CS536: Project Proposal
Speech Emotion Recognition

Harshini Bonam (NetID: sdb202) **Sai Mounica Pothuru** (NetID: sp1912)
Chaitanya Sharma Domudala (NetID: cd817)

April 15 2020

Introduction

Speech Emotion Recognition, abbreviated as SER, is a task of processing and classifying speech signals to recognize the embedded emotions. While this is efficiently performed by humans as a natural part of speech communication, the ability to conduct it automatically using programmable devices is still an ongoing subject of research. This is capitalizing on the fact that tone and pitch in the voice reflects underlying emotion.

SERs objective is to create efficient, real-time methods of detecting the emotions of diverse human-machine communication users. Adding emotions to machines has been recognized as a critical factor in making machines appear and act in a human-like manner (André et al., 2004). Machines capable of understanding emotions could provide appropriate emotional responses and exhibit dynamic personalities. Success in this field is defined by the machine's capability to conduct very natural and convincing conversations by appealing to human emotions.

What to Solve

The objective of this project is to build a model that recognizes emotion from speech using using Python library librosa over RAVDESS dataset to analyze audio and music. This dataset consists of 7356 files rated by 247 individuals 10 times on emotional validity, intensity, and genuineness. The entire dataset is 24.8GB from 24 actors.

How to Solve

Model Development:

- **Feature Extraction:** The Mel-Frequency Cepstral Coefficients (MFCC) are extracted from the the audio samples on a per-frame basis with a window size of a few milliseconds. The MFCCs summarise the frequency distribution across the window size, which allows to analyse both the frequency and time characteristics of the sound. These audio representations will allow us to identify features for classification.
- **Classifier Model:** A Deep Neural Network is trained with the training data set to generate predictions. In this regard, an MLP Classifier(Multilayer Perceptron Classifier) [4] and a CNN (Convolution Neural Network) [1] would be built as they optimizes the log-loss function using LBFGS or stochastic gradient descent. MLP Classifier is preferred over SVM [2, 3] or Naive Bayes as it has an internal neural network for classification and is a feedforward ANN model.

How to Evaluate

- The evaluation metric for this project is the classification accuracy which is defined as,

$$accuracy = \frac{Numberofcorrectclassification}{TotalNumberofclassification} \tag{1}$$

  Other metrics such as Precision, Recall (or combined as the F1 score) may also be used depending on the balance of the data set.

- The target is to beat the baseline SVM and Continuous Wavelet Transform (CWT) reported a maximum accuracy of 60.1%. [5]

References

[1] Reza Chu. Speech emotion recognition with convolutional neural network, https://towardsdatascience.com/speech-emotion-recognition-with-convolution-neural-network-1e6bb7130ce3, 2019.

[2] Shen L Pan Y, Shen P. Speech emotion recognition using support vector machine. int j smart home, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.390.8138rep=rep1type=pdf, 2012.

[3] Gaikwad VB Praksah C. Analysis of emotion recognition system through speech signal using knn, gmm svm classifier, https://d1wqtxts1xzle7.cloudfront.net/50967341/52$_i$*jecs.pdf*?1482159805 = *response* − *content* − *disposition* = *inline*%3*b* + *filename*%3*danalysis*$_o$*f*$_e$*motion*$_r$*ecognition*$_s$*ystem*$_t$*.pdfexpires* = 1617681689*signature* = *qiwttvhbusmdxn*1*ovs*9*ztlzzat*5*f*2*gbwwdowv*5*n*4*gcphxfibhopwjwvgku*23*y*2*jkpdfg* − *ngtnj*1*cvyyq*5*faevrajt*5*rkxie*1*jynpu*66*ku*5*wkzafo*3*ejcfk*8*jyt*68*ljvzyiejhbkts*9*y*6*qjjfodzzvfjswyjehbuxisvv*4*bcgyv*5*kubg* 4*jy*2*c*1*rdgnjecqefn*8*addpdwfrhrk*4*r gvk*25*vfc*9*ngaqe*05*uualhjnv*1*yie*17*vdx*4*iyujqhdelz*3*tuv*0*bce*1*a mrdyf*2*urkimtwxp*1*v* *decox*1*tlvpa* − 8*ygeg*, 2015.

[4] Neethu Sundarprasad. Speech emotion detection using machine learning techniques, https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1647context=etd$_p$*rojects*, 2018.

[5] Dezhong Peng Zhang Yi Yuni Zeng, Hua Mao. Spectrogram based multi-task audio classification, https://link.springer.com/article/10.1007/s11042-017-5539-3, 2017.