Auxiliary Attention Pooling Network-Based Recording Device Detection System

Devansh Chowdhury

Supervisor: Dr. Vinal Patel

ABV-IIITM, Gwalior

September 26, 2022

विश्वजीवनामृतं ज्ञानम्

Table of Contents

1. Introduction
2. Literature Review
3. Objective
4. Novelty
5. Methodology
6. Experimental Results and Conclusions
7. Conclusion
8. References

Introduction

- Speech content conveyed by a speech recording is usually the most important information.
- Although some other information carried by a speech recording can also be useful.
- Speech recording also embed the information about the recording device, such as the microphone, that is used to record the speech.
- 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 gave rise to two branches of research.
- First, where the whole signal was regarded as being useful feature.
- Second was where the noises that arose from mechanical imperfections from signals were considered.

Introduction

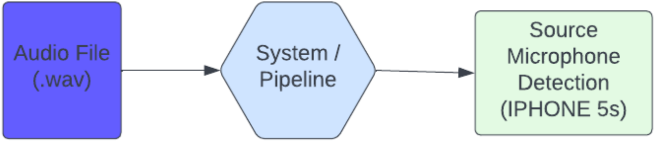


Figure 1: Classification of source microphone

Motivation

- The ability to correctly identify the source microphone based on the audio file can open doors to multiple segments of real-life applications.
- In criminology and forensics, determining the audio recording device can help determining whether a certain record is from a proper device and thus determining its validity.
- In copyright disputes, finding out the actual ownership of a certain record may help deal with multiple claims of ownership.
- Different pieces of one recording must show different recording devices, else we can infer that the record may have been modified.

Research work flow

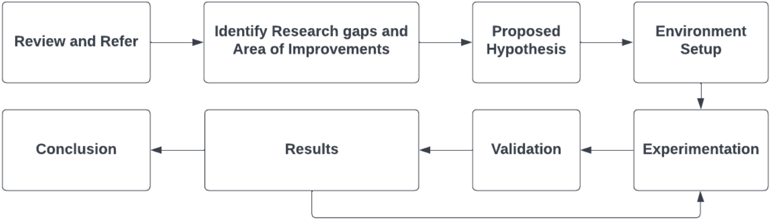


Figure 2: Research flowchart

Literature Review

- C.Kotropoulos and S. Samaras discussed the extraction of MFCC features from audio file and using gaussian mixture model for the classification [1].
- B. Singhal, AR. Naini, and PK. Ghosh discussed the idea of using attention network for classification of audio files for neutral speech and whispered speech by making a customized dataset [1].
- C.Jin, R.Wang, D.Yan, B. Tao, Y. Chen and A. Pei discussed the importance of targeting silent segments in the audio file to focus solely on noises [2].

Research Gap

- 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.
- There was a huge problem with having open-sourced good quality audio data.
- Mobiphone dataset only had 24 audio files for each device which was too low to use any model up to its maximum efficiency [3].
- Considering all the factors two decisions were set in stone, first one was the use of dual implementation of CNN based and attention-based models and the second being finding a way to overcome the over fitting encountered by other researchers.

Objective

- 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.
- We would be focusing more on the modeling and training side of the problem to improve the quality of predictions.

Novelty

- The novel idea here is to avoid over fitting we will be tasking our model to predict the gender of the user also.
- This additional prediction is known as auxiliary task.
- No other research paper has used this important metadata information to increment the model performance
- This way our system would be multitasking which would make it difficult to start over fitting the dataset which was the main problem in all the research papers as the open-sourced dataset for such a problem is very small in size.

Proposed model Workflow

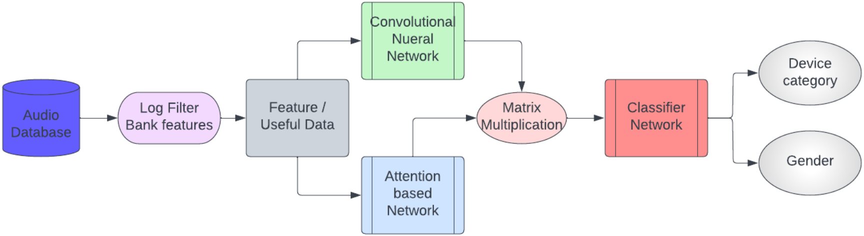


Figure 3: Proposed Model Workflow consists of three blocks - CNN, LSTM and Classifier Network.

