DESIGN AND DEVELOPMENT OF DEEP-LEARNING ENABLED AUDIO SPOOF DETECTOR

A MAIN PROJECT REPORT

Submitted by

GOURAV GOPAL 1905015
NALIN SURIYA S 1905031
VISHAL KARTHIK S 1905060
YOKESH R S 1905062

In partial fulfilment for the award of the degree

Of

BACHELOR OF ENGINEERING

In

COMPUTER SCIENCE AND ENGINEERING

COIMBATORE INSTITUTE OF TECHNOLOGY

(Government Aided Autonomous Institution Affiliated to Anna University)

COIMBATORE-641 014

ANNA UNIVERSITY - CHENNAI 600 025

OCTOBER 2022

COIMBATORE INSTITUTE OF TECHNOLOGY

(Government aided Autonomous Institution Affiliated to Anna University)
COIMBATORE – 641014

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project **"Design and development of Deep-Learning enabled audio spoof detector"** is the bonafide work of **GOURAV GOPAL (1905015), NALIN SURIYA S (1905031), VISHAL KARTHIK S (1905060), and YOKESH R S (1905062)** under my supervision during the academic year 2022-2022.

Dr.G.Kousalya, M.E.,Ph.D	**Dr.M. Mohanapriya M.E., Ph.D**
HEAD OF THE DEPARTMENT,	**ASSOCIATE PROFESSOR,**
Department of CSE	Department of CSE,
Coimbatore Institute of Technology,	Coimbatore Institute of
Coimbatore	Technology, Coimbatore

Submitted for **19CS74 MAIN PROJECT** held on

_____

Internal Examiner **External Examiner**

Place:

Date:

TABLE OF CONTENTS

ACKNOWLEDGEMENT

We express our sincere thanks to our Secretary **Dr.R.Prabhakar** and our Principal **Dr. Rajeswari A** for providing us a great opportunity to carry out our work. The following words are a small part to express our gratitude to them. This work is the outcome of their inspiration and product of plethora of their knowledge and rich experience.

We record the deep sense of gratefulness to **Dr.G.Kousalya**, Head of the Department of Computer Science and Engineering, for his encouragement during thistenure.

We equally tender our sincere thankfulness to our project guide **Dr.M. Mohanapriya M.E.,Ph.D**, Associate Professor, Department of Computer Science and Engineering, for hervaluable suggestions and guidance during this course.

During the entire period of study, the entire staff members of the Department ofComputer Science and Engineering have offered ungrudging help. It is also a great pleasure to acknowledge the unfailing help we have received from our friends.

It is matter of great pleasure to thank our parents and family members for theirconstant support and co-operation in the pursuit of this endeavor.

ABSTRACT

Authentication has become very important aspect of our day to day lives starting from normal lock screen pin to human retina based authentication systems. One among those popular and complex authentication systems are the audio based authentication systems where in people use certain words to unlock devices and objects like mobiles, doors etc., Audio Authentication generally involves authentication based on words and voices. Issue in the existing system is that the system verifies and extracts the features of words and voices but it does not classify human voice and recorded human voices.

The above issue can be overcome by using models like RNN and LSTM for classification.

