we reproduce the results similar to those listed in the paper by running the code using the provided trained model and dataset. After that, we created a new PF subset based on the real and entirely fake RFP subsets, resulting in a total of 50,650 PF speech files for this experiment. The location of fake segments is similar to the original PF subset in the RFP dataset, which was explained earlier in Section 5.2 (Table 1). The only difference is that each segment length is equal to 640 ms, which is the segment length that exists in Partial spoof experiments. Finally, we trained the model using our custom data, the 50,650 PF speech files.

Result. After testing the proposed CM using training, validation, and evaluation sets of the RFP dataset, the EER result was 3.70%. While the PartialSpoof proposed CM model achieved a low EER 0.49% when trained on the proposed PS dataset, the detection efficiency was affected when using the RFP dataset. Additionally, in the PartialSpoof Database and Countermeasures paper [72], the authors demonstrated that when evaluating ASVspoof 2019 LA dataset using a pre-trained model based on PartialSpoof dataset, the result was 0.90% EER. We carried out the same test using the RFP evaluation subset, and the result was as high as 64.63% EER. This suggests that the pre-trained model on the PartialSpoof dataset is not capable of dealing with unseen PF data. Furthermore, we believe that the low EER of 0.90% achieved by the proposed CM model on unseen ASVspoof 2019 LA data is due to the identical fake segment generation methods used in the Partialspoof dataset, which the CM model was trained on.

8.2. SSL_Anti-Spoofing

SSL_Anti-spoofing [73] is a novel approach for identifying spoofing attacks and deepfakes in automatic speaker verification. The proposed method leverages Wav2vec 2.0, a self-supervised representation learning model, to extract trustworthy speaker embeddings. Additionally, the method employs data augmentation techniques to enhance the model's ability to accurately detect spoofing attacks and deepfakes. The model yielded a state-of-the-art (SOTA) EER (Equal Error Rate) of 0.82% on the ASVspoof 2021 LA track.

Procedure. To verify the effectiveness of the model and ensure that it produces results similar to those listed in the SSL_Anti-spoofing paper, we first utilised the provided trained model and dataset and ran the code to generate results. Subsequently, we trained the model on the RFP dataset and then evaluated the trained model on the RFP evaluation subset.

Result. The EER for the SSL model trained on the RFP dataset was 50.16% while the EER for the SSL model as trained on the ASVspoof 2021 listed in SSL_Anti_Spoofing paper was 0.82%. Although the model achieved state-of-the-art (SOTA) when using entirely fake audio, it was unable to detect partially fake speech. We believe that the model can be improved to detect both entirely and partially faked speech.

8.3. Betray Oneself: A Novel Audio DeepFake Detection Model (M2S-ADD)

Betray Oneself [74] is a new method for detecting audio deepfakes. The method focuses on identifying deepfakes via the conversion of mono (single-channel) audio into stereo (dual-channel) audio. This conversion aims to expose any discrepancies or artefacts that may have been introduced during the deepfake generation process. The researchers tested their proposed model (M2S-ADD) on the ASVspoof 2019 dataset and achieved better results than all listed baselines in their paper.

Procedure. To ensure that the model is functioning correctly and producing the same outcomes as outlined in the Betray Oneself paper, we initiated the code using the pre-trained model and dataset offered by the authors. After that, we trained the model on the RFP dataset.

Result. The EER obtained by applying the M2S-ADD model on the evaluation set of the RFP dataset was 18.69%. In contrast, the EER achieved in the Betray Oneself paper on the ASVspoof2019 dataset was only 1.34%. Based on these results, we suggest that with further improvement, the M2S-ADD model has the potential to yield better outcomes for PF speech detection.

9. Discussion

The importance of ground-truth data. Our experiments show that even when participants know the speaker's voice, they are fooled by partial fake speech. As a result, it is necessary to develop detection models that efficiently detect fake and partial fake speech, The availability of real speech samples for all target individuals is essential so the attacker can launch fake and partial fake speech attack. At the same time, a detection mechanism having the target individual speech can improve the protection against fake and partially fake speech attack [75]. For example, if an organisation provides a service via a call centre, they have to have a short speech of an individual during the enrollment process. Later, the organisation will have a real recording of all enrolled users, and when someone impersonates any enrolled user, the system can identify the caller based on the voice. All existing audio datasets do not contain both real and fake speech from the same speaker [77].

