2020

Sending Mixed Signals: Feature Extraction for Gendered Emotional Speech Classification and Modeling

David McCulloch Pfeffer

The College of Wooster, dpfeffer20@wooster.edu

Sending Mixed Signals: Feature Extraction for Gendered Emotional Speech Classification and Modeling

Independent Study Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Bachelor of Arts in Computer Science in the Department of Mathematical & Computational Sciences at The College of Wooster

by
David Pfeffer
The College of Wooster
2020

Advised by:

Sofia Visa, Ph.D. (Computer Science)

Abstract

This study seeks to analyze waveforms of gendered, emotional human speech to extract acoustic features from the RAVDESS emotional speech dataset. After a phase of preprocessing using min-max normalization, qualitative and quantitative signal analysis is performed on the samples in which maximum amplitude, average amplitude, and summed amplitude features are extracted. We find that the summed amplitude feature is the most informationally-rich and compelling of the features extracted, and move forward with its use in the classification phase. We deploy two clustering algorithms to perform classification on the speech samples: *k*-means and agglomerative clustering. The results of the clustering show similarities between some gendered emotion samples, but fail to cluster along gender or emotion type. Model prototypes are then created through the inclusion of more samples, and through further qualitative analysis performed on a variant of the summed amplitude feature. These reveal similarities in the volume shifts within each emotion variant, and a marked increase in volume for happy male speech and sad female speech specifically. Finally, we propose a framework for an automated speech classification algorithm.

Acknowledgments

It is near impossible to properly acknowledge my family and the people in my life whom I love for all that they have done for me up to this point. Regardless, here are a few lucrative examples:

I would like to acknowledge my parents for their endless support throughout my developing years, but particularly through college. Whether it be in the audience of a performance, through a phone call from multiple continents away, or checking in to make sure that I am not buckling under the pressures of the professional world, I have felt it wholeheartedly. I would also like to acknowledge my brother for his unwavering sense of comedic support, and for compelling me to succeed both academically and professionally.

I would like to acknowledge my advisor, Dr. Visa, whom has been my handler and offered me unparalleled support for all four years from being a nervous first-year in a class of upperclassmen, to being an experienced (yet still nervous) computer scientist. Additionally, I would like to acknowledge Dr. Lisa Wong. The Wooster Chorus has been a creative refuge for me throughout my college experience. It is not a coincidence that through one ensemble I've grown as an artist, and through another organization I have grown as a leader, that she is the advising faculty.

Finally, I would like to acknowledge my dear friends, new and old. Thank you for inspiring me, for challenging me, and for caring for me. Thank you for the laughs, the coffee klatches, the introspection, the sing-alongs, and the *endless* advice. We have grown together through our experiences, and will continue to grow.

Contents

List of Figures

Figure Page

List of Tables

CHAPTER 1

Introduction

Human beings are *natural* language processors. We constantly consume language and attempt to interpret the layers of meaning within it, whether it be in relation to content, context, or intent. In this vein, emotion plays a key factor. Determining the emotion that underlies language can be very important in determining context and intent, specifically. Human beings most often consume language via spoken utterances (a word or statement) through their sense of hearing. Therefore, human beings are able to determine the emotion behind language much easier when consuming spoken utterances versus written language. Since spoken language is easier to derive emotional value from, there must be latent acoustic (audio-based) features within the utterances that allow us to make these determinations. This is especially significant considering that many people can determine emotion from hearing a single utterance, without context. Thus, our first research question is simple: how do we convey emotions within our speech?

From here, there is another factor to be investigated that may influence the manner in which emotions are conveyed acoustically: gender. People of different gender identities may speak in different (pitch) registers, and thus may employ different patterns in their speech to convey emotion. Some possible acoustic features include variance in volume, pitch, and rate of speech. The dataset utilized in this study contains denoted *gendered* speech between two genders: male and female.

While this dataset is limited in its inclusion of only two genders, it provides us still with another research question: how do males and females convey their emotions differently within their speech?

