PARALINGUISTIC FEATURE EXTRACTION FOR AUDIO DEEPFAKE DETECTION METHODS

ISABEL VRIELINK

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2022806

COMMITTEE

prof. dr. Eric Postma
dr. Lara Mentink

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 21, 2024

WORD COUNT

6325

ACKNOWLEDGMENTS

I would like to thank Eric Postma for guiding me through this thesis.

PARALINGUISTIC FEATURE EXTRACTION FOR AUDIO DEEPFAKE DETECTION METHODS

ISABEL VRIELINK

Abstract

This thesis explores the implementation of paralinguistic feature extraction in audio deepfake detection models. The proposed model extracts paralinguistic features from audio clips and converts them into 1024-dimensional vector embeddings. The audio clips are sourced from the ASVspoof2019 dataset, existing of both real and deepfake audio samples. The vector embeddings serve as the input for the logistic regression model, the employed machine learning method to perform the binary classification task within this thesis. The evaluation tools, such as Equal Error Rate (EER) and accuracy assess whether the incorporation of paralinguistic feature extraction is an effective tool for audio deepfake detection systems by putting the model's performance in perspective with state-of-the-art methods. The proposed model achieved an EER of 3.04% and an accuracy of 97.9%. The results indicate that the implementation of paralinguistic feature extraction indeed is a promising approach for future audio deepfake detection methods.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

1.1 *Data Source*

The ASVspoof2019 Logical Access (LA) database serves as the sole dataset for this thesis. The dataset is owned by the ASVspoof consortium. This multi-speaker speech synthesis dataset consists of speech data, either spoofed or bona fide, collected from 107 anonymized human speakers (46 male, 61 female). It is publicly available on *https://www.asvspoof.org* [1].

1.2 *Figures*

All figures included in this thesis are designed and generated by the author of the thesis.

1.3 *Code*

The Python code for this thesis is primarily written by the author. An exception is the segment of code where the extraction of the embeddings is done. For this step, the Tensorflow2 framework of the TRILLsson model is used. This framework, developed by Google Research [2] is sourced from Kaggle [3]. A certain part of the code for the extraction of the embedding vectors is written by the thesis supervisor. Inspiration for the design of the logistic regression architecture is drawn from the Introduction to Machine Learning course of the Cognitive Science and Artificial Intelligence program [4]. Stack Overflow [5], GitHub [6] and ChatGPT-4O [7] were utilized for debugging purposes. Reused code is explicitly documented in the Python scripts.

1.4 *Technology*

Execution of the Python script is done on Google Colab Pro+ and on Visual Studio Code. The thesis is assembled using Overleaf. For checking of grammar and spelling, ChatGPT-4O [7] is utilized. No additional tools were used to construct this thesis.

2 INTRODUCTION

Today, we as a society are familiarising ourselves with all recent, groundbreaking technologies that are released, particularly in the field of artificial intelligence (AI). Every day, new, even more advanced techniques appear from all over the world. The latest developments around artificial intelligence brings benefits to us all; studies show, for example, that AI-driven technologies revolutionize the process of environmental monitoring to combat climate change and enable robust solutions to enhance data protection in cybersecurity [8, 9]. However, providing advanced technologies such as artificial intelligence with the public can initiate abuse of the technology. Artificial intelligence already brought several challenges to the surface, mainly in the field of privacy, security, ethics and mutual trust [10, 11, 12]. With our data being accessible online, including samples of our voice, the availability of 'big data' feeds an increasing problem. By utilizing deep

learning techniques, it is possible to mimic the voice of someone when available data of that person speaking is present. Some AI-generated deepfakes are being used for fraudulent cases as voice phishing or biometric spoofing [13]. The availability of different methods for these techniques keeps increasing and most importantly, distinguishing real from fake is getting increasingly harder [14]. It is for that reason essential to continue developing up-to-date technologies for the detection of deepfakes. In that way, trust and security can be maintained in digital media and communication systems. A large part of the state-of-the-art audio deepfake detection methods [15, 16, 17, 18] approach detection by extracting verbal aspects of speech like the actual words spoken or the syntax of the sentences. For example, this extraction is done with techniques like automatic speech recognition (ASR) [19, 20]. The features are visualized with the use of spectrograms or waveforms and used as input for machine or deep-learning models to train deepfake detection methods. However, by extracting the verbal cues that make speech, subtle hints that make human voice different from advanced synthesized audio may be overlooked [21, 19]. Those hints, the non-verbal part of speech, such as intonation, emotional tone, rhythm and stress are cues that add that extra layer to the uniqueness of speech. The non-verbal aspects of voice are referred to as paralinguistic features. Paralinguistic feature extraction may capture inconsistencies in AI-generated speech that are missed by the current methods. To research into this new subject, this thesis is built on the following research question:

> *To what extent is paralinguistic feature extraction an effective approach for audio deepfake detection methods?*

