Audio Deepfake Detection: A Survey (Summarized)

IEEE Journal vol. 14, No. 8

August 2023

Index Terms: Audio, deepfake detection, survey, features, classifiers.
https://arxiv.org/pdf/2308.14970

It's my first time reading and summary paper. If there is anything improper, please contact me. I'll update it on GitHub now and then. Since I'm a freshman of research, I haven't learned most of ML knowledge, therefore, many parts are to be filled.

1 Highlight of this paper

learning common discriminative audio features; computing methodologies to build a generalized automatic system.

2 Background & Concerns to research recently

Current studies on audio deepfake detection: pipeline, end-to-end detector
1. **pipeline**: front-end feature extractor + back-end classifier. (standard form over last decades)
2. **end-to-end**: employ a model to jointly optimize two things above via operating directly upon raw audio waveform.

2.1 five types of deepfake audio

TTS = text analysis + speech waveform generation modules
SWGM has 2 methods: concatenative + statistical parametric TTS(=acoustic model + vocoder)
–¿ Some end-to-end model: Variational Inference with adversarial learning for end-to-end TTS (VITS) + FastDiff-TTS
Voice Conversion: change timbre and prosody while content remains the same;
3 main approaches = statistical parametric + frequency warping + unit-selection
Emotion Fake = parallel data based + unparallel one

Scene Fake
partially fake

2.2 Competitions

2.3 Benchmark Datasets

How to compare different spoofing methods? Set up datasets.
Replay is considered a low-cost and challenging attack.
Not very important for this part. Just briefly introduce the outcome of these years.
citing list: [6], [25] - [28], [39] - [47]

2.4 Evaluation Metrics

$$P_{fa}(\theta) = \frac{\#\{fake\ trials\ with\ score > \theta\}}{\#\{total\ fake\ trials\}}$$

$$P_{miss}(\theta) = \frac{\#\{genuine\ trials\ with\ score < \theta\}}{\#\{total\ genuine\ trials\}}$$

with θ increasing, $P_{fa}(\theta)$ decreasing while $P_{miss}(\theta)$ increasing. And both of them $P \in [0, 1]$

$$\exists\theta, s.t. P_{fa}(\theta) = P_{miss}(\theta)$$

There are 2 rounds and derived 2 EER

$$W_{EER} = \alpha ERR_{R1} + \beta ERR_{R2}$$

So what is the meaning of EER here?
https://www.zhihu.com/question/37436914/answer/384679638

3 Discriminative Features

4 categories
short- and long-term **based on** Digital Signal Processing Algorithm

3.1 short-term spectral features

short-term: 20-30 ms
inadequate in capturing temporal characteristics of speech feature trajectories
mainly by Applying **short-time Fourier transform (STFT)** [52]
the speech signal $x(t)$ could be represented as

$$X(t, \omega) = |X(t, \omega)| e^{j\phi(\omega)}$$

where $|X|$ is the magnitude spectrum and ϕ is the phase spectrum, $|X|^2$ is the power spectrum
Most of **magnitude based features** derived from the power spectrum; **Phase based features** derived from the phase spectrum.

3.1.1 Short-term magnitude based features

The statistical averaging inherent in parametric modelling of the magnitude spectrum may introduce artefacts, such as over-smoothed spectral envelopes. (This sentence would be important, although I don't know what it means right now.)

Magnitude spectrum features
log magnitude spectrum (LMS) containing the formant information, harmonic structure and all spectral details of speech signal.
formant information: important for **speech recognition** but not for **fake detection**, especially for TTS or VC
Therefore, residual LMS (RLMS) = inverse linear predictive coding (LPC) filter to reduce the impact of formant information but better analyse the details of spectrum such as harmonics. (Why?)