Various techniques for fake generation speech exist, including TTS and VC, online commercial services, and open-source models. Yet, all detection methods target specific techniques for fake generation speech. We believe there is a significant need for a universal method that can detect all types of fake speech regardless of generation methods, as well as a need for robust detection models that have the ability to detect data that is not included in the training model.

Real-world proactive defense against synthesis attacks. Content creators, educational platforms, and similar media content providers can increase the protection of the voices in their media before publishing them, like the proactive defense approach leveraging adversarial examples to disrupt unauthorized speech synthesis proposed in the Anti-Fake study [75]. That approach ensures that the synthesised DeepFake audio does not sound like the victim's voice to either humans or machines.

10. Conclusions

Our study aims to investigate the potential threat posed by partially fake speech generated through deep learning techniques. Our findings indicate that publicly available fake speech generation methods are capable of deceiving both humans and machines. Furthermore, we observed that existing defenses against completely fake speech are less effective against partially fake speech. Therefore, our study highlights the urgent need for new defense mechanisms against partially fake speech attacks and calls for further exploration of the associated challenges and opportunities. Our study provides a reliable benchmark for future research in this area.

Funding:

Data Availability Statement:

Conflicts of Interest:

References

1. Boone, D.R. *Is Your Voice Telling on You?: How to Find and Use Your Natural Voice*; Plural Publishing: San Diego, CA, USA, 2015.
2. Kumar, P.; Jakhanwal, N.; Bhowmick, A.; Chandra, M. Gender classification using pitch and formants. In Proceedings of the 2011 International Conference on Communication, Computing & Security, Rourkela Odisha, India, 12–14 February 2011; pp. 319–324.
3. Casanova, E.; Weber, J.; Shulby, C.D.; Junior, A.C.; Gölge, E.; Ponti, M.A. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2022; pp. 2709–2720.
4. Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv* **2023**, arXiv:2301.02111.