This research introduces an existing approach, paralinguistic feature extraction, from the research field of non-semantical speech tasks into deepfake detection methodologies, which is never done before. The traditional detection methods, while proven to be effective in the past [16, 17, 22, 15, 18, 23]may not always keep on level with the AI-generation advancements. Because there is primarily a focus on the linguistic content of audio, the research gap of incorporating audio details from the non-verbal layer of speech into detection methods is interesting to dive into. The emotional intonation of the words said, the variations in speech rhythm or even the variation of pitch our voice holds, while all being subtle, can probably illustrate a more complete picture of the audio sample needed to be identified. By utilizing the TRILLsson method, developed by Google Research [2], paralinguistic features are extracted from each input audio sample. Those features are defined into a 1024-dimensional embedding vector, which will serve as the input for the logistic regression model. TRILLsson is primarily designed for non-semantic speech tasks and

excels in the paralinguistic feature extraction making it an ideal fit for this thesis. The Automatic Speaker Verification and Spoofing Countermeasures Challenge database [20], especially the 2019 version; ASVspoof2019 Logical Access (LA) has been used in a number of deepfake detection studies [15, 16, 17, 22, 23, 24]. The publicly accessible dataset contains over 125.000 audio samples which are either bona fide (real) or spoofing, facilitating the the possibility to train advanced deepfake detection models. The spoofing audio samples are generated by several spoofing methodologies to keep variation in the dataset and introduce state-of-the-art generation techniques. To benchmark my thesis' method against existing methods, results will be put in perspective with the performance of six other, recent, audio deepfake detection models that have all used ASVspoof2019 to train and develop their detection model on. The equal error rate (EER) is a prominent performance metric used in the state-of-the-art spoofing detection methods. The metric, one single value, reflects the point of where the balance between false acceptance and false rejection is equal [24]. When methods are being compared under somewhat equal conditions, the EER can offer a clear insight in how effective a model is. To enable the best performance analysis, next to the standard performance metrics, EER will be the main evaluation method in this thesis. To perform the most effective binary classification on this input data, a robust model is required. Since logistic regression implements supervised learning techniques and is commonly used for binary classification tasks, it is an ideal fit for this research. To handle an extensive dataset as ASVspoof2019, a computational efficient model works best[25]. The model architecture is carefully designed to perform well on the ASVspoof2019 dataset. In summary, this thesis addresses the research gap in audio deepfake detection and the integration of paralinguistic feature extraction by leveraging TRILLsson. The performance of this model is put in perspective by analysing the performance of this technique to the state-of-the-art models in audio deepfake detection which all have utilized the ASVspoof2019 dataset. This research aims to contribute to new techniques, generalizability and robustness of audio deepfake detection models and sheds a light on the demand for constant innovation in the field to be an advanced countermeasure against the creation of it.

3 RELATED WORK

3.1 *Introduction to Deepfake Technologies and Detection Challenges*

With the explosive growth of possibilities in artificial intelligence, the evolution of deepfake technologies draw a lot of attention. Deepfakes, created through various advanced deep learning methods, make it possible to

convincingly impersonate individuals if the necessary data for it is available. The prevalent technique to do so, Generative Adversarial Networks (GANs), has made it possible to easily manipulate multimedia content, as described by Masood et al. [14]. In the early days of deepfake detection, the vast majority of the detection models were machine learning based. Now, a deep learning approach is more common [26] and the recent standard to combat deepfake creation. While a lot of different approaches pop up from all over the world, each improvement in deepfake creation needs a countermeasure. As mentioned in the study of Ranjan, Vatsa and Singh [26], there is need for more robust and generalizable methods while also still considering the ethical, sustainable and privacy implication within audio deepfake detection systems. This generalization of deepfake detection options was examined in cross-modal research conducted by Müller et al. [27] where it is stated that models performing well on the ASVspoof benchmark of spoofing attacks are often too closely tailored to the data and perform less outside of this particular set. The need of variety in detection models is increasing. While detection models based on a deep learning approach reach impressive performance, less computational expensive options should remain available in order to be accessible for the public [2] and more sustainable [25].

The Automatic Speaker Verification and Spoofing Countermeasures Challenge database (ASVspoof2019) [1] appears in a lot of detection models as the main data source. The dataset is based on the Voice Cloning Toolkit (VCTK) corpus, which is a multi-speaker English speech database. ASVspoof2019 contains real and spoofed audio clips from 107 different speakers (46 male, 61 female). The part of ASVspoof2019 used and discussed in this thesis is the Logical Access (LA) set [20]. This dataset contains labels to provide for supervised learning techniques. The origin of this database is the need for reliable sources to train detection models on. Every few years the ASVspoof consortium releases an up-to-date database reflecting the state-of-the-art examples of spoofing techniques so detection techniques can try to keep up.

3.2 *State-of-the-Art Approaches in Audio Deepfake Detection*

In the past two years, several different approaches for deepfake detection methods were proposed. To review the state-of-the-art in the field of audio deepfake detection the methods most relevant to this study are discussed below. These papers correspond to each other in the way that they all have utilised ASVspoof2019 as the dataset to build their detection model upon.

In recent work of Yadav et al. [16], the ASVspoof2019 dataset was employed to fine-tune a pre-trained transformer. This transformer, devel-

oped to identify the origin of a speech signal, combined a self-supervised learning algorithm with the capabilities of mel spectrograms to identify that signal. The method is proposed as Synthetic Speech Attribution Transformer (SSAT). The transformer architecture achieved a notable accuracy of 90.2% on the ASVspoof2019 dataset. With such performance, this type of identification systems using attribution models offer a robust framework to build research like this thesis upon.