CHAPTER 1

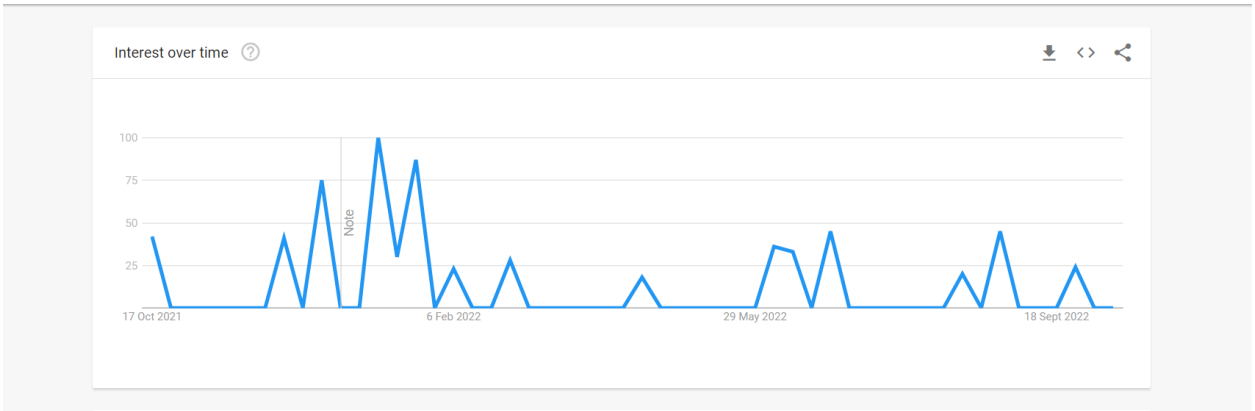
INTRODUCTION

1.1. INTRODUCTION:

Development of VCD's, have boosted the realization of smart homes, voice-controlled authentication systems etc.,.These VCDs are vulnerable to different spoofing attacks .Audio Authentication is becoming very essential part of our lives. Audio Authentication spoofing is becoming an issue. High-quality audio recorders enable bypassing this audio authentication system by just recording the human voice and reusing them for accessing the same system. Thus, there exists a need to develop a voice anti-spoofing framework capable of detecting multiple audio spoofing attacks.

1.2. Google trend analysis:

The diagram shown below displays the google trend analysis about the domain of this project.



CHAPTER 2

LITERATURE SURVEY

2.1 Fusion of BiLSTM and GMM-UBM Systems for Audio Spoofing Detection (2019)

Authors: Ivan Rakhmanenko , Alexander Shelupanov , Evgeny Kostyunchenko

Description

In this study, Bidirectional Long Short Term Memory (BiLSTM) networks with constant Q cepstral coefficients (CQCC) are used to classify real audio from fake audio in anti-spoofing systems.

By fusing the BiLSTM and GMM-UBM systems, a fusion mechanism is used to increase the variability of the systems' decision-making processes and their accuracy.

Over the baseline systems, these proposed systems significantly improved performance.

Advantages

- Uses time domain dependency relation in audio and gives better results.
- Uses fusion mechanism to improve accuracy.

Disadvantages

- It doesn’t provide classification model for classification of various attacks.

2.2 Improving anti-spoofing with octave spectrum and short-term spectral statistics information (2020)

Authors: Jichen Yang , Rohan Kumar Das

Description:

In this paper, feature level exploration is done and a new feature is formed that can capture improved discriminative information between real and fake audio from that of the existing features that are already explored.

Similar to most of the systems being used nowadays, the spoofing detection system used for this work contains a front end feature extractor and a back-end classifier.

The feature extractor extracts effective information and whereas the classifier identifies whether an audio is genuine or spoofed.

Advantages

- The systems combines linear power spectrum and Octave power spectrum to produce better to form better novel feature eCQCC indeed produces good results.

- Works for Synthetic and Playback Spoof attacks.

Disadvantages

- Does not prove to be much effective in classification of audio spoofed data.

2.3 Spoofed Speech Detection with Weighted Phase Features and Convolutional Networks (2022)

Authors: Gökay DISKEN

Description:

In this paper, features are extracted using a frame-wise weighted magnitude spectrum and these are found to be effective on replay attacks. Along with frame-wise weighted spectrum, a cosine normalized phased spectrum is used to attain better performance as phase-based features have shown decent performance for this particular task.

These extracted features are then sent to the Convolutional Neural Network as input and then are classified as either synthetic or replay attacks.

Advantages:

- Frame –wise weighted magnitude spectrum results in more effective method for spoof detection.
- Discusses various Attacks like synthetic, replay attacks and S10.

Disadvantages:

- Takes much computational cost for training.

2.4 Detection of Various Speech Forgery Operations Based on Recurrent Neural Network (2020)

Author: Diqun Yan and Tingting Wu.

Description:

In this paper, feature extraction methods like LFCC and MFCC are used to extract audio features and then these are sent as input to the Recurrent Neural Network (RNN) frame with two-layer LSTM to detect four common audio forgery operations.

These are experimented mainly on TIMIT and UME databases and various evaluations like intra-database evaluation as well as cross-database evaluation are done and the detection accuracies of each of the above are identified.

Advantages:

- Feature extraction techniques like MFCC and LFCC are used to extract audio features..
- In this work, RNN is used as it can capture the correlation between the frames in a speech recognition application.
- Hence it is considered better than CNN as it does not capture the sequential correlation well.

