Man vs Machine: Discerning Between In-Person and Electronic Audio

Emmanuel Adeniran
Yale University
emmanuel.adeniran@yale.edu

Nicholas Georgiou
Yale University
nicholas.georgiou@yale.edu

Debasmita Ghose
Yale University
debasmita.ghose@yale.edu

ABSTRACT

Determining whether human voices are being played through electroacoustic transducers, like a television's or computer's loudspeakers, or are being originated from a person without replay, is an important distinction for voice-activated devices to make. Ideally, these devices should be able to discern whether audio is originating from people who are contemporaneous and collocated with the robot, versus audio that is originating from an electronic device. The ability of a robot to discern these "here and now" scenarios is critical to enabling a robot to understand circumstances surrounding human social interactions. This body of work addresses this conundrum by implementing a method to differentiate attributes that are characteristically and consistently fundamental to voices replayed through loudspeakers (electronic) versus originated by a person (in-person). Therefore, we attempt to build a model that can distinguish between attributes in natural and electronic audio, recorded on a given microphone. Additionally, we attempt to extend this to build a generalized model, that can effectively generalize across audio recorded from multiple microphones. In order to do that, we explore two different techniques: The first technique uses a method proposed by [3], to calculate the sub-base over-excitation of an audio, which has been found to be a distinguishing feature between natural and electronic audio. The second technique extracts spectrograms from fixed length audio signals, and uses a CNN to classify them into natural and electronic audio. Results from the signal processing technique indicate that these attributes can be detected using the method developed to detect over-excitation in the sub-bass region of audio, reinforcing the results obtained in the the work [3]. We also show similar results from the deep learning technique which shows that we are able to discern between natural and electronic audio recorded using the same microphone and recorded using different microphones with reasonable classification accuracy. The implications of a robust method of differentiating mediated human voices are numerous. There are potential applications to human-robot interactions, particularly in academia, security, and commerce.

KEYWORDS

audio classification, signal processing, neural networks

ACM Reference Format:
Emmanuel Adeniran, Nicholas Georgiou, and Debasmita Ghose. . Man vs Machine: Discerning Between In-Person and Electronic Audio. In *Proceedings of* . ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

The main goal of this project is to be able to accurately classify human voices relayed through and emitted by an electronic device (electronic), versus voices directly spoken by a human (in-person).

This classification of audio signals, which has proven to be quite difficult, is an increasingly important problem to address, especially with the growing number of Internet of Things (IoT) devices that are present in our everyday lives. Many of these IoT devices have voice-activation commands, which in most circumstances should only be triggered by a human collocated with the device, as opposed to by a voice emitted by electronic devices such as phones, radios, or televisions. Even in cases where voice recognition is used for extra security, a device can still be vulnerable to replay attacks. The ability to classify audio being emitted by an electronic device may help thwart unauthorized or unwanted access to voice-controlled devices.

Discerning between in-person and electronic audio is critical to determining when a device should respond to a prompt or not. One approach to solving this problem attempts to exploit the fundamental differences between audio produced by a human or an electronic device. Some work already done in security, as illustrated in [3], shows that in-person audio is characteristically and consistently distinguishable from electronic audio. Different electronic microphones and speakers have acoustic properties that are sensitive to varying frequency ranges. Because of this, handcrafted features used in signal processing methods might not generalize very well to different speakers and microphones, and may require extensive tuning for it to work for a given use case. So, in order to have a system that generalizes to different microphone types and noise levels, using a neural network might be useful.

This can also be applied in situations that can help people on the Autism Spectrum. This audio classification can play a part in an experiment that gives participants opportunities to practice social skills that are useful in workplace environments. Based on a given time interval's audio classification, a robot will decide whether to interrupt the participant, if he/she is watching TV or listening to music, or not, if he/she is talking with a friend or on the phone. These interruptions can help the participant learn to adapt to such situations, which can help increase the participant's employability, as seen in [7].

We also attempt to generalize this method to be able to classify in-person and electronic audio recorded from multiple microphones. This can be useful in scenarios where multiple smart devices are connected together to a smart device hub, and all of the smart devices have different kinds of microphones in-built. So, taking advantage of such networks, an attacker may attempt to perform malicious activities using a replay attack on any microphone. Therefore, a generalized model should be still be able to detect this anomaly, and prevent such replay attacks.

