Audio deepfakes: A survey

# Zahra Khanjani; Gabrielle Watson; Vandana P. Janeja

## 2023

This section first presents the critical discussion, analysis, and summarization regarding the compiled works focusing on audio deepfake generation

# Scholarcy Highlights

* We found that generative adversarial networks (GANs), convolutional neural networks (CNNs), and deep neural networks (DNNs) are common ways of creating and detecting deepfakes
* We found that for text deepfakes, there are more generation methods but very few robust methods for detection, including fake news detection, which has become a controversial area of research because of the potential heavy overlaps with human generation of fake content
* Text-to-speech takes text as an input; when a synthesis request is sent to TTS, a voice should be specified to speak the words; we can say that the SS-TTS models have been trained using audio samples of actual speeches
* Due to the importance of choosing a proper dataset, we briefly introduce the popular datasets in English that are highly used in audio deepfake detection: 1. ASVspoof datasets:
* This section first presents the critical discussion, analysis, and summarization regarding the compiled works focusing on audio deepfake generation

# Scholarcy Summary

## Preliminaries

To deeply understand different categories of deepfakes, their attacks, and their detection methods, we need to know some of the concepts that are the basis of deepfake technology.

These concepts include understanding different networks as well as some necessary foundational definitions.

## Networks used in deepfake generation and detection

Deepfakes are generated using combinations of four typical networks: encoder–decoder networks (ED), convolutional neural networks (CNN), generative adversarial networks (GAN), and recurrent neural networks (RNN).

Like other ANNs, there are hidden neurons in the first hidden layer that receive the input vector Xt and calculate the linear combination of the individual components, in addition to performing a nonlinear transformation (F, activation function).

There is a shortcoming of the equation above: in the early phase of the training, when the generative network has not generated enough and proper fake data, and the fake data is significantly different from the real data, the discrimination network refuses the fake samples with high probability.

The purpose of these figures is to provide a quick and summarized look at different audio deepfake framework architectures

These frameworks’ diagrams are color-coded, so orange and green refer to fake and real, respectively, blue means using neural networks.

We describe each of these types of deepfakes and ways by which they can be detected and created

## Deepfake categories

Deepfakes are categorized as audio, text, video, and image deepfakes.

For each category, related papers are surveyed and the technology trends and frameworks are briefly discussed.

Audio deepfakes have been ignored in the surveys related to deepfakes.

To the best of our knowledge, this paper is the first survey focusing on generating and detecting audio deepfakes.

In the audio deepfake section, we discuss some important frameworks in detail and provide readers with sufficient guidance for audio deepfake tools, some of which are shown in Supplementary Table 1 in the Appendix.

## Non-AI generated

Replay attacks are defined as replaying the recording of a target speaker’s voice.

The two subtypes are far field detection and cut and paste detection attacks (Pradhan et al, 2019).

In far field detection replay attacks, the test segment is a far field microphone recording of the victim that has been

## Audio deepfakes

Speech synthesis is the artificial speech that may be created by different technologies such as an audio deepfake.

Defense: Some advantages of this category are that to defend against replay attacks one can use text dependent speaker verification (Villalba and Lleida, 2011).

Some of the replay attack detection systems have been proposed by working on the features which are fed into the network (Witkowski et al, 2017).

Before the ASVspoof Challenge 2017 (Kinnunen et al, 2017; Lavrentyeva et al, 2017), there were only a couple of research papers done on replay attack, and after this challenge, more approaches for this attack were researched (Tom et al, 2018; Pradhan et al, 2019).

It means that the performance of the detection technique was really better than the previous ones; the best EER was 12% in the development set and 2.76% on the evaluation as stated in other literature (Tom et al, 2018)

## AI-generated audio fakes

Speech synthesis is one of the most important audio deepfake principles, and defined as artificially producing human speech by means of software or hardware system programs.

It can copy a voice fast, be adapted quickly to create what the creators want the phrase to be, and is language-agnostic

It can be used in the radio industry, traffic reports with auto voice overs, and streaming news bulletin systems, and the options are endless (Descript, n.d.).

