Evaluating Acoustic Parameters for DeepFake Audio Identification

Albérick Euraste Djiré, Aminata Sabané, Abdoul-Kader Kabore, Rodrique Kafando, Tégawendé F. Bissyandé

Evaluating Acoustic Parameters for DeepFake Audio Identification

Albérick Euraste Djiré
Centre d'Excellence en IA (CITADEL)
Université Virtuelle du Burkina Faso
eurasted@gmail.com

Aminata Sabané
UFR Sciences Exactes et Appliquées
Université Joseph Ki-Zerbo (UJKZ)
aminata.sabane@citadel.bf

Abdoul-Kader Kabore
Centre d'Excellence en IA (CITADEL)
Université Virtuelle du Burkina Faso
abdoulkader.kabore@citadel.bf

Rodrique Kafando
Centre d'Excellence en IA (CITADEL)
Université Virtuelle du Burkina Faso
rodrique.kafando@citadel.bf

Tégawendé F. Bissyandé
Centre d'Excellence en IA (CITADEL)
Université Virtuelle du Burkina Faso
tegawende.bissyande@citadel.bf

***Abstract*—The progress made in the field of machine learning applied to signal processing offers interesting perspectives in terms of technological evolution but also causes some troubles in terms of ethics and security. For example, we are witnessing the emergence of audio deepFakes used to orchestrate scams. However, although the tools used in the generation of these deepFake audios show good results which can sometimes produce audios that seem to be confused with real audio, it is not impossible to dissect them. In order to detect them, many methods exist, in particular the analysis of the acoustic parameters which can attest to the authenticity of an audio extract. These parameters include energy, power, pitch, signal spectrum, cepstral coefficients, etc. However, these acoustic parameters are numerous and not all of them are suitable for detecting deepFake audio. This paper presents a comparative review of acoustic parameters useful in detecting DeepFake audio. Among them, we highlight the relevance of the study of cepstral parameters such as MFCC compared to other acoustic parameters such as mel-spectograms. The objective is to provide reliable leads in the detection of deepFake audio.**

***Index Terms*—deepFake audio, detection, mel-spectogram, MFCC**

I. Introduction

Artificial intelligence has contributed to major advances in many areas of digital technology, with audio signal processing being one of the primary beneficiaries. From voice assistants to audio transcription services, AI has significantly transformed the audio landscape. Specially, speech synthesis systems have been extensively employed in various applications such as audiobooks, despite their increasing misuse in generating DeepFake audios.

Most of the time, these deepFake audios are generated from a speech synthesis model (Text to speech), using neural networks pre-trained with the target person's voice in order to produce audio messages with the voice of the latter from submitted text messages.

Also, advanced methods in terms of voice synthesis are based on the prediction of acoustic parameters. It generally consists of three modules, including the word processing module, the acoustic parameter prediction module and a vocoder as shown in Figure 1. The first module aims at the prediction of phonetic units from the text provided. These are most often phonemes because they are the units that best translate the pronunciation of a word or a sentence. The second module uses these phonetic units to predict the acoustic parameters of the synthesis message. Among these acoustic parameters, we have the mel-spectogram, the fundamental frequencies, the energy, the pitch of the signal etc. The vocoder is responsible for converting the predicted acoustic parameters into a waveform useful for producing the audio message [20].

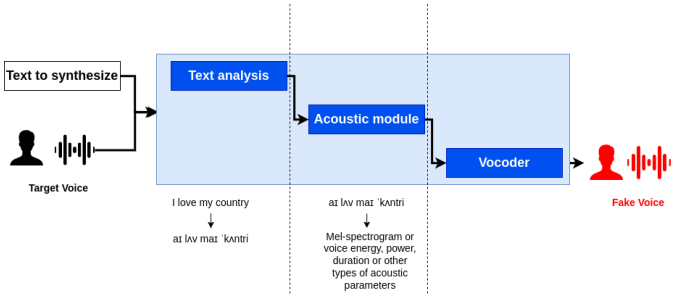

Fig. 1. Advanced models of vocal syntesis process

However, these voice synthesis models have shortcomings. In addition, we have problems relating to taking into account the context of the message which, even if we have a large corpus, is difficult to overcome [9]. Also, these models present inferior results compared to authentic speech extracts by relying on the Mean Opinion Score (MOS) as illustrated in Table I [18]. This subjective measure is the average of these scores between subjects and is widely used to determine the quality of a transmission or the synthesis of an audio signal [15].