Disadvantages:

- The cross-database evaluation accuracy could be improved in this model.

2.5 Fake Audio Speech Detection (2020)

Author: Shilpa Lunagaria, Mr. Chandresh Parekh

Description:

In this paper, deep fake audio forgery is identified using Deep Learning algorithms. Audio files are taken as input and model is trained to uniquely identify features for voice creation and voice detection. The model could then classify between whether the audio is real or fake.

The accuracy obtained for this model during training and validation phases are pretty high but the testing accuracy could be improved more by extracting more features and using different algorithms.

Advantages:

- The real and fake voices can be identified.
- The accuracies obtained for training, validation are considerably high (99%, 95% respectively).

Disadvantages:

- This work only focuses on deep fake audio forgery and it doesn't detect or identify other audio forgery operations.
- The testing accuracy in this model could be improved with better algorithms (just 85% accuracy).

2.6 Deep Learning based DFWF Model for Audio Spoofing Attack Detection (2020)

Author: Kottilingam Kottursamy

Description:

In this paper, fake audios are detected without continual-learning based methods and fakes are also detected without forgetting so as to develop a model that can detect spoofing attacks in an incremental fashion.

To retain the original model memory, a method known as knowledge distillation is introduced. This model using Deep Learning techniques reduces the error rate up to 80%.

Advantages:

- Overcomes the demerit of loss of previous data by introduction of knowledge distillation loss.

Disadvantages:

- Poor performance in detection of unseen data.
- Requires both original and new data.
- Storage of data.

2.7 Voice spoofing countermeasure for voice replay attacks using deep learning. (2022)

Author: Jincheng Zhou, Tao Hai1, Dayang N. A. Jawawi, Dan Wang, Ebuka Ibeke and Cresantus Biamba

Description:

The paper discusses about the immense usage of Automatic Speaker Verification (ASV) system which verifies users with their voices and it's susceptibility to voice spoofing attacks - logical and physical access attacks.

A secured voice spoofing countermeasure to detect voice replay attacks is proposed. This has enhanced the ASV system security by building a spoofing countermeasure dependent on the decomposed signals that consist of prominent information. It uses two main features— the Gammatone Cepstral Coefficients (GCC) and Mel-Frequency Cepstral Coefficients (MFCC) — for the audio representation. For the classification of the features, Bi-directional Long-Short Term Memory Network in the cloud, a deep learning classifier.

Numerous audio features and respective feature's capability to obtain the most vital details from the audio for it to be labelled genuine or a spoof speech is examined. Furthermore, it uses various machine learning algorithms to illustrate the superiority of the system compared to the traditional classifiers.

The results of the experiments were classified according to the parameters of accuracy, precision rate, recall, F1-score, and Equal Error Rate (EER). The results were 97%, 100%, 90.19% and 94.84%, and 2.95%, respectively.

Advantages:

- Avoids replay attacks in ASV.
- Voice biometrics.
- More accurate with the method of Speech Decomposition.

Disadvantages:

- Does not discuss about many audio spoof attacks, focuses on Replay attacks only.

2.8 LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems (2022)

Author: Mohit Dua, Chhavi Jain, Sushil Kumar

Description:

In this paper , the system that is proposed tries to address the problem of classifying legitimate user and the malicious attacks using deep learning (DL) methods and ensemble of different neural networks. The first model that is discussed is a combination of time-distributed dense layers and long short-term memory (LSTM) layers.

The other two deep neural networks (DNNs) are based on temporal convolution (TC) and spatial convolution (SC). Finally, an ensemble model comprising of these three DNNs has also been analysed. All these models are analysed with Mel frequency cepstral coefficients (MFCC), inverse Mel frequency cepstral coefficients (IMFCC) and constant Q cepstral coefficients (CQCC) at the frontend, where the proposed ensemble performs best with CQCC features.

The proposed work uses ASVspoof 2015 and ASVspoof 2019 datasets for training and testing, with the evaluation set having speech synthesis (SS) and voice conversion (VC) attacked utterances. Performance of proposed system trained with ASVspoof 2015 dataset degrades with evaluation set of ASVspoof 2019 dataset, whereas performance of the same system improves when training is also done with the ASVspoof 2019 dataset.

Advantages:

- Various feature extraction techniques are used which extracts audio quality features.
- LSTM models are used which increases performance.
- Ensemble method is used to get a consolidated decision from models.

