# Audio Deepfake Detection : A Survey

## Introduction

* Deepfake technologies for malicious purposes
* A few survey of audio deepfake
* The contributions of this paper with comprehensive survey

## Overview

### Types of Deepfake Audio

* The definition of deepfake audio and referring five kinds of it

#### Text-to-Speech

* Synthesis intelligible and natural speech, given any arbitrary text, using machine learning models

#### Voice Conversion

* Cloning a person’s voice digitally to change the speaker’s speech remaining the content same

#### Emotion Fake

* Changing the audio in such a way that the emotion of the speech while other information remains the same

#### Scene Fake

* Tempering of the acoustic scene of the original utterance with another scene via speech enhancement technologies while the speaker identity and speech content remain unchanged

#### Partially Fake

* Changing several words in an utterance

### Competitions

* The series of competitions such as the ASV spoof and ADD challenges have played a key role in accelerating the development of audio deepfake detection
* The ASVspoof challenges : Task(LA, PA, DF)
* The ADD challenges : Task(LF, PF, FG, RL, AR)

### Benchmark Datasets

* Being largely dependent on well-established datasets with various fake types and diverse acoustic conditions for the development of audio deepfake detection techniques
* Insufficiency of the varieties of spoofing techniques and the datasets conducted repeatable and comparable spoofing detection in earlies
* The datasets which have played a pivotal role in accelerating the development of audio deepfake detection in recently

### Evaluation Metrics

* Equal error rate(EER), the evaluation metrics for audio deepfake detection
* The relationship between the values in EER metrics
* Two rounds of evaluation in the detection task of audio fake game track in ADD

## Discriminative Features

* The goal of feature extraction : learning discriminative features via capturing audio fake artifacts from speech signals
* Four categories : short-term spectral features, long-term spectral features, prosodic features and deep features

### Short-term Spectral Features

* Computed mainly by applying the short-time Fourier transform(SFTF) on a speech signal
* Mainly composed of short-term magnitude and phase based features

#### Short-term magnitude based features

* Usefulness for detecting generated speech
* Magnitude spectrum features directly derived from the magnitude spectrum : LMS, RLMS
* Power spectrum features derived from the power spectrum : LPS, Cep, LFCC, RFCC, MFCC, IMFCC, LPCC

#### Short-term phase based features

* Being used to discriminate between human and generated speech
* GD based features : GD, MGD, MGDCC, APGD
* Other phase features : IF, BPD, RPS, PSP, CosPhase

### Long-term Spectral Features

* Short-term spectral features not good at capturing temporal characteristic
* Proposing long-term spectral features to capture long-range information from speech signals, and four types : STFT, CQT, HT, WT

#### STFT based features

* Four kinds of STFT
* Modulation feature(ModSpec, Global M), SDC, FDLP, LBP

#### CQT based features

* Long-term window transform providing higher frequency resolution at lower frequency, but higher temporal resolution at higher frequencies in contrast to the STFT
* CQT, CQCC, eCQCC, ICQCC, CQTMGD

#### HT based features

* Computing from the analytical signal obtained by the HT, such as mean Hilbert envelope coefficients(MHEC)

#### WT based features

* Being derived mainly by performing WT(wavelet transform) on speech signals
* MWPC, CFCC, CFCCIF

### Prosodic Features

* Introduce of prosodic features, spanning over longer segments and important parameters : fundamental frequency(f0), duration, energy distribution, speaking rate etc
* Trials of capturing f0(pitch pattern) and phoneme duration

### Deep Features

* Because of the flaws by biases due to limitations of handmade representations in the aforementioned features, deep feature are motivated to fill the gap.

#### Learnable spectral features

* Using learnable neural layers to estimate the standard filtering process, being categorized to partially and fully learnable spectral features
* Partially learnable spectral features extracted by training a neural network based filterbank matrix with a spectrogram obtained by applying STFT on a speech signal, and introducing the studies of it(FBCC, ConvRBM, nnAudio, FastAudio)
* Fully learnable spectral features learned directly from raw waveforms to approximate the standard filtering process, extracted by training a filterbank matrix with a spectrogram(TD-FBanks, SincNet, RawNet2, LEAF)

#### Supervised embedding features

* Involving the extraction of deep embeddings from deep neural network via supervised training and four kind of it
* Spoof embeddings : extracted from a neural network based model trained on the bonafide and spoofed data
* Emotion embeddings : using a supervised speech emotion recognition model trained with emotion labelled data
* Speaker embeddings : using a supervised speech recognition model using training data with speaker identity label
* Pronunciation embedding : extracted from a speech recognition model trained with labelled data

