A Comparative Study on Physical and Perceptual Features for Deepfake Audio Detection

Menglu Li
menglu.li@ryerson.ca
Department of Electrical, Computer and Biomedical Engineering
Toronto Metropolitan University
Toronto, ON, Canada

Yasaman Ahmadiadli
yahmadiadli@ryerson.ca
Department of Electrical, Computer and Biomedical Engineering
Toronto Metropolitan University
Toronto, ON, Canada

Xiao-Ping Zhang
xzhang@ee.ryerson.ca
Department of Electrical, Computer and Biomedical Engineering
Toronto Metropolitan University
Toronto, ON, Canada

ABSTRACT

Audio content synthesis has stepped into a new era and brought a great threat to daily life since the development of deep learning techniques. The ASVSpoof Challenge and the ADD Challenge have been launched to motivate the development of Deepfake audio detection algorithms. Currently, the detection models, which consist of front-end feature extractors and back-end classifiers, utilize the physical features mainly, rather than the perceptual features that relate to natural emotions or breathiness. Therefore, we provide a comprehensive study on 16 physical and perceptual features and evaluate their effectiveness in both Track 1 and Track 2 of the ADD Challenge. Based on results, PLP, as a perceptual feature, outperforms the rest of the features in Track 1, while CQCC has the best performance in Track 2. Our experiments demonstrate the significance of perceptual features in detecting Deepfake audios. We also seek to explore the underlying characteristics of features that can distinguish Deepfake audio from a real one. We perform statistical analysis on each feature to show its distribution differences on real and synthesized audios. This paper will provide a potential direction in selecting appropriate feature extraction methods for the future implementation of detection models.

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; Artificial intelligence; Neural networks; • **Applied computing** → Investigation techniques.

KEYWORDS

ADD 2022, Anti-spoofing, Deepfake audio, Feature extraction, Countermeasures, Feature selection

ACM Reference Format:
Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2022. A Comparative Study on Physical and Perceptual Features for Deepfake Audio Detection. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia (DDAM '22), October 14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3552466.3556523

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DDAM '22, October 14, 2022, Lisboa, Portugal.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9496-3/22/10…$15.00
https://doi.org/10.1145/3552466.3556523

1 INTRODUCTION

The emergence of deep learning methods has resulted in various applications in video editing, speech recognition and biometrics. Although these applications are built to optimize the workflow in every context, there have been several online frauds and data forgeries within the applications. The term, Deepfake, came to the spotlight after multiple online users tried to alter different multimedia such as videos, images and audio using deep learning methods. Videos and voices of politicians and actors were used to replace the original speaker. As audio synthesis techniques grow in popularity, especially after the development of deep learning techniques, anti-spoofing methods are proposed in the literature to detect bonafide audio from Deepfake ones [1, 2, 31, 36].

In order to motivate researchers in the speech processing field to better distinguish the fake and real audio in the Automatic Speaker Verification (ASV) systems, the ASVSpoof challenge was first proposed in 2015 [33]. The original ASVSpoof dataset consists of two categories of spoofing attacks, which are Logical Access (LA) and Physical Access (PA). Another category named Speech Deepfake (DF) has been added to the ASVspoof2021 challenge . Both LA and DF sets consist of synthesized audio samples generated by Text-to-Speech (TTS) and Voice Conversion (VC) algorithm, while the DF task involves the compressed audios without ASV systems [11]. However, in the ASVSpoof dataset, background noises contained in fake audios are ignored, and also the dataset does not consider the case that small fake segments are hidden inside a real speech. Therefore, the Audio Deep synthesis Detection (ADD) challenge is proposed to overcome these issues. The ADD dataset contains three main tracks: low-quality fake audio detection (LF) as Track 1, partially fake audio detection (PF) and audio fake game (FG) as Track 2 and Track 3 respectively [35]. In this paper, we will focus on Track 1 and Track 2 of the ADD challenge dataset.

With the development of anti-spoofing challenges, various Deepfake audio detection algorithms have been proposed. The overall architecture of detection models can be categorized as End-to-End structure and Two-Part structure. In End-to-End architectures, the network accepts raw audio as input and the utterances are processed through the network to produce the classification outcome [7, 27]. In Two-Part architecture, a front-end feature extractor is utilized to convert the audio waveform into a parametric feature representation, which is to be analyzed by the back-end classifier[32]. In this paper, we focus on both physical and perceptual features to represent the characteristics of the utterances and compare their performance on the detection results. Physical features indicate the mathematical computations on audio waves such as the spectrum,

the cepstral coefficients and the energy, while the perceptual features focus on human perception of audio namely pitch, loudness and timbre. For the back-end classifiers, we will adopt both traditional machine learning methods such as Gaussian Mixture Models (GMMs), and the state-of-the-art deep neural networks-based models, which are RawNet2 and SE-Res2Net50.

