Fall 2023 – CS460G Machine Learning – Project Proposal

Group Members	Email
Habib Unluoglu	hun222@uky.edu
Sania Eskandari	ses235@uky.edu
Brooke McWilliams	brmc230@uky.edu
James Birch	jabi236@uky.edu

Proposed Subject: Speech Emotion Recognition

1 INTRODUCTION

Numerous forms of communication exist, but the speech signal stands out as one of the fastest and most instinctive means of human interaction. Consequently, speech can also serve as a rapid and effective mode of communication between humans and machines (Ayadi et al. 2011). Expressing emotions and attitude through language is also important. Specifically, detecting emotional cues in speech signals and determining the underlying emotions within spoken expressions is a significant and valuable endeavor (Swain et al. 2018). Therefore, due to the digitally growing era and the importance of emotions within spoken expressions, speech emotion recognition holds substantial importance in a wide array of applications, including Human-Computer Interaction (SCI), lie detection, steering assistance in automotive systems, intelligent tutoring, audio mining, security, telecommunications, and human-machine interaction in various settings like homes, hospitals and shops (Koduru et al. 2020).

We, as a group, have chosen to embark on our group project in the domain of "***speech emotion recognition***" for specific reasons. One of the primary motivations behind our choice is the growing importance and relevance of this field in various applications and industries. Additionally, we believe that our exploration of speech emotion recognition will not just deepen our comprehension of the technology but will also grant us valuable perspectives into the ever-changing realm of human-computer interaction, affective computing, and the prospects for enhancing user experiences across diverse scenarios. By embarking on this project, we aim to contribute to the body of knowledge in this area and develop practical applications that can

make a positive impact in fields. A challenge for this project will be assuring that the data is not skewed and stays uniform after extracting the features from each file. Male voices could give higher feature values while female voices could give lower feature values but result in the same class label (emotion).

2 DATASETS

2.1 Overview

We are using four datasets collected from “Kaggle”. These datasets include Crema, Ravdess, Savee, and Tess. Each dataset is labeled in a unique way to convey the emotion the speaker is using.

2.2 Crema

Starting with Crema the third component of the file name is describing the emotion of the speaker. There are four emotion labels associated with this dataset and they are disgust (DIS), fear (FEA), happy (HAP), and neutral (NEU). The Crema dataset includes a total of 7442 audio files.

2.3 Ravdess

Next is the Ravdess dataset. These files are labeled in such a way that each set of numbers represents a different identifier. The first set of numbers describes the modality of the file (full-AV, video-only, or audio-only). The second set of numbers describes the vocal channel (speech or song). We will only use the speech for purposes of this project and to keep the datasets equivalent. The third set of numbers represents the emotion (normal, calm, happy, sad, angry, fearful, disgust, and surprised). The fourth set of numbers describes the statement being made (kids are talking by the door, or, dogs are sitting by the door). The fifth set of numbers represents the first or second repetition of the statement. Finally the sixth set of numbers describes which actor made the statement, odd numbers being male and even numbers being female. This dataset has 24 total actors and each actor has a total of 60 audio files. The total size for all actors is 1440 audio files.

2.3 Savee

The Savee dataset contains a prefix in the file names to describe the emotion of the speaker. There are 7 different emotions in this dataset; anger (a), disgust (d), fear (f), happiness (h), neutral (n), sadness (sa), and surprise (su). There are a total of 480 files and they are not evenly split between emotions so one class may have more occurrences than another.

2.4 Tess

Lastly is the Tess dataset. This set contains audio files from 2 actresses expressing the emotions of fear, pleasant surprise, sad, angry, disgust, happy, and neutral. Each actress has a total of 1400 audio files so the total size for the dataset is 2800 audio files to use.

2.5 Conclusion

Since these datasets all contain more or less information than the other, we will carefully examine the data to determine which ones to use and potentially, if we can, combine some data to get more training examples for our model.

3 EVALUATION METRICS

Given that each of the four datasets varies in the amount of information it contains, we will thoroughly examine the data to determine which one to use potentially. Additionally, if we can, we may combine some dataset to get more training examples for our model. There are various potential evaluation metrics we can use in our dataset. We will remove some of the evaluation metrics listed below and select the most suitable ones once we have finalized our preferred dataset. Potential evaluation metrics can be listed as:

Word Error Rate (WER): WER measures the percentage of words that are incorrectly recognized by the system. It quantifies the number of word insertions, deletions, and substitutions in the recognized text compared to the reference text.

Character Error Rate (CER): Similar to WER, CER calculates the percentage of character-level errors in the recognized text as compared to the reference text. It's more fine-grained than WER and is often used in applications where individual characters matter.

Phone Error Rate (PER): PER assesses the accuracy of recognizing individual phonemes (speech sounds) within the speech signal. It quantifies the number of errors in phoneme recognition.

Frame Error Rate (FER): FER evaluates the performance of automatic speech recognition (ASR) systems in the context of continuous speech. It measures the accuracy of the phonetic or acoustic modeling by considering frames of audio.

Accuracy: Accuracy is a simple metric that measures the percentage of correctly recognized words or characters in the transcription. While easy to understand, it may not account for the severity of errors.

Precision and Recall: Precision measures the proportion of correctly recognized positive instances (correctly recognized words) out of all instances recognized as positive. Recall measures the proportion of correctly recognized positive instances out of all actual positive instances.

F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives.

Confusion Matrix: A confusion matrix provides a more detailed breakdown of true positives, true negatives, false positives, and false negatives, allowing for a deeper understanding of error types.

Perplexity: In the context of language modeling, perplexity measures how well the model predicts a sequence of words. Lower perplexity indicates better language model performance.

Mean Opinion Score (MOS): MOS is a subjective evaluation metric where human listeners rate the quality and intelligibility of recognized speech. It's often used to assess the user experience.

The choice of evaluation metrics will depend on our preferred dataset including specific goals of the speech recognition system and the nature of the task. For example, in applications like transcription services or voice assistants, WER and CER may be critical, while in acoustic modeling or phonetic analysis, metrics like PER or FER could be more appropriate. It is essential to select metrics that align with the objectives of the system and the intended use cases. We will solidify our choice of evaluation metrics as we delve deeper into the project over time.

4 BASELINE MODEL

Recurrent Neural Networks (RNNs) are a specialized type of neural network architecture specifically designed to process sequential data (Goodfellow et al., 2016). RNNs excel in tasks involving natural language processing, speech recognition, and time series analysis. They incorporate recurrent connections that enable the network to retain information from previous time steps and integrate it with the current prediction, allowing them to capture dependencies and long-term relationships within sequential data.

While RNNs have demonstrated success in modeling sequential data, they encounter challenges such as the vanishing or exploding gradient problem, which can impede their ability to learn long-term dependencies. To mitigate these issues, variants of RNNs, including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), have been developed. These variants introduce gating mechanisms that regulate the flow of information, enabling better preservation of relevant information over longer sequences (Goodfellow et al., 2016).

The baseline model we have decided to use for comparison is the **Recurrent Neural Network (RNN) model.** RNN is made to handle input sequences of different lengths, making it great for

varying audio files. Another baseline model that might be good for comparison is the Convolutional Neural Networks (CNN) model. CNN is great for audio files, being able to learn and extract features from spectrograms and other audio representations. Either of these neural networks would be great for processing and classifying audio files.

6 REFERENCES

Ayadi, M.M., Kamel, M.S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognit., 44, 572-587.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. Chapter 10.

Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. International Journal of Speech Technology, 23(1), 45-55.

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology, 21, 93-120.