# Audio deepfakes : A survey

## Introduction

* Being more prevalent and impacting society in various ways of deepfakes and other internet-based misinformation
* Uses in political campaigns of deepfakes and the trend might continue
* Deepfakes for fraud and exponentially increasing attention of the importance and impact of deepfakes
* Four broad categories of deepfakes : Audio, Text, Video, Image deepfakes
* The aims of this article : Recent trends in each of the deepfake categories and shortcomings of defenses against them, A guide to generation and detection of audio deepfake architectures, The countermeasures and future research directions in this field
* The recent trend in Section 2, Systematic review in Section3, Datasets in Section 4, Intuition of audio generative networks in Section 5, Discussion and future direction in Section 6, and Conclusion in Section 7.

## Preliminaries

* Fundamentals of deepfake technology

### Deep learnings vs. machine learning and artificial neural networks

* The meanings of ML, DL, and ANN
* The concept of ANN architecture and two major categories : shallow/deep ANNs, The significance of DL

### Networks used in deepfake generation and detection

* Commonly, deepfakes generated using combinations of four typical networks : ED, CNN, GAN, RNN
* ED : Encoder-decoder networks
* CNN : Convolutional neural networks
* GAN : Generative adversarial networks
* The explanation of GAN equation and its shortcoming and structure
* RNN : Recurrent neural networks
* The explanation of RNN equation and widely used for speech recognition and sequential ones
* The full forms of the acronym
* In the following section : Deepfake categories

## Deepfake categories

* The technology trends and frameworks of deepfake categories

### Audio deepfakes

* The outline of audio deepfakes

#### Non-AI generated : Replay attacks

* The definition of replay attack and the two subtypes of it : far field detection, cut and paste detection attacks
* Attack : Threat to speaker verification system because of low-cost, also can be used against voice assistants
* Defense : Text dependent speaker verification and a current technique that detects end-to-end replay attacks by using deep convolutional networks

#### AI-generated audio fakes

* The most important audio deepfake principles, Speech synthesis(SS) and its leading company named, Lyrebird-Descript
* The disadvantages of SS : a lot of processing power and data storage, sparsely spoken language, not recognizing periods or special characters, ambiguities with homographs, the intelligibility and naturalness of prosody, hard to imitate accents, no breathing/laughter/pauses/sighs

##### Speech synthesis (Text-to-speech)

* Text-to-speech(TTS), analyzing the text and making the speech sound in line with text inputted using the rules of linguistic description of the text, also its advantages(human like, offering different accents and voices), and configuring(samples of actual voices, speaking rate, pitch, volume, sample rate hertz)
* Attack : Char2Wav(an end-to-end speech synthesis generation framework)
* WaveNet : The explanation of its formula and architecture
* The overall structure of WaveGlow, Tacotron 1(end-to-end text-to-speech generative model), Tacotron 2(recurrent seq-to-seq feature prediction network and modified WaveNet vocoder)
* Deep Voice3 : encoder, decoder, converter
* MelNet(Time-delayed stack, Centralized Stack, Frequency-delayed stack)
* Using NN TTS synthesis can make the speech audio not in training, ex. Char2Wav, Baidu 3 Voice, Deep Voice 1, 2, Lyrebird-Descript, WaveNet, Voco, Tacotron 1, 2, WaveGlow, Melnet
* Defense : review in the VC Defense section

##### Voice conversion and impersonation

* The helpful benefit of VC
* Attack : pushing the anti-spoofing capabilities to improve in normal speaker verification systems
* Gaussian mixture model, GANs, CycleGAN, StarGAN-VC, ASSEM-VC, Cotatron-VC
* Audio deepfake tools : Supplementary Table 1 in the Appendix
* Defense : ML(SVM, Random Forest, K-Nearest Neighbors), DL(ResNet, TCN, biLSTM), the features which are fed to the models are really challenging in the field of audio deepfake
* DeepSonar : Layer-wise neuron activation patterns, simple binary classification

### Text deepfake

* Exposed fabrications
* Humorous fakes, large hoax

### Video deepfake

* Reenactment
* Video synthesis and editing
* FaceSwap
* FaceApp, varying policies in other countries and with no legal jurisdiction, Zao, WeChat

### Image deepfakes

* Faceswap(Snapchat, Fakeapp, Zao), Synthesis(NVIDIA’s 112, StyleGANS2, MRF Markov random field, GANS and auto-regressive networks, Conditional GANS), Editing(FaceApp)

## Audio deepfake datasets

* A significant impact on the performances of dataset
* ASV spoof datasets : 2015(TTS, VC), 2017(replay), 2019(replay, TTS, VC with LA, PA), 2021(LA, PA with additional scenario : speech deepfake database)
* FoR(Fake or Real dataset) : Multiple version, a phone call or voice message, deep voice 3, Google Wavenet, no VC algorithm
* WaveFake dataset : English, Japanese, no VC, various TTS algorithm

## Intuition behind the AI-generated audio

* Brief explain some of the intuitive logic behind AI-generated audio, Vocoder which synthesizes waveforms based on acoustic or linguistic features obtained from previous steps in the network.
* Wavenet based on CNN
* Wave-Glow based on Glow
* MelGAN, Hifi\_GAN based on GAN

## Discussion and future directions

* The critical discussion, analysis and summarization regarding the compiled work focusing on audio deepfake generation and a summarization of the current techniques, future directions

### Deepfake generation

* Deepfake quality, trade-off between quality and some of the other aspects regarding audio deepfakes
* Data vs. Quality(MOS) : MelNet, single speaker dataset as well as VoxCeleb2 multi-speaker dataset
* Sampling Frequency(kHz) vs. Quality(MOS) : the higher sampling rate may give way to higher audio quality
* Availability vs. Quality : the more the availability and reproducibility, the more development the technology will have and more likely to be used for nefarious purposes or research
* Using other deepfake types for a certain type : needs of using non-parallel data

### Future defense against audio fakes

* More powerful and faster-developing of deepfake generation methods than detection methos
* Prevention : Blockchains and distributed ledger technologies(DLTs), uses of perceived emotions
* Mitigation : Keeping some detection tools proprietary only to people who need it like fact checkers for reporters
* Detection : Working on the features which are fed into the network, “generalization” by changing or improving both of the networks and features as well as defining different loss functions
* References regarding audio deepfakes Table 2
* References regarding the other deepfakes Table 3
* More references of visual deepfakes

## Conclusion

* Necessity of awareness of deepfakes and the harm and questioning what we see and hear online since the content can be misleading, more research in the field of audio deepfakes.