Our contributions in this paper are three-fold. First, we select 16 different physical and perceptual features, and compare their detection performance on both Track 1 and Track 2 in the ADD challenge. Second, we perform a statistical analysis to evaluate the ability of the selected features in capturing underlying characteristics that distinguish between Deepfake and real audios. Third, we are the first to provide a concise review on a series of perceptual features, and demonstrate their significance in detecting Deepfake audios.

The remainder of this paper is organized as follows. Section 2 provides a detailed summary of the related work on audio feature extraction. Section 3 presents the description of the selected features. Section 4 describes the experiment and our results. Section 5 analyzes the experiment results and presents the evaluation. Section 6 concludes the paper.

2 RELATED WORK ON FEATURE SELECTION

For the related work, we focus on the features that are adopted in the current Deepfake audio detection algorithms with the Two-Part architecture. Features are used for extracting the discriminative information between real and fake audio. Therefore, a good choice of features will benefit Deepfake detection performance. Currently, there are two major categories of widely used features. The first type is traditional acoustic features, and the other type is to apply deep learning techniques to the acoustic features to extract the embedding representation.

The traditional acoustic features work as the backbone to describe the characteristics of audio clips, which mostly involve Fourier Transform and Constant-Q Transform. This type of feature includes phase features, power spectrum features, and cepstral coefficients. Das et al. [9] derive seven long-range acoustic features based on long-term Constant-Q Transform and find that the long-range features outperform the features obtained from short-time transforms. Sahidullah et al. [23] compare 19 acoustic features based on first-order Fourier coefficients and second-order spectral features. Results indicate the selected features that contain high-frequency regions, detailed spectral information, and dynamic characteristics, receive a better performance in the spoofing attack detection. Albadawy et al. [2] propose to apply the third-order spectral correlations revealed by bispectral analysis to detect synthesized speeches which are generated by the GAN-based algorithms. Alzantot et al. [4] propose a fusion extractor that takes the weighted average of multiple acoustic features and converts them into spectrograms.

The deep learning-based features have the ability to maximize the difference between real and spoof speeches. Wu et al. [32] use the real audio data only to train a CNN-based feature genuinization transformer. The trained transformer can transform the testing speech to amplify the difference between the real and spoof audio. Chen et al. [6] apply three variants of ResNet structures to extract deep feature representation of the raw audio signals. The detection performance of the ensemble system that combines these three deep embedding features is competitive. Some research works also combine acoustics features and embedding features. Balamurali et al. [5] concatenate the traditional acoustic features such as MFCCs, and CQCCs, with the deep embedding produced by an autoencoder to form the final set of features, which ensures the robustness of detection.

The previous works of these two categories mainly focused on the physical features, which are the mathematical measurements computed directly through the audio soundwave. However, there is not much attention to perceptual features, even though the perceptual features are important in identifying real and spoofing audios because Deepfake audios may lack natural emotions, pauses or breathiness. Therefore, we explore the perceptual features in this paper and compare their performance in detection Deepfake audios with the physical features. The proposed analysis results can provide a convincing reference to select features while developing audio spoofing detection algorithms.

3 DESCRIPTION OF THE SELECTED FEATURES

In this section, we provide detailed descriptions of the selected features for both categories of physical and perceptual. We also include references for the features that are adopted or discussed in the previous literature. As Table 1 indicates, there are several features that have not been used in the field of Deepfake audio detection, to the best of our knowledge.

3.1 Physical features

Mel-Frequency Cepstral Coefficients (MFCC): MFCC [10] indicates the energy of a speech signal in the frequency domain. This feature is based on the logarithm of the Mel-scale filter bank and the decorrelation of Discrete Cosine Transform (DCT), which gathers information representing low-frequency regions of the utterance. MFCC is inspired by imitating the human hearing system and is widely used due to its robustness in presence of noise. Librosa library [20] is used to extract this feature in our experiment.