Power spectrum features
Most well studied in ADD.
includes:

1. log power spectrum (LPS): compute directly on raw power spectrum

2. cepstrum (Cep): apply discrete cosin transform (DCT)

   (Dimensions of these two above are too high)

3. filter bank based cepstral coefficients (FBCC): address the aforementioned issue, and include

   1. Rectangular Filter Cepstral Coefficients (RFCC): using linear scale rectangular filters
   2. Linear Frequency Cepstral Coefficients (LFCC): extracted with linear triangular filters
   3. mel frequency cepstral coefficient (MFCC): derived from mel scale triangular filters, with denser placement in lower frequencies to simulate human
   4. inversed MFCC (IMFCC): utilizes triangular filters that are linearly spaced on inverted-mel scale, higher emphasis to the high-frequency region

4. Mel-frequency principal coeffitients features (MFPC): similar to MFCC, using principal component analysis (PCA)

5. all-pole modelling based cepstral coefficient (APCC): all-pole modeling representation of signal converted to LPCC

6. subband spectral features (SS) = subband spectral flux coefficients (SSFC) + spectral centroid magnitude coefficients (SCMC) + subband centroid frequency coefficients (SCFC) + discrete Fourier mel subband transform (DF-MST)

3.1.2 Short-term phase based features

The phase spectrum itself does not have stable patterns for fake audio detection due to phase warping.
Post-processing methods are instead utilised to generate useful short-term phase based features including Group Delay (GD) based and other phase features.

GD based features including:

1. GD: the derivative of phase spectrum along the frequency axis, which is referred as to a representation of filter phase response
2. Modified GD (MGD): compute from spectrum after cepstral smoothing frame-by-frame (variation of GD); extract a more clear phase pattern than GD.
3. MGD cepstral coefficients (MGDCC): from MGD, using both phase and magnitude info. e.g. [62]
4. All-pole GD (APGD): all-pole modelling.

other phase features

1. instantaneous frequency (IF): phase spectrum along the time axis. With GD, provides complementary info for spoofed speech detection.
2. baseband phase difference (BPD): extracted from baseband STFT, provides more stable time-derivative phase information compared to IF
3. relative phase shift (RPS): reflects "phase shift" of harmonic components about the fundamental frequency.
4. pitch synchronous phase (PSP)
5. cosine-phase (CosPhase) based feature: apply cosine function to unwrapped phase spectrum following by DCT
6. CosPhase principal coefficients (CosPhasePC): means of PCA; reduce the dimensionality of CosPhase features

3.2 long-term spectral features

Rather than computing frame-by-frame like STSF, proposed to capture long-range info from speech signals.

3.2.1 STSF based features

1. modulation features
    1. modulation spectrum (ModSpec): contains long-term temporal characteristics

   2. global modulation (Global M): combine spectral (e.g. MFCC) and temporal modulation info

2. shifted delta coefficients (SDC): 1. capture long-term speech info; 2. compute by augmenting delta coefficients of multiple speech frames

3. frequency domain linear prediction (FDLP): DCT on speech signal by linear prediction analysis on different subbands

4. local binary pattern (LBP): obtain long-span info upon spectral features

3.2.2 CQT based features

very useful!!! Long-term window transform. Provides higher frequency resolution at lower frequencies, but higher temporal resolution at higher frequencies in contrast to the STFT. The center frequencies of each filter and the octaves are geometrically distributed for CQT.

1. CQT spectrum (CQTgram): logarithm on raw power magnitude spectrum obtained via CQT [70][71]

2. CQ cepstral coefficient (CQCC): DCT of the log power magnitude spectrum via CQT

3. extended CQCC (eCQCC): combination from octave power spectrum with CQCC features computed from Linear power spectrum

4. inverted CQCC (iCQCC): inverted linear power spectrum of long-term

5. CQT-based modified group delay (CQTMGD) [69]

3.2.3 HT based features

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

3.2.4 WT based features

performing WT, which includes:

1. mel wavelet packet coefficients (MWPC): wavelet-packet transform on speech signals [57]

2. cochlear filter cepstral coefficients (CFCC): wavelet transform-like auditory transform, relevant mechanism occurring in human cochlea

3. CFCC +Instantaneous Frequency (CFCCIF)

   . The IF and phase of the envelope of the cochlear filter are vital features for speech perception of human listeners.