This project attempts to exploit the characteristic differences in audio produced by humans and electronic devices, similar to [3]. This project will also explore alternatives to the algorithm used in the paper such as potentially using convolutional neural networks on spectrograms of the audio signal to solve this problem. Ideally, audio data will be an input to the designed system, which will make a classification to either accept (in-person) or decline the input (electronic).

So, our work has two contributions:

- We are able to successfully discern between in-person and electronic audio sources, recorded on one microphone using two different methods.
- We are able to discern between in-person and electronic audio with reasonably high confidence, agnostic of the hardware it is recorded on.

2 RELATED WORK

Signal Processing. Blue et.al. [3] distinguish between human and electronic speakers to prevent the user against replay attacks on smart home devices, where an electronic voice can play malicious commands to the smart speakers, to trigger some unwanted activities. They use a property inherent to the design of modern speakers called sub-bass over-excitation. Sub-bass over-excitation is the presence of significant low frequency signals (20-80 Hz) that are outside the range of human voice. This is one of the features of the audio that differs between the human vocal tract and the construction of the encasing of modern electronic speakers. They also show that sub-bass over-excitation is a phenomena present in all classes of speakers, whether high or low quality. The specifics of this method is explained in greater detail in Section 3.3.

Through our experiments, we realised that this method does not generalize well to audio recorded through different kind of microphones. We also found that the results are sensitive to the cutoff frequency for what we define the sub-bass over-excitation region to be based on the hardware we use.

Audio Classification using Spectrograms. One popular methodology to understand the time-frequency characteristics of audio is to break the audio signal into small time-windows and take the Short Time Fourier Transform of each of the time windows. This allows us to associate an amplitude with each part of the signal for different frequencies. A visual representation of this for the entire signal over all windows is called the spectrogram. A spectrogram is a 2D heatmap of the signal, where one of the axes represent the frequency and the other axis represents the time. So, for each each time, the amplitude distribution of different frequencies is color-coded, to construct the visual representation of the audio signal, as shown in Figure 3.

This is a commonly used representation of audio signals, for different audio classification applications. and CNNs are commonly used for this kind of classification. Multiple different For example, [5] use CNNs over spectrograms of music to classify the genre of the music and [1] use a similar methodology for speech emotion classification. [**?**] use a siamese style CNN to perform audio search in databases containing a sound and a vocal imitation of the sound, using mel spectrogram of the audio signal. [1],[2] and [4] show some other applications of CNNs on spectrograms for audio classification.

To the best of our knowledge, there is no work that attempts to solve the problem for classifying in-person and electronic audio using a deep learning technique, so as a starting point, we decide to use the basic CNNs used for audio classification on spectrograms. We believe that this method can be effective in classifying in-person and electronic audio, since the CNN will be able to learn features that are characteristic to the source of the audio.

3 TECHNICAL APPROACH

3.1 Hardware Description

We record audio using the UMA-8-SP USB Microphone Array [6] which has an Embedded Digital Signal Processor and a Stereo Digital Amplifier. It has seven microphones arranged in a circular configuration to capture audio signals coming in from different directions.The microphone has a sampling rate of 48kHz and has a flat frequency response.

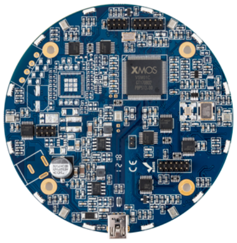


Figure 1: UMA-8 USB Microphone Array

The second microphone we used was a Yeti Professional Multi-Pattern USB Microphone manufactured by Blue. Its relevant technical specifications are: Sample Rate: 48kHz; Bit Rate: 16bit; Capsules: 3 Blue-proprietary 14mm condenser capsules; Polar Patterns: Cardioid, Bidirectional, Omnidirectional, and Stereo; Frequency Response: 20Hz - 20kHz; Sensitivity: 4.5mV/Pa (1 kHz); and its Max SPL: 120dB (THD: 0.5% 1kHz). The Yeti's on-board amplifier has an impedance: >16 ohms; THD: 0.009%, a Frequency Response: 15Hz – 22kHz, and a Signal to Noise: 100dB. The other two microphones are built-in OEM (Apple (Model Identifier:MacBookPro14,3) and Dell Spectre) microphones.