Constant-Q Cepstral Coefficients (CQCC): Constant-Q transform (CQT) is a perceptually motivated time-frequency analysis of a speech signal [29]. CQCC can be computed using CQT and spectral analysis. Frequency bins represented by constant Q are in a geometric scale different from the linear scale of DCT. Uniform sampling is applied to the constant Q power spectrum log followed by DCT on the resultant log to obtain CQCC [21, 34].

Δ^2 - MFCC / Δ^2 - CQCC: Static features do not take time-variant aspects of the speech into consideration. To obtain the time derivatives of spectrum-based features, deltas (Δ) and delta-delta (Δ^2) coefficients, called differential and acceleration respectively, are utilized [8]. The Δ^2 feature is the second-order derivative of the static feature through a few consecutive frames over time. In our experiment, we extract the Δ^2 feature from the static one and concatenate it to the original static feature as the input.

Mel-spectrogram: Mel spectrogram is a time-frequency representation. When an audio signal gets divided into windowed segments and the fast Fourier transform is applied to each segment,

Table 1: Summary of the selected features along with their configuration parameters, dimensions and the references of the related studies in anti-spoofing models

Type	Feature Name	Configuration Parameters	Dimension	Applied / discussed in Deepfake audio detection models
Physical features	MFCC	Hop length = 160, Number of filter = 20	(20, 301)	[4, 5]
	Δ^2 - MFCC	Delta window size = 3	(60, 301)	[8, 23]
	CQCC	Number of filter = 19	(20, 352)	[29, 34]
	Δ^2 - CQCC	Delta window size = 3	(60, 352)	[8, 23]
	Mel-spectrogram	Hop length = 160, Hanning window	(129, 301)	[4]
	ZCR	Hop length = 160	(1, 301)	-
	RMSE	Hop length = 160, STFT window	(1, 301)	-
	Spectral based feature	Hop length = 160, Hanning window	(3, 301)	-
	Spectral flatness	Hop length = 160, Hanning window	(1, 301)	-
Perceptual features	Pitch based feature	Computed with Parselmouth software	(2, 105)	-
	Onset Strength	Hop length = 160, time lag = 1	(1, 301)	-
	HNR	Computed with Parselmouth software	(1, 107)	-
	Intensity	Computed with Parselmouth software	(1, 129)	-
	Jitter-shimmer based feature	Computed with Parselmouth software	(1, 11)	[14]
	Chromagram	Hop length = 160, Number of chroma bins = 12, Hanning window	(12, 301)	[24]
	PLP	Window length = 0.025	(19, 306)	-

the outputs are called a spectrogram. Mel scale is a logarithmic scale that shows equal distances on the scale, have the same perceptual distance, which aims to scale the frequency so that it matches with the human hearing system. Mel-spectrogram is obtained by converting the spectrograms to the Mel scale, where the Librosa library is used.

Zero-crossing Rate (ZCR): The ZCR is the rate at which the sign of a signal changes from positive to zero to negative or from negative to zero to positive, which is a temporal feature of a speech signal computed in the time domain [15]. Its value is being used as a feature to classify percussive sounds, music genre classification, and speech analysis. The ZCR is also a significant indicator of the frequency of the audio signal. We apply Librosa to obtain this feature.

Root-Mean-Square Energy (RMSE): The energy of a speech utterance corresponds to its amplitude. After summing up the energy of all samples in an audio frame and dividing it by the total number of samples, the RMSE is obtained by taking the square root of the resulting value. In our experiment, we compute the RMSE value based on spectrograms.

Spectral based features: We combine three types of spectral measurements as the spectral-based features, which consist of the spectral centroid, the spectral bandwidth, and the spectral roll-off. The spectral centroid measures where the center of mass for a spectrum is, which shows where most of the energy is focused. The spectral bandwidth is obtained by computing the variance of the spectral centroid. The spectral roll-off is the threshold frequency for a spectrogram bin such that 85% of the spectral energy is lower than that value. We obtain these three measurements using Librosa with the same set of configuration parameters and concatenate the resulting vectors together as the spectral-based features.

Spectral flatness: Spectral flatness is also called the tonality coefficient and it is used to measure the amount of correlation in speech signals. This feature is utilized to indicate the pureness tone of an audio sample opposed to being noisy. Spectral flatness is usually converted to decibels (dB). The meaning of tonal in this feature is the number of peaks in the power spectrum of the utterance. Higher spectral flatness value indicates a similar amount of power in all spectral bands such as white noise. Pure tones have a value of 0 for spectral flatness which will result in a spiky power spectrum, focused on a small number of bands [12]. This feature can be extracted by Librosa library.