We already know that humans are adept at determining the emotion conveyed by speech. Can we train a computer to do the same? Machine learning algorithms, particularly in the field of natural language processing, have been resourceful and powerful in breaking down language into machine-interpretable data. This data can then be analyzed to create a machine learning model which will serve to form predictions on future data of the same type. For our purposes, the data to be focused on are audio samples of emotional speech, which come to us first as an audio waveform. From here, the language will be broken down into numerical features, which could be derived from volume, pitch, and rate of speech as described earlier, and then fed into machine learning algorithms which will (a) organize the data mathematically by emotion and/or gender, or (b) form predictions of the emotion and/or gender of the speech uttered. Regardless, it is the automated nature of these machine learning algorithms that is compelling, as they closely resemble the seemingly automatic processes that occur in the human brain when human beings determine emotion from speech.

We propose that human speech contains latent acoustic features that allow for the distinction and categorical approximation of the utterance's emotional value as well as the gender of the utterer, which can be discovered through feature extraction and then processed into discrete judgements by machine learning classification algorithms. Specifically, we model three emotions (neutral, happy, sad) from the same utterance spoken by three males and three females.

The next chapter provides a brief introduction into the dataset and tools used in this study, and begins to perform analysis on the raw data as supplied by the RAVDESS dataset [8].

The remaining parts of this thesis are organized as follows: Chapter 2 provides insight into the tools and dataset used in this study and discusses the qualitative and quantitative signal analysis performed to extract features. Chapter 3 describes machine learning classification and the clustering algorithms used in this study, and then discusses the results of these clustering algorithms. Chapter 4 discusses further emotional speech analysis that is performed using k-means clustering to model the data. Chapter 5 describes the prototypical models that are built for the emotional speech previously analyzed. Finally, Chapter 6 examines our conclusions and proposes an automatic emotional speech classification algorithm, and also discusses future work to extend this study.

CHAPTER 2

Signal Analysis of Emotional Speech

The first phase in the process of signal analysis begins with purely the raw audio signal waveform. Before we move to performing analysis on the frequencies and energy found within the audio sample, the amplitude of the sound wave can be studied. The amplitude is representative of the volume of the sound itself, or how loud or soft the sound is. Other properties of the audio sample like the aforementioned can be found via mathematical computation and other signal analysis techniques. This chapter will breakdown qualitative and quantitative signal analysis performed on the original waveform, the breakdown of the wave into different word/sound frames, and the simple mathematical preprocessing technique of normalization. We define qualitative analysis as similar to visual analysis, in which purely the graphical representation of the waveform is analyzed, and quantitative analysis as the analysis of numerical representations of the waveform. Analysis techniques such as these are representative of human analysis of the human voice. We associate different volumes with different emotions. For example, one might attribute loudness to anger or happiness, and softness to sadness or nervousness. Machine analysis serves to bolster and/or inform these basic arguments.

For this chapter, three emotions are analyzed for two voice actors, one male and one female. The emotions analyzed are ‘neutral’ (a baseline), ‘happy,’ and ‘sad.’

This results in the analysis of six speech audio signals. Before this, however, the dataset and various Python libraries used are described.

2.1 Data Corpus and Tools

2.1.1 RAVDESS

The dataset used for this thesis is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), created by Steven R. Livingstone [8]. This dataset is quite large: it contains 1,440 audio-visual files, wherein 24 voice actors record 60 different trials of emotional speech in the form of regular speech or song. Half of the voice actors are male, and half are female. The voice actors record two different statements in 8 different emotions: neutral (a baseline), calm, happy, sad, angry, fearful, disgust, and surprised. Additionally, each of the emotions (besides neutral) are recorded twice, once with a normal emotional intensity, and another time with a strong emotional intensity.

The first statement performed by the voice actors, and used here, is "the kids are talking by the door." For the beginnings of this study, only the emotions "neutral," "happy," and "sad" are considered from two voice actors, one male and one female. Once qualitative and quantitative analysis is performed on these samples, and algorithmic work is performed, then the rest of the dataset can be analyzed. This dataset was downloaded from Kaggle, on online data science community created by Google that houses datasets and data models.