Figure 2: Yeti Microphone by BLUE

We also recorded using a Yeti mic, a MacBook Pro's OEM microphone and an HP Spectre OEM Microphone. The following table shows the IDs assigned by us to the different microphones, for simplicity in reporting and interpreting results:

ID	Microphone
1	HP Spectre
2	UMA-8SP
3	Yeti USB Mic
4	MacBook Pro

Table 1: List of microphones used to record audio

3.2 Data Collection Protocol

During the final aspects of the project, collected more data in the form of audio pairs using different microphones. The microphones included the Yeti, MacBook Pro's and Dell's built-in microphones, and a miniDSP VocalFusion UMA-8-SP USB Microphone Array. The audio pairs were collected Waveform Audio File Format. Audio recorded from these devices were trimmed to five-second bits and operated on by the signal processing and NN models. Audio recorded were from different settings including conversations, classrooms, and podcast recordings. We will then generate the spectrograms for each of these audio signals and label the spectrograms based on whether they are generated from natural human voice or human voice that is played back from a speaker. Emmanuel will be responsible for collecting, preprocessing and creating the data pipeline for both the signal processing and deep learning methods.

(1) "O.K. Google, Browse to evil.com."
(2) "O.K. Google, call grandma."
(3) "O.K. Google, record a video."
(4) "Hey Google, text John buy spam today."
(5) "Hey Google, post I'm so evil on Twitter."
(6) "Alexa, call grandpa."
(7) "Alexa, text mom what was my social security number again?"
(8) "These puffy tarantulas cross bravely shepherding homeless grouper through explosions."

Figure 3: Command Phrases

Additional data was collected using an automated procedure where, the subject was asked to speak random phrases and a fixed length audio was recorded from a given microphone. Then, the subject was asked to stop speaking, and the recorded audio was played by a speaker, and while the audio was playing, it was re-recorded by the same microphone. We repeated this procedure for all the specified recording hardware in Table 1.

3.3 Signal Processing Technique to Classify In-Person and Electronic Audio

For both in-person and electronic audio files, the analysis was the same. This approach and implementation were very similar to the ones taken in [3]. Only one channel from the microphone array that did the recording was used in the analysis. The audio was then sliced up into windows, each with the duration of one tenth of a second.

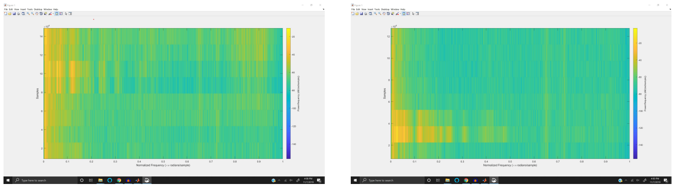
For each window, the discrete Fourier transform (DFT) was computed for the frequency range of 20-250 Hz, using a fast Fourier transform (FFT) algorithm in MATLAB. Fourier transforms find frequency components of a signal by converting the signal from the time to frequency domain. Next, the single-sided amplitude spectrum and the power spectral density is calculated, and normalized, for each window in the frequency range of 20-250 Hz. Taking the integral of this power spectral density within a certain frequency range gives the amount of energy in the audio file within this range. This integral was taken for the sub-bass region (20-90 Hz) and the region of 20-250 Hz for the normalized power spectral energy curves. The energy balance metric (EBM) for every window as calculated by taking the energy in the sub-bass region and dividing it by the energy in the 20-250 Hz region.

$$\textit{energy balance metric} = \frac{E_{\textit{Sub-bass Region}}}{E_{\textit{Total Evaluated Region}}}$$

Each window had a different EBM value, and the final value of the EBM for an audio file was equal to the median EBM value, after removing outliers.

