speakertest

December 17, 2023

0.1 Abstract

In this project, we deal with the problem of detecting whether a person you are talking to on the phone has placed you on speakerphone or not. We run a series of experiments on data collected via speakerphone and non-speakerphone calls. We show that when background noise is isolated from speech, we are able to perform this detection with high accuracy, but the same models do not work when speech is involved. We conclude that the way forward for this problem is to segment out audio regions where the person is not talking, and run the speakerphone detection on these regions.

0.2 Introduction

Detecting whether or not you are placed on speakerphone is a challenging and fun problem. Its use cases can range from relatively innocent – this can serve as a clue for when you are being prank-called– to perhaps more serious, in the cases of kidnapping, etc. In fact, this Reddit post has a request for such a feature.

Surprisingly, we find almost no work done has been done on this problem.

However, if one looks at a broader problem of when one inadvertently reveals information that they don't mean to reveal about themselves in audio and video contexts, there are several interesting works. To list a couple:

1. In *"A Practical Deep Learning-Based Acoustic Side Channel Attack on Keyboards"(1)*, researchers are able to predict what is being typed by the audio emited by the typing via a Zoom call.
2. In *"The visual microphone: Passive recovery of sound from video" (2)*, researchers are able to detect audio being played from vibrations of objects in the same room.

For the speakerphone problem, our approach is to collect data via speakerphone and non-speakerphone, and see if a machine learning model can learn the difference to then predict this.

0.3 Data Collection

From a basic survey, it seems:

1. Phone call datasets with and without speaker are not available. The usual reason to collect such data is detection of various speech properties, which is different from our usecase.
2. Another idea would have been to collect video datasets and use the existence or non-existence of a headset to conclude whether the audio is on speaker or not. However, even such datasets are not readily available.

Thus, we had to collect data ourselves. We used a Google Pixel 4a Android phone, and made phone calls to our laptop using Google Voice. We collected samples with and without speakerphone on. We then converted these mp3 files to Mel-frequency cepstral coefficients(MFCC's) which is a standard way of obtaining features from audio files.

0.4 Experiments

First, we test our set up (conversion from MP3 files to MFCC's, creating the dataset, and running a basic SVM classifier on some dummy data. As expected, when both class-labels are from the same MP3 file, the classifer performs horribly (Accuracy shown below:)

```
0.25
```

And when the two classes are different songs, a simple classifier performs well:

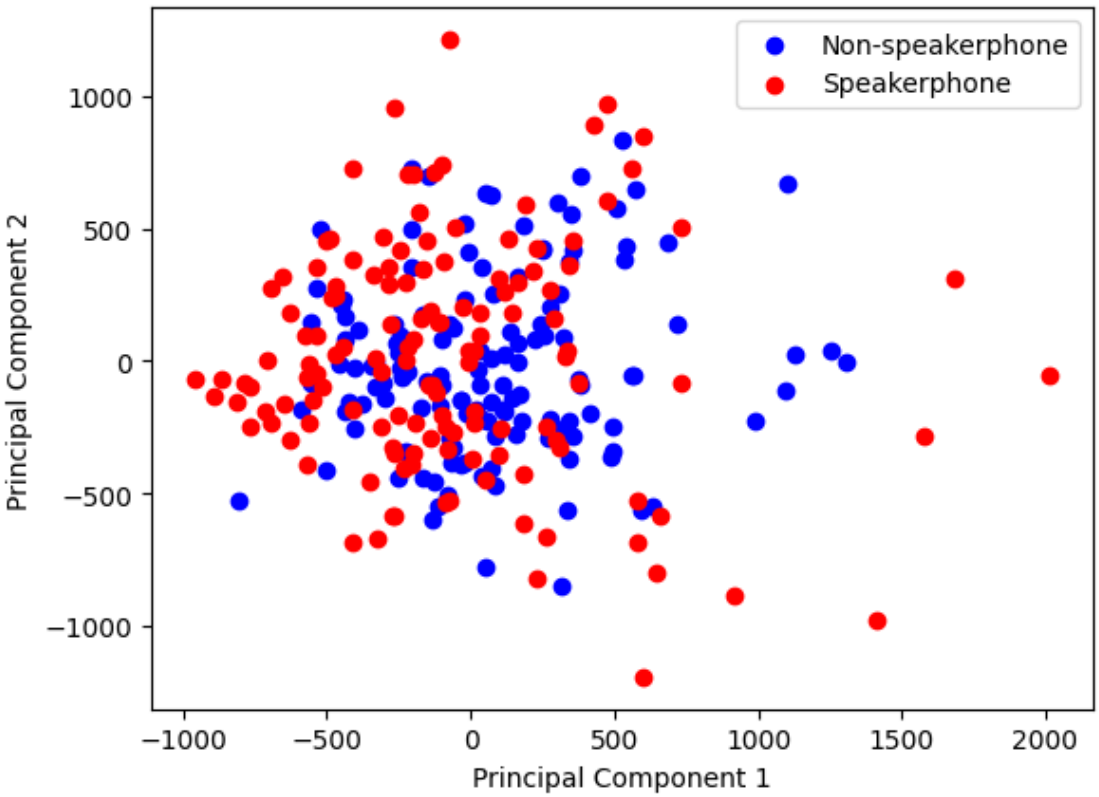
```
0.9743589743589743
```

0.4.1 Experiment 1

We play a recording of a Youtube video with and without speaker. This is about 7 minutes long, divided into segments of 3 seconds. We then run a basic SVM classifier on this data to get the following accuracy:

```
0.7857142857142857
```

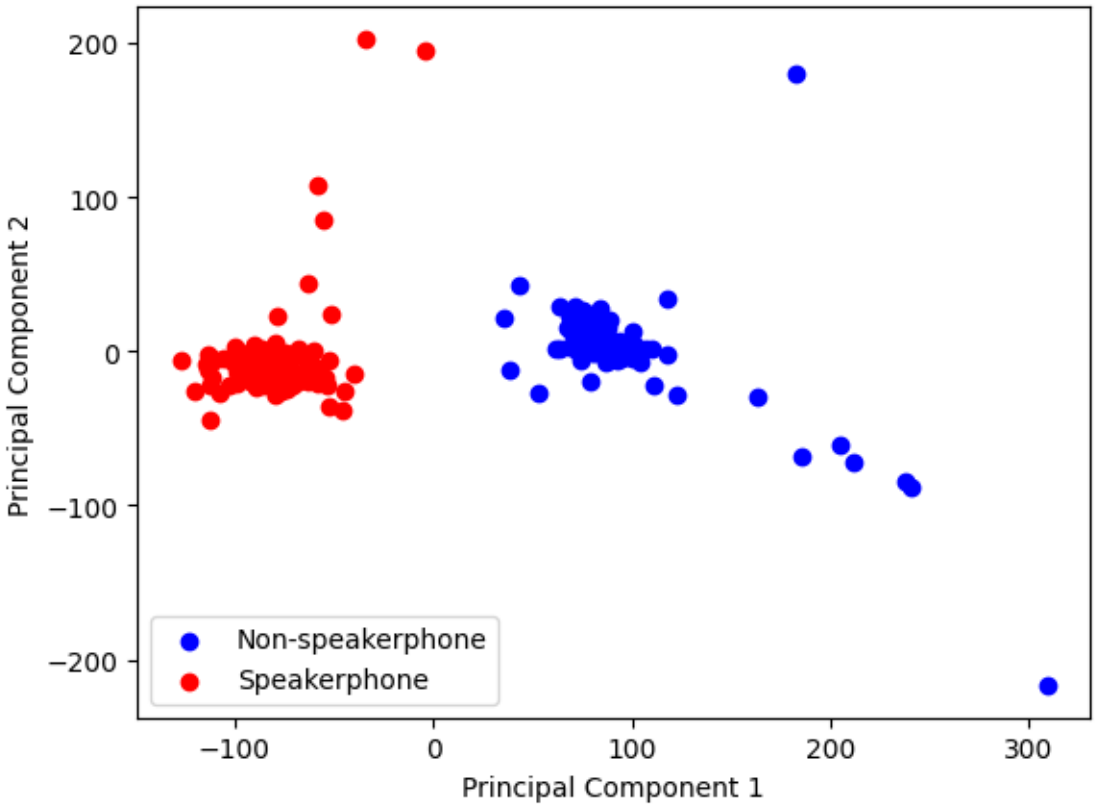
The accuracy is okay, but still not great for 2-class classification. We try to plot the principal components of both the speakerphone and non-speakerphone data points, and we find that there does not seem to be much separation as shown below. It seems like the speech of the Youtube video is interfering too much with the detection of speakerphone/ non-speakerphone.