3.2 Perceptual features

Pitch based features: Pitch is one of the major perceptual properties of sounds, which is related to how high or low it sounds when humans hear them. Pitch can be represented by the frequency numerically, which is called pitch frequency. However, two sounds with the same pitch frequency may have different pitch sensations because of pitch strength. Pitch strength is independent of the pitch frequency and is measured by the height of the first peak in the stimulus waveform auto-correlation [25]. We apply the pitch analysis tool in the Parselmouth library [17] to obtain the numerical value of pitch frequency and pitch strength, which are concatenated together to represent the pitch characteristics of the audio clips.

Onset Strength: Onsets happen when there is a sudden rise of energy across the spectrum, which indicates the beginning of a sound event. The detection of onsets has been proved useful in the auditory scene analysis [26] and tempo estimation [30]. In the context of speech analysis, the strength of onsets may extract the edge information for the periods of speech and eliminate the effects of the noise. Librosa library is adopted to obtain the onset strengths for each audio clip, and we set the hop length as 160 to make the window length 0.025 seconds.

Harmonics-to-Noise Ratio (HNR): The HNR describes the ratio between the periodic and non-periodic segments of a sound, which quantifies the amount of additive noise of the speech in the unit of dB [13]. This feature is widely adopted to diagnose pathological voices in the vocal acoustic analysis [13]. The Deepfake speech synthesized by TTS systems may not have the current phonation, so the HNR could be useful in detecting synthesized speeches. Parselmouth library is used to extract the value of HNR.

Intensity: The feature of intensity represents the power carried by a sound wave per unit area. In human perception, the sound intensity usually is described by the loudness of sound in the unit of dB. In our experiment, the intensity values are also fetched using Parselmouth library.

Jitter-shimmer based feature: Jitters measure variations in the fundamental frequency of the audio signals, while shimmers show

the amplitude variations of the sound waveform [28]. Parselmouth library provides five measurements for jitters and six measurements for shimmers in total. Jitter measurements include local jitter, local absolute jitter, rap jitter, ppq5 jitter, and ddp jitter. The measurements in shimmers consist of local shimmer, local shimmer in dB, apq3 shimmer, apq5 shimmer, apq11 shimmer, and dda shimmer. We concatenate these 11 measurements into a one-dimension vector and regard it as a set of jitter-shimmer-based perceptual features.

Chromagram: Chromagram is a perceptual feature that relates to the different pitch classes within the time window. It maps the spectral audio information into one octave based on the short-time Fourier Transform. In our implementation using Librosa, one octave includes twelve chroma bins, and the hop length is set to be 160.

Perceptual Linear Predictive Coefficient (PLP): The PLP is a low-dimensional representation of speech, which combines spectral analysis and linear prediction analysis. In order to compute the PLP coefficients, the critical-band spectral resolution, equal loudness pre-emphasis, and intensity-to-loudness compression are performed on the weighted windows of audio samples [16]. Compared to Linear Predictive Coefficients (LPCs), PLP is more consistent with human auditory and is robust to noise [3, 16]. We apply the Spafe library [19] to extract the PLP coefficients of the speech data.

4 EXPERIMENT

4.1 Dataset

For the experiment, we use ADD 2022 Dataset [35], which is a high-fidelity multi-speaker Mandarin speech corpus and is provided by the 2022 Audio Deep synthesis Detection Challenge. The training set of the ADD 2022 dataset contains 3012 real speeches and 24072 synthesized speeches. For the purpose of training, we resample all audio samples to 16kHz and fix all the audio lengths to 3 seconds by either truncating the longer audio clips or concatenating the shorter audio clips repeatedly. The adaptation sets of ADD 2022 Dataset are used for performance testing, since we do not have access to the official test datasets. We mainly focus on two tracks of the provided adaptation sets. The first track (Track 1) consists of real speeches and fully fake audios generated using the TTS and VC algorithms with various background noises. The second track (Track 2) comprises partially fake speech generated by manipulating the original real utterances with real or synthesized audio.

4.2 Classifiers

To compare the effectiveness of various features, we select three common baseline or state-of-the-art Deepfake audio detection classifiers.