3.4 Spectrograms

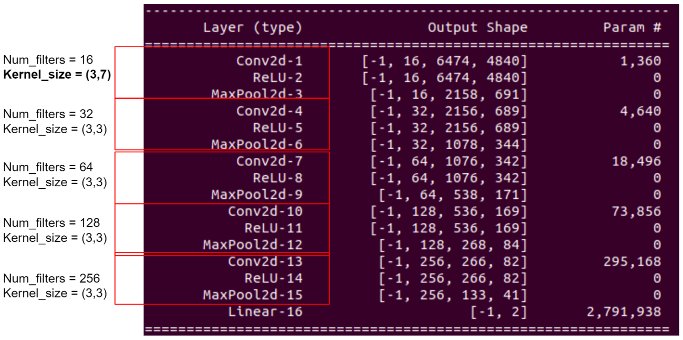
As explained in Section 2, a spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. A spectrogram is represented with two geometric dimensions: one axis represents time, and the other axis represents frequency; a third dimension indicating the amplitude of a particular frequency at a



a) Electronic

b) In-Person

Figure 4: Spectrograms of the phrase "hello" recorded by a microphone(in-person) and re-recorded by a microphone (electronic) forms



```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1       [-1, 16, 6474, 4840]           1,360
              ReLU-2       [-1, 16, 6474, 4840]               0
         MaxPool2d-3        [-1, 16, 2158, 691]               0
            Conv2d-4        [-1, 32, 2156, 689]           4,640
              ReLU-5        [-1, 32, 2156, 689]               0
         MaxPool2d-6        [-1, 32, 1078, 344]               0
            Conv2d-7        [-1, 64, 1076, 342]          18,496
              ReLU-8        [-1, 64, 1076, 342]               0
         MaxPool2d-9         [-1, 64, 538, 171]               0
           Conv2d-10        [-1, 128, 536, 169]          73,856
             ReLU-11        [-1, 128, 536, 169]               0
        MaxPool2d-12         [-1, 128, 268, 84]               0
           Conv2d-13         [-1, 256, 266, 82]         295,168
             ReLU-14         [-1, 256, 266, 82]               0
        MaxPool2d-15         [-1, 256, 133, 41]               0
           Linear-16                    [-1, 2]       2,791,938
================================================================
```

Figure 5: Network Architecture

particular time is represented by the intensity or color of each point in the image. There are different types of spectrograms, depending on how the time, frequency and amplitude axes are scaled. The two important types are Log Spectrograms, where the frequency and amplitude axes are scaled logarithmically and Mel spectrogram, where the frequency axis is scaled according to a precomputed non-linear function. We decide to use the Mel Spectrogram, because it is most commonly used in the Deep learning literature for audio.

Figure 3 shows sample spectrograms of the phrase "hello" in both natural and electronic versions. One of the key things we can observe from these spectrograms is that they are visually different, which means that we can possibly use a CNN to classify them as natural or electronic audio, possibly irrespective of the hardware.

3.5 Deep Learning Method to Classify In-Person and Electronic Voices

We extract Mel spectrograms of fixed length audio signals, with 2048 fft points, hop length of 512 and 128 mels to construct the spectrogram. We then passed these through a Convolutional Neural Network. whose architecture is shown in Figure 4.

4 EXPERIMENTS

4.1 Classification using Sub-Bass Over-Excitation

Our implementation and preliminary results showed that the discernment of electronic audio versus in-person audio can be successful. At least 22 audio recordings were made using the miniDSP VocalFusion UMA-8-SP USB Microphone Array. Each audio file was recorded and compressed in MP3 format at a sampling rate of 48 kHz. Each of the audio files was then played back and re-recorded using the same microphone array, resulting in pairs of the same audio; recorded from a playback (electronic), and recorded without playback (in-person). Each pair of recordings was passed through the algorithm to obtain an Energy Balance Metric (EBM). This was also done for all of the Yeti microphone recordings, as well.

Some differences in the characteristics of the electronic and in-person audio files can be seen in Figures 6 and 7, as well as 8 and 9. Figure 6 shows the single-sided amplitude spectrum for an electronic file, while Figure 7 shows the same graph, but for the in-person audio. Figure 8 shows the normalized Power Spectral Density for an electronic file, while Figure 9 shows the same graph, but for the in-person audio.

A plot of the Amplitude Spectra, on a logarithmic scale within the given frequency bins, can be seen below. They depict how the amplitude of the audio signal is distributed over the frequency domain window. Figure 5, below shows clear visual differences within the sub-bass and overall regions compared to the same sample of in-person voice, depicted in Figure 6.

