Ethan Bokelberg, Calli Fuller, Garrett Faucher, Finn Dority
CS 254 Project Final Report
12/7/2020

1) Introduction

Right now, machine learning applications to speech are really good at recognizing *what* is being said, but there are a lot of nuances in communication that go beyond the actual words being spoken. Things like intonation, accent, and emphasis help inform context to give us a better understanding of what someone is saying and their motivations. Take the example of a keynote speaker doing a virtual presentation. Perhaps this person starts their presentation with the following sentence: 'I hope you're all having a great year'. For someone following along through a closed captions feature, this could be read in a number of different ways. Perhaps the intent is to communicate genuine good wishes to an audience. It's also possible that the speaker is being sarcastic because this year has been full of a lot of turmoil and hardship. In a world where movement toward digitalization has been accelerated, it's becoming easier and easier to lose these nuances and pieces of context. The goal of our project is to try to help bridge that gap. We have decided to use a dataset of labeled voice data and use machine learning algorithms to identify certain characteristics about the speaker. These characteristics may include gender, age range, accent, and mood. We found that the best model for this audio classification was the Convolutional Neural Network, followed closely by an MLP Neural Network. We were able to get near perfect classification by gender, and fairly reasonable emotion classification as well, especially compared to human classification of emotion.

2) Problem Definition and Algorithm

a) Task Definition

We will be preprocessing our data to extract the Mel-Frequency Cepstral Coefficients (MFCC) from the audio data we have collected and will be training our model to identify the gender and mood of the speaker based on a short audio clip. MFCC feature extraction is a method that is used to process signalling and features using DFT and are then warped on a Mel scale to represent the relationship between perceived frequencies of tone and the actual measured frequency. This task is interesting and important because it seeks to bridge the gap between the nuances of speech recognition, like tone and intonation, and the concreteness of what a computer can process and understand.

b) Algorithm Definition

For speech recognition, Viterbi Search, Neural Networks, and Discrimition Training models are the most popular choices. For our project, we decided to implement a "divide and conquer" sort of approach, with each member trying out running the data through different algorithms for comparison. Ethan worked with a Random Forest classifier, Calli worked with MLP Neural Networks, and Garrett worked on a Convolutional Neural Network (CNN). We decided on these algorithms because we felt that SVM and Random Forest were two of the best options for non neural network

classification, and the two levels of Neural Networks were an obvious choice based on what we learned about them in class and in seeing their frequent use in published papers about audio classification.

Convoluted Neural Networks are typically used when modeling audio data because CNNs are best suited for processing image data and therefore are a powerful tool to leverage when classifying spoken sounds using spectrogram images. Multilayer Perceptron Neural Networks are also known to perform well in speech recognition and categorization. These models utilize numeric representations of audio data such as the Mel-Frequency Cepstral Coefficients (MFCC) which represents the sound of a wav audio file by utilizing distinct hop length and HTK-style mel frequencies, as well as Chroma which is a condensed feature that represents the tonal substance of melodic sound signals.

3) Experimental Evaluation

a) Methodology

We will be evaluating our model based on its ability to correctly categorize speakers into the following categories: speaker male or female and the mood being expressed in each segment of speech. With precision defined as the rate of true positives relative to the sample size and recall defined as the rate of true positives relative to the predicted results, we believe that precision is a better metric to evaluate the performance of our models rather than recall. It is more important to our application purposes that we be able to measure and optimize the rate of true positives relative to the sample rather than simply the proportion of true positives relative to the relevant samples as our goal is correct categorization across a spectrum.

The data we used to train and test (found here, also cited in bibliography as (Livingstone & Frank, 2018)) is a well known and popular data set from the Ryerson Audio-Visual Database of Emotion Speech and Song. The data was collected by taking audio and recordings of 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, disgust expressions, and a song that contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

This method of data collection means that the recordings represent a more ‘induced’ state of each emotion as opposed to recording clips of more natural speech and later going back to label specific segments where emotion was expressed. This data set operates well for a preliminary analysis of classifying emotion in speech, but to train a model to identify emotion in a real-time application setting like a keynote speaker doing a virtual presentation, it would likely require data that is more representative of natural speech patterns.

It's important to note that this data set consists of 1440 data points which makes it a relatively small sample. As a result, our models are likely prone to overfitting. This can be seen in our Training and Validation Loss plots. To combat this, we regularized the data in the Neural Network models and erred on the simplistic end by keeping the number of neurons and layers minimal.

b) Results

Using the Neural Network module from sklearn, we were able to get good results classifying gender using MFCC and reasonable results for classifying emotion using gender and MFCC.