To examine the performance of ASVspoof2019 on certain spoofing attacks where the model was performing less effectively on, research conducted by Hu and Zhou [17] presented the Online Hard Example Mining (OHEM) algorithm. This algorithm is built to be able to detect unknown voice spoofing attacks in datasets in order to be more robust against new data. This approach achieved an exceptional performance on the EER of 0.77% on the ASVspoof2019 (LA) dataset. This algorithm is designed to capture the most challenging spoofing attack samples present in a dataset and supports the model's ability to improve performance against this type of samples. As mentioned, releasing robust methods that have impact and possibilities of implementation on multiple approaches is crucial in advancing detection methods.

In a fully automated end-to-end deepfake audio detection method, proposed by Wang et al. [22], a wav2vec pre-trained model is used to create an automated detection system. This wav2vec pre-trained model is combined with a modified differentiable architecture search, light-DARTS. This system is designed to automate the process of learning speech representations and simultaneously learning and optimizing complex neural structures such as convolutional operations and residual blocks. The performance of this architecture achieved an equal error rate (EER) of 1.08% on the ASVspoof2019 dataset. With the automation of feature extraction this model explores the efficiency of implementing such tasks instead of doing the feature extraction and hyperparameter tuning more manually, aiming to leave less space for development mistakes and increase time efficiency.

In research, proposed by Wang et al. [15], challenges among traditional feature extraction are discussed. The study mentions that the state-of-the-art detection methods lose crucial information for identification of audio during feature extraction. To tackle this issue, direct feature extraction from raw audio signals is proposed to be utilised in, for example, RawNet. RawNet [28], a deep neural network architecture that operates directly on raw audio signals, enhances the ability to extract specific audio characteristics in samples used for tasks like spoofing detection. Wang et al. [15] designed an architecture in which both orthogonal convolution is implemented into RawNet and temporal convolutional networks are introduced

to capture long-term dependencies in audio speech signals. With these advancements, the study proved to be able to minimize loss of information, subsequently leading to a more effective detection system. This architecture, termed TO-RawNet, is trained and tested with the ASVspoof2019 (LA) dataset and presented a 66,09% decrease in EER compared to RawNet.

In a recent study conducted by Zhang et al. [18], the AASIST2 architecture was proposed. The study discusses how performance in short utterance evaluation in speech anti-spoofing can be improved by changing residual block to Res2Net blocks in the proposed AASIST systems. Short utterances are speech samples that are significantly shorter than commonly used samples in research, often less than a few seconds [18]. These improvements in short utterance anti-spoofing makes it possible to extract multi-scale features from audio signals in order to increase performance in short utterance evaluation. This new approach of deepfake detection is employed by utilizing the ASVspoof2019 as the main dataset. The performance of AASIST2 showed improvements on the performance of short utterance evaluations in anti-spoofing achieving a decrease in EER to 8.36%.

In research of Conti et al. [23], the aim was training deepfake detection methods on emotional behaviour. The motivation for this approach rose from the observation that deepfake generators failed to accurately recreate natural emotional behaviour. This led to release of a new approach incorporating high-level features from the state-of-the-art Speech Emotion Recognition (SER) systems into audio deepfake detection methods. These high-level features being analysed, such as pitch variations and speech rate, form a well-sourced base for detection models. This method was able to capture semantic audio information for enhancing detection and distinguishing emotional behaviour in audio deepfake detection research and proved to be robust in different datasets including the ASVspoof2019 (LA) dataset. The proposed method achieved an accuracy of 94% and forms an important baseline for this thesis.

3.3 *Paralinguistic Feature Extraction for Detection Methods*

With the state-of-the-art methods being able to capture emotional behaviour, a wide range of spoofing techniques perform well on short utterance evaluations, the capabilities of audio deepfake detection models are rapidly growing. In this thesis, the focus is on paralinguistic feature extraction. Paralinguistic features include elements such as stress, intonation, emotional tone and rhythm [29]. To extract these paralinguistic features, a product developed by Google Research is employed. The product, TRILLsson, created and released by Shor and Venugopalan [2] has not yet been

applied in audio deepfake detection methods but may have a lot of potential. TRILLsson, named after its foundational TRILL framework, which stands for Triplet Loss on Learned Representations, leverages knowledge distillation to create methods that can be applied on smaller devices, making paralinguistic feature extraction possible for the public. While most research on audio are of large size and done on private secured data, this research group aimed to make it a public product which could be scalable from large devices to small ones while still being powerful. The product's architecture was inspired by its parent model, the Conformer model applied to Paralinguistics (CAP12) [30, 31]. By distillation, Shor et al. [2] tried to mimic the well-performing output of the CAP12 model and compressing it into smaller models which needed to maintain that same performance. The paralinguistic embedding vectors, generated by the distilled version of the conformer model contain a rich representation of the non-verbal aspects of the audio samples. The product offers a robust framework as a base for audio deepfake detection. The model achieved an accuracy of 96% on 6 of 7 task in the Non-Semantic Speech Benchmark (NOSS) which assess for performance on a variety of speech related tasks. Next to that, TRILLsson outperformed the Wav2vec 2.0 model which is a notable performance [2]. The second research from Google Research on paralinguistic speech by Shor et al. [32] focusses on making advancements in capturing speech representations that highlight the paralinguistic aspects. With the use of a conformer-based architecture, the model was able to capture both contextual relationships in speech and local acoustic features. The model proved to be as accurate for longer time frames as for shorter ones, retaining 96% accuracy on several speech tasks [32]. To effectively capture the paralinguistic features from the ASVspoof2019 audio samples, Shor and Venugopalan [2] released the publicly available TRILLsson on the TensorFlow framework [3]. The TRILLsson model generates 1024-dimensional embedding vectors which are the vectors implemented in this thesis. They reflect a wide range of paralinguistic features and are ready to be employed for various machine and deep learning tasks.