Proposed model architecture

The proposed model workflow in Figure 3 is explained below -

- The model architecture comprises the feature being extracted from the audio file and are passed as inputs to both CNN and LSTM-based Networks.
- The CNN network as shown in Table 1 is made up of 3 blocks consisting of a 2D convolutional layer, batch normalization followed by average pooling.
- The LSTM network as shown in Table 2 is made up of two simple LSTM layers followed by average pooling.
- These two outputs are multiplied with each other and applying average pooling again.
- The result as shown in Table 3 is passed onto the Classifier network to predict the mobile device category and gender of the user.

Proposed convolutional network

Table 1: Implementation details of proposed convolutional architecture

No.	Layer	Filters/Pooling
1	Conv 2D 3x3, ReLu	20
2	Batch Normalizaton	–
3	Average Pooling 2D ReLu	(2,1)
4	Conv 2D 3x3, ReLu	20
5	Batch Normalizaton	–
6	Average Pooling 2D ReLu	(1,2)
7	Conv 2D 3x3, ReLu	20
8	Batch Normalizaton	–
9	Average Pooling 2D ReLu	(2,1)

Proposed attention and classifier network

Table 2: Implementation details of proposed Attention Architecture

No.	Layer	Filters/Pooling
1	LSTM, ReLu	24
2	LSTM	1
3	Activation, Sigmoid	–
4	Average Pooling 1D ReLu	–

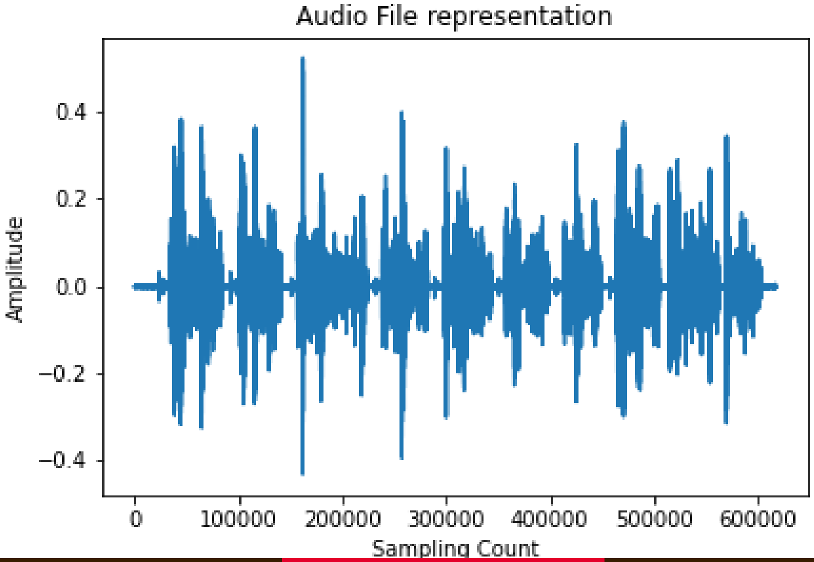
Table 3: Implementation details of proposed Classifier Architecture

No.	Layer	Filters/Pooling
1	Global Average Pooling 1D	24
2	Dense, ReLu	30
3	Dense, Softmax, (Device Classification)	21
4	Dense, Sigmoid, (Gender Classification)	1