Before diving in, it is necessary that the dataset be organized into a data structure within Python. The dataset's filenames are organized using a 7-part numerical format. The full bounds of this format are as follows:

- Modality (01 = full audio-visual, 02 = video, 03 = audio)

- Vocal channel (01 = speech, 02 = song)
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised)
- Emotional intensity (01 = normal, 02 = strong)
- Statement (01 = "the kids are talking by the door," 02 = "the dogs are sitting by the door")
- Repetition (01 = first repetition, 02 = second repetition)
- Actor (01-24, odd numbers are male, even numbers are female)

An example filename used in this study is "03-01-03-01-01-01-01.wav," which is a file for happy male speech from the first voice actor.

2.1.2 Pandas

Pandas (Python Data Analysis) is a Python library that provides a multitude of data analysis and visualization tools. Pandas is widely used for its data structure and visualization tools in the data science industry as well as academia, and is accredited for its intent to remain an open-source tool [3]. The files are first imported using Python's OS (operating system) library. Then, the `DataFrame` data structure from Pandas is leveraged. A `DataFrame` is a two-dimensional tabular structure whose size is mutable. Additionally, the rows and columns are labeled. As such, the RAVDESS data is organized into different columns for gender and emotion.This is useful both for ease of access, and for attaching a label to the filename instead of having to parse for the numeric identifier later. This automatic, algorithmic approach is more robust than simply loading the files via their name, one at a time.

2.1.3 NumPy

NumPy is "the fundamental package for scientific computing" [2]. It is a Python library that contains a variety of powerful, robust tools. The most iconic of which is its `ndarray`, or an N-dimensional array structure. It is very similar to Python's regular list, but it is heavily utilized as input for other scientific computing Python libraries, like Matplotlib, SciPy, and Scikit-learn. However, `ndarray`s have their own functionality further attached to them in the form of mathematical operations. Specifically utilized in this thesis are various statistical functions such as `max()`, `mean()`, and `sum()`. Additionally, `ndarray`s can be converted to and from Python lists. Similarly to Pandas, NumPy also commits itself to remaining an open-source library [2].

2.1.4 Matplotlib

Matplotlib is a Python library that allows for the creation of plots, histograms, bar charts, and other figures. The ease in which plotting is performed truly is the hallmark of this library. The figures created are often used in academic writing. All figures for this thesis are created using Matplotlib, whether they be regular plots, bar charts, or another type. The `plot()` function simply takes in `ndarray`s containing the x and y values to be plotted, and users can change the various aspects of the graph including the title, axis labels, font, etc. to their liking. Like Pandas and NumPy, Matplotlib also is a sponsored project of NumFOCUS, a nonprofit which ensures that Matplotlib remain open-source for users [1].

2.1.5 Wavio

Wavio is a Python module created by Warren Weckesser that allows for easy reading and writing of .wav files. Wavio leverages the `wave` module in Python, but creates an abstraction in that there are less limitations for input .wav files [9]. Also, the simplicity of Wavio's two sole functions (`wavio.read` and `wavio.write`) make it an excellent choice for a beginner in audio signal analysis. For the purposes of this thesis, only Wavio's `read()` function is used. This function takes in a .wav file, and returns the sampling rate, the sample width, and a numpy array containing the data (the amplitude of the audio signal). At the time of writing, Wavio is an open-source tool.

2.2 Qualitative Analysis of Three Different Emotions

Qualitative analysis involves making observations and estimations of the audio signal based purely on viewing the waveform itself. This form of analysis is one of the most simple and natural processes that we as humans perform. It is a good tool in forming hypotheses and observations on data, and audio signals are no exception. Characteristics of the wave itself such as its amplitude, length of peaks and troughs, and general topography can all be viewed and analyzed. For this section, the speech signals between two genders for each emotion are compared.