3.4 *Implementation of Techniques for Audio Deepfake Detection*

In order to effectively handle the paralinguistic embedding vectors, a robust machine or deep learning model is required. Logistic regression, renowned for its simplicity yet effectiveness and interpretability in performing a binary classification task [25, 33], is a suitable approach for this thesis. This machine learning model employs a supervised learning paradigm and is an ideal fit with the labels provided by ASVspoof2019. Logistic regression uses the sigmoid function to calculate the probability that a given input belongs

to a particular class, in this case, either genuine or spoofed. In the training process, the model aims to minimize the difference between the predicted probabilities and the actual classifications by using optimization methods such as gradient descent to adjust the model's parameters. The final output of the model is a probability, which thresholds at 0.5, categorizing the output into the binary classes [34, 25, 33].

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

The sigmoid function, mathematically defined in Equation 1 above, maps the output of the logistic regression, a linear decision into probabilistic outputs ([0, 1]) to fit the binary classification task [35]. Evaluating and comparing different deepfake detection methods requires a robust metric that reflect the performance of each model in a concise way. The Equal Error Rate (EER) is a commonly used metric for this task. It represents the point of where the false acceptance and false rejections are equal. This is a single value, expressed in percentages where a low score reflects a better model effectiveness, accuracy and reliability [24]. In addition to the EER, the standard performance metrics such as accuracy, precision, recall and the F1-score also provide insights [36] into the effectiveness of this model compared to the state-of-the-art models in deepfake detection. With this information, a trade-off can be interpreted in what the strengths and weaknesses of a certain technique are, delivering a complete overall picture for this comparative study. The rise of deepfake technologies has initiated the creation of advanced detection methodologies. While machine learning remains a part of the process, the deep learning technologies have now become the standard for detection [37]. With the use of ASVspoof2019 [1] dataset, the implementation of paralinguistic feature extraction can be explored.

The recent studies discussed in this background, such as wav2vec models, TRILLsson and the OHEM algorithm reflect significant improvements in performance compared to the older studies or within their paper itself. The need for diverse research and variation in methodologies in detection systems keeps existing.

4 METHOD

This section outlines all the applied methods in this thesis to perform a paralinguistic feature extraction on the input data and consequently perform a binary classification task with the logistic regression architecture. Additionally, it elaborates on the pre-processing steps and the evaluation metrics used to assess the performance of the proposed model.

4.1 *Experimental Setup*

The foundation of this thesis is the ASVspoof2019 Logical Access (LA) database which is employed as the sole dataset. The size of the dataset is 80.000 audio samples stored in .flac format. The ASVspoof2019 database is designed for spoofing detection architectures [1] and with the provided labels, indicating the either genuine or spoof samples, the dataset is a strategic choice for training the logistic regression model employed in this thesis. The audio samples are processed by the TRILLsson paralinguistic feature extraction which resulted in 1024-dimensional feature embedding vectors. The embedding vectors are the input for the logistic regression model, used for the binary classification task. Evaluation metrics are EER, accuracy, precision, recall and F1-score. Next to that, the evaluation metrics are visualized together with the ROC Curve and the confusion matrix of the test set.

4.2 *Dataset*

Each audio clip within the dataset is standardized to a length of one second at a 16.0 kHz sample rate, which aligns with what the state-of-the-art methods use as average input length. This standardization of audio time length ensures consistent input data which is crucial for an effective feature extraction and classification of the audio samples. While being short utterance, the sample retains enough audio to reflect the paralinguistic information and make analysis possible. The ASVspoof2019 dataset is of large size so by this extra step the data is maintained at a manageable size.

4.3 *TRILLsson*

To extract paralinguistic features from the audio clips, the TRILLsson Tensorflow2 framework [3] is employed. The model, code provided by the thesis supervisor extracts the embeddings from each audio clip. These embedding are 1024-dimensional feature embeddings containing the paralinguistic information about all the individual audio samples. The choice of using TRILLsson [2] aligns with the aim of this thesis, incorporating paralinguistic feature extraction for audio deepfake detection methods.

4.4 *Preprocessing*

The preprocessing stage of the logistic regression architectures includes pairing the TRILLsson embeddings with their labels to do supervised

learning in the model. All label-embedding vector pairs are placed in a data frame to form one coherent input item. Due to size of the dataset and issues with pairing the labels and vectors, the input size of the dataset is 80,000 embedding vectors where the pre-defined train, validation and test set are merged in one set. Next for the model architecture, the division of the merged dataset is 70/20/10 and implemented for training (70%), validation (20%) and the test set (10%) respectfully. This division will provide the model with enough training data and comprehensive insights afterwards, leaving an amount of around 8000 test samples, which is a sufficient amount of unseen data, to assess the model's performance.

4.5 *Logistic Regression*

A logistic regression model is selected for this thesis because of its efficiency in binary classification tasks. The architecture of the logistic regression model in the provided Python script is built to suit the input data well and optimize performance, enhancing simplicity and interpretability. The input for the model is the pre-processed data frames containing the embedding vectors together with its corresponding labels being either 0 or 1, spoofed or genuine (real) audio. The model has a high iteration limit set at 10,000, allowing to correctly handle the large volume of data and ensuring adequate learning. Consequently, the input data is scaled, and the model is put in training. A step of normalization is done using 'StandardScaler', this scaling tool is meant to improve the algorithm's performance [38]in stabilizing the numerical input of the embedding vectors by standardizing the range of feature values.