Thus, these may look like human voices, but they will remain synthesized voices. Even if these sounds are ineradicable to the human ear, which perceives sounds as a whole, the

Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Développement (CITADEL)

Model	MOS
Tacotron 2	3.82 ±0.085
Deep Voice 3	3.75 ±0.03
Multispeech	3.65±0.14
Fastspeech	3.83 ±0.08
Real Voice	4.30 ±0.07

TABLE I
PERFORMANCE OF VOICE SYNTHESIS MODEL

analysis of certain acoustic parameters will undoubtedly make it possible to classify the audio extracts according to their authentic or fake nature. For this, the analysis of the acoustic parameters proves to be a credible solution to the resolution of this problem. However, maybe not all of these parameters are effective in determining audio deepfakes. Between spectograms, scepstral and spectral parameters, waveforms, fundamental and harmonic frequencies, there are a multitude of acoustic parameters that can help determine the authenticity of an audio. This paper focuses on examining these key acoustic parameters and their effectiveness in detecting DeepFake audio.

II. State of art

Our present work has been carried out on the basis of various articles dealing with techniques for detecting audio deepfakes and in particular those based on Artificial Intelligence.

Regarding the useful acoustic parameters we have the basic components of an audio signal such as the waveform, the fundamental frequency and any harmonics. Besides that, we have the mel-spectrogram, which is a representation of a signal in both the time and frequency domain following the Mel scale used by some vocoders, cepstral parameters such as Mel-Frequency Cepstral Coefficients (MFCC), Linear-Frequency Cepstrum Coefficients (LFCC), constant-Q Cepstrum Coefficients (CQCC) and spectral parameters.

A. Spectrograms

A spectrogram is a representation of a sound signal in the time and frequency domain. Indeed, the spectrograms represent sequences of the spectrum of the following sound signal as a function of time, the pitch of the signal being represented by a colored scale as shown in Figure 2 [19]. There are several types of spectrograms including the mel spectrogram using the mel scale. This scale makes it possible to locate the pitch of a sound depending on whether it is low or high. Mel is related to hertz by the equation 1.

$$m = 2595 * \log_{10}(1 + f/700) \quad (1)$$

The mel spectrogram is widely used in modern speech synthesis models. Indeed, in models such as tacotron, deepvoice, multispeech or fastspeech, the acoustic parameter prediction models output mel-spectograms that the vocoder converts into a waveform, even though the last two models are able to synthesize waveforms directly.