5. Ngoc, P.P.; Quang, C.T.; Chi, M.L. Adapt-Tts: High-Quality Zero-Shot Multi-Speaker Text-to-Speech Adaptive-Based for Vietnamese. *J. Comput. Sci. Cybern.* **2023**, *39*, 159–173.
6. Saeki, T.; Maiti, S.; Li, X.; Watanabe, S.; Takamichi, S.; Saruwatari, H. Learning to speak from text: Zero-shot multilingual text-to-speech with unsupervised text pretraining. *arXiv* **2023**, arXiv:2301.12596.
7. Peng, K.; Ping, W.; Song, Z.; Zhao, K. Non-autoregressive neural text-to-speech. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2020; pp. 7586–7598.
8. Kawamura, M.; Shirahata, Y.; Yamamoto, R.; Tachibana, K. Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time fourier transform. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
9. Betker, J. Better speech synthesis through scaling. *arXiv* **2023**, arXiv:2305.07243.
10. Jiang, Z.; Liu, J.; Ren, Y.; He, J.; Zhang, C.; Ye, Z.; Wei, P.; Wang, C.; Yin, X.; Ma, Z. Mega-tts 2: Zero-shot text-to-speech with arbitrary length speech prompts. *arXiv* **2023**, arXiv:2307.07218.
11. Jing, X.; Chang, Y.; Yang, Z.; Xie, J.; Triantafyllopoulos, A.; Schuller, B.W. U-DiT TTS: U-Diffusion Vision Transformer for Text-to-Speech. In Proceedings of the Speech Communication; 15th ITG Conference, Aachen, Germany, 20–22 September 2023; VDE: Offenbach, Germany, 2023; pp. 56–60.
12. Ning, Z.; Xie, Q.; Zhu, P.; Wang, Z.; Xue, L.; Yao, J.; Xie, L.; Bi, M. Expressive-vc: Highly expressive voice conversion with attention fusion of bottleneck and perturbation features. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
13. Meftah, A.H.; Alashban, A.A.; Alotaibi, Y.A.; Selouani, S.A. English emotional voice conversion using StarGAN model. *IEEE Access* **2023**, *11*, 67835–67849.
14. Fernandez-Martín, C.; Colomer, A.; Panariello, C.; Naranjo, V. Choosing only the best voice imitators: Top-K many-to-many voice conversion with StarGAN. *Speech Commun.* **2024**, *156*, 103022.
15. ElevenLabs. 2024 Available online: https://elevenlabs.io/ (accessed on).
16. Descript. 2024. Available online: https://www.descript.com/ (accessed on).
17. HSBC Voice ID. 2024. Available online: https://ciiom.hsbc.com/ways-to-bank/phone-banking/voice-id/ (accessed on).
18. WeChat VoicePrint. 2024. Available online: https://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/ (accessed on).
19. Alipay Sound Wave Payment. 2024. Available online: https://opendocs.alipay.com/open/140/104601 (accessed on).
20. What Is Alexa Voice ID. 2024. Available online: https://www.amazon.co.uk/gp/help/customer/display.html?nodeId=GYCXKY2AB2QWZT2X (accessed on).
21. Klepper, D.; Swenson, A. AI-Generated Disinformation Poses Threat of Misleading Voters in 2024 Election. PBS NewsHour. 2023. Available online: https://www.pbs.org/newshour/politics/ai-generated-disinformation-poses-threat-of-misleading-voters-in-2024-election/ (accessed on).
22. Rudy, E. Don't Watch Sinister Taylor Swift Video or Risk Bank-Emptying Attack That Just Takes Seconds. The Sun. 2024. Available online: https://www.thesun.co.uk/tech/25342162/taylor-swift-fans-ai-attack-dangerous-video/ (accessed on).
23. Stupp, C. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. WSJ. 2019. Available online: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402/ (accessed on).
24. Brewster, T. Fraudsters Cloned Company Director's Voice in $35 Million Heist, Police Find. Forbes. 2023. Available online: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/ (accessed on).
25. Karimi, F. Mom, These Bad Men Have Me': She Believes Scammers Cloned Her Daughter's Voice in a Fake Kidnapping. CNN. 2023. Available online: https://edition.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html/ (accessed on).
26. Wenger, E.; Bronckers, M.; Cianfarani, C.; Cryan, J.; Sha, A.; Zheng, H.; Zhao, B.Y. "Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, 15–19 November 2021; pp. 235–251.
27. Bilika, D.; Michopoulou, N.; Alepis, E.; Patsakis, C. Hello me, meet the real me: Voice synthesis attacks on voice assistants. *Comput. Secur.* **2024**, *137*, 103617.
28. Gao, Y.; Lian, J.; Raj, B.; Singh, R. Detection and evaluation of human and machine generated speech in spoofing attacks on automatic speaker verification systems. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 544–551.
29. Simmons, D. BBC Fools HSBC Voice Recognition Security System. 2017. Available online: https://www.bbc.co.uk/news/technology-39965545/ (accessed on).
30. Mehrish, A.; Majumder, N.; Bharadwaj, R.; Mihalcea, R.; Poria, S. A review of deep learning techniques for speech processing. *Inf. Fusion* **2023**, *99*, 101869.
31. Tan, C.B.; Hijazi, M.H.A.; Khamis, N.; Nohuddin, P.N.E.B.; Zainol, Z.; Coenen, F.; Gani, A. A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction. *Multimed. Tools Appl.* **2021**, *80*, 32725–32762.