4.6 *Evaluation*

The performance of the logistic regression model with the by TRILLsson generated embedding vectors as input need several evaluation metrics to assess the quality of the model. The method's performance is evaluated with accuracy, precision, recall, the F1-score, the Equal Error Rate (EER), the ROC curve and a confusion matrix. The EER is a common evaluation metric used in studies about deepfake detection. This metric projects the balance between false acceptance and rejection rate where both are equal [24], it reflects the robustness and reliability of this detection model. Next, accuracy is measured separately for the validation set as well for the test set to assess for performance of the model on that set and sets an indication for cases of overfitting. Subsequently, the precision, recall and F1-score are calculated for both sets separately to illustrate the overall performance of the model for different scenario's and provide insights into the balance of

predictions made. To visually assess the model performance, the Receiver Operating Characteristic curve (ROC) and confusion matrix are extracted for the test set, providing a detailed overview of the model's classification efficacy and error rate performance [39, 40, 41].

4.7 *Methodological Details*

The Python libraries and modules utilized for this thesis are essential to ensure robust data manipulation, visualization, processing and training. These modules and libraries are employed for all machine learning and data analysis tasks. Libraries as Librosa and Pydub are essential for audio processing the ASVspoof2019 dataset. Pandas and Numpy are employed for the pre-processing of the label – embedding vector pairs. TensorFlow with Tensorflow.keras are needed for the TRILLsson paralinguistic feature extraction process. Additionally, the libraries as Sklearn and Matplotlib are essential for the training and evaluation of the logistic regression model.

- Librosa
- Pydub
- Pandas
- Numpy
- Tensorflow
- Os
- Glob
- Kagglehub
- Pydot
- Cartopy
- Matplotlib
- Tensorflow_hub
- Libarchive
- Sklearn
- Tensorflow.keras

5 RESULTS

In this section the results of the logistic regression performing a binary classification task on the ASVspoof2019 dataset are discussed. First, the result of the model performance will be discussed. This includes the Equal Error Rate (EER), accuracy, precision, recall, the F1-score, the ROC curve and a confusion matrix. These evaluation metrics illustrate how well the logistic regression performed the binary classification task on the dataset, therefore indicating to which extent this paralinguistic feature extraction is a sound approach for this task. Next, the state-of-the-art models are shortly being reviewed to sketch the landscape which is needed to determine if the proposed model's performance aligns with the state-of-the-art performance on deepfake detection.

5.1 *Model Performance Overview*

In this result section, the performance of the proposed model is reviewed. The results presented in table 1 indicate that the logistic regression model employed for audio deepfake detection using the ASVspoof2019 dataset is a well-performing model. The model exhibits high accuracy rates, the validation set achieved an accuracy of 97.6% and the test set exhibits even a slightly higher accuracy rate at 97.9%. This indicates that the model predictions on the test set somewhat were slightly more accurate. These results reflect the model's ability to generalize well on unseen data and it suggests that the model was in fact capable of learning the features relevant for distinguishing real audio samples from fake samples. This is a positive outcome for performing a binary classification task on deepfake detection.

The precision, suggesting whether the model correctly identifies a genuine sample as genuine, is a relevant metric for this research. In this context, the occurrence of a false positive, where a spoofed sample is incorrectly identified as genuine, can have undesirable consequences, therefore, a higher precision performance is preferred in a detection model. The model has a precision of 88.6% on the validation set and 90.5% on the test set. Once again, the test set reflects scores indicating that the model is well performing.

In terms of recall, the model achieved a score of 88.6% on the validation set and 89.4% on the test set. This metric indicates the percentage of genuine audio being captured by the model, explicitly said, reflecting the proportion of all true positives correctly identified. Thus, the proposed model captured almost 90% of the genuine samples existing in the dataset. With the wide variety of spoofing techniques present in the dataset, the

model was effectively filtering out the most of spoofing techniques while being exposed to unseen data.

The scores indicate that correctly classifying the genuine and spoofed samples became slightly easier for the model after training on almost 65.000 samples. To gain insights about the balance between positive predictions of the model and the actual captured positive cases, the F1-score is used. The score, based on a mean of the precision and recall only results in a high outcome when both mentioned scores are. In this case, the F1-score for the validation score is 88.6% and resulted for the test set in 89.9%.

The model presents slightly an increase of performance between the validation set to the test set. However, these differences are not notably large. In brief, the model performed well on the validation set and showed to be well-trained and prepared for the test set as seen in the performance rates.

Additionally, the Equal Error Rate (EER) reflects the point where false positives and false negatives are equal. Comprehensively, the rate of where the number of spoofed samples being incorrectly classified as genuine and genuine samples being incorrectly classified as spoofed are equal. Both cases can have undesirable consequences and therefore a low EER is preferred. The model achieved an EER of 3.04%. The closer the EER is to 0%, the more reliable the model is considered to be. With an EER of 3.04% the model implies to be reliable and effective as a detection method.