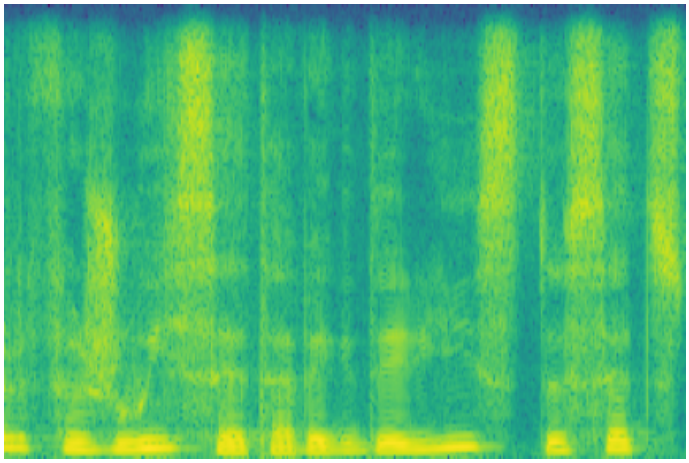

Fig. 2. Sample spectrogram of a sound signal

B. Cepstral parameters

A cepstrum designates the transformation of a signal from the time domain to another domain analogous to the time. The extraction of the cepstrum from an audio signal involves a series of well-defined steps, each adhering to established methodologies in signal processing. The process commences with Sampling, typically using the Nyquist-Shannon sampling theorem as a guideline. This is followed by Pre-emphasis, where a first-order filter is commonly applied. The next stage involves Framing the signal into short frames, followed by Windowing each frame using a window function such as a Hamming window. The final step is the computation of the Fast Fourier Transform (FFT), utilizing algorithms like the Cooley-Tukey radix-2 algorithm.
Then according to the characteristic of the studied signal, we must apply a filter which according to its nature will give us a different cepstrum. The most used being the filter bank on the Mel scale giving, after taking the logarithm of that spectrum, and computing its inverse Fourier transform, the MFCC [11] and Linear filter bank giving the LFCC [6] as shown in Figure 3.

C. DeepFake audio detection models

There are several methods and models aimed at detecting fake voice. Much work has also been done on the analysis of acoustic parameters using deep learning methods in order to classify audios according to their authentic or fake nature. For example, we can cite the use of DNNs, CNNs and their variants applied to the classification of acoustic parameters such as MFCCs, LFCCs and spectrograms. However, most of these models had limitations and inadequacies in solving the problem [2].

Another work on this subject consisted in the analysis of histograms for the detection of fake voices [4]. This analysis uses computer vision with CNN to classify histograms according to their nature [5]. However, although the approach is interesting, the model turns out to be non-scalable and very affected by the data transformation process [2].

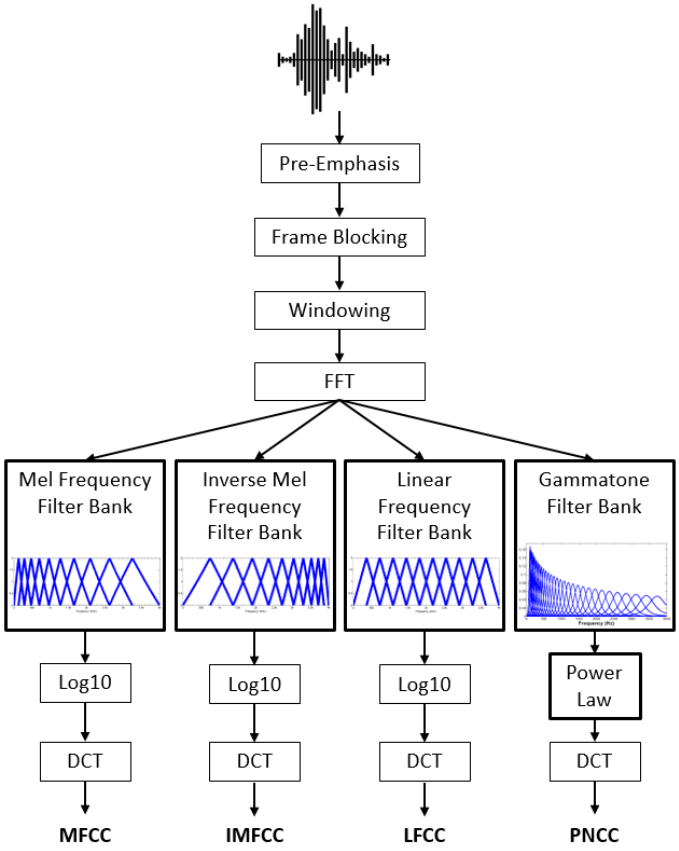

Fig. 3. Cepstral parameters computing process

There are also other successful works in this field, in particular Deepsonar which monitors the behavior of the neurons of a DNN during the analysis of the different MFCCs [16]. Although it presents good results in terms of classifying audios according to their nature, it is very sensitive to noise and therefore requires high quality audios for efficient processing.