2.2.1 Neutral Wave Form Analysis

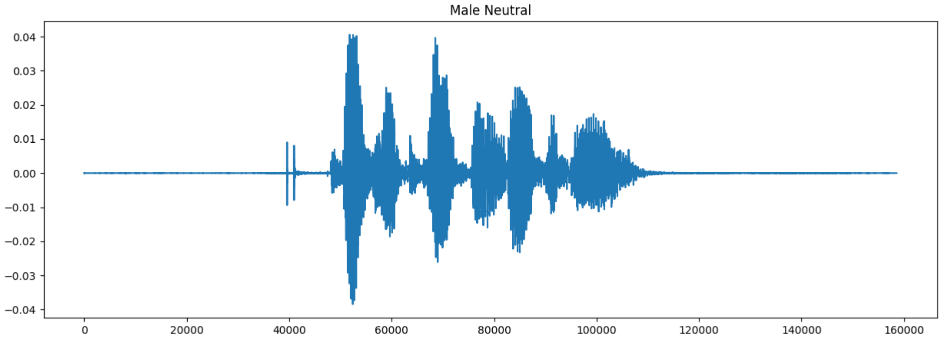

Figure 2.1: Audio signal of neutral male speech

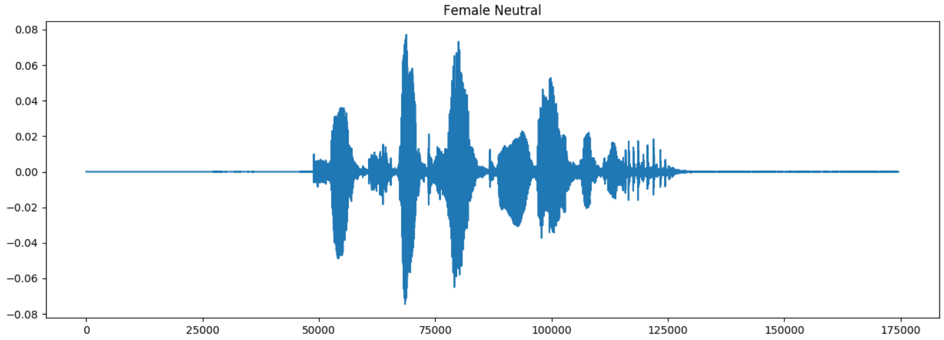

Figure 2.2: Audio signal of neutral female speech

As Figures 2.1 and 2.2 show, the ’neutral’ emotion does not elicit large variances in amplitude in the wave form. This is perhaps due to the monotonic tendency of truly neutral speech. In both the male and female cases, around 5-6 peaks in amplitude are observed. The male case exhibits softer/lower speech, wherein the highest peak in amplitude measures around 0.05. The female case exhibits slightly louder/higher speech, where the highest peak in amplitude measures closer to 0.075. In both cases there is considerably more silence/less noise between words. This can be viewed as the low amplitude regions between the peaks.