#### Self-supervised embedding features

* Being costly and technically demanding of supervised embeddings, self-supervised embedding is needed
* Wav2vec based features : extracted from the pretrained Wav2vec or Wav2vec2.0 models
* XLS-R based features : extracted from the pretrained XLS-R, a variant Wav2vec2.0 models
* HuBERT based features : extracted from the pretrained HuBERT models
* When the pre-trained model is fine-tuned with anti-spoofing data, a simple neural network with an average temporal pooling and linear layer is sufficient

## Classification Algorithms

* Importance of the back-end classifier, which aims to learn high-level feature representation of the front-end input features and model excellent discrimination capabilities

### Traditional Classification

* LR, PLDA, RF, GBDT, ELM, KNN, SVM, GMM

#### SVM based classifiers

* Excellent classification capabilities and proposing a one-class SVM classifier only trained using genuine utterances to classify real and fake voices

#### GMM based classifiers

* Conventional classifier, widely used in fake audio detection as it is an effective generative model employed as the baseline model in a series of competitions

### Deep Learning Classification

* Significantly outperforming the traditional methods due to their powerful modelling capabilities

#### CNN based classifiers

* Good at capturing spatially-local correlation
* LCNN
* Difficult to train and result in performance degradation → ResNet
* Not considering channel relationship → SENet

#### ResNet based classifiers

* Employing a residual mapping
* AFN
* Limitation to unseen fake attacks

#### Res2Net based classifiers

* Improving the generalization to unseen fake utterances

#### SENet based classifiers

* Squeeze-and-Excitation network : adaptively modelling interdependencies between channels
* ASSERT

#### GNN based classifiers

* Learning underlying relationship among data
* The aforementioned studies not focusing on learning the relationship between neighbouring sub-bands or segments
* GAT, GCN

#### DART based classifiers

* A particular variant of neural architecture known as differentiable architecture search, automatically optimize the operations contained within architecture blocks, including convolutional, pooling, residual connections operation
* PC-DARTS

#### Transformer based classifiers

* Good at modelling local and global artifacts and relationship because of the partially fake utterances contains some discontinuity between concatenated audio

## End-to-End Models

* Machine learning based classifier can vary greatly when combined with different features
* Integrating feature extraction and classification in an end-to-end manner upon the raw speech waveform

### CNN based classifiers

* Attempting the CNN based models to end-to-end fake audio detection
* CLDNN, CRNNSpoof : not perform well in cross-dataset evaluation
* TSSDNet/Inc-TSSDNet : capability to unseen datasets

### ResNet2 based classifiers

* Motivated by the power in text-independent speaker verification
* TO-RawNet

### ResNet based classifiers

* Becoming easy to train and achieve promising performance
* Res-TSSDNet, RW-ResNet

### GNN based classifiers

* Inspired by the success of GAT to model complicated relationship among graph representations
* RawGAT-ST, AASIST, RawBoost, Orth-AASIST

### DART based classifiers

* The aforementioned end-to-end methods can only automatically learn feature and network parameters rather than network architecture
* Operating directly upon the raw speech signal
* Jointly optimizing of both the network architecture and network parameters
* Raw PC-DARTS

### Transformer based classifiers

* Modelling local and global artefacts and relationship directly on raw audio
* Rawformer, SE-Rawformer

## Generalization Methods

* Dropping sharply at performance when dealing with out-of-domain dataset in real-life scenarios, the generalization ability

### Loss Function

* LMCL, SAMO

### Continual Learning

* Continuous training and adaptation of models on new information, aiming to overcome catastrophic forgetting existing in fine-tuning
* DFWF, RAWM

## Performance comparisons

### Top-performing methods in typical competitions

* The top-performing systems in the ASVspoof and ADD with their performance evaluated in terms of the EER
* Features
* Classifiers
* Superficially decrease as seen from the increase in EER, due to the increasing difficulty of the tasks

### Evaluation of Features

* Evaluating the discriminative performance of different handcrafted features in fake audio detection
* Performing worse in out-of-domain
* More being robust with features from pretrained models in out-of-domain, and that concatenation of features may be a viable strategy to improve the robustness of detection systems

### Evaluation of Classifiers

* Being not robust to out-of-distribution evaluation, and degrading significantly when the testing data is out-of-distribution

### End-to-End Models

* Degrading on the In-the-Wild test set by trained on ASVspoof 2021, but performing better on that test set by trained on ADD 2023

## Future directions

* Collecting audio datasets in the wild
* Designing large-scale multilingual datasets
* Improving generalization ability and robustness of detection models
* Dealing with rapid development of deepfake technologies
* Improving the interpretability of detection results
* Exploring more reasonable evaluation metrics

## Conclusions

* Proposing robust and general algorithms with valid and reliable samples in order to make the detection of deepfake audio applicable to real situations