III. Methodology

This work was done according to methodology showed by the Figure 4.

A. Extraction of useful parameters

The second part of our work consisted in the extraction of the acoustic parameters useful for our experiments. These acoustic parameters are the Mel spectrogram, the Mel frequency cepstral coefficients (MFCC).
Another aspect of this present work was to study the impact of the duration of the signal to be analyzed on the performance of the models. To do this, we used the audio from the dataset previously created, which we segmented into a sequence of 10s and 02s. We thus obtained 02 datasets, derived from the initial dataset, which we used to extract the acoustic parameters.

Concerning the extraction of the MFCCs, we used the librosa library which offers us several tools for audio analysis and in particular for the calculation of the MFCCs [10]. The results were stored in matrices of dimension **128*y** where **y** is proportional to the duration of the extract. For audios of 10 seconds **y=250** and for audios of 02 seconds **y=48**. Then, concerning the 10 seconds extracts, we homogenized the dimensions of these matrices in order to obtain matrices of dimension 128*250. To do this, we used the zero padding method to complete the lower dimensional matrices and for the higher dimensional matrices, we just extracted the first 250 columns.

Concerning the obtaining of the mel-spectrograms, we also used the librosa library which allows a synthesis of the mel-spectrograms of an audio signal. The results were stored as an image of size 700px*700px.

B. Detection models

At the end of the extraction of the useful parameters, we trained various models of neural networks in order to evaluate on the one hand the relevance of the parameter in the detection of DeepFake audio and on the other hand the performance of the models in the detection of DeepFake audio. To do this, we submitted each of these datasets to a convolutional neural network, a recurrent neural network and a Resnet model [8]. The models used were structured as follows:

- The CNN whose intermediate layers consisted of four convolutional layers whose number of filters are respectively 32, 64, 64, 128. At the output of each convolution layer, we have a pooling layer of size *2*2* and using Max Pooling. These layers are each followed by a dropout layer to limit overfitting. After these intermediate convolution layers, we have two FC layers respectively of size 512 and 2. Of course between the two layers we have a Dropout layer and the last layer provides us with the result of the prediction. Regarding the correction functions, we used a ReLu function for all convolution and FC layers. However we applied a softmax function on the last layer of our convolution network.
- The bidirectional LSTM recurrent network whose intermediate layers consisted of a one-dimensional convolution layer whose role was to reduce the size of the data to be processed for the downstream layers, of 02 bidirectional recurrent layers each consisting of 125 LSTM cells and 02 FC layers. Like the previous network, each Convolution, Recursive, and FC layer is followed by a Dropout layer. To the recurring layers we applied as a correction layer, a tanh type layer. The convolutional layer has at its output a correction layer of the ReLu type and the last two correction layers applied to the FC layers are respectively of the ReLu and softmax type.
- The Resnet-50 network in which all layers can be trained. At the end of this network we added 02 FC layers respectively of size 64 and 02 with the correction layer respectively of ReLu and softmax type.

IV. Experiments

During this work, we chose to compare the performances of two types of acoustic parameters: spectrograms and cepstral parameters. In addition, these two types of parameters will be

