CS 584 Machine Learning - FINAL PROJECT REPORT

December 2nd, 2022

Speech Classification & Emotion Recognition Using Advanced Machine Learning Techniques

Team Members

Evelyn Martin - A20495420

Jasleen Bhatia - A20495939

Sri Hari Sivashanmugam - A20480978

ABSTRACT

The most basic and typical method of human communication is speech. With the intention of developing a human-computer interface that would enable both humans and machines to converse similarly, speech recognition systems were developed. Covid has encouraged individuals to participate in online psychological sessions. It may occasionally be challenging for the other person to understand who they are speaking to since a small number of them are reluctant to turn on their cameras. They may effectively monitor the speaker and draw conclusions by using this program to determine the speaker's gender and mood. We are employing a neural network model to decode this information utilizing a number of important audio properties.

I. OBJECTIVE

The main goal of this project is to build a machine learning model that is capable of taking in a voice as input and to be able to classify it into different classes. Any speech recognition algorithm that extracts voice characteristics and performs speech recognition must be taught on the computer a succinct summary of the suggested strategy. This involves training a system so that voice recognition can be carried out. The machine has to be trained using any of the speech recognition algorithms which would extract the features of the voice and perform speech recognition.

Our proposed approach will be going to focus on three important issues, and we will be implementing-

- The features used to characterize different emotions.
- The classification techniques used in research.
- The important design criteria of emotional speech databases. Moreover, we are going to increase the average classification accuracy of speaker-independent speech emotion recognition systems.

II. PROBLEM DESCRIPTION

The following factors make speech emotion recognition a highly difficult endeavor. It is initially unclear which speech qualities are effective at differentiating. The second issue is that the speaker's culture, setting, and environment in general all affect how a speech is delivered. The majority of the research has been on categorizing monolingual speech with the presumption that there are no cultural differences among speakers. The goal of this project is to create a machine-learning model that can take a voice as input and categorize it

into several groups. We categorize the audio by labeling positive and negative comments by identifying their attitudes using sentiment analysis after classifying the inputs as male or female.

III. DATA PROCESSING

2.1 Dataset Description

LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech. This data was recorded for audiobooks from the LibriVox project. From this we are only bothered about the 100 hours of cleaned audio data. We have 251 audio files and labels alongside them marking if the speaker of the audio file is a male or a female. On an average each of these audio files lengths about 25 minutes.

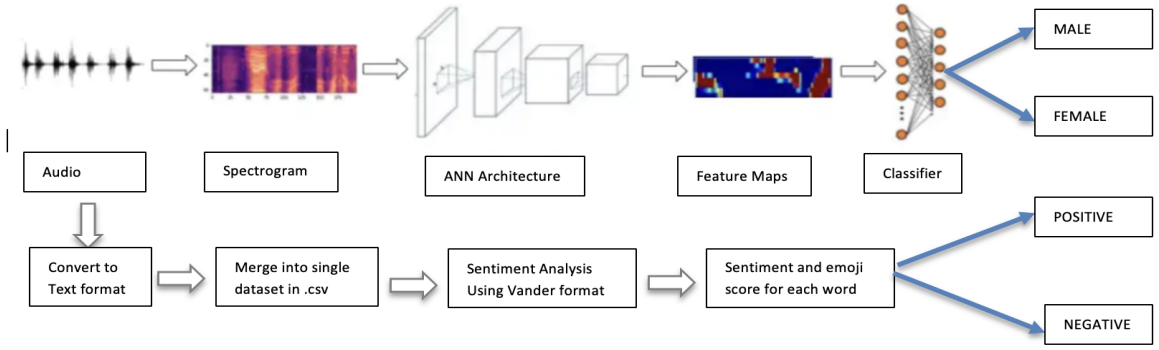
2.2 Issues in data and changes made

The data at hand is very large and to preprocess it everytime and use it for training it will take a lot of time, for us it took almost 6 hours. So we preprocessed the data for once and stored this preprocessed data as a pickle file, so that this pickle file can be loaded to train and test various models. For sentiment analysis, we use the text or the close captioned data, which means we have a hindrance of punctuations, long sentences and stop words. To handle this, we remove or replace the punctuations and stop words. We converted the long sentences into a list of words using the concept of tokenization. We then used these tokens into root words using a concept called lemmatization/stemming.