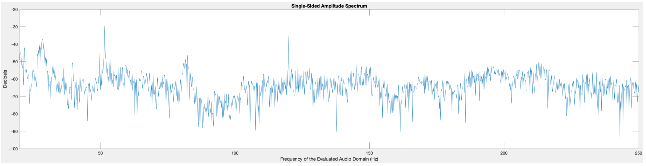


Figure 6: Electronic Audio Amplitude Spectrum Example

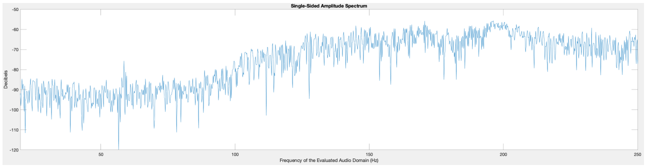


Figure 7: In-Person Audio Amplitude Spectrum Example

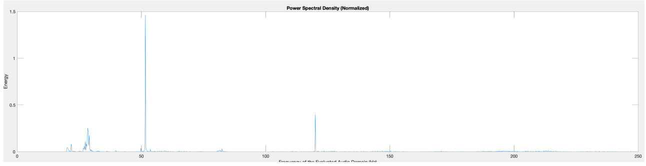


Figure 8: Electronic Audio Power Spectral Density Example

Figure 8, above, depicts the Power Spectral Density (PSD) for an electronic audio file, and shows the spectral power distribution decreasing as frequency increases from 20 to 250 Hz. High energy levels are found within the sub-bass region of the window (20-80 Hz), indicative of the energy contributed by the resonance of the loudspeaker encasement at those frequency ranges. Below, however, in Figure 9, is the PSD for the same sample of in-person human voice, indicating low energy levels within the sub-bass region.

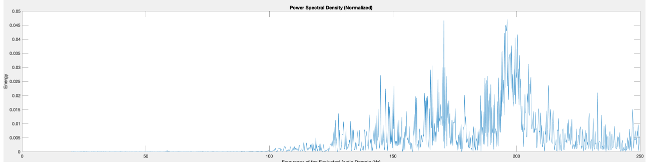


Figure 9: In-Person Audio Power Spectral Density Example

A composite set of results illustrate consistency across 22 additional audio samples. EBMs for 22 of the audio recordings were obtained with the upper band of the sub-bass region set to 80 and 90 Hz, sequentially. Results indicate that setting the upper band of the sub-bass region to 90 Hz resulted in significantly larger area of separability between the clusters of electronic and in-person audio, as shown in Figures 10 and 11, below.

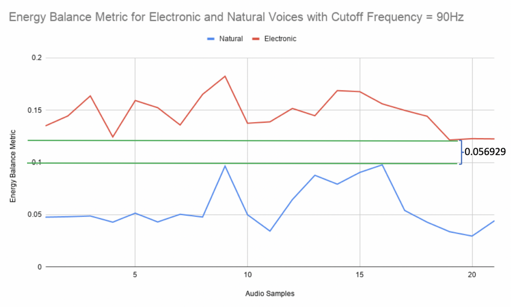


Figure 10: Line Plot of EBM calculated for different corresponding In-Person and Electronic Audio with the Sub-Bass region described as 20-90 Hz and EBM Width of Demarcation of 0.056929

The line plot showing EBMs' width of demarcation of 0.056929 for an upper band set at 90 Hz as shown in Figure 10 above is significantly larger than the EBMs' width of demarcation of 0.028092 for an upper band set at 80 Hz, as shown below in Figure 11.