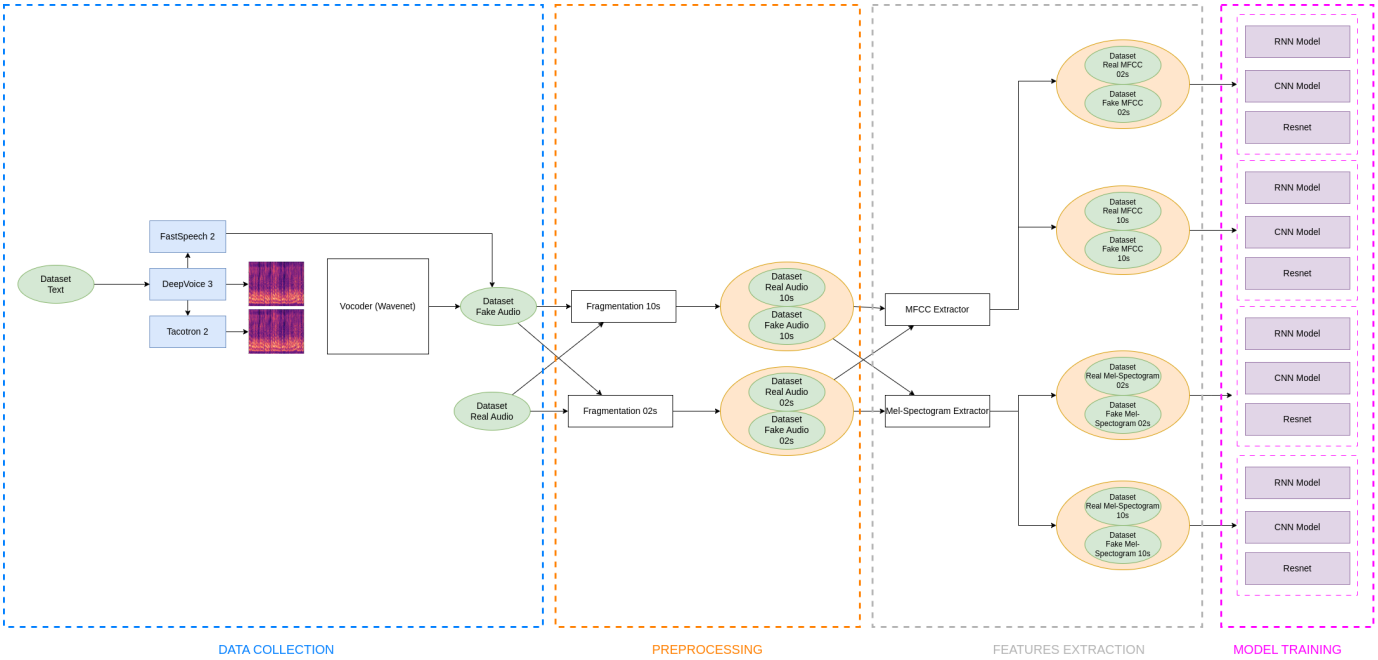

Fig. 4. Proposed workflow

treated according to a common scale. We chose the mel scale which is a scale which best transcribes the human perception of sounds. Thus, the parameters studied will respectively be the spectograms applied to the mel scale or mel spectrograms and the MFCC for the cepstral parameters.

A. Data

The first part of this work consisted of data collection. To do this, we generated from a speech synthesis model including Tacotron 2 [17], Deepvoice 3 [3] and FastSpeech 2 [12] audio messages to constitute a false sound of dataset extraction. The first two models are speech synthesis models based on the prediction of acoustic parameters, in this case the mel spectrogram, and whose overall structure is as shown in Figure 1. Concerning the vocoder used for them, it is wavenet which is a vocoder based on the dilated causal convolution [1]. As for the last model, it is based on the use of transformers and unlike previous models, the vocoder is optional as the model can directly synthesize waveforms. Also, concerning the acoustic parameter prediction module of FastSpeech 2, it consists of several sub-modules each managing a specific parameter. Thus, there is a signal duration prediction sub-module, a pitch prediction sub-module, an energy prediction sub-module, which makes this module quite scalable [13].

To the data generated from the various models listed representing the fake audio datasets, we added audio snippets from various recordings and datasets like synplaflex which is a corpus of audiobooks in French composed of 87 hours of good quality speech [14] and other recorded audio messages mainly in French language thus representing the authentic audio snippets dataset. Table II provides a comprehensive breakdown of the data that has been collected.