2.3 Tools

- Programming Language – Python
- Framework – Google Colab
- Libraries - keras, nltk, sklearn, numpy, matplotlib, pandas
- Development - GitHub

IV. PROJECT WORKFLOW



In Recent days it is hard to recommend customers because of the fake data Rather than using text files which has a lot of grammatical errors and punctuation it might lead us to a different understanding of the person. So we use audio files where we can analyze the tone, and frequency of the person's vocal expressions combined with the emotional data.
These audios are then passed onto the spectrogram, spectrogram is mainly used here to see the timely changes of the voice when an individual speaks. As per the above-mentioned picture, it is used to present the audio file's waveform in a visual representation. As the audio has been analyzed in a spectrogram we can know where there is a low pitch of voice and where there is a high pitch of voice, which might help in detecting the gender of the person easily.

The ANN classifier is given huge datasets to train the algorithm with different types of voice data to classify whether it is male or female.
In addition to this, The audio recordings were also evaluated, and text files were created from them. After converting each audio recording, we combined all the text into a single file. To accomplish our secondary goal of determining the speaker's attitude, mood, and opinion, we use the VANDER pre-trained model (Valence Aware Dictionary for Sentiment Reasoning). Vander is a lexical rule and rule-based sentiment analysis tool specifically tailored to the speaker's emotions. This method determines the compound score for each phrase by summing the lexical ratings of -1 (Extremely Negative) and +1 (Extremely Positive).

V. METHOD

For this prediction, we've utilized the Python soundfile, librosa, and sklearn packages. To build a model, we employed an artificial neural network. Stochastic gradient descent is used by the ANN to optimize the log-loss function. In essence, artificial neural networks are just computational neural networks that are trained via backpropagation and supervised

learning. After the data has been analyzed and depending on mistakes, learning takes place by altering connection weights. The sound file from the library will be extracted,

- The short-term power spectrum of a sound is represented by the MFCC (Mel Frequency Cepstral Coefficient).
- Chroma: Applies to pitch class 12
- Mel: Frequency of the Mel Spectrogram.

The initial stage of an automatic speech recognition system is to detect audio signal components that are useful for recognizing linguistic content and removing background noise, according to MFCC.

Chroma: The amount of energy in each pitch class contained in the signal is indicated by a chroma vector, which is generally a 12-element feature vector.

Mel: The frequency scale underwent some sort of non-linear alteration to produce the Mel Scale. This Mel Scale is designed such that noises that are equally spaced apart on the Mel Scale also "sound" to people since they are equally spaced apart.

In contrast to the Hz scale, where there is a clear distinction between 500 and 1000 Hz and a barely audible difference between 7500 and 8000 Hz.The input layers in this case take 193 characteristics, which are subsequently sent to the dense layers (3 hidden layers). We added dropout layers to neural nets after each dense layer to speed up processing and time. The output is then covered with the SoftMax activation function after the input has undergone a series of changes utilizing the hidden layer. Due of the somewhat big amount of data, we have also employed the Adam optimizer here.

Additionally, we examined the audio recordings and turned them into text files. We integrated all the text into one file after converting each audio recording. We employ the VANDER pretrained model to help us reach our secondary aim of assessing the speaker's attitude, mood, and opinion (Valence Aware Dictionary for Sentiment Reasoning). Vander is a tool for sentiment analysis that is especially customized to the speaker's emotions and is lexical rule and rule based. By adding the lexical ratings of -1 (Extremely Negative) and +1 (extreme Positive), this approach calculates the compound score for each sentence

VI. DATA MODELLING

An artificial neural network is built with sequential arrangement of layers. In this project we have used dense layer and dropout layer. The output is carried on to the dense layer. For the first three layers we have used relu as our activation function. The final layer has the softmax activation function to label the class that has passed through the output layer. There are dropout layers after each dense layer in the NN intentionally to improve the