2.2.2 Happy Wave Form Analysis

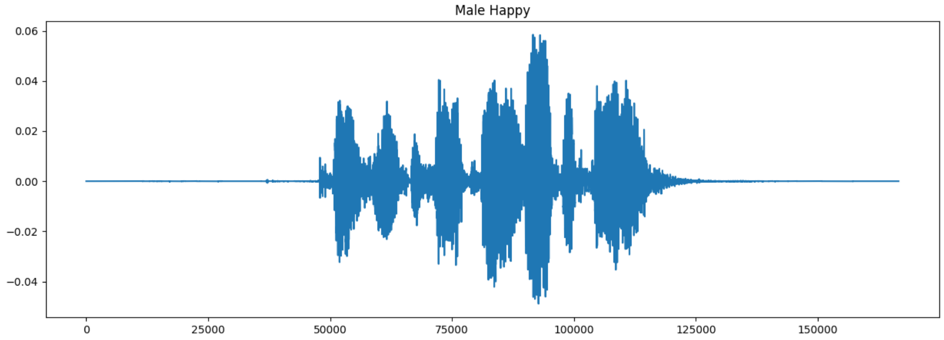

Figure 2.3: Audio signal of happy male speech

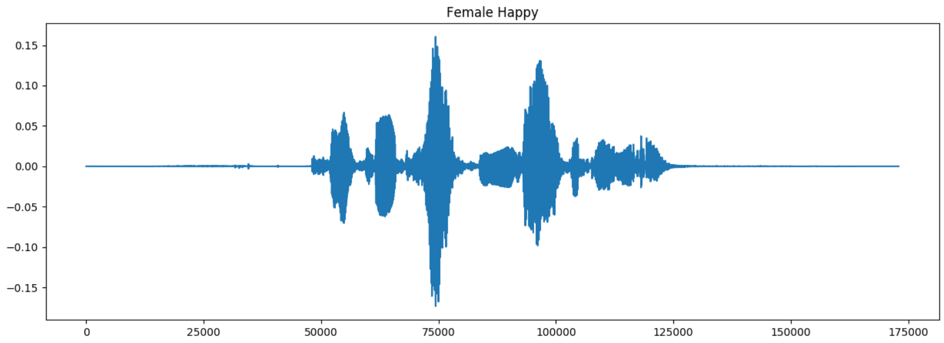

Figure 2.4: Audio signal of happy female speech

The ’happy’ emotion shown in Figures 2.3 and 2.4 elicits more excited speech than the ’neutral’ emotion. In the male case, the highest peak in amplitude still reaches roughly 0.05. However, these peaks are sustained for longer intervals of time. This results in fewer instances of silence/lower noise between words. The overall topography between the male cases of ’neutral’ and ’happy’ are quite similar, except for the held out words in the latter. This is not the case with the female voice. In the female case for ’happy,’ the peaks in amplitude are much larger than in ’neutral,’ measuring close to 0.16. Additionally, there are less peaks that are spaced out differently than the female case in ’neutral.’ Viewing the topographical differences between the two cases, we can see that the pitch wave contour for the female case of ’happy’ is dramatically different than all of the other cases, showing that this is the loudest/highest emotional speech.

2.2.3 Sad Wave Form Analysis

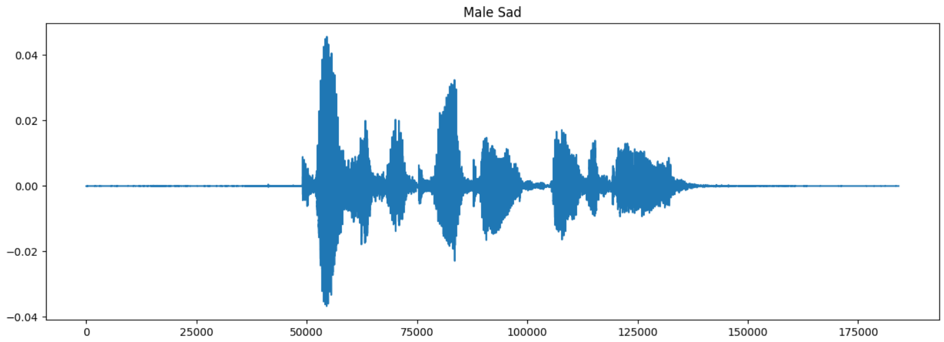

Figure 2.5: Audio signal of sad male speech

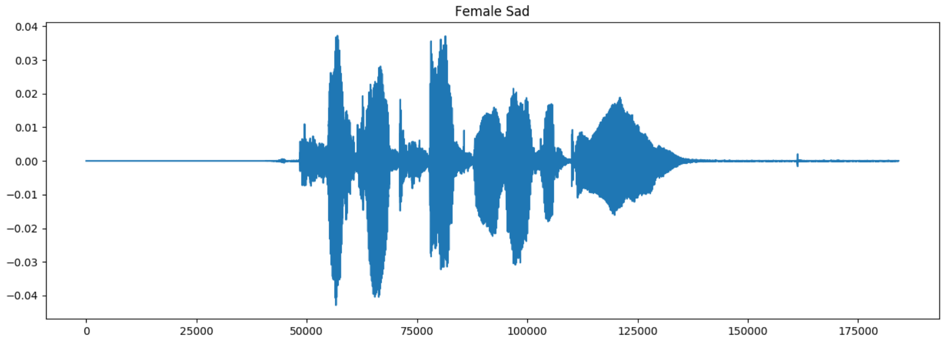

Figure 2.6: Audio signal of sad female speech

The 'sad' emotion elicits much softer speech than the other emotions (see Figures 2.5 and 2.6). Looking generally at both the male and female cases, the peaks in amplitude are much lower than 'happy,' and even lower than in 'neutral.' It is a little ambiguous as to what are peaks and what is general noise within the audio sample, but some peaks can be made out of the topography. In the male case, the highest peak amplitude is measured at just under 0.05, whereas the rest of the peaks are measured closer to 0.03. In the female case, all of the peaks are measured uniformly around 0.04. There is noticeably more silence/background noise between words in the 'sad' emotion. Comparing the topography of these two cases to the other emotion cases, we see that the ’sad’ emotion has ultimately the softest and lowest emotional speech.