	Number of audio	Average time	Speakers
Audio Fake	30000	11.6s	6
Authentic Audio	30000	12.9s	13

TABLE II
AUDIO DATASET VOLUME

	Segmentation time	Number of extracts
Mel Spectogram	10s	60000
MFCC	10s	60000
Mel Spectogram	02s	100000
MFCC	02s	100000

TABLE III
ACOUSTIC PARAMETERS DATASET VOLUME

Then, from this dataset, we segmented the audios into 10 seconds snippets and 02 seconds snippets. From these two sets of data obtained, we proceeded to the extraction of the acoustic parameters useful to our experiments, namely the mel-spectrograms and the MFCCs in accordance with the stated methodology. Thus, at the end of this step, the result is the creation of 04 new datasets that we use in training and evaluating the performance of our analysis models. The details of the different datasets are summarized in Table III .

B. Results

We evaluate the performance of our models with the F1_Score metric and the accuracy calculated from the confusion matrix. The F1_Score is defined as the harmonic average between precision and recall. It is calculated using the formula 2 [7].

$$F1_score = 2 * 1/(1/Precision + 1/Recall_{F1score}) \quad (2)$$

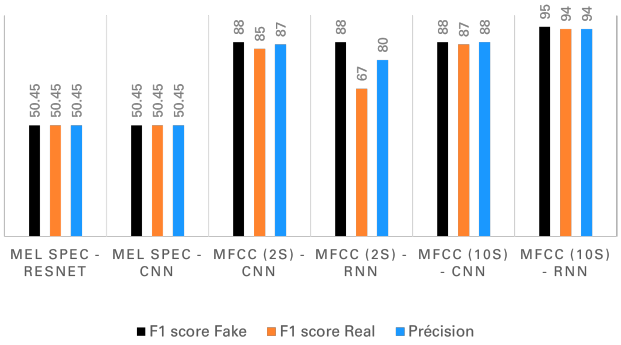

Fig. 5. Experiment results

The results obtained from these experiments are shown in Figure 5.

Thus, according to this result, it appears that the analysis of the cepstral parameters, in this case the MFCC, makes it possible to obtain a vocal DeepFake detection model with fairly decent performance. Whether for long or short extracts or on the nature of the type of model used for the analysis, the results are relatively similar even if the recurrent network constructed and applied to the classification of short MFFCs has lower performance than the others. However, the performance of these models can be improved by modifying their parameters such as the depth of the network, the size of the intermediate layers and the dropout applied between the different layers. However, besides the analysis of cepstral parameters, that of mel-spectrograms did not provide good results, regardless of the technology used. This can be explained by the excessive similarity between the representations of fake and authentic mel-spectrograms. Thus, it would perhaps be possible to work on the numerical data obtained after the calculation of the constituent values of the mel-spectrogram rather than on its graphic representation. We can also explore other acoustic parametres such as waveforms, power spectrums, the other cepstral parameters, the succession of formants which designates a frequency at which we observe a maximum of energy of the sound spectrum of a sound or a word, etc.

C. Discussion

In order to detect deepfake audio, several techniques and methods have been developed. Among them, one of the most interesting to date remains audio analysis using deep learning techniques. However, various works have shown that raw audio analysis produces unsatisfactory results in terms of detecting deepFake audio. Subsequently, the experiments focused on the analysis of acoustic parameters, always using deep learning techniques. The results provided following this vary depending on the acoustic parameters and the model used in the processing of these parameters. The objective of this present work is to establish a comparative assessment of the relevance of acoustic parameters in order to reveal the effective acoustic parameters in the detection of deepfake audio. In this sense, the experiments revealed the effectiveness of the analysis of MFCCs compared to the Mel spectrogram. However, there remain several acoustic parameters to explore such as the log-frequency spectrograms, the Frequency Mask or the Large Margin Cosine Loss, the Fast Fourier Transform, the Short Time Fourier Transform and the other cepstral parameters. Continuing work in this area in order to provide an exhaustive assessment of the performance of acoustic parameters in the detection of deepfake audio. Still with the aim of producing more efficient systems in terms of detecting audio deepfakes, it would be interesting to consider combining the analysis of several credible acoustic parameters. Another point to address is also the problem of the language used. Given that most of our work has focused on French audio messages, it would be interesting to consider the use and, if necessary, improvement of current models to take into account the multilingual parameter.