Spoken languages that do not have a standardized writing system make it hard to make a good speech synthesizer and linguistic components not available in all languages of the world make it hard (Kuligowska et al, 2018)

Another disadvantage is that SS systems do not recognize periods or special characters (Kuligowska et al, 2018).

One can sometimes tell it is not human-like because there is no breathing, laughter, pauses, and sighs among other things in human speech (Kuligowska et al., 2018)

## Speech synthesis (Text-to-speech)

Audio deepfake includes text-to-speech (TTS), which analyzes the text and makes the speech sound in line with text inputted using the rules of linguistic description of the text.

Text-to-speech takes text as an input; when a synthesis request is sent to TTS, a voice should be specified to speak the words; we can say that the SS-TTS models have been trained using audio samples of actual speeches.

WaveNet. The input text is presented to causal convolutions, the output comes to dilated convolutional layers, and goes to gated activation units, and the activation functions.

Using neural network text-to-speech synthesis can make the speech audio in the voice of many speakers even those not in the training

This only needed 5 s (Jia et al, 2019).

The output of this phase goes to the decoder which contains: prenet, attention blocks, causal convolutions, a fully-connected layer, and a binary final frame prediction

It can use one of the existing vocoders for producing audio (WORLD, Griffin-lim and Wavenet).

## Voice conversion and impersonation

The last subcategory of audio deepfakes is voice conversion, which takes the speech signal by the first speaker, the source, and modifies it to sound like it was spoken by the second speaker, i.e., the target speaker.

Used a neural network framework to impersonate voices from different genders well with reconstructing time domain signals with the Griffin Lim method

This led to the model creating very convincing samples of impersonated speech.

A singing voice conversion (SVC) framework using GAN is proposed (Sisman et al, 2019)

They tried to convert a source singer’s voice to sound like that of the target singer, without changing the lyrical content with the use of a GAN-based model (Sisman et al, 2019).

Using text-to-speech networks in the structure of voice conversion may generate high-quality audio like ASSEM-VC (Kim et al, 2022), which is the state-of-the-art voice conversion system in terms of its naturalness.

Some noise in the input speech may corrupt the alignment estimated using Cotatron

## Text deepfake

The text deepfake field is teeming with papers and techniques to create deepfakes; detection methods are catching up but not fast enough.

One of the subcategories of a textual deepfake is exposed fabrications, which are those that are being fraudulently reported, like tabloids and yellow press that use sensationalism and eye-catching headlines to get more profit/traffic.

The video was not proven to be fake by experts

This can make people disbelieve true facts because it is uncomfortable.

A benefit of reenactment deepfakes is if one can not dance, one can transpose a dancer’s moves onto ones own prerecorded video to look like one can dance (Chan et al, 2019).

The story was disproved by experts so it was retracted from the news sites (NBC News, 2019)

If these hoaxes are not fact checked by mainstream news outlets or those spreading information on social media, it can be hard to know what information out there is fake.

There are some tools coming out that will help journalists and front line workers fact check images, such as Google Assembler (Assembler, n.d.)

## Video deepfake

Video editing has been around since 1997 like in the movie Forrest Gump to digitally put in archival footage of JFK and manipulate his mouth movements (O’Sullivan, 2019).

Writing in new laws into policy can take a while and may not catch up to how fast deepfake technology is changing

These laws can protect people impersonating others that could ruin their reputation like pornographic deepfakes, which is 96% of deepfake videos according to research by DeeptraceLabs (Ajder et al, 2019).

An app called Zao (Murphy and Huang, 2019) has become very popular less skilled users can faceswap their bodies of movie stars and put themselves into well-known movies and TV clips.

It was the most downloaded app among the Chinese apps over the weekend of 30 August.

## Image deepfakes

The last category discussed for deepfake technology is image deepfakes. Faceswap: One of the subcategories of image deepfakes is faceswap.

