Burhan Bhatti
Isaiah Glymour
Ron Nafshi
Lanyi Wang

Customer: Henrique Potter

Explainable AI for Voice Privacy

Problem Statement

The essence of this project is based on the textbook *Profiling Humans from their Voice* by Dr. Rita Singh. The term profiling from voice refers to the deduction of personal characteristics of the person and their environment solely from audio recording. Dr. Singh demonstrates that voice is a rich and dangerous data type, and information such as the speaker's age, height, gender, weight, emotion, injuries, and even information such as if they are sitting or standing and the material of the room they are in can be accurately gauged. This is possible by decomposing sound waves into Mel-frequency cepstral coefficients (MFCC) using the Fourier transform; each decomposed MFCC feature represents components of the audio signal that identify some linguistic content. Since voice is such a data rich data type that is so easily obtained, our project is interested in exploring feature selection in the MFCC to audio that the user may want to keep private. We explore the development of MFCC feature extraction machine learning schema, including Random Forest, Decision Trees, and Neural Networks.

Communication

Our group will use Slack to communicate. Our code and files will be organized via a github repository. Large files will be shared via OneDrive or a physical flashdrive. Furthermore, we will all meet with our sponsor at least once a week - every Wednesday at 5PM. We will also meet separately as needed, depending on the stage of the project.

Language and Frameworks

Our project will be constructed using PyTorch and TensorFlow. We will make use of a variety of standard Python data science packages such as Numpy, Pandas, and Scikit-Learn, as well as Librosa for audio feature extraction and SHAP for feature selection. Furthermore, we will use Google Collab notebooks to work collaboratively and train our models, as well as access to Google's GPU's which will be critical to managing the size of the data.

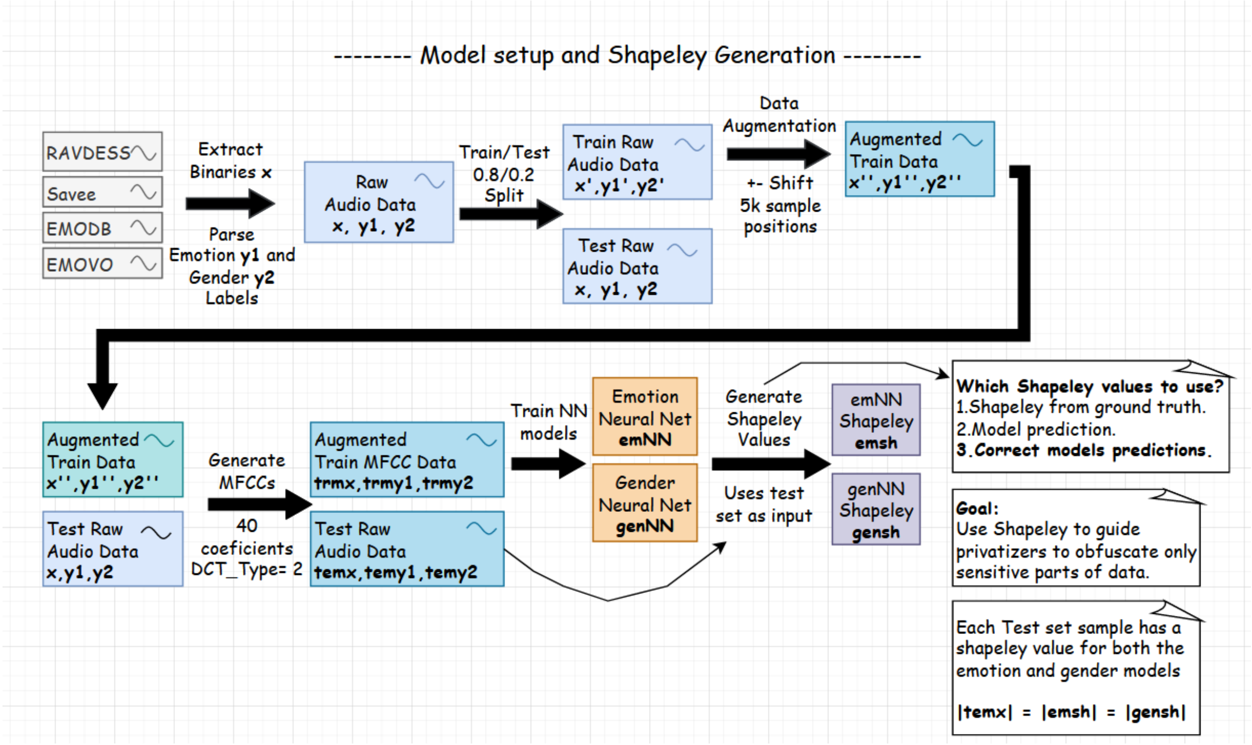
Testing

We will evaluate our models with respect to several datasets provided by our customer, including at least one from each of English, German, and Italian, as well as a combined dataset featuring multiple languages. Our goal will be to remove features so that our models can no longer distinguish certain aspects of the speaker's voice, such as gender. In particular, we hope that our models are robust to the removal of certain features from the MFCC input.

Model Architecture and Division of Labor

We plan to test three machine learning models: decision trees, random forests, and neural networks. We will attempt to find high performing systems within each of these three regimes, with the hope of proving that even very strong models can be foiled by our removal of particular features from the data. Separating our work into these three categories provides an excellent way to "parallelize" the project: one person will work on each of decision trees and random forests, while two people will investigate neural networks (which is the broadest of the categories). Moreover, we have two group members who have taken both CS 1675 and CS 1699, making this division of labor particularly amenable considering the group members' skills.