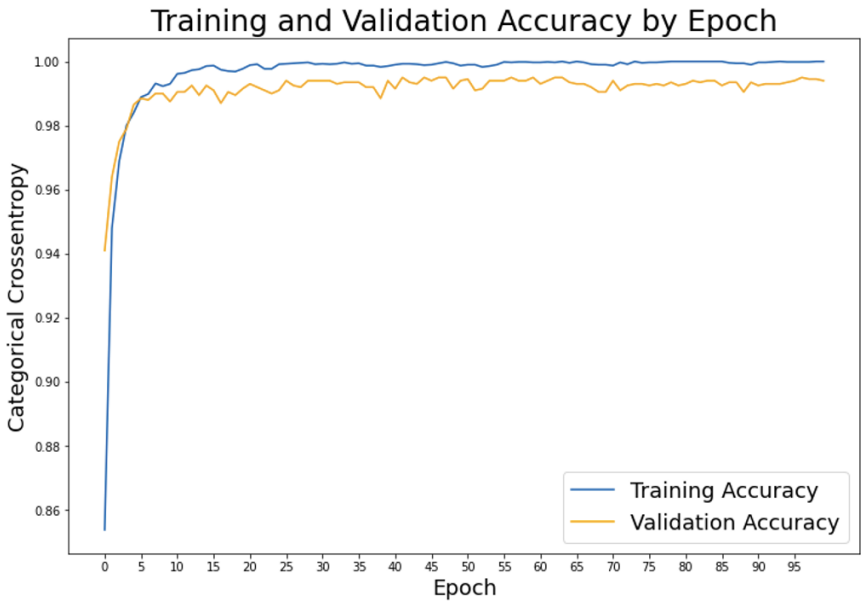
processing speed and time to results. The model is compiled using the adam optimizer as to update the weights consecutively which is executed by the training data.

Below are the stats of the model:

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_4 (Dense)              (None, 193)               37442

 dropout_3 (Dropout)          (None, 193)               0

 dense_5 (Dense)              (None, 128)               24832

 dropout_4 (Dropout)          (None, 128)               0

 dense_6 (Dense)              (None, 128)               16512

 dropout_5 (Dropout)          (None, 128)               0

 dense_7 (Dense)              (None, 2)                 258

=================================================================
Total params: 79,044
Trainable params: 79,044
Non-trainable params: 0
_________________________________________________________________
```

VII. MODEL EVALUATION

We use several assessment criteria to measure the prediction outcomes of these diverse strategies in handling the trip prediction problem in order to evaluate their performance. During the training phase, we record the training accuracy and validation accuracy for each epoch (totally 100 epochs). We use the test data to predict the gender using the trained model. We observe a test accuracy of 99.2%. Below is an image of the accuracy score and the list of misclassified audio files. Below mentioned are the statistics -



	file	label	preds
9038	3259-158083-0078.flac	0	1
9140	696-93314-0001.flac	0	1
9173	4406-16882-0066.flac	1	0
9281	250-142276-0051.flac	0	1
9598	2893-139310-0018.flac	1	0
9620	6064-56168-0008.flac	0	1
9633	3259-158083-0016.flac	0	1
9705	6531-61334-0096.flac	0	1
9860	7780-274562-0037.flac	0	1
10247	289-121665-0028.flac	0	1
10335	7402-90848-0012.flac	1	0
10344	3214-167602-0003.flac	1	0

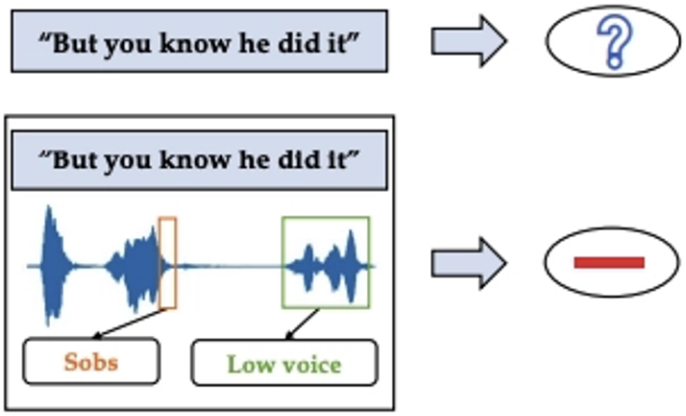
```
[ ]  1-round(len(df_test[df_test['label'] != df_test['preds']])/len(df_test),3)
```