The Receiver Operating Characteristics (ROC) curve, as shown in Figure 1 is a graphical representation of the model's performance. It illustrates the true positive rate against the false positive rate to give insights of the accuracy and recall performance of the model doing a binary classification task [40]. For the test set, the area under the ROC curve (AUC) shows a value of 0.99, which indicates that the logistic regression model performs exceptionally well on the classification task and is not neighbouring the probability that it is just guessing. The confusion matrix, shown in Figure 2, is a tool that displays the number of correct and incorrect predictions made by the model by categorizing the predictions into true positives, true negatives, false positives and false negatives. The proposed model correctly identified 756 true positives and 7104 true negatives, indicating strong capabilities of correctly classifying both categories.

Overall, the performance review of the logistic regression model demonstrates robust and promising capabilities. With an accuracy of 97.9% on the test set the model shows to be generalizable to unseen data and gives confidence to perform well on the binary classification task. Beside that, the high precision and recall rates reflect the model's ability to capture the deepfakes effectively. Additionally, the EER of 3.04% underscores the

reliability of the model. These promising results highlight the model's proficiency as an effective and innovative deepfake detection approach.

Table 1: Performance of the Model

Metric	Validation Set	Test Set
Accuracy	97.6	97.9
Precision	88.6	90.5
Recall	88.6	89.4
F1-Score	88.6	89.9

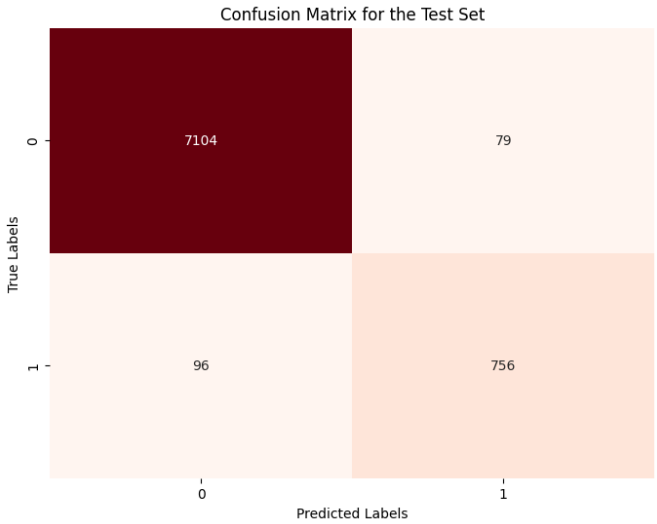

Figure 1: Confusion Matrix

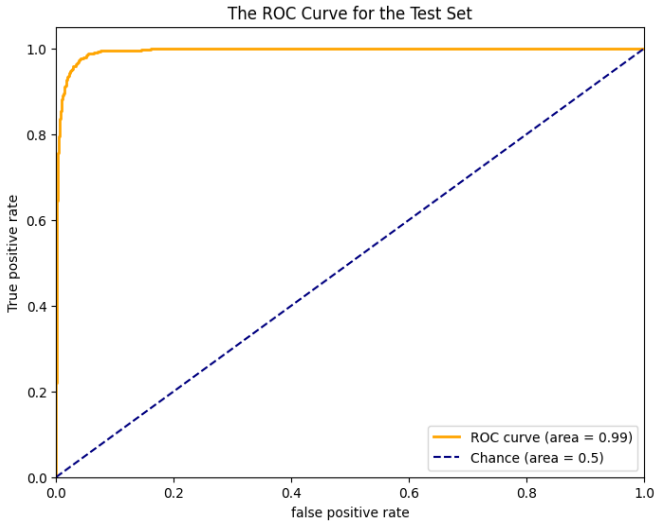

Figure 2: ROC Curve

5.2 *Comparing Results With State-of-the-Art Models*

When compared to state-of-the-art models [16, 17, 22, 15, 18, 23], the performance of the logistic regression offers this research interesting insights. In Table 2 below, all information regarding performance of all models is projected. Worth noting is that the size and division of the dataset for the proposed model is modified while most of the state-of-the-art methods mentioned are not, leading to an invalid total comparison but these models are used to put the proposed model in perspective. While most models utilised robust deep learning techniques, they achieved varying performances compared to this proposed machine learning model. For example, the light-DARTS model [22] achieved an EER of 1.08% and the OHEM algorithm approach [17] achieved an EER of 0.77% both being exceptionally high. On the contrary, the AASIST2 [18] model achieved an EER 8.36% meaning that not all state-of-the-art deep learning models are outperforming machine learning approaches. This brief review tells us that the model has not performed below standard.

5.3 *In Summary*

The performance of the model proposed in this thesis, a logistic regression approach performing binary classification on a deepfake detection dataset, exhibits strong and promising results. With an accuracy of 97.9%, precision of 90.5% and an EER of 3.04%, the model is neighbouring the state-of-the-art methods and suggests that paralinguistic feature extraction in

Table 2: Performance of the State-of-the-Art Models

ASVspoof2019	Model	Score	
		EER	Accuracy
Logical Access	SSAT	-	99.8
Logical Access	OHEM	0.77	-
Logical Access	Light-DARTS	1.08	-
Logical Access	TO-RawNet	1.58	-
Logical Access	AASIST2	8.36	-
Logical Access	SER network	-	94
Logical Access	Proposed Model	3.04	97.9

combination with a logistic regression model is an effective and reliable method.