From this basic qualitative analysis, we can spot trends in emotional speech between two genders. Overall, female speech across the three different emotions is generally louder and higher in amplitude. Additionally, between emotions, the ’happy’ emotion was the loudest and highest in amplitude readings, whereas the ’sad’ emotion registered the softest. The ’sad’ emotion also had the most silence and space between words. In order to begin quantitative analysis on the emotional speech and the words themselves, preprocessing steps need to take place.

2.3 Preprocessing the Waveform

Data preprocessing within data mining and machine learning involves taking in raw data, and transforming it into a viable data format for further use and/or analysis. Raw data is often inconsistent between samples, and thus analysis and comparisons between the samples can be skewed. Lacrose et al. declare the objective of this process as minimizing "GIGO," or the **G**arbage that gets **I**nto the model, so that the **G**arbage that the model gives **O**ut is also minimized [7]. The garbage in this case being outliers, missing values, and the like. For our purposes, we preprocess the audio signal data so that it may be easily analyzed and compared against itself in an equally-weighted format. This allows for unbiased conclusions that are independent of the amplitude of other speech samples. Such preprocessing techniques are normalization and sound-frame construction.

2.3.1 Normalization

Normalization is a preprocessing technique that organizes and scales the data. This standardizes the scale in which the variable in question affects the results [7]. There are a variety of normalization scales and techniques such as z-score normalization, standard scaling, and power transformation. For our purposes we use min-max scaling, which looks at how much larger the value is than the minimum [7]. This scales the data into a fixed range between 0 and 1 inclusive. The minimum values are scaled to 0, while the maximum values are scaled to 1, with everything else in between. Min-max scaling is usually performed using the following equation:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This is useful in making comparisons between audio signals that have varying maximum and minimum amplitudes. Specific to this study, normalization helps when comparing the speech of many different individuals. For instance, someone may be very loud across all emotions, while another may be more subdued across all emotions. Normalization is performed in Python using the Scikit-learn library's preprocessing package and the `normalize()` function, making sure to use the parameter `norm='max'` for min-max scaling. In the following Figure 2.7, the data for the neutral male audio signal is scaled accordingly:

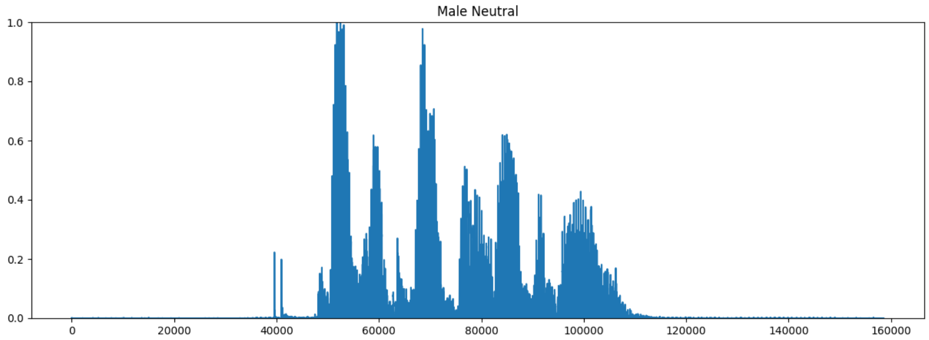

Figure 2.7: Normalization of neutral male speech

For ease of computation later, the negative values from the signal are dropped off. The effect of this is negligible due to the symmetric nature of the wave form itself. It can be seen that the general topography of the wave form is still intact from before normalization. It is important to note that only the y-values are scaled, so that the time-positioning of the x-values remain the same. As such, the peaks and troughs are still in their same places on the x-axis, but have now been scaled from 0.0 to 1.0. Now that all of the samples are within this bound, they can be equally compared to each other. Before, these comparisons were performed at the surface level, looking at the entirety of the signal. However, more information can be gathered when we look even more close up: at the word-level.