To classify gender, we first had to reshape the MFCC data into 1440 rows-1 for each file read in initially, and properly assign whether the speaker of that MFCC was male or female. Once this was done, the X and Y data was split into test and training using a 1:4 split, and X_train and X_test were regularized using the StandardScaler() function from the sklearn preprocessing package. The training data was then passed through the MLPClassifier() function from sklearn.neural_network package with three hidden layers, each of size 5, and eta of 1000. When tested, this model returned a confusion matrix of:

Gender Classification

	Male	Female
0	145	3
1	0	140

Using the classification_report() function from the sklearn, we found that the precision of this model was approximately 99%, compared to the 98.95% precision calculated by our own precision() function.

For emotion classification, the most accurate results were found using a two dimensional Convolutional Neural Network. The most successful CNN used four convolutional layers with filter sizes ranging from 24 to 80. Leaky-ReLU with an alpha of 0.1 was applied after each convolutional layer which was followed by a varying amount of dropout. Max-pooling of 2x2 was included after the first convolutional layer but not included thereafter. Lastly, the layers were flattened and next two dense output layers with 64 for the first and 8 for the last. There was one Softmax output layer with 8 units for each of the respective emotions that were being classified. The algorithm utilized the Adam optimizer and categorical cross-entropy.

The chosen architecture of the CNN was determined by automating hyperparameter tuning with the Python library Keras Tuner. Hyperparameters were randomly generated for a given range of values and models based on those hyperparameters were compiled and tested. This process was viable due to access to a GPU and TenorFlow utilizing the GPU's power which led to very quick iteration. The hyperparameters generated by Keras Tuner were scrutinized and the range of values were narrowed until an optimal algorithm was ready.

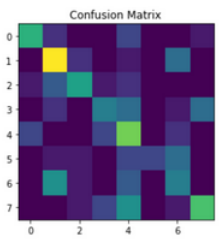
c) Discussion

In our initial pre-proposal results, we got a precision value of 93% with a basic SVM classifier, which indicated that for this classification by gender of the MFCC from the WAV files, 93% of the predicted genders were the correct gender. For running through the basic SVC function, this was a reasonably good precision, however, other machine learning algorithms, particularly the neural networks, showed a significant improvement, getting our precision for gender classification to approximately 99%.

There was a noticeable difference in results between gender classification and emotion classification present from the Random Forest, which gave a precision of 97% for gender classification and 47% for emotion classification. Like in the neural networks, separating the data by gender prior to classifying by emotion showed an improvement in performance, with precision for male samples increasing to 50% and increasing to 54% in female samples. However, this was still notably worse performance than the neural networks, so we elected to focus on optimizing those for further improving results.

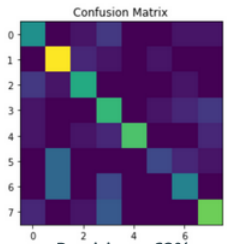
The MLP Neural Network had excellent results for classifying gender using MFCC, operating with a precision of 99%. When classifying emotion, however, the neural network struggled. When using just MFCC, the model functioned with a precision of 45%. This was slightly improved to 51% when gender was included, and then slightly improved again when the data was segmented by gender to start and then trained on both male and female audio data. In the models that were trained on the split data, the results for the female emotion classification were noticeably better than those for the male emotion classification. The male classification returned a precision that averaged on the lower end of 50-55%, while the female emotion classification averaged a precision between 60% and 65%. Both models struggled to identify fear, sadness and neutral emotion. Individually, the male model was best at identifying anger and the female model was best at identifying happiness.

Male Emotion Classification:



Precision = 54%

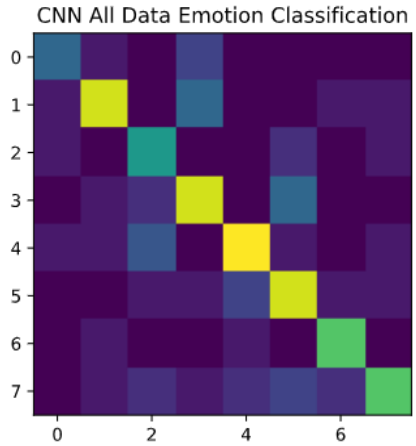
Female Emotion Classification:

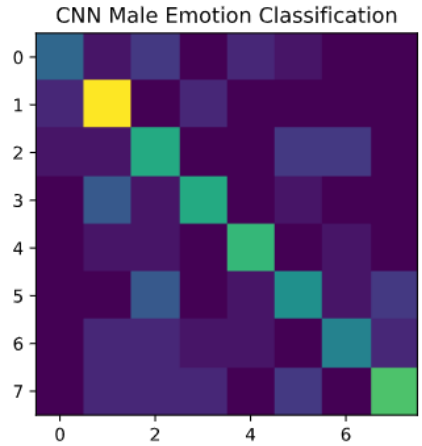


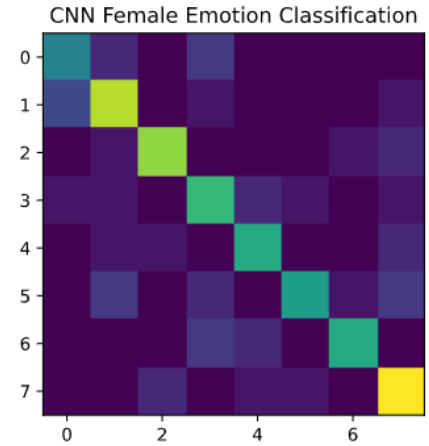
Precision = 62%

0 = Neutral	4 = Angry
1 = Calm	5 = Fearful
2 = Happy	6 = Disgust
3 = Sad	7 = Surprised

The CNN showed improvement in emotion classification and similar results for gender classification. Gender classification yielded precision of 99.6%. For the entire dataset, emotion classification gave a precision of 63.8%. After splitting the dataset by gender there were notable results. The male split dataset provided a lower precision 61.1% and the female split dataset provided a precision of 71.1%. Despite the gains from the CNN architecture, the algorithm was still prone to overfitting despite efforts to mitigate overfitting with dropout. Overfitting may be attributed to the small sample size of data that we have access to as dropout could have handled the issue. Like the MLP Neural Network, female emotion classification was more accurate than the male emotion classification. The model that included all gender data did well with classifying surprise and disgust but struggles with all other classifications. The male model excelled in identifying anger and surprise, but struggled with calm. The female model excelled in identifying happy, but struggled with calm, happy, and surprise.







4) Related Work

A common theme that was noticed in the cited articles was the focus on a single characteristic that was being identified. The approach we propose differs in that our goal is to identify multiple aspects about the audio file in order to build a comprehensive impression of characteristics that the speaker is exemplifying. Examples of preprocessing audio data are abundant and as such it will not be the majority of this project, but instead building a system to accurately identify a diverse set of characteristics. For example, in the paper "Voice-Based Recognition System for Non-Semantics Information by Language and Gender" the goal was to categorize speakers by language and gender rather than adding additional parameters. Additionally researchers were able to identify facial and vocal expressions to understand a speaker (Livingstone & Frank, 2018). We believe that it is possible to improve and merge research developments and provide a comprehensive system that perceives a clearer image of the speaker.

In consulting others' work, we've found several examples of both classification on gender and emotion. With regard to gender classification, we found an article (Becker, 2016) that hit the same level of accuracy as ours, while using a multi-layer perceptron instead of our CNN.

At the precision of ~99%, there's not much room for improvement, which shows that we were able to tune our results to near perfect.

We also found a paper (Joy et al., 2020) that followed very similar procedures to us in emotion classification. They used an MLP neural network on MFCC features, which was one of the neural networks we worked with. Their results gave them an accuracy of 58% on classification, which is very comparable to our value of ~60% from the MLP neural network, and something we were able to notably surpass with the CNN reaching ~70%.

In terms of how humans do with emotion classification, we found a study done on how well humans do for emotion classification (Laussen & Hammerschmidt, 2020), which resulted in an approximate 80% accuracy for humans. This is only a 10% increase in accuracy compared to our machine learning algorithms - and humans are even worse at classifying "surprise" in a voice sample than most documented models. This is important to note, as although at first glance a precision of ~70% for the emotion classification model does not seem great, when we compare it to what it is attempting to simulate, that is, human emotion classification, it is really not all that far off, and works very well as a starting point for further ventures into voice classification.

5) Code and Data Set

File Repo: https://github.com/ebokelbe/CS254-Final-Project

We decided to implement a "divide and conquer" sort of approach, with each member trying out running the data through different algorithms for comparison. Ethan worked with a Random Forest classifier, Calli worked with MLP Neural Networks, and Garrett worked on a Convolutional Neural Network (CNN)

Our data csv is too large to upload to github, so we used the uvm file transfer system to share it between ourselves.