6 DISCUSSION

In this thesis, a comparative study is done on audio deepfake detection all using the ASVspoof2019 [20] dataset. The dataset contains a large set of audio clips either being bona fide or spoof. The aim of the dataset is to provide for research in spoofing. The thesis is based on paralinguistic feature extraction in order to create new insights in how that technique performs on a dataset a lot of detection models are using. For detection, logistic regression is leveraged as model for this thesis. The detection of audio deepfakes is a binary classification task by supervised learning with the embedding vectors either representing genuine or spoofing audio clips. The performance was measured with the Equal Error Rate (EER) and the standard performance metrics such as accuracy, precision, recall and the F1-score. The performance is visualized with the ROC curve and the confusion matrix.

In the preprocessing phase, issues with uploading and processing items from the ASVspoof2019 input audio samples arose. Due to these issues is the number of genuine and spoofing samples were slightly out of balance, as for example shown in Figure 1 1 for the test set. When considering the model's performance, especially on accuracy and precision, this subtle imbalance does not invalidate the conclusion about the proposed method. Furthermore, it was not possible to perform a total comparison with other methods due to the changes in input data. However, as stated before, the aim of this thesis was to research whether the incorporation of paralinguistic feature extraction may or may not be an effective approach.

This is examined on the model's own terms and that provided sufficient information to be able to assemble a sound and reliable conclusion.

While the ASVspoof2019 dataset is well presented in this branch of research, even in a lot of recent studies [16, 17, 22, 15, 18, 23] it is not the most recent resource available. The deepfake generation field is developing rapidly [14] and with training data from 2019, it will not be as up to date as deepfakes generated nowadays. For future research, it is crucial to keep incorporating newer training sets to be able to keep up with the advancements within deepfake generation and keep pace with possible threats. The ASVspoof consortium [1] is releasing new databases every few years to facilitate for that. Second, research has shown [27] that models training on ASVspoof2019 present slight overfitting to that dataset and perform less on other datasets. This indicates that ASVspoof2019 on itself is not sufficient in some cases and multiple dataset options should be considered.

Logistic regression proved to be an effective model to perform this research on, the capabilities of the model in capturing information for the classification task was essential for a well-functioning machine learning model handling those paralinguistic embedding vectors. With the emerging developments in artificial intelligence it is, however, crucial to implement the newest advancements in machine and deep learning modelling in detection systems too. Other architectures like linear SVMs, Convolutional Neural Networks, transformers and more are interesting alternatives that have promising characteristics for deepfake detection methods. So, while the generation techniques are evolving, the available datasets and machine and deep learning techniques must also advance.

Moreover, the need for privacy and ethical considerations in big data, which is used for the generation and detection of deepfakes, is of all-time high demand [13]. Next to that, sustainable alternatives need to be considered [25] to minimize the footprint artificial intelligence models leave nowadays.

7 CONCLUSION

This thesis explored the implementation of paralinguistic feature extraction in audio deepfake detection methods. By use of the audio samples from the ASVspoof2019 dataset, the model used a well-sourced dataset containing of 80.000 genuine and spoofing audio samples. The application of the TRILLsson model, developed by Google Research, provided the research with the paralinguistic 1024-dimensional vector embeddings. With these embeddings, the logistic regression model was able to perform a binary classification on the dataset. In this process, an in-depth analysis is done

to examine whether this new approach may or may not be effective. Prior to this research, paralinguistic feature extraction was only utilised in non-semantic tasks. Incorporating this approach as a new method for audio deepfake detection methods proved to be the right step. The paralinguistic embedding vectors, generated by Google's TRILLsson [2] method achieved an impressive performance while being in the early phase of, hopefully, a long lasting aid to deepfake detection methods. Notably, the logistic regression model achieved an impressive accuracy of 97.9% and an Equal Error Rate (EER) of 3.04%, suggesting that this approach is indeed a promising and reliable approach for future deepfake detection methods. Highlighting and extracting the paralinguistic features from someone's voice will be a powerful tool to operate with. While deepfake generating models are rapidly evolving, recreating a consistent and cohesive pitch, rhythm or emotional tone will be a challenging task. Future research should explore the integration of more diverse datasets, different machine or deep learning methods and examine a variety of combinations between both. For now, we might consider taking a lead on this deepfake generation and detection sprint and begin to implement the paralinguistic feature extraction technique on a larger scale.

REFERENCES

[1] ASVspoof Consortium, "Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge," https://www.asvspoof.org/index2019.html, 2019, accessed: 14-May-2024.

[2] J. Shor and S. Venugopalan, "Trillsson: Distilled universal paralinguistic speech representations," *arXiv preprint arXiv:2203.00236*, 2022.

[3] Google, "Trillsson: Distilled universal paralinguistic speech representations," Kaggle, 2024, [Accessed: 14-May-2024]. [Online]. Available: https://www.kaggle.com/models/google/trillsson

[4] S. Ong, "Introduction to machine learning," 2023, course number: 822047-B-6, Spring semester 2023-2024, Tilburg University.

[5] S. Overflow, "Equal error rate in python," https://stackoverflow.com/questions/28339746/equal-error-rate-in-python, 2015, accessed: 14-May-2024.

[6] G. Research, "How to get whole embedding of trillsson model. #1442," https://github.com/google-research/google-research/issues/1442, 2022, accessed: 14-May-2024.

[7] OpenAI, "Chatgpt-4," https://www.openai.com/chatgpt, 2023, accessed: 14-May-2024.