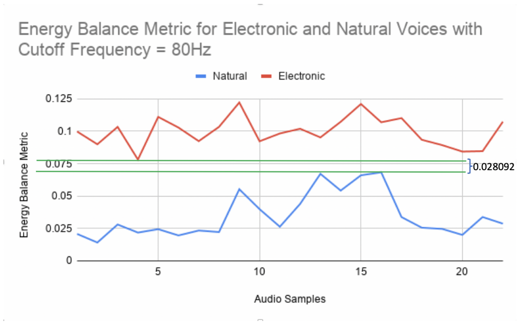


Figure 11: Line Plot of EBM calculated for corresponding In-Person and Electronic Audio with Sub-Bass region of 20-80 Hz and EBM Width of Demarcation of 0.028092

The two Figures above show that there is clear difference between the electronic and in-person, natural audio files. The electronic EBM value is always higher than the in-person EBM value, and the lowest electronic EBM is higher than the highest in-person EBM. This can lead to the selection of a threshold, based on the hardware, in which any EBM value above it is classified as electronic, and any value below is classified as in-person. Additionally, Figures 8 and 9 shows that selecting an appropriate cutoff frequency for the sub-bass region can lead to a better demarcation of the between the two classes, and in our case, the cutoff frequency of 90Hz worked better than 80Hz, which was used by the paper[3]. More data was collected and tested to further solidify this classification. Below are test data results for two microphones, the built-in mic on a MacBook Pro, and BLUE's Yeti USB Microphone-Platinum. Figure 12, contains a plot across 68 data pairs, while figure 13 has a plot across 412 data pairs of electronic versus in-person voices. Figure 12 illustrates a possible limitation to the signal processing method, as the EBMs were not consistently separable between the audio pairs, whereas figure 13, using a professional-grade microphone like the Yeti, resulted in a differentiation that was consistent approximately 95 percent of the time.

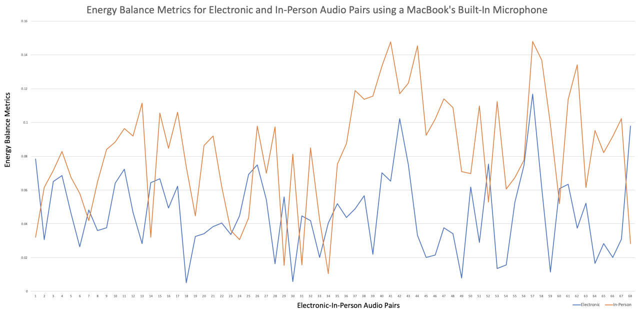


Figure 12: Energy Balance Metrics for Electronic and In-Person Audio Pairs using a MacBoo Pro's Built-In Microphone

Mic ID	Classification Accuracy
1	83.33%
2	88.88%
3	75%
4	85%

Table 2: Classification Accuracy when trained and tested on different audio spectrograms recorded from the same microphone

Mic ID	Classification Accuracy
1, 2	61.53%
2, 3	82.14%
3, 4	89.69%
2, 3, 4	86.11%
1, 2, 3, 4	79.03%

Table 3: Classification Accuracy when trained and tested on different audio spectrograms recorded from combination of different microphones

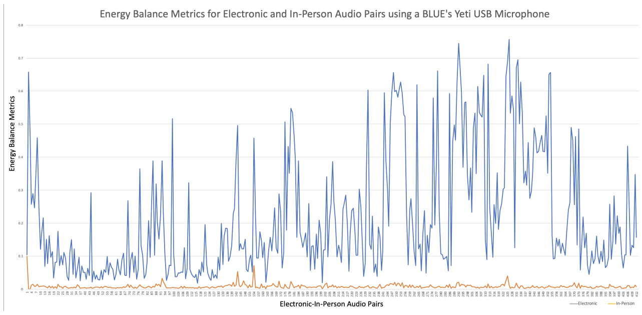


Figure 13: Energy Balance Metrics for Electronic and In-Person Audio Pairs using a BLUE's Yeti USB Microphone

4.2 Classification using CNNs on the Spectrogram of the Audio Signal

4.2.1 Training Details. We train the CNN using the Adam optimizer with a learning rate of 1e-5 and a batch size of 1 on 2 NVIDIA Tesla M40 GPUs with 12GB memory. We added a dropout layer to the last linear layer of the CNN to prevent overfitting, and the network was trained for 6 epochs. 80% of the data was used for training and rest of the 20% were used for testing.