2.3.2 Frame Construction

A good method for analyzing speech waves at the word-level is by creating "frames" of sound. The frames are actually time interval divisions of the audio signal itself, in which each interval comprises one word. In other words, these frames are not arbitrary, as they relate directly to each word. For the audio signals looked at in this study, there are seven frames of sound for the seven words in the example sentence, "the kids are talking by the door." These word frames can be constructed manually by hand-selecting the points on the x-axis to splice into frames.

Manual frame construction was performed for each speech signal by estimating points on the x-axis where words that were uttered start and end. While these estimations are performed via qualitative analysis, they are heavily influenced by the topography of the wave form, specifically in regard to where peaks occur. Figure 2.8 below displays the manually constructed frames:

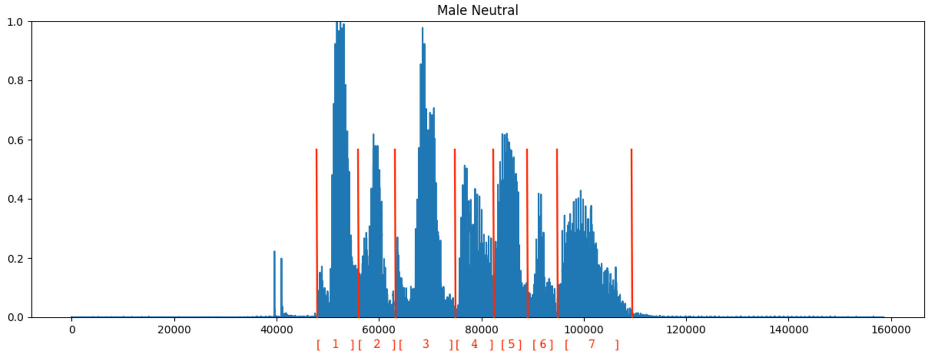

Figure 2.8: Manual frame division and construction for neutral male speech

The red lines in Figure 2.8 mark the interval points for each frame. These interval points are indexed directly into the array holding the audio signal data. For example, the frames for Figure 2.8 are constructed in Python with the following line of code:

```
male1_frames = [norm_ys[0][48000:56000], norm_ys[0][56001:64000],
                norm_ys[0][64001:74000], norm_ys[0][74001:82500],
                norm_ys[0][82501:90000], norm_ys[0][90001:96000],
                norm_ys[0][96001:110000]]
```

While these manually indexed frames are not perfect in their division of the word frames, they are accurate in their representation of the qualitative analysis performed prior on the audio signals. Here, norm_ys is a two-dimensional array of the normalized y-axis values for all of the speech samples. So as to not get frames that overlap, each consecutive frame starts one index after the previous. These frames are then converted into ndarrays from their original list format. Frame construction is vital in creating a usable data format for the analysis of speech signals. These word frames can be easily tested on and used within computation and especially within feature extraction. In the next section, quantitative analysis is performed on the speech signals using the word frames as a corpus, and features are extracted.

2.4 Quantitative Analysis and Feature Extraction

Quantitative analysis is the process of making observations and conclusions strictly from numerical data. This form of analysis is most often used in empirical studies, and is much more formal than qualitative analysis. Quantitative analysis could not be performed on the wave form itself, but can be performed after the preprocessing phase of frame construction. This process can be performed on non-normalized data, but normalization ensures an equal weighting between samples, as described earlier. Quantitative techniques such as measuring the maximum, average, and summed amplitude of the words themselves are performed for each frame. These operations can be viewed as discrete features extracted from the speech signal. Then, using these new features, comparisons between two genders and the three emotions are analyzed.

2.4.1 Maximum Amplitude Feature

The maximum amplitude for a wave form is the highest volume reading of the sound wave. Specifically in regard to the word-level, the maximum amplitude is the highest recorded volume for the word itself. This measure gives insight into how loud each word is uttered, and therefore the speech pattern and volume contour for an emotion. The maximum amplitude for each word is calculated using the Python library NumPy's `max()` function. For each figure, male and female speech is compared against each other within the confines of each emotion. Recall that each emotional speech instance is normalized against itself, independent of gender. As such, there may be multiple instances of a recording of 1.0 (the highest possible).