[8] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial intelligence for cybersecurity: Literature review and future research directions," *Information Fusion*, p. 101804, 2023.

[9] O. N. Chisom, P. W. Biu, A. A. Umoh, B. O. Obaedo, A. O. Adegbite, and A. Abatan, "Reviewing the role of ai in environmental monitoring and conservation: A data-driven revolution for our planet," *World Journal of Advanced Research and Reviews*, vol. 21, no. 1, pp. 161–171, 2024.

[10] T. O. Oladoyinbo, S. O. Olabanji, O. O. Olaniyi, O. O. Adebiyi, O. J. Okunleye, and A. Ismaila Alao, "Exploring the challenges of artificial intelligence in data integrity and its influence on social dynamics," *Asian Journal of Advanced Research and Reports*, vol. 18, no. 2, pp. 1–23, 2024.

[11] A. D. Sontan and S. V. Samuel, "The intersection of artificial intelligence and cybersecurity: Challenges and opportunities," *World Journal of Advanced Research and Reviews*, vol. 21, no. 2, pp. 1720–1736, 2024.

[12] B. T. Familoni, "Cybersecurity challenges in the age of ai: theoretical approaches and practical solutions," *Computer Science & IT Research Journal*, vol. 5, no. 3, pp. 703–724, 2024.

[13] A. Khan and K. M. Malik, "Securing voice biometrics: One-shot learning approach for audio deepfake detection," in *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2023, pp. 1–6.

[14] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023.

[15] C. Wang, J. Yi, J. Tao, C. Zhang, S. Zhang, R. Fu, and X. Chen, "Torawnet: Improving rawnet with tcn and orthogonal regularization for fake audio detection," *arXiv preprint arXiv:2305.13701*, 2023.

[16] A. K. S. Yadav, E. R. Bartusiak, K. Bhagtani, and E. J. Delp, "Synthetic speech attribution using self supervised audio spectrogram transformer," *Electronic Imaging*, vol. 35, pp. 1–11, 2023.

[17] C. Hu and R. Zhou, "Synthetic voice spoofing detection based on online hard example mining," *arXiv preprint arXiv:2209.11585*, 2022.

[18] Y. Zhang, J. Lu, Z. Shang, W. Wang, and P. Zhang, "Improving short utterance anti-spoofing with aasist2," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 636–11 640.

[19] E. Pastor, A. Koudounas, G. Attanasio, D. Hovy, and E. Baralis, "Explaining speech classification models via word-level audio segments and paralinguistic features," *arXiv preprint arXiv:2309.07733*, 2023.

[20] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[21] T. Liu and X. Yuan, "Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 23, 2023.

[22] C. Wang, J. Yi, J. Tao, H. Sun, X. Chen, Z. Tian, H. Ma, C. Fan, and R. Fu, "Fully automated end-to-end fake audio detection," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022, pp. 27–33.

[23] E. Conti, D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, A. Sarti, M. C. Stamm, and S. Tubaro, "Deepfake speech detection through emotion recognition: a semantic approach," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8962–8966.

[24] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "Range-based equal error rate for spoof localization," *arXiv preprint arXiv:2305.17739*, 2023.

[25] S. Saha, M. Sahidullah, and S. Das, "Exploring green ai for audio deepfake detection," *arXiv preprint arXiv:2403.14290*, 2024.

[26] R. Ranjan, M. Vatsa, and R. Singh, "Statnet: Spectral and temporal features based multi-task network for audio spoofing detection," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2022, pp. 1–9.

[27] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *arXiv preprint arXiv:2203.16263*, 2022.

[28] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[29] D. Crystal and R. Quirk, *Systems of prosodic and paralinguistic features in English*. Walter de Gruyter GmbH & Co KG, 2021, vol. 39.

[30] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, and J. Qadir, "Transformers in speech processing: A survey," *arXiv preprint arXiv:2303.11607*, 2023.

[31] D. Aguirre, N. G. Ward, J. E. Avila, and H. Lehnert-LeHouillier, "Comparison of models for detecting off-putting speaking styles." in *INTERSPEECH*, 2022, pp. 2303–2307.

[32] J. Shor, A. Jansen, W. Han, D. Park, and Y. Zhang, "Universal paralinguistic speech representations using self-supervised conformers," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3169–3173.

[33] B. Bhavitha, A. P. Rodrigues, and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," in *2017 International conference on inventive communication and computational technologies (ICICCT)*. IEEE, 2017, pp. 216–221.

[34] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic regression model optimization and case analysis," in *2019 IEEE 7th international conference on computer science and network technology (ICCSNT)*. IEEE, 2019, pp. 135–139.

[35] A. ZAIDI, "Mathematical justification on the origin of the sigmoid in logistic regression," *Central European Management Journal*, vol. 30, no. 4, pp. 1327–1337, 2022.

[36] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE access*, vol. 10, pp. 25 494–25 513, 2022.

[37] D. A. Coccomini, R. Caldelli, F. Falchi, and C. Gennaro, "On the generalization of deep learning models in video deepfake detection," *Journal of Imaging*, vol. 9, no. 5, p. 89, 2023.

[38] M. M. Ahsan, M. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, p. 52, 2021.

[39] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[40] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[41] J. Liang, "Confusion matrix: Machine learning," *POGIL Activity Clearinghouse*, vol. 3, no. 4, 2022.