GMM: Gaussian Mixture Model is a widely used audio spoofing classification model based on traditional machine learning. It implements the expectation-maximization algorithm to fit the provided data into a mixture of a finite number of Gaussian distributions. This classifier usually plays the role of a base method to compare with other more advanced deep learning-based classifiers [32]. For instance, it has been selected as a baseline for both ASVSpoof 2019 and ASVSpoof 2021 challenges. We adopt the official code[1] provided by ASVSpoof 2021 to implement the GMM classifier, setting the number of mixture components as 100, the type of covariance parameter as diagonal, and the convergence threshold as 0.01.

RawNet2[27]: The RawNet2 is a deep learning-based Deepfake audio detection algorithm, which consists of CNN and ResNet structures. The model consists of two residual blocks, and each one contains two batch normalization layers, two leaky-Relu, two convolutional, a feature map scaling layer, and a max-pooling layer. The key point of the feature map scaling layer is that this layer can act as an attention mechanism to extract a more discriminative representation of audios. This algorithm produced the second-best performance in ASVSpoof 2019 challenge and has been selected as a baseline for ASVSpoof 2021 challenge.

Originally, the RawNet2 is proposed as an End-to-End algorithm. Its first set of SincNet layers works as a customized filter bank to learn low-level speech representations from raw speech waveforms [22]. In our experiment, we replace the SincNet layers with the various front-end feature extractors and feed the extracted feature vectors directly to the following residual blocks of the RawNet2 model. This classifier is trained using a learning rate of 0.0001, 100 epochs and a batch size of 10.

SE-Res2Net50[18]: This classifier is an upgraded ResNet detection architecture, which enables multiple feature scales during the training process. After the first convolutional layer, the input feature maps are split into multiple subsets, and then the subsets are connected using a hierarchical residual-like structure. A squeeze-and-excitation (SE) block is stacked onto each ResNet block to assign different weights to different feature subsets. This state-of-the-art achieved an EER of 1.99% in the LA track of ASVSpoof2019 challenge. In our experiment, we set the number of feature scales as 8, and train this model using a learning rate of 0.0001, 50 epochs and a batch size of 10.

4.3 Evaluation metric

We adopt Equal Error Rate (EER) as the indicator to measure the detection performance. An EER corresponds to a threshold where the false positive rate (False Alarm) and false negative rate (Miss) are equal. A detection result with a lower EER score is regarded to be more accurate. The EER is also used as the official evaluation metric for the ADD 2022 Challenge [35].

4.4 Results

Table 2 shows the detection results of each selected feature for Track 1. The best performance is produced by the PLP-RawNet2 and Mel-spectrogram-RawNet2 model with an EER score of 19%. Comparing the 16 features, PLP has the lowest EER score for all three back-end classifiers in Track 1. This result emphasizes the strength of PLP in being consistent with the human hearing system and the robustness to noises.

Table 3 indicates the detection result for Track 2, which is partial Deepfake detection. The majority of the selected features receive a poor EER score after feeding to the three back-end classifiers, except for CQCC-related features. The CQCC-SE-Res2Net50 model achieves a 22% EER score.

Motivated by these observations, we perform further experiments by combining two single features to train as a new feature.

[1] https://github.com/asvspoof-challenge/2021

Table 2: The EER results of the selected features for Track 1

Feature	EER		
	GMM	RawNet2	SE-Res2Net50
MFCC	0.52	0.22	0.35
Δ^2 - MFCC	0.44	0.23	0.26
CQCC	0.49	0.38	0.42
Δ^2 - CQCC	0.47	0.41	0.41
Mel-spectrogram	0.52	**0.19**	0.46
ZCR	0.56	0.33	0.33
RMSE	0.56	0.36	0.36
Spectral based feature	0.35	0.26	0.29
Spectral flatness	0.40	0.42	0.44
Pitch based feature	0.41	0.3	0.37
Onset Strength	0.36	0.32	0.43
HNR	0.33	0.29	0.48
Intensity	0.55	0.28	0.29
Jitter-shimmer based feature	0.42	0.38	0.43
Chromagram	0.54	0.34	0.32
PLP	**0.29**	**0.19**	**0.23**
Pitch based + MFCC	0.63	0.22	0.24
CQCC + PLP	0.32	0.39	0.41

