CANDIDATES DECLARATION

I hereby authenticate that the work being presented in this report **Auxiliary attention pooling network-based recording device detection system**, in partisan fulfillment of the requirement for the award of the Degree of **Bachelor of Technology** is an authentic record of my research work conducted during the period *June 2022* to *September 2022* under the supervision of **Dr.Vinal Patel**. All the imported references in the figure, papers and tables have been cited.

Date: Signatures of the Candidates

This is to certify that the above declaration is true to the best of my knowledge.

Date: Signatures of the Research Supervisor

CHAPTER 1

Introduction and Literature Review

This chapter includes the introduction, motivation and literature review of the related work and the objective of this thesis.

1.1 Introduction

Identification of the audio device is the most important piece of information that could be extracted from an audio file [1] . In theory, each audio device is unique based on the number of parts, their quality, and their architecture, and every small detail has a contribution to making that device unique. This makes it possible to classify devices using the audio file recorded on that system. But in practice to identify these small irregularities and inaccuracy is a difficult task. The quality of audio microphones is increasing day by day. With the new advancements in science and technology, it is almost unreal to see how clear and crisp the audio recorder is by microphones. It is impossible for a mere human to even try to differentiate microphones based on their recorded audio file. This advancement in microphone technology makes our task of classifying audio based on their source microphone very difficult [2]. The only way to accurately predict the microphone is to identify what types of noises and imperfections they part to the sound when they are recorded. Also, there is constant research going on the opposite side on improving the quality of Deepfakes which aims at making changes in forging the audio files and still look as authentic as the real ones [3].

1.2 Problem/Motivation

The ability to correctly identify the source microphone based on the audio file can open doors to multiple segments of real-life applications. It could be used in audio forensics [11] to check the credibility of an audio signal to make the system more robust to forged or tampered audio files. They could also be applied to detect audio bots [4]

.Contribute to Improvement in the field of speech-to-text conversion [5]. The quality of the recorded audio signal can also be improved by removing the unique noise. In the future, this work could be expanded to distinguish among phones of the same model but with different serial numbers.

1.3 Literature Review

Several Studies have been conducted throughout the decade to identify a microphone source from the sound it recorded. Past research has hinted that all mechanical devices have certain inaccuracies or slight irregularities and exploitation of this shortcoming is the best path forward. To add to this uphill battle no sizeable audio corpus has yet been officially open-sourced.As discussed in [6] Deep learning approaches have fared far better in this area but they are still not performing at their potential due to limited data bringing us back to the old-age tradeoff between variance and bias [7]. The most common wall here is the problem of a very simple corpus. Our system aims to classify audio samples to the audio devices they were recorded on. The simplest workflow to acquire this information is to process the audio files, extract features from them and then classify them based on feature of their respective class. There have been several attempts in this area. As discussed in [8], proposed that to fully focus on noise imparted by the audio device is to carefully extract particular segments of silence as these would have the least influence on speech signals and help the model to classify audio based solely on the noise signals.As discussed in [9], worked on the problem of identifying if two audio files were from the same device or not. They used MFCC features and K-SVD algorithm for the identification along with the KISS metric.

As discussed in [10], proposed that an audio signal is made up of two signal noise signals and a speech signal. Although the speech signal is generally accepted as the feature to move forward with. So multiple techniques have been used aimed at removing noise from the audio sample to make the speech clearer. But noise signals can be viewed as a fingerprint of an audio device as these are unique to each device.The noise traces if and can be isolated would be incredibly useful for audio device identification. They removed the original speech from the audio and extracted the Fourier coefficient histogram of the signal as the feature vector, which has a powerful descriptive capability for audio signals, Then multiple models were trained upon this data like MLP, CNN, and Softmax Regression model. These were then averaged. Also to avoid the penalty of having one model being a misfit,voting was used.As discussed in [11], they concentrated on the preprocessing aspect of the data. Rather than cleaning the data and removing the noise they used alternately approached and focused on eliminating the speech from the audio. The idea is that this noise that is introduced while recording is unique and attempted to identify this transfer function. Features were extracted

using MFCC. Then using Kmeans clustering to segregate all points to the k devices. As discussed in[6], they proposed a system to improve the efficiency of MFCC parameters. As MFCC is the most common and widely used preprocessing algorithm to retrieve features from audio files. More interest was poured into frequency domain features over time domain features, as it has been observed that frequency domain features have performed better. Audio recordings are split into frames of 20-30 ms and MFCC coefficients are used as features. The database used was the same as ours the [12] mobiphone dataset. They were able to achieve a 66% accuracy.They used the GMM-UBM model for the classification part. As discussed in [13], their work proposed using an alternate attention layer for audio recording classifications. They also used filter bank features to extract information from the audio files and then pass it to the classification model. They also experimented using the data of whispered tone speech which was self-recorded and developed by them. These helped in understanding the practicality of real-life situations. They were able to achieve 84% accuracy on the [10] Mobiphone dataset.

As discussed in [16], their work was also aimed at using a distinctive specific signature that an audio device gives to an audio signal while recording. Using this distinctiveness along with a CNN classification model which would learn this difference was their proposed plan. They used DFT for frequency domain representation of the audio signal. The model was made using multiple simple Conv 1D layers followed by multiple dense layers to give the final output. As researched in [14], recently attention implementation using transformer in IWSLT 2022 has given recommendable results on speech-to-text conversions,proving the ability of attention to correctly process the sequential information present in the audio signal. As researched in [15],their work was aimed at identifying the phone model used to record a video. They made two different detectors both based on CNN which jointly exploited the audio and visual information present in the frames of the video and analyzed them. The first one applied a voting system to detect the best features of two CNN-based models. The second detector made the combined decision and was given both models' outputs. It is performed by jointly analyzing video and audio data. They successfully showed that using different detectors/models gave a huge advantage over just a single model.

1.4 Objective

Recorded audio can be of any type - mono or stereo, indefinite period, recorded on any microphone. Our objective here is to make a system/pipeline workflow that can accurately identify the identity of the source microphone from the recorded audio sample. The paper focuses more on the modeling and training side of the problem to improve the quality of research. We will also be predicting the gender of the user through the

voice in the audio sample.

1.5 Research Gaps

The main limitation that could be found in all the above approaches was that the system was always able to perform well on the training part but was not able to generalize it to new test samples. Also, there was a huge problem with having open-sourced good-quality audio data.Mobiphone [10] dataset only had 24 audio files for each device which was too low to use any model up to its maximum efficiency. Weighing in all the factors two decisions were set in stone, first one was the use of dual implementation of CNN-based and attention-based models,the second wall was how to overcome the overfitting encountered by other researchers. We proposed a supervised auxiliary attention pooling and CNN-based neural network.