The dataset we are using (found here, also cited in bibliography as (Livingstone & Frank, 2018)) is extremely well labeled, with each audio file following a naming convention of the following format: Modality-Vocal Channel-Emotion-Emotional Intensity-Statement-Repetition-Actor. Because of this, we will not have any need to go through the data and label it, which gives us more time to just get started on work with the algorithms. The contents of the dataset, in the exact words of the dataset creators, is "24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, disgust expressions, and the songs contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only, Audio-Video, and Video-only" (Livingstone & Frank, 2018). For the purposes of our project, we will only focus on the audio only non-singing subset of the dataset, which should provide us with 1440 data points. This dataset is publicly available and we will not need any special hardware for processing. We will be using a python package called Librosa and pydub to process the audio data in order to pull out the MFCC values to feed into our model. Librosa is able to take WAV files provided from our dataset

and turn them into numpy arrays in a format called MFCC. Since these WAV files were of differing lengths, pydub was used to pad the audio samples in order to have MFCCs of equal sizes for manipulation in the future. Once padded, all of the sound data can be combined into a large numpy file for manipulation that is no longer bound to the WAV file format. When the data was received, the data's labels were stored in the individual WAV files names. The labels were extracted from the file names via Python into a numpy array which was additionally stored as a separate numpy file.

6) Conclusion

By working on this less-focused aspect of voice recognition, we not only get to work in a less-explored field (along comes all the fun of finding things out for yourself) but we also get to do something that has the potential for "real" effects. Online interaction typically lacks the context of a real-world conversation, and making algorithms that can determine and re-inject this context lost is important, especially for those who might not otherwise get what's going on. Our work with the neural networks was able to get extremely good performance on the gender classification, which is very useful in aiding in identifying a speaker, and also proved useful as a trait to split the dataset by when classifying by emotion. As previously noted, we noticed improvements in the performance of our algorithms when we separated the data into male and female samples and ran them through separate algorithms, so with our near perfect precision in identifying gender, performance will also be improved in the emotion classification by separating the data. Additionally, as we previously mentioned, our emotion classification is not too far off from how the average human classifies emotions, which makes our model an ideal starting point for further development of machine learning models. Perhaps with this level of clarity, algorithms and machine learning systems akin to ours will be able to pave the way for the deaf or socially underdeveloped to understand deeper meanings behind context-filled conversations, leading towards a more inclusive and accurate form of online discourse.

Bibliography

Ertam, F. (2019). An effective gender recognition approach using voice data via deeper LSTM networks. *Applied Acoustics, 156*, 351-358. doi: 10.1016/j.apacoust.2019.07.033.

W. Li, D. Kim, C. Kim and K. Hong, "Voice-Based Recognition System for Non-Semantics Information by Language and Gender," 2010 Third International Symposium on Electronic Commerce and Security, Guangzhou, 2010, pp. 84-88, doi: 10.1109/ISECS.2010.27.

J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," in IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 74-99, Nov. 2015, doi: 10.1109/MSP.2015.2462851.

K. Zvarevashe and O. O. Olugbara, "Gender Voice Recognition Using Random Forest Recursive Feature Elimination with Gradient Boosting Machines," 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Durban, 2018, pp. 1-6, doi: 10.1109/ICABCD.2018.8465466.

H. Suzuki, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda and T. Kitamura, "Speech recognition using voice-characteristic-dependent acoustic models," 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., Hong Kong, 2003, pp. I-I, doi: 10.1109/ICASSP.2003.1198887.

Livingstone, Steven & Russo, Frank. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE. 13. e0196391. 10.1371/journal.pone.0196391.

McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, Oriol Nieto,... (2020) Librosa Python Package (Version 0.8.0)[Python Package] https://zenodo.org/record/3955228

Lausen, A., Hammerschmidt, K. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanit Soc Sci Commun* **7,** 2 (2020). https://doi.org/10.1057/s41599-020-0499-z

Joy, J, Kannan, A, Ram, S, & Rama, S. “Speech emotion recognition using Neural Network and MLP classifier.” International Journal of Engineering, Science, and Computing, vol 10, iss. 4, pp. 25170-25172, Apr. 2020. https://ijesc.org/upload/17015f34daa6e925c92ce026adabfca9.Speech%20Emotion%20Recognition%20using%20Neural%20Network%20and%20MLP%20Classifier.pdf

Becker, K. “Identifying the gender of a voice using machine learning” *primaryobjects.com*, June 2016. http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/#Update-Narrowing-Acoustics-to-Within-Human-Vocal-Range