Table 3: The EER results of the selected features for Track 2

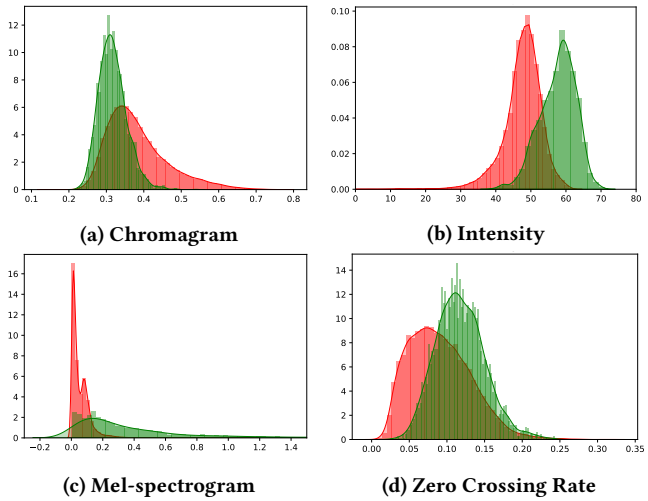
Feature	EER		
	GMM	RawNet2	SE-Res2Net50
MFCC	0.94	0.96	0.97
Δ^2- MFCC	0.74	0.86	0.97
CQCC	0.50	0.35	**0.22**
Δ^2 - CQCC	0.49	0.36	0.24
Mel-spectrogram	0.96	0.93	0.72
ZCR	0.64	0.92	0.95
RMSE	0.95	0.76	0.89
Spectral based feature	0.72	0.92	0.92
Spectral flatness	0.65	0.76	0.44
Pitch based feature	0.45	0.42	0.39
Onset Strength	0.49	0.80	0.70
HNR	0.23	0.64	0.53
Intensity	0.89	0.98	0.93
Jitter-shimmer based feature	0.81	0.48	0.38
Chromagram	0.68	0.89	0.88
PLP	0.53	0.99	0.94
Pitch based + MFCC	0.99	0.91	0.95
CQCC + PLP	**0.07**	**0.32**	0.35

For the first combination, we choose MFCC and pitch-based features, because they are the most widely used and fundamental features in the categories of physical and perceptual respectively. We select CQCC and PLP as the second combined feature because of their excellent performance in Track 1 and 2. However, the combination of two features does not necessarily improve the detection results in Track 1 because of overfitting, especially for pitch-based features + MFCC, even though their training loss is lower than training by a single feature. A different case has been shown for Track 2. The combination of CQCC and PLP achieves a significant improvement in Track 2 detection, which receives the EER score of 7% for the GMM model and 32% for the RawNet2 model.

We do not perform a comparison between our observations to ADD 2022 official baseline systems, because in [35], the EER scores are obtained by evaluating baselines using the test sets. Our observations show the detection performance of selected features and classifiers on the adaptation sets.

5 ANALYSIS AND EVALUATION

ADD 2022 dataset covers realistic situations in audio clips, such as with background noises and with several small fake segments hidden in a real speech audio. Some of the features we selected for



(a) Chromagram **(b) Intensity**

(c) Mel-spectrogram **(d) Zero Crossing Rate**

Figure 1: Distribution plots of the selected features. Green is for real speech, and red represents the synthesized speech

experimenting such as PLP, CQCC perform well on this dataset regarding the EER value due to their robustness to noise and their consistency with human hearing system.

Besides the experiment in detection performance, we also compute the distribution of feature values for the training set of ADD dataset. For the real audio set and the synthesized audio set, we extract all selected features separately and calculate the mean value of each features. To show the difference between real audio and synthesized audio, we plot the distribution of these calculated mean value for each features. We select the distribution plots of four features to show in Figure 1, which are chromagram, intensity, Mel-spectrogram, and ZCR. In the distribution plots, the green shape indicates the real audio data, while the red shape represents the synthesized audios. The x-axis stands for the mean value, and the y-axis represents the distribution density.

According to Figure 1, we observe that there are significant statistical differences between the real speeches and the synthesized speeches, which can be extracted by the selected features and be utilized to train the back-end classifiers to distinguish the synthesized audio from the real one. Therefore, the distribution differences of the features in Figure 1 give us an aspect of explanation for their satisfying performance on fully fake audio detection tasks (Track 1), as Table 2 indicates. However, since there is an overlap in the distribution of real and synthesized data, it provides potential difficulties in detecting partial synthesized segments in the real audio. The classifiers may be fooled by the real segments within the partial fake audio clips, which results in poor detection performance in the partial detection tasks (Track 2). For instance, for Mel-spectrogram feature with RawNet2 back-end classifier, the EER scores for Track 1 and Track 2 are 19% and 93%, respectively. Similar observations are found for the rest of the select features. Another reason that the most of classifiers receive an EER higher than 50% in Track 2 is because the adaptation sets of Track 2 only contain partial Deepfake audio samples. There is no fully real audio in the dataset for testing.