Below is a proposed workflow we plan to use for the project. We plan to first work on the common pipeline to read in audio data and augmented training data for MFCC's, and then split into groups to work on the individual models. Our customer has provided us with comprehensive documentation and examples to build the pipeline and work with the three desired models.


-------- Model setup and Shapeley Generation --------
RAVDESS
Savee
EMODB
EMOVO
Extract Binaries x
Parse Emotion y1 and Gender y2 Labels
Raw Audio Data x, y1, y2
Train/Test 0.8/0.2 Split
Train Raw Audio Data x',y1',y2'
Test Raw Audio Data x, y1, y2
Data Augmentation
+- Shift 5k sample positions
Augmented Train Data x'',y1'',y2''
Augmented Train Data x'',y1'',y2''
Test Raw Audio Data x,y1,y2
Generate MFCCs
40 coeficients DCT_Type= 2
Augmented Train MFCC Data trmx,trmy1,trmy2
Test Raw Audio Data temx,temy1,temy2
Train NN models
Emotion Neural Net emNN
Gender Neural Net genNN
Generate Shapeley Values
Uses test set as input
emNN Shapeley emsh
genNN Shapeley gensh
Which Shapeley values to use?
1.Shapeley from ground truth.
2.Model prediction.
3.Correct models predictions.
Goal:
Use Shapeley to guide privatizers to obfuscate only sensitive parts of data.
Each Test set sample has a shapeley value for both the emotion and gender models
|temx| = |emsh| = |gensh|

Additional Tools and Possible Issues

Training our models will require good graphics cards. To this end, we will use a variety of resources: cards owned privately by our team members, computing resources on Google Collab, and potentially, depending on time and availability, compute time on Pitt's CRC GPUs. Moreover, the success of each model we plan to investigate is unknown; perhaps one of the three will prove to be relatively weak while another model provides strong insights and demands more focus, shifting our group work plans unfairly.

Timeline

A brief timeline provided by our customer is below:

1. Read our brief summary of the book "Profiling humans from their voice" describing voice privacy issues and watch the video by Dr. Singh for a fast introduction.

2. Implement three ML models to classify emotions. Read and understand the software environment (what are the inputs, outputs, how inputs are read, what configurations exist, what are the different options for utility, defenses, and attacks).
3. Modify the software to also collect the following metrics: Accuracy, F1, false positives, and false negatives.
4. Search and implement 2 more Datasets, run the software, produce results similar to existing results but with new datasets (for all metrics).
5. Isolate the results per dataset (only evaluate with single datasets and compare the results).
6. Use the Shap library to explain the results from step 2 and rank the features in order of importance.
7. Retrain the models from step 2 while removing the most important features of a particular emotion (based on step 6) for each sample.
8. Process and plot the results.
9. Deploy the experiment in Occam.
10. Analyze the outputs and write a final report.

Scrum Epic

As a data scientist, I want to protect user privacy using audio data.

User Stories:

1. As a data scientist, I want to extract MFCC's from audio files to train a decision tree classifier which can predict both gender and emotion from voice data.
    - Acceptance Criteria: Given an arbitrary sound file of variable length and volume in English, German, or French, create an MFCC decomposition with m features,

where m can be increased or decreased to change the scope of the features the models uses.

2. As a data scientist, I want to train an effective Random Forest model to further predict gender and emotion from voice data.
    - Acceptance Criteria: The Random Forest Classifier is robust and able to classify emotions in the test set with acceptable results in the F-norm and strong Precision-Recall.
3. As a data scientist, I want to train Neural Networks to further predict gender and emotion from voice data.
    - Acceptance Criteria: The Random Forest Classifier is robust and able to classify emotions in the test set with acceptable results in the F-norm and strong Precision-Recall.
4. As a data scientist I want to collect following metrics from the predictions of the models: Accuracy, F1, false positives, and false negatives for each model
    - Acceptance Criteria: Collected data for accuracy, F1 test, false positives and false negative for all models.
5. As a data scientist I want to find two more datasets where we can extract MFCC data so we can run our software on the datasets and then compare it with our existing results.
    - Acceptance Criteria: All data metrics for previous results are collected from the new dataset.
6. As a data scientist I want to isolate the results from each dataset so I can compare each dataset independently.

- Acceptance Criteria: Data is segregated neatly by dataset so any similarities or differences are more easily identified.

7. As a data scientist I will use the sharp library to analyze the results from all of our different models
   - Acceptance Criteria: After the data is analyzed, I have the features ranked in order of their importance and the degree to which they factor in various classification decisions.
8. As a data scientist, I want to retrain all of our previous models while removing the most important feature of a particular emotion for each sample to further predict gender and emotion from voice data and allow us to have a better understanding of how that feature interacts with our models
   - Acceptance Criteria: All models are robust and able to classify emotions in the test set with acceptable results in the F-norm and strong Precision-Recall.
9. As a data scientist I will process and plot all the results from all of our different models
   - Acceptance Criteria: All the data from our model is compiled to make interpreting our results easy and intuitive.
10. As a data scientist I will run upload our experiment onto Occam so that the models can be more easily reproduced by other data scientists
    - Acceptance Criteria: Reproducing the experiment is straightforward through Occam.
11. As a data scientist I will analyze all the outputs and turn these results into a final report where our interpretations of the data is explained

- Acceptance Criteria: A report is made which contains my interpretations of the data I collected and if I found any significant features which do indeed help predict both gender and emotion data.