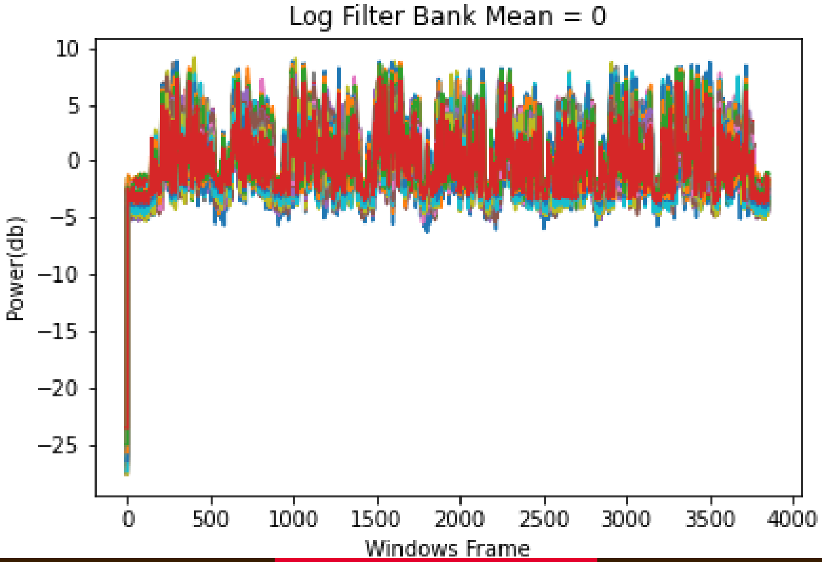
Data preprocessing and Feature extraction

- Mobiphone dataset comprises 21 different mobile devices with 24 audio files each comprising 12 males and 12 females.
- We divide each audio signal into mutiple window segments as in Figure 4.
- Compute the power spectrum for each of the window frame using short term fourier transformation.
- Compute Mel-filterbank coefficients.These are numerical values extracted from important part of audio and has most value. It mimic cochlea type for human.
- Multiply the results obtained in last two steps and taking log, these are my log filter bank features. They are compact representation of an audio signal as shown in Figure 5.

Feature extraction



Feature extraction



Experimental Setup

1. Open sourced dataset named mobiphone is used.
2. Data is divided into 3 fold.Each one would be use for training, validation and testing respectively with zero overlap.
3. Total number of fold for each experminent is six and each fold ran for 30 epoch. Best validation accuracy model was used each time.
4. Adam optimizer is used.
5. Categorical crossentropy loss function is used.
6. Window Frame size = 25ms, Window Frame step = 1ms
7. Epoch = 30

Experiment 1

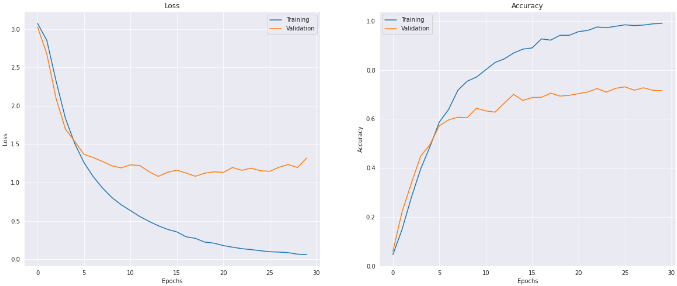


Figure 6: Baseline model made up of convolutional and attention network.