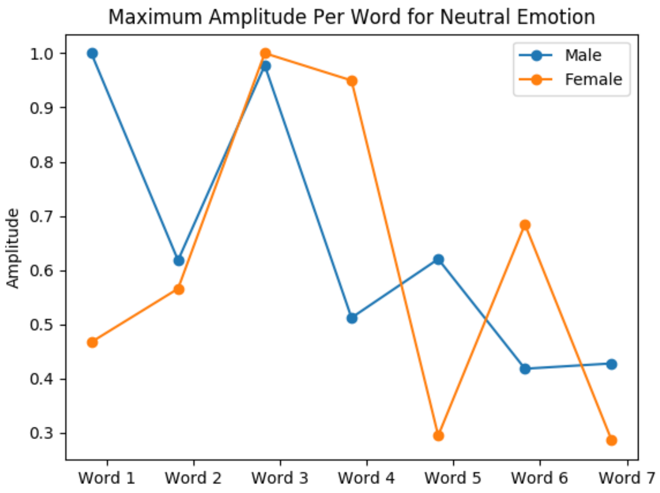

Figure 2.9: Maximum amplitude for male and female neutral speech

For the "neutral" emotion, or the baseline, the results in Figure 2.9 are split between the genders. For words 1, 2, 5, and 7, the male case has higher recordings of amplitude. Consequently, words 3, 4, and 6 have higher recordings of amplitude for the female case. This may signify that within a neutral tone, males tend to start their speech with a higher volume and then drop out later on, while females may start softer and get higher in volume later in their speech patterns. Within the "neutral" emotion, it appears that male and female speech have equal distribution in volume. Readings as high as 1.0 and as low as 0.3-0.4 can be seen for both male and female speech. This is especially compelling due to the neutral speech's ideal as a baseline in emotion recognition. The main takeaway from this experiment can be summarized as follows.

Remark on Neutral speech for the Max feature: Male and female vary in volume throughout the sentence (5/7 words).

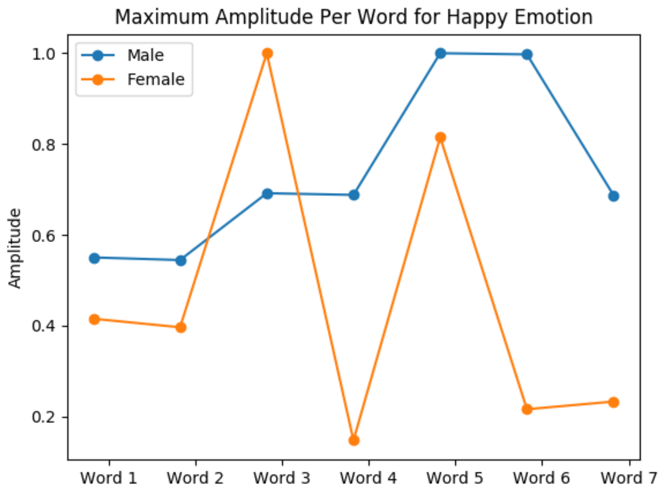

Figure 2.10: Maximum amplitude for male and female happy speech

For the "happy" emotion shown in Figure 2.10, quantitative analysis shows that the male case has higher recordings of maximum amplitude. This is observed for words 1, 2, 4, 5, 6, and 7. Solely word 3 has a higher recording for female speech. The male speech has multiple recordings of 1.0, and its lowest recordings are at 0.5 for normalized amplitude. Female speech trends lower in amplitude, with most recording around 0.2-0.4 (however, there is a recording for 1.0). This would suggest that male speech is generally louder overall when using a happy tone, or when affected by happiness. Comparing to Figures 2.3 and 2.4, male speech was seen as recording *lower* in amplitude than female speech. However, this evidence from the maximum amplitude of the normalized speech shows that male speech is louder for each word, not merely for one or two peaks. Here we observe the benefit of normalization: if a person is louder than the rest, it would dominate an analysis of maximum amplitude such as this one. With normalization, we bring all subjects within the same range, and then look for similar trends across the sentence. The main takeaway from this experiment can be summarized as follows.

Remark on Happy speech for the Max feature: Male is louder than female throughout most of sentence, with varying intensity (5/7 words).