0.4.2 Experiment 2

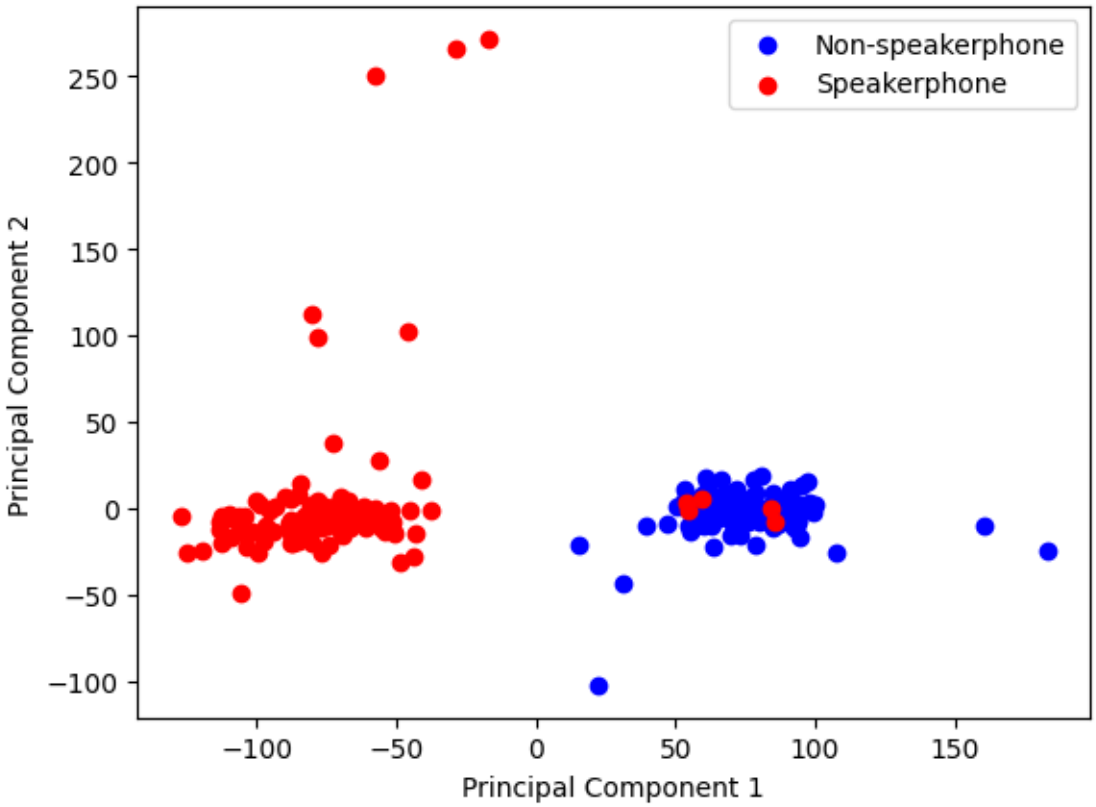
We try another strategy. This time we record only the background noise of a room both on speaker and without speaker. While the simple classifier still performs badly (accuracy shown below) when we plot the principal components, they seem to be further apart from each other.

`0.5853658536585366`



The previous samples were taken with a 3 hour gap. We confirm that the same thing holds when the samples are taken at nearly the same time (definitely the same background noise in both speakerphone/ nonspeakerphone cases). Again the SVM accuracy is bad, but the principal components plot makes us hopeful.

```
0.391304347826087
```



0.4.3 Experiment 3

Looking at the above plot, it seems like a Linear Discriminant Analysis (LDA) classifier would perform well. Indeed, it turns out to be the case, and the above performs well on test data with excellent accuracy:

```
1.0
```