Experiment 2

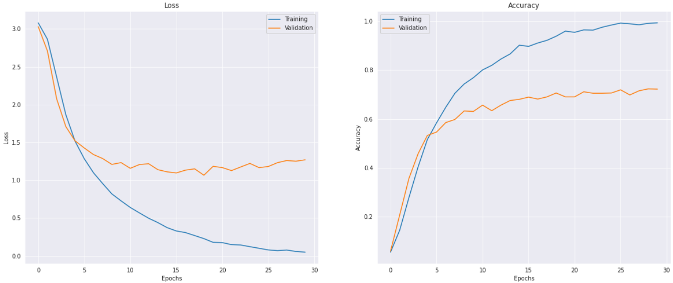


Figure 7: 3 fold cross-validation based on Speaker

Experiment 3

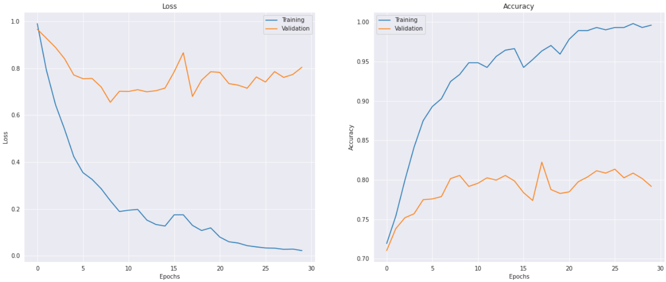


Figure 8: Auxiliary output architecture

Experiment 4

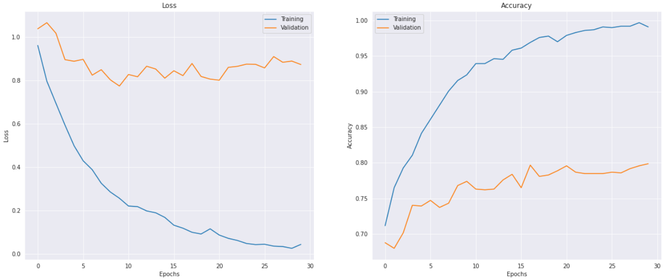


Figure 9: White noise augmentation

Experiment 5

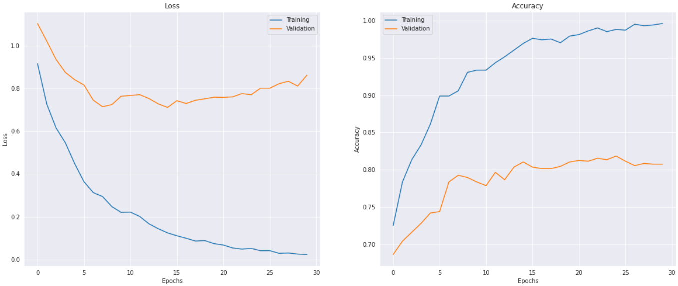


Figure 10: Kernel size hyper tuning

Experiment 6

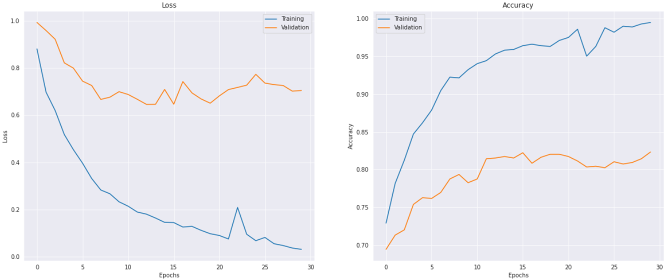


Figure 11: Loss weightage ratio hyper tuning

Experimental results

Table 4: All Experiments Result Comparison on basis of Validation & Test Accuracy

Model	**Validation**	**Test**
Simple CNN + LSTM based network	73.4%	70%
Speaker based Cross Validation	68.6%	64.8%
Auxiliary Model Architecture	75.9%	77.7%
White Noise Augmentation	77.38%	81.1%
Kernel size optimization	82.89%	81.01%
Loss Weightage Optimization	82.6%	80.2%
Best 3 out of 6 Model	87.5%	86.9%

Experimental Conclusion

- Cross-validation performed better on the device category.
- Auxiliary model architecture also gave a boost to our accuracy score.
- White noise addition also helps us push the score to 80%+.
- Tweaking kernel size and loss weightage proportion didn't yield any improvement.
- We then used the best three models out of the six-fold models which gave us 86.9% accuracy.

Conclusion

- We have successfully implemented a Multitasking based Auxiliary Attention Pooling Network for recording device classification.
- We used the top three models out of six that had the highest validation accuracy and were able to successfully train a system that gave an 86.9% accuracy on test data.
- We achieved a better accuracy score higher than any other published results by 2.9%.
- In the future, this could be used to improve the efficiency of speech-to-text conversion, individual smartphone device identification,etc.

References I

[1] B. Singhal, A. R. Naini, and P. K. Ghosh, "wspire: A parallel multi-device corpus in neutral and whispered speech," in *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 146–151, 2021.

[2] C. Jin, R. Wang, D. Yan, B. Tao, Y. Chen, and A. Pei, "Source cell-phone identification using spectral features of device self-noise," vol. 10082, pp. 29–45, 02 2017.

[3] C. Kotropoulos and S. Samaras, "Mobile phone identification using recorded speech signals," in *2014 19th International Conference on Digital Signal Processing*, pp. 586–591, 2014.

Thank You