```
0.992
```

VIII. SENTIMENT ANALYSIS

We listened to the recordings, analyzed them, and converted them to text files. After converting each audio recording, we combined all the text into a single file. To accomplish our secondary goal of determining the speaker's attitude, mood, and opinion, we use the VANDER pretrained model (Valence Aware Dictionary for Sentiment Reasoning). A model for text sentiment analysis called VADER (Valence Aware Dictionary for Sentiment Reasoning) is sensitive to both the polarity (positive/negative) and intensity (strong) of emotion. It can be used right away on unlabeled text data and is included in the NLTK package.

For instance, the words "love," "enjoy," "glad," and "like" all express a good feeling. Additionally, VADER is wise enough to comprehend the underlying meaning of these terms, such as the negative connotation of the phrase "did not love." Additionally, it recognizes how to emphasize words with capitalization and punctuation, like "ENJOY."



We won't make an effort to distinguish between a sentence's subjectivity, objectivity, or truthfulness. We are simply interested in whether the text communicates a favorable, unfavorable, or neutral opinion. Think on the text being studied first. It's possible that a model that was developed using reviews that were paragraph-long would be ineffective. Use a model that is suited for the job at hand. Next, choose the kind of analysis to do. Some basic sentiment analysis algorithms take it a step further and consider bigrams, which are two-word combinations. We'll be working on entire sentences, and to do this, we'll import VADER, a trained NLTK lexicon. This method determines the compound score for each phrase by summing the lexical ratings of -1 (Extremely Negative) and +1 (Extremely Positive).

	audio_folder	audio_name	text	scores	compound	sentiment
0	7800	283478	AS THE FRONT DOOR WAS FLUNG OPEN AND AN EXCITE...	{'neg': 0.092, 'neu': 0.855, 'pos': 0.053, 'co...	-0.2023	Negative
1	7800	283478	BUT WE DIDN'T BREAK HIS OLD WINDOW YOU KNOW EX...	{'neg': 0.068, 'neu': 0.855, 'pos': 0.077, 'co...	0.1236	Positive
2	7800	283478	FRANK YOU TALK WITH HIM I'D BE APT TO GET SASS...	{'neg': 0.138, 'neu': 0.785, 'pos': 0.076, 'co...	-0.2500	Negative
3	7800	283478	HE HELD UP SOMETHING HE HAD IN HIS HAND SO THA...	{'neg': 0.094, 'neu': 0.842, 'pos': 0.064, 'co...	-0.1280	Negative
4	7800	283478	AS THOUGH WONDERING WHETHER THE MISSILE THAT H...	{'neg': 0.209, 'neu': 0.791, 'pos': 0.0, 'comp...	-0.8176	Negative

IX. OBSERVATIONS & CONCLUSION

In the internet age where the information flow has grown rapidly, there is an increase in digital communication especially after the COVID. The listener finds it extremely challenging to comprehend the speaker.

Through this work, we developed a model which will help to detect the the gender of the speaker with the sentiments behind it. The model is implemented using deep learning sequential algorithm which is 99.2 % accurately predicting.

Furthermore, the analyzed input audio measures the attitude and sentiments of the speaker.

X. PLANNED FUTURE WORK

Currently, our data is not enough to detect the emotions of the the speaker. To do this, we need to manipulate the data by analyzing the speaker pitch, intensity and frequency and creating a model which will capture the correct prediction of speech emotion recognition. We believe this model will give us more clarity regarding the speaker that can explain in real world. The data exploration and analysis so far has given us valuable insights regarding the bottlenecks of computational resources and analysis process.

XI. REFERENCES

- https://en.wikipedia.org/wiki/Multilaye_r_perceptron
- https://ijcrt.org/papers/IJCRT2105931.pdf
- https://research.aimultiple.com/
- J. Ma, H. Jin, L. Yang, J. Tsai, “Ubiquitous Intelligence and Computing”, Third International Conference, UIC 2006, Vol. 4159.
- J. Nicholson, K. Takahashi, R. Nakatsu, “Emotion recognition in speech using neural networks”, Neural Computing & Applications, 2000, Vol. 9, No. 4, pp. 290-296.

Project Link:
[Github]