V. Conclusion

At the end of our various experiments, it appears that it is possible to detect the audio deepFake thanks to the analysis of the acoustic parameters. However, as expected, not all acoustic parameters are reliable for this, as we have seen with mel spectrograms. Regarding the analysis of MFCCs, the results are conclusive and a slight improvement in the performance of the models is observed when they are subjected to long extracts. Also, it would be good to salute the various works that have been carried out in this field, even if their components have limits, their results are sometimes usable for the detection of deepFake audio and can direct research in the right direction. Several other acoustic parameters are to be considered such as the Frequency Mask or the Large Margin Cosine Loss [6], the other types of spectrum and cepstral coefficient. It should also be noted that the Deep Learning models used in this work can always be improved either by optimizing their parameters or by enriching the dataset used.

Acknowledgment

This research was conducted at the *Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Développement* and was supported by a research grant from Canadian International Development Research Centre (IDRC) and the Swedish International Development Cooperation Agency (SIDA) research grant within the Artificial Intelligence for Developement (AI4D) initiative.

References

[1] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[2] Zaynab Almutairi and Hebah Elgibreen. A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. *Algorithms*, 15(5):155, 2022. Publisher: MDPI.

[3] Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, and Jonathan Raiman. Deep voice: Real-time neural text-to-speech. In *International Conference on Machine Learning*, pages 195–204. PMLR, 2017.

[4] Dora M. Ballesteros, Yohanna Rodriguez, and Diego Renza. A dataset of histograms of original and fake voice recordings (H-Voice). *Data in Brief*, 29:105331, April 2020.

[5] Dora M. Ballesteros, Yohanna Rodriguez-Ortega, Diego Renza, and Gonzalo Arce. Deep4SNet: deep learning for fake speech classification. *Expert Systems with Applications*, 184:115465, 2021. Publisher: Elsevier.

[6] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury. Generalization of audio deepfake detection. In *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, pages 132–137, 2020.

[7] Leon Derczynski. Complementarity, f-score, and nlp evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266, 2016.

[8] Riaz Ullah Khan, Xiaosong Zhang, Rajesh Kumar, and Emelia Opoku Aboagye. Evaluating the performance of resnet model based on image recognition. In *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, pages 86–90, 2018.

[9] Simon King. An introduction to statistical parametric speech synthesis. *Sadhana*, 36(5):837–852, October 2011.

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

[11] P Prithvi and Dr T Kishore Kumar. Comparative Analysis of MFCC, LFCC, RASTA-PLP. 4(5):4, 2015.

[12] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[13] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, Robust and Controllable Text to Speech, November 2019. arXiv:1905.09263 [cs, eess].

[14] Aghilas Sini, Damien Lolive, Gaëlle Vidal, Marie Tahon, and Élisabeth Delais-Roussarie. Synpaflex-corpus: An expressive french audiobooks corpus dedicated to expressive speech synthesis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[15] Robert C Streijl, Stefan Winkler, and David S Hands. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227, 2016.

[16] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1207–1216, 2020.

[17] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, and Samy Bengio. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[18] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*, 2017.

[19] Lonce Wyse. Audio spectrogram representations for processing with convolutional neural networks. *arXiv preprint arXiv:1706.09559*, 2017.

[20] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19):4050, 2019. Publisher: Multidisciplinary Digital Publishing Institute.