Image synthesis can allow someone to make a new AI-generated image for personal reasons or for entertainment.

Another advantage is that neural textures can allow one to resynthesize new views of static objects and edit the scene along with re-rendering dynamic animated surfaces.

The results are very realistic and it uses data-driven unconditional generative image modeling (Karras et al, 2020).

This opens an opportunity for artists because of all the different image variations can bring their ideas to the forefront.

FaceApp is a newer mobile application that allows one to alter the age, smile, and change

## Audio deepfake datasets

Dataset has a significant impact on the performance as well as generalizability of an audio deepfake detection system.

Due to the importance of choosing a proper dataset, we briefly introduce the popular datasets in English that are highly used in audio deepfake detection: 1.

ASVspoof 2019: The ASVspoof 2019 (Wang X. et al., 2020) edition is the first audio spoof detection challenge that considered all three spoofing attack types.

They have separated the existing scenarios as follows: Spoofing attacks within a logical access (LA) scenario generated with the latest TTS-SS and VC technologies.

Replay spoofing attacks within a physical access (PA) scenario

This dataset is useful in terms of peforming different types of analysis based on the different types of attacks.

This scenario is similar to the LA task, but there is no speaker verification

## Fake or real dataset

FoR dataset (Reimao and Tzerpos, 2019) is a new dataset which contains multiple versions: version one is original synthesized files.

The last version is a rerecorded version of the third one

This type of rerecording can allow for testing scenarios where speech is received over a phone call or voice message (Reimao and Tzerpos, 2019).

FoR dataset uses some high quality text-to-speech algorithms such as deep voice 3 (Ping et al, 2018) and Google wavenet (Oord et al, 2016).

In this dataset no VC algorithm is used

## WaveFake dataset

The dataset (Frank and Schönherr, 2021) consists of 117,985 generated audio clips (196 h total time).

This dataset includes both English and Japanese samples.

This dataset does not include any VC algorithms.

One advantage of this dataset is that they have used various state-of-the-art text-tospeech algorithms

## Discussion and future directions

This section first presents the critical discussion, analysis, and summarization regarding the compiled works focusing on audio deepfake generation.

A summarization of the current techniques as well as future directions against deepfake is presented.

## Deepfake generation

The most significant aspect is how believable it is to the victim, that means “deepfake quality.”.

Performance, the Mean Opinion Score (MOS) of the generated audio is better when the framework is trained using single speaker datasets (Oord et al, 2016; Ping et al, 2018; Kumar et al, 2019; Kong et al, 2020).

Parallel VC datasets refer to the datasets with utterances of the same linguistic content, but uttered by different people (Zhang J.-X. et al, 2020), but in practice, non-parallel VC which is more challenging is needed

It seems that more work should be done regarding audio deepfake frameworks using non-parallel data, and in this way, researchers may use image deepfake frameworks as the base of their proposed framework.

CycleGAN and StarGAN are two frameworks for image deepfake generation that are used as the base of two audio deepfake frameworks (Fang et al, 2018; Kameoka et al, 2018), which can work with

## Future defense against audio fakes

Researchers demonstrate that deepfake generation methods are more powerful and faster-developing than the prevention, mitigation and detection methods.

It is astute to have these tools open source so there is more research generated about these topics and collaboration but, on the flipside, it might be better to keep some detection tools proprietary only to people who need it like fact checkers for reporters.

This is so those making the generation models, perhaps for nefarious purposes, would not know exactly what features make it easier to detect a deepfake like, for example, someone pointed out that deepfakes do not blink well (Li et al, 2018).

All authors contributed to the article and approved the submitted version

## Findings

27% of respondents from the survey say they do not factcheck articles they share.

These laws can protect people impersonating others that could ruin their reputation like pornographic deepfakes, which is 96% of deepfake videos according to research by DeeptraceLabs (Ajder et al, 2019)

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary material of their affiliated organizations, or those of the publisher, the editors and the reviewers.

Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/