   . TTS and VC generate frame by frame, human speech generate in continuum [77]

3.3 prosodic features

Prosody refers to non-segmental info, including **syllable stress, intonation patterns, speaking rate and rhythm**.

* it spans over longer segments.

The important prosodic parameters include:

1. fundamental frequency (F0)
   - . is known as pitch
   - . pitch pattern statistics [79]
   - . dividing short-range autocorrelation function [61]
   - . not very useful for VC
   - . pitch extraction algorithm: capture discriminative features of F0 [81] (unreliable in noisy environ & requires a large amount of training data)
2. duration (e.g. phone duration, pause statistics)
3. energy distribution
4. * speaking rate

[82] fuse F0, phoneme duration (extracted from pre-trained model HuBERT) and energy

These features are less sensitive to channel effects compared to spectral features. Provide complementary info to spectral features.

3.4 deep features

* via deep neural network-based models
* make up the flaws of biases due to the limitations of handmade representations.

3.4.1 Learnable spectral features

learnable neural layers to estimate the standard filtering process

Partially learnable spectral features are extracted by training a neural network-based filterbank matrix with a spectrogram obtained by STFT
achievements these years: [84] - [89]
Fully learnable spectral features are learned **directly** from raw waveform to approximate the standard filtering process.
achievements these years: [51], [90] - [95]

3.4.2 Supervised embedding features

involves the extraction of deep embeddings from deep neural networks via supervised training.

Spoof embeddings are extracted from a neural network based model trained on the bonafide and spoofed data.
achievements these years: [49], [96] - [99]

Emotion embeddings are learnt using a supervised speech emotion recognition model trained with emotion labelled data.
– directly used to detect fake utterances
achievement these years: [100]

Speaker embeddings are trained using a supervised speaker recognition model using trained data with speaker identity label.
– auxiliary feature
achievement: [101]

Pronunciation embeddings
achievement: [82]

3.4.3 Self-supervised embedding features

extract deep embedding features from a self-supervised speech model trained using any bona fide speech data.

1. Wav2vec based features [106]-[107]
2. XLS-R based features [105][108][109]
3. HuBERT based features [82][83][102]

** Learning deep embedding features using self-supervised training is suggested as a potential direction to improve the generalization of fake audio detection.

4 Classification algorithm

4.1 Traditional Classification

1. Logistic Regression (LR) [137][138]
2. probabilistic linear discriminant analysis (PLDA) [114][139]
3. random forest (RF) [114]
4. gradient boosting decision tree (GBDT) [114]
5. extreme learning machine (ELM) [140]
6. KNN [138], SVM [141], GMM [142], etc.

4.1.1 SVM based classifers

[110]
However, it's very difficult to know the exact nature of spoofing attacks in practical scenarios.
Therefore, [111] using genuine utterances to classify.
[143] use i-vectors as the input features of SVM to discriminate.

4.1.2 GMM based classifiers

[79]
[112] train a GMM classifier fed with MFCC features
[144] log-scale likelihood ratio based on GMM
[113] i-vectors trained with GMM mean supervector to jointly perform VC attacks detection and speaker verification obtaining promising performance.
[48] choose GMM classifiers for benchmarking various features
[115] GMM in standard 2-class classifier in which the classes correspond to genuine and spoofed speech.

4.2 Deep Learning Classification

mostly based on this

4.2.1 CNN based classifiers

CNNs are good at capturing spatially-local correlation
[7] proposed a feature genuinization transformer with CNN trained only using genuine speech, and the outputs of this transformer are then fed into the Light CNN based classifier.
[116][117][118][146]

4.2.2 ResNet based classifiers

[147] employing a residual mapping
[95][120][121]
[122] propose an Attentive Filtering Network, based on dilated residual network, using convolution layers and modifying the residual units by adding a dilation factor
[148] employ a standard 34-layer ResNet with multi-head attention pooling layer
[149] compact network ResMax combining MFM activation and ResNet

4.2.3 Res2Net based classifiers

generalizability to unseen fake attacks are less limited compared to ResNet
[123] the feature maps within one ResNet block are split into two multiple-channel groups linked by a residual-like connection.
[124] combine Res2Net and Phase network, fed with phase and magnitude features.