These statistical differences between real and fake audio distribution occur in the perceptual features as well, which causes the perceptual difference in the human hearing system. This finding suggests that besides the physical feature, the effectiveness of perceptual features is also substantial in the development of Deepfake audio detection algorithms. For example, feeding intensity and pitch-based feature to RawNet2 model directly receives the EER score of 28% and 30% respectively. These experiment results demonstrate that even a simple extraction of perceptual features can have a significant detection performance.

In evaluating the detection performance of static MFCC, static CQCC and their Δ^2 coefficients, we observe that the effectiveness of the delta operation does not have a significant improvement in detection performance, especially when training with deep learning-based classifiers. Relatively, the Δ^2 coefficients work better on GMM model, which outperforms the corresponding static features in both Track 1 and Track 2.

We also observe good performance for some features that are able to extract the noise information, such as zero-crossing rate and harmonics-to-noise ratio. This characteristic is reflected in the detection performance of Track 1 which focuses on detecting fake audios with background noises or music effects. The ZCR-RawNet2 model achieves 33% EER, while the HNR-RawNet2 model has an EER score of 29%.

In terms of back-end classifiers, the two selected deep learning-based classifiers outperform the GMM model for most of the features in Track 1. It demonstrates the advance of deep learning-based classifiers in revealing the underlying information provided by the feature, and in being robust to the presence of noise, especially for large-scale features like MFCC and chromagram. For instance, the EER result of Chromagram-GMM, Chromagram-RawNet2, and Chromagram-SE-Res2Net50 are 54%, 34%, and 32%, respectively. Deep learning-based classifiers also have a better computational performance in terms of training time. However, the deep learning-based state-of-the-arts do not perform well on Track 2. They may lack the ability to examine the partial fragments of an audio clip.

6 CONCLUSION

In this paper, we present a comprehensive study comparing the performance of multiple features, including physical and perceptual acoustic features, for both Track 1 and Track 2 of ADD challenges. To the best of our knowledge, this is the first work to explore a series of perceptual features and demonstrate their substantial effectiveness in spoofing audio detection. For Track 1, PLP, as a perceptual feature, receives the lowest EER score of 19% compared to the rest of the selected features. For Track 2, most of the selected features perform poorly. Strikingly, the combination of CQCC and PLP brings a significant improvement in detection performance, which receives the EER score of 7% with the GMM classifier and 32% with the RawNet2 classifier for Track 2. The comparison results suggest that the perceptual features are useful to contribute to a satisfying detection performance, and they should also be utilized in the process of detecting Deepfakes.

Furthermore, this paper is an effect in investigating the underlying characteristics of Deepfake audios. We perform statistical analysis for each selected feature to emphasize the statistical differences in its distribution between the real speech and Deepfake speech. The distribution differences are also discovered in perceptual features, which provides the explanation for the ability of perceptual features to distinguish Deepfake speech from real speech. This analysis presents a potential direction of feature selection for the future implementation of Deepfake detection models. For future work, we will continue validating this characteristics analysis onto different spoofing datasets. More work remains to be done to integrate the signature characteristics that can distinguish real and Deepfake audios into the structure of detection classifiers.

REFERENCES

[1] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. 2020. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.

[2] Ehab A AlBadawy, Siwei Lyu, and Hany Farid. 2019. Detecting AI-Synthesized Speech Using Bispectral Analysis.. In *CVPR workshops*. 104–109.

[3] Sabur Ajibola Alim and N Khair Alang Rashid. 2018. *Some commonly used speech feature extraction algorithms*. IntechOpen London, UK:.

[4] Moustafa Alzantot, Ziqi Wang, and Mani B Srivastava. 2019. Deep residual neural networks for audio spoofing detection. *arXiv preprint arXiv:1907.00501* (2019).

[5] BT Balamurali, Kinwah Edward Lin, Simon Lui, Jer-Ming Chen, and Dorien Herremans. 2019. Toward robust audio spoofing detection: A detailed comparison of traditional and learned features. *IEEE Access* 7 (2019), 84229–84241.