32. Tayebi Arasteh, S.; Weise, T.; Schuster, M.; Noeth, E.; Maier, A.; Yang, S.H. The effect of speech pathology on automatic speaker verification: A large-scale study. *Sci. Rep.* **2023**, *13*, 20476.
33. Mandasari, M.I.; McLaren, M.; van Leeuwen, D.A. The effect of noise on modern automatic speaker recognition systems. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 4249–4252.
34. Radha, K.; Bansal, M.; Pachori, R.B. Speech and speaker recognition using raw waveform modeling for adult and children's speech: A comprehensive review. *Eng. Appl. Artif. Intell.* **2024**, *131*, 107661.
35. Zhang, W.; Yeh, C.C.; Beckman, W.; Raitio, T.; Rasipuram, R.; Golipour, L.; Winarsky, D. Audiobook synthesis with long-form neural text-to-speech. In Proceedings of the 12th Speech Synthesis Workshop (SSW) 2023, Grenoble, France, 26–28 August 2023.
36. Kim, M.; Jeong, M.; Choi, B.J.; Kim, S.; Lee, J.Y.; Kim, N.S. Utilizing Neural Transducers for Two-Stage Text-to-Speech via Semantic Token Prediction. *arXiv* **2024**, arXiv:2401.01498.
37. Vecino, B.T.; Gabrys, A.; Matwicki, D.; Pomirski, A.; Iddon, T.; Cotescu, M.; Lorenzo-Trueba, J. Lightweight End-to-end Text-to-speech Synthesis for low resource on-device applications. In Proceedings of the 12th ISCA Speech Synthesis Workshop (SSW2023), Grenoble, France, 26–28 August 2023; pp. 225–229. https://doi.org/10.21437/SSW.2023-35.
38. Donahue, J.; Dieleman, S.; Bińkowski, M.; Elsen, E.; Simonyan, K. End-to-end adversarial text-to-speech. *arXiv* **2020**, arXiv:2006.03575.
39. Tiomkin, S.; Malah, D.; Shechtman, S.; Kons, Z. A Hybrid Text-to-Speech System That Combines Concatenative and Statistical Synthesis Units. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 1278–1288. https://doi.org/10.1109/TASL.2010.2089679.
40. Zhang, M.; Zhou, Y.; Zhao, L.; Li, H. Transfer learning from speech synthesis to voice conversion with non-parallel training data. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1290–1302.
41. Yoon, H.; Kim, C.; Um, S.; Yoon, H.W.; Kang, H.G. SC-CNN: Effective speaker conditioning method for zero-shot multi-speaker text-to-speech systems. *IEEE Signal Process. Lett.* **2023**, *30*, 593–597.
42. Lian, J.; Zhang, C.; Yu, D. Robust disentangled variational speech representation learning for zero-shot voice conversion. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6572–6576.
43. Wang, Z.; Chen, Y.; Xie, L.; Tian, Q.; Wang, Y. Lm-vc: Zero-shot voice conversion via speech generation based on language models. *IEEE Signal Process. Lett.* **2023**, *30*, 1157–1161.
44. Wu, Y.; Tan, X.; Li, B.; He, L.; Zhao, S.; Song, R.; Qin, T.; Liu, T.Y. Adaspeech 4: Adaptive text to speech in zero-shot scenarios. *arXiv* **2022**, arXiv:2204.00436.
45. Google Cloud Text-to-Speech API Now Supports Custom Voices. 2017. Available online: https://cloud.google.com/blog/products/ai-machine-learning/create-custom-voices-with-google-cloud-text-to-speech/ (accessed on).
46. Real-Time State-of-the-art Speech Synthesis for Tensorflow 2. 2021. Available online: https://github.com/TensorSpeech/TensorFlowTTS/ (accessed on).
47. An Open Source Text-to-Speech System Built by Inverting Whisper. 2024. Available online: https://github.com/collabora/WhisperSpeech/ (accessed on).
48. Van Niekerk, B.; Carbonneau, M.A.; Zaïdi, J.; Baas, M.; Seuté, H.; Kamper, H. A comparison of discrete and soft speech units for improved voice conversion. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6562–6566.
49. Li, J.; Tu, W.; Xiao, L. Freevc: Towards high-quality text-free one-shot voice conversion. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
50. Liu, S.; Cao, Y.; Wang, D.; Wu, X.; Liu, X.; Meng, H. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1717–1728.
51. Yi, J.; Tao, J.; Fu, R.; Yan, X.; Wang, C.; Wang, T.; Zhang, C.Y.; Zhang, X.; Zhao, Y.; Ren, Y.; et al. Add 2023: The second audio deepfake detection challenge. *arXiv* **2023**, arXiv:2305.13774.
52. Yi, J.; Fu, R.; Tao, J.; Nie, S.; Ma, H.; Wang, C.; Wang, T.; Tian, Z.; Bai, Y.; Fan, C.; et al. Add 2022: The first audio deep synthesis detection challenge. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 9216–9220.
53. Martín-Doñas, J.M.; Álvarez, A. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 9241–9245.
54. Li, J.; Li, L.; Luo, M.; Wang, X.; Qiao, S.; Zhou, Y. Multi-grained Backend Fusion for Manipulation Region Location of Partially Fake Audio. In Proceedings of the DADA@ IJCAI, Macau, China, 19 August 2023; pp. 43–48.
55. Cai, Z.; Wang, W.; Wang, Y.; Li, M. The DKU-DUKEECE System for the Manipulation Region Location Task of ADD 2023. *arXiv* **2023**, arXiv:2308.10281.
56. Liu, J.; Su, Z.; Huang, H.; Wan, C.; Wang, Q.; Hong, J.; Tang, B.; Zhu, F. TranssionADD: A multi-frame reinforcement based sequence tagging model for audio deepfake detection. *arXiv* **2023**, arXiv:2306.15212.
57. Ryan, P. Deepfake' Audio Evidence Used in UK Court to Discredit Dubai Dad. The National. 2020. Available online: https://www.thenationalnews.com/uae/courts/deepfake-audio-evidence-used-in-uk-court-to-discredit-dubai-dad-1.975764/ (accessed on).