4.2.4 SENet based classifiers

[125][126] Squeeze-and-Excitation network: interdependencies between channels
[127] Anti-Spoofing with Squeeze-Excitation and Residual neTworks (ASSERT)
[150] self0attention layer to detect partially fake audio
[151] self-distillation for ADD

4.2.5 GNN based features

Graph neural network: like graph attention network (GAT) and Graph Convolutional Network (GCN): used to learn underlying relationships among data
* The fake artefacts used to detect spoofing attacks are often located in specific temporal segments or spectral subbands
[128] GAT to modelling segments and subbands
[153] utilize GCN incorporation prior knowledge to learn spectro-temporal dependency information.

4.2.6 DARTS based classifiers

Differentiable Architecture Search (DARTS): a variant of neural architecture search, automatically optimizes the operations contained with architecture blocks, including convolutional, pooling, residual connections operations
[129] partial channel connections (PC-DARTS)
[155] combine DARTS and MFM activation

4.2.7 Transformer based classifiers

The transformer is good at modelling local and global artefacts and relationship [156]

5 End-to-end models

Fuse extraction and classifiers together via a deep neural network, jointly optimized directly upon the raw waveform
avoid limitations introduced by the use of knowledge-based features, rather than generic decomposition

5.1 CNN based models

[158]
[159] raw waveform convolution long short term neural network (CLDNN): reduce time and spectral variations, as well as long-term temporal memory layers to model long-term temporal info
[130] 5 1D convolution layers, a bidirectional LSTM layer and 2 fully-connected layers
[160] tackle cross-dataset evaluation: time-domain synthetic speech section net (TSSDNet), Inception parallel convolutions structures (Inc-TSSDNet)

5.2 RawNet2 based models

[94][131][132]
CNN with residual blocks, operates directly on raw audio through time-domain convolution

5.3 ResNet based models

[160][162]

5.4 GNN based models

[128][133] inspired by GAT the model complicated relationships among graph representations
Also a very important part. Learn it from paper directly.

5.5 DARTS based models

[135] automatic approach not only operates directly upon raw speech signal but jointly optimizes of both the network architecture and network parameters based on partially-connected differentiable architecture search from the raw audio waveform.

5.6 Transformer based models

[136]
Learn it from paper directly.

6 Generalization methods

their performance drops sharply when dealing with out-of-domain datasets in real-life scenarios
(The generalization ability of ADD systems is still poor.)

6.1 Loss Function

[169] large margin cosine loss (LMCL) function and online frequency masking augmentation
[170] one-class learning, construct a compact representation of genuine audio representation and utilize an angular margin to separate the fake utterances in the embedding space.
[171] speaker attractor multicenter one-class learning (SAMO)
core idea: real utterances are clustered around a number of speaker attractors and the method pushes away fake voices from all the attractors in a high-dimensional embedding space.

6.2 Continual Learning

continuous training and adaptation of models on new info, aiming to overcome catastrophic forgetting existing in fine-tuning.
[172] regularization-based continual learning method, Detecting Fake Without Forgetting (DFWF)
doesn't need to access old data but can remember previous info.
but may result in error accumulation
[173] regularized adaptive weight modification (RAWN) to tackle the issue above.

7 Performance Comparisons

8 Future directions

9 Terminologies

Deepfake: the usage of deep learning methods to seamlessly swap faces in videos. Any audio or video in which important attributes have been either digitally altered or swapped, with the help of AI.
ADD: Audio Deepfake Detection
ASV: Automatic Speaker Verification

10 Follow up materials

[12] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: Challenges and future directions," Algorithms, 2022.