[6] Tianxiang Chen, Elie Khoury, Kedar Phatak, and Ganesh Sivaraman. 2021. Pindrop Labs' Submission to the ASVspoof 2021 Challenge. In *Proc. ASVspoof 2021 Workshop*. 89–93.

[7] Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. 2020. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing* 14, 5 (2020), 1024–1037.

[8] Pronaya Prosun Das, Shaikh Muhammad Allayear, Ruhul Amin, and Zahida Rahman. 2016. Bangladeshi dialect recognition using Mel frequency cepstral coefficient, delta, delta-delta and Gaussian mixture model. In *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*. IEEE, 359–364.

[9] Rohan Kumar Das, Jichen Yang, and Haizhou Li. 2019. Long range acoustic and deep features perspective on ASVspoof 2019. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 1018–1025.

[10] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing* 28, 4 (1980), 357–366.

[11] Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, et al. 2021. ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *arXiv preprint arXiv:2109.00535* (2021).

[12] Shlomo Dubnov. 2004. Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Signal Processing Letters* 11, 8 (2004), 698–701.

[13] Carole T Ferrand. 2002. Harmonics-to-noise ratio: an index of vocal aging. *Journal of voice* 16, 4 (2002), 480–487.

[14] Yang Gao, Jiachen Lian, Bhiksha Raj, and Rita Singh. 2021. Detection and evaluation of human and machine generated speech in spoofing attacks on automatic speaker verification systems. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 544–551.

[15] Fabien Gouyon, François Pachet, Olivier Delerue, et al. 2000. On the use of zero-crossing rate for an application of classification of percussive sounds. In *Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy*, Vol. 5. Citeseer, 16.

[16] Hynek Hermansky. 1990. Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America* 87, 4 (1990), 1738–1752.

[17] Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics* 71 (2018), 1–15.

[18] Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. 2021. Replay and synthetic speech detection with res2net architecture. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 6354–6358.

[19] Ayoub Malek. 2020. *spafe/spafe: 0.1.2.* https://github.com/SuperKogito/spafe

[20] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8. 18–25.

[21] Zeyan Oo, Longbiao Wang, Khomdet Phapatanaburi, Meng Liu, Seiichi Nakagawa, Masahiro Iwahashi, and Jianwu Dang. 2019. Replay attack detection with auditory filter-based relative phase features. *EURASIP journal on audio, speech, and music*

processing 2019, 1 (2019), 1–11.

[22] Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 1021–1028.

[23] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. 2015. A comparison of features for synthetic speech detection. (2015).

[24] Summra Saleem, Aniqa Dilawari, Muhammad Usman Ghani Khan, and Muhammad Husnain. 2019. Voice Conversion and Spoofed Voice Detection from Parallel English and Urdu Corpus using Cyclic GANs. In *2019 International Conference on Robotics and Automation in Industry (ICRAI)*. IEEE, 1–6.

[25] William P Shofner and George Selas. 2002. Pitch strength and Stevens's power law. *Perception & psychophysics* 64, 3 (2002), 437–450.

[26] Elyse S Sussman. 2017. Auditory scene analysis: an attention perspective. *Journal of Speech, Language, and Hearing Research* 60, 10 (2017), 2989–3000.

[27] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6369–6373.

[28] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. 2013. Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology* 9 (2013), 1112–1122.

[29] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. 2017. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language* 45 (2017), 516–535.

[30] George Tzanetakis and Graham Percival. 2013. An effective, simple tempo estimation method based on self-similarity and regularity. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 241–245.

[31] Ahmet Semih Uçan, Fatih Mustafa Buçak, Mehmet Ali Han Tutuk, Halis İbrahim Aydin, Ertuğrul Semiz, and Şerif Bahtiyar. 2021. Deepfake and Security of Video Conferences. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 36–41.

[32] Zhenzong Wu, Rohan Kumar Das, Jichen Yang, and Haizhou Li. 2020. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. *arXiv preprint arXiv:2009.09637* (2020).

[33] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. 2015. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Sixteenth annual conference of the international speech communication association.*

[34] Jichen Yang, Rohan Kumar Das, and Haizhou Li. 2018. Extended constant-Q cepstral coefficients for detection of spoofing attacks. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1024–1029.

[35] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. 2022. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9216–9220.

[36] Teng Zhang, Lirui Deng, Liang Zhang, and Xianglei Dang. 2020. Deep learning in face synthesis: A survey on deepfakes. In *2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET)*. IEEE, 67–70.