58. Zhang, Z.; Zhou, L.; Wang, C.; Chen, S.; Wu, Y.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv* **2023**, arXiv:2303.03926.
59. Casanova, E.; Shulby, C.; Gölge, E.; Müller, N.M.; De Oliveira, F.S.; Junior, A.C.; Soares, A.d.S.; Aluisio, S.M.; Ponti, M.A. SC-GlowTTS: An efficient zero-shot multi-speaker text-to-speech model. *arXiv* **2021**, arXiv:2104.05557.
60. Li, Y.A.; Han, C.; Mesgarani, N. Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models. In Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 9–12 January 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 920–927.
61. Kang, W.; Hasegawa-Johnson, M.; Roy, D. End-to-End Zero-Shot Voice Conversion with Location-Variable Convolutions. *arXiv* **2022**, arXiv:2205.09784.
62. Resemblyzer. 2023. Available online: https://github.com/resemble-ai/Resemblyzer/ (accessed on).
63. Jia, Y.; Zhang, Y.; Weiss, R.; Wang, Q.; Shen, J.; Ren, F.; Nguyen, P.; Pang, R.; Lopez Moreno, I.; Wu, Y.; et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
64. Amazon Connect Voice ID. 2023. Available online: https://aws.amazon.com/connect/voice-id/ (accessed on).
65. AlAli, A.; Theodorakopoulos, G. An RFP dataset for Real, Fake, and Partially fake audio detection. In Proceedings of the 1International Conference on Cyber Security, Privacy in Communication Networks, 9-10 December 2023; pp. 1–15.
66. Demirsahin, I.; Kjartansson, O.; Gutkin, A.; Rivera, C. Open-source Multi-speaker Corpora of the English Accents in the British Isles. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), Marseille, France, 1 May 2020; pp. 6532–6541.
67. Yamagishi, J.; Veaux, C.; MacDonald, K. *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92)*; The Centre for Speech Technology Research (CSTR), University of Edinburgh: Edinburgh, UK, 2019.
68. Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. *arXiv* **2016**, arXiv:1609.08675.
69. Köbis, N.C.; Doležalová, B.; Soraperra, I. Fooled twice: People cannot detect deepfakes but think they can. *Iscience* **2021**, *24*, 103364.
70. Kaate, I.; Salminen, J.; Jung, S.G.; Almerekhi, H.; Jansen, B.J. How do users perceive deepfake personas? Investigating the deepfake user perception and its implications for human-computer interaction. In Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter, Torino, Italy, 20–22 September 2023; pp. 1–12.
71. Ahmed, M.F.B.; Miah, M.S.U.; Bhowmik, A.; Sulaiman, J.B. Awareness to Deepfake: A resistance mechanism to Deepfake. In Proceedings of the 2021 International Congress of Advanced Technology and Engineering (ICOTEN), Taiz, Yemen, 4–5 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–5.
72. Zhang, L.; Wang, X.; Cooper, E.; Evans, N.; Yamagishi, J. The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *31*, 813–825.
73. Tak, H.; Todisco, M.; Wang, X.; Jung, J.w.; Yamagishi, J.; Evans, N. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv* **2022**, arXiv:2202.12233.
74. Liu, R.; Zhang, J.; Gao, G.; Li, H. Betray Oneself: A Novel Audio DeepFake Detection Model via Mono-to-Stereo Conversion. *arXiv* **2023**, arXiv:2305.16353.
75. Yu, Z.; Zhai, S.; Zhang, N. Antifake: Using adversarial audio to prevent unauthorized speech synthesis. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, Copenhagen, Denmark, 26–30 November 2023; pp. 460–474.
76. Hashmi, A., Shahzad, S.A., Lin, C.W., Tsao, Y. and Wang, H.M Understanding Audiovisual Deepfake Detection: Techniques, Challenges, Human Factors and Perceptual Insights *arXiv* **2024**, arXiv:2411.07650
77. Akhtar, Z., Pendyala, T.L. and Athmakuri, V.S Video and audio deepfake datasets and open issues in deepfake technology: being ahead of the curve Forensic Sciences, 4(3), pp.289-377
78. Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M., and Wei, J. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. *arXiv* **2021**, arXiv:2109.13821.