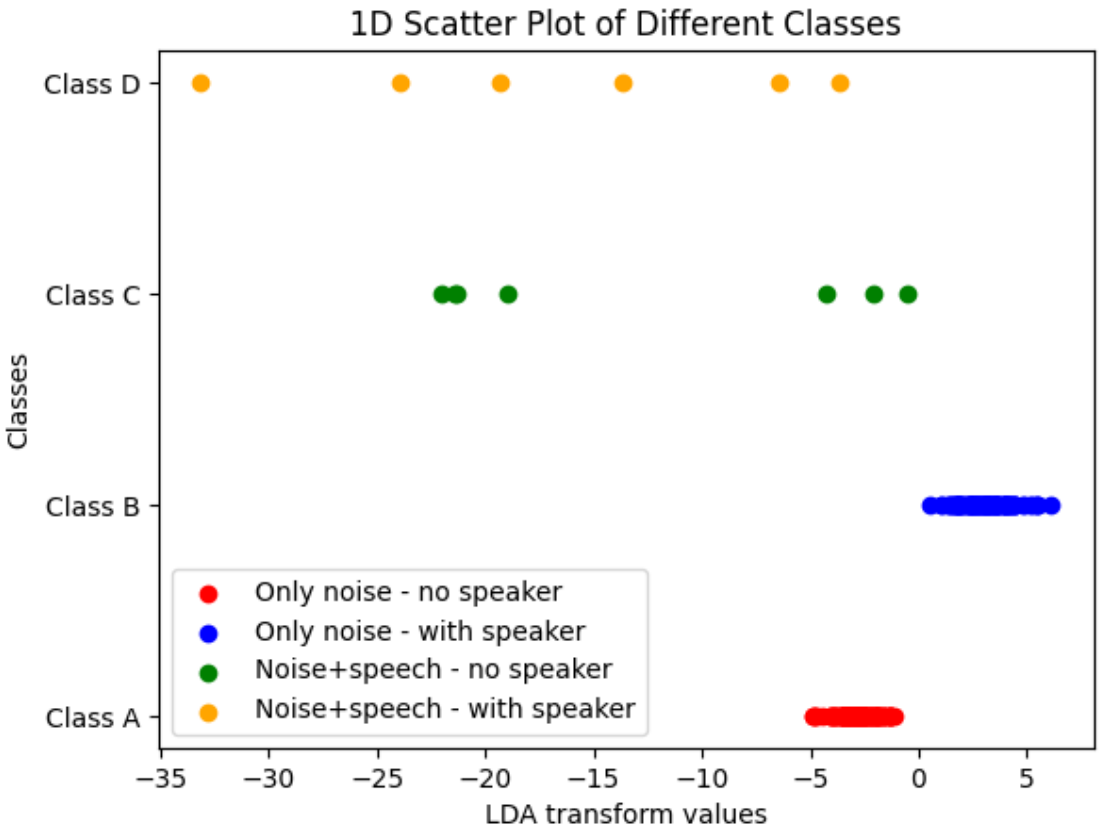
0.4.4 Experiment 4

Now that we have a trained model from the previous data, we try to apply this same model with speech. I record two speech segments in the same room where the background noise was obtained with and without speakerphone, and run the classifier. However, this time the accuracy is horrible:

```
0.5384615384615384
```

0.4.5 Experiment 5

We try to study what happens when the speech is added back. We plot the values after LDA for all 4 classes of noise and speech, with and without speakerphone. There is a clear separation between the noise classes (speakerphone and without). But the same separator clearly puts all the speech samples in the speakerphone category.



0.5 Conclusion

It seems that the best way forward is to segment out the speech portion from the audio obtained from the speakerphone and then apply this analysis. The above experiments have been done on a small number of data samples (around 100), and without much generalization (all data collected in the same room). Thus, both the segmentation and generalization are immediate next steps from this project. From these small experiments, it does seem like machine learning can be used to do this detection, but will require more work and fine tuning. All of the experiments above, and all the data collected are clearly recorded in the following github repository: speakerphone_project.

0.6 References

1. J. Harrison, E. Toreini and M. Mehrnezhad, "A Practical Deep Learning-Based Acoustic Side Channel Attack on Keyboards," 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Delft, Netherlands, 2023, pp. 270-280, doi: 10.1109/EuroSPW59978.2023.00034.
2. Davis, Abe & Rubinstein, Michael & Wadhwa, Neal & Mysore, Gautham & Freeman, William & Durand, Fredo. (2014). The visual microphone: Passive recovery of sound from video.