4.2.2 Results and Analysis. The following observations can be made from the training procedure and results of the CNN:

- The results are highly dependent on the quality of the microphone used to record audio. For example, the classification accuracy of the data recorded from Microphone 3, which is the Yeti mic, is not as high compared to the other microphones.
- The model generalizes reasonably well across data recorded from different microphones. Table 3 shows that when the model was trained and tested with data from different microphones, which had different audio quality, the classifier was able to give classification accuracies comparable to the ones shown in Table 2, which show the classification accuracy for spectrograms extracted from the same source.
- The network trains in 4-6 epochs, beyond which it shows signs of overfitting. This means that the spectrograms of the in-person and electronic audio are highly distinguishable in most cases in the embedding space.

5 CONCLUSION AND FUTURE WORK

This work is successfully able to discern between in-person and electronic audio recorded using different microphones with a reasonably high confidence. Here, we propose two methods, one being the sub-base over-excitation of the audio signal, which can be used to make the two audio sources discernable, as shown in [3]. The second method is to extract spectrograms from these audio signals, and use a CNN to classify the spectrograms into in-person and electronic audio. Both these methods were able to classify electronic and in-person audio with reasonably high accuracy, when they were recorded from the microphone. For the case where data was recorded from different microphones, the method which used a CNN to classify spectrograms, performed better than the signal processing technique, because the signal processing technique uses a hand crafted feature representation, specific to one particular microphone. However, the signal processing technique performed better than the deep learning technique where data was recorded using a good quality directional microphone(Microphone 2).

Analysis of the output of the network shows that it gets confused between In-Person audio from one microphone and Electronic audio from another microphone, since some of their spectrograms might look similar to the network. So, the network might need more data to learn. Another approach that might help the network learn these differences between the classes, is to use a Siamese network. The intuition behind that is that the siamese network will look at pairs of similar and dissimilar images at a time, and the triplet loss function used to train the network will reinforce these differences. Also, siamese network is a technique used for few-shot learning, which means that we might be able to classify audio with fewer examples, which will make it deployable real time. Therefore, we plan to explore this technique in the near future.

6 SUMMARY OF CONTRIBUTIONS

Nick wrote the MATLAB code to implement most of the related work paper. This involved audio analysis that calculated Fast Fourier Transforms (FFTs) and Power Spectral Density graphs for each window. Nick also worked on computing energy balance metrics (EBMs) for each audio file, which were later used to compare in-person and electronic audio files. He also wrote the code to continuously record and wait for audio files to be placed in a folder in MATLAB.

Emmanuel collected and processed data. Analyzed the MATLAB code and structured the approach to fine tuning the parameters of the MATLAB code in order to determine optimal computational thresholds for the sub-bass regions. Compiled and analyzed preliminary results. In conjunction with Nicholas, implemented a structured signal processing approach in code that returns near real-time feedback on the determination of the processed audion signals.

Debasmita collected data and analyzed the results of the signal processing algorithm for its shortcomings and for opportunities and areas of improvement. She then experimented using CNNs on spectrograms to detect other salient features amenable to distinguishing between electronic versus in-person voices. Multiple experiments were performed by her by training the CNN on spectrograms from different microphones and then validating the generalization capabilities of the neural network on both in-sample as well as out-of-sample spectrograms.

REFERENCES

[1] Badshah, A. M., Ahmad, J., Rahim, N., and Baik, S. W. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)* (2017), IEEE, pp. 1–5.

[2] Bae, S. H., Choi, I., and Kim, N. S. Acoustic scene classification using parallel combination of lstm and cnn. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)* (2016), pp. 11–15.

[3] Blue, L., Vargas, L., and Traynor, P. Hello, is it me you're looking for?: Differentiating between human and electronic speakers for voice interface security. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks* (2018), ACM, pp. 123–133.

[4] Dörfler, M., Bammer, R., and Grill, T. Inside the spectrogram: Convolutional neural networks in audio processing. In *2017 International Conference on Sampling Theory and Applications (SampTA)* (2017), IEEE, pp. 152–155.

[5] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (2017), IEEE, pp. 131–135.

[6] MiniDSP. Uma-8-sp usb microphone array spec sheet, 2018.

[7] Orsmond, G. I., Krauss, M. W., and Seltzer, M. M. Peer relationships and social and recreational activities among adolescents and adults with autism. *Journal of autism and developmental disorders 34*, 3 (2004), 245–256.