Disadvantages:

- Performance on combined dataset is low.
- Deep CNN can be used to improve performance.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

This chapter explains the entire methodology that has been used in the system in detail. The methods used are Recurrent Neural Networks and Long Short Term Memory.

3.2 METHODS USED

3.2.1 Recurrent Neural Network:

A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes can create a cycle, allowing output from some nodes to affect subsequent input to the same nodes. This allows it to exhibit temporal dynamic behaviour. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. Recurrent neural networks are theoretically Turing complete and can run arbitrary programs to process arbitrary sequences of inputs.

The term "recurrent neural network" is used to refer to the class of networks with an infinite impulse response, whereas "convolutional neural network" refers to the class of finite impulse response. Both classes of networks exhibit temporal dynamic behaviour. A finite impulse recurrent network is a directed acyclic graph that can be unrolled and replaced with a strictly feedforward neural network, while an infinite impulse recurrent network is a directed cyclic graph that cannot be unrolled.

3.2.2 Long Short Term Memory:

Long short-term memory (**LSTM**) is an artificial neural network used in the fields of artificial intelligence and deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. Such a recurrent neural network (RNN) can process not only single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition, machine translation, robot control, video games, and healthcare. LSTM has become the most cited neural network of the 20th century.

The name of LSTM refers to the analogy that a standard RNN has both "long-term memory" and "short-term memory". The connection weights and biases in the network change once per episode of training, analogous to how physiological changes in synaptic strengths store long-term memories; the activation patterns in the network change once per time-step, analogous to how the moment-to-moment change in electric firing patterns in the brain store short-term memories. The LSTM architecture aims to provide a short-term memory for RNN that can last thousands of timestamps, thus "**long** short-term memory".

REFERENCES

- Dua, M., Jain, C. & Kumar, S. LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems. *J Ambient Intell Human Comput* **13**, 1985–2000 (2022). https://doi.org/10.1007/s12652-021-02960-0
- Yamagishi, Junichi; Todisco, Massimiliano; Sahidullah, Md; Delgado, Héctor; Wang, Xin; Evans, Nicolas; Kinnunen, Tomi; Lee, Kong Aik; Vestman, Ville; Nautsch, Andreas. (2019). ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2555.
- https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn
- https://colah.github.io/posts/2015-08-Understanding-LSTMs/
- Sak, H., Senior, A.W., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *INTERSPEECH*.
- Ankur, Tanjemoon & Kundu, Bipasha & Foysal, Md & Ortiz, Bengie & Chong, Jo. (2022). LSTM-Based COVID-19 Detection Method Using Coughing. 10.21203/rs.3.rs-2106413/v1.
- Akyol K, Şen B. Automatic Detection of Covid-19 with Bidirectional LSTM Network Using Deep Features Extracted from Chest X-ray Images. Interdiscip Sci. 2022 Mar; 14(1):89-100. doi: 10.1007/s12539-021-00463-2. Epub 2021 Jul 27. PMID: 34313974; PMCID: PMC8313418.
- Ivan Rakhmanenko Fusion of BiLSTM and GMM-UBM Systems for Audio Spoofing Detection August 2019 International Journal of Advanced Trends in Computer Science and Engineering 6(4):1741-1746.
- Jichen Yang, Rohan Kumar Das,Improving anti-spoofing with octave spectrum and short-term spectral statistics information,Applied Acoustics, Volume 157,2020,107017,ISSN 0003-682X,
- https://ijirt.org/master/publishedpaper/IJIRT149877_PAPER.pdf
- Mittal, Aakshi & Dua, Mohit. (2022). Automatic speaker verification systems and spoof detection techniques: review and analysis. International Journal of Speech Technology. 25. 10.1007/s10772-021-09876-2.
- https://irojournals.com/aicn/article/pdf/4/3/4
- https://journals.pan.pl/dlibra/publication/141648/edition/123487/content/archives-of-acoustics-2022-vol-47-no-2-spoofed-speech-detection-with-weighted-phase-features-and-convolutional-networks-br-disken-gokay?language=en
- Zhou, J., Hai, T., Jawawi, D.N.A. *et al.* Voice spoofing countermeasure for voice replay attacks using deep learning. *J Cloud Comp* **11**, 51 (2022). https://doi.org/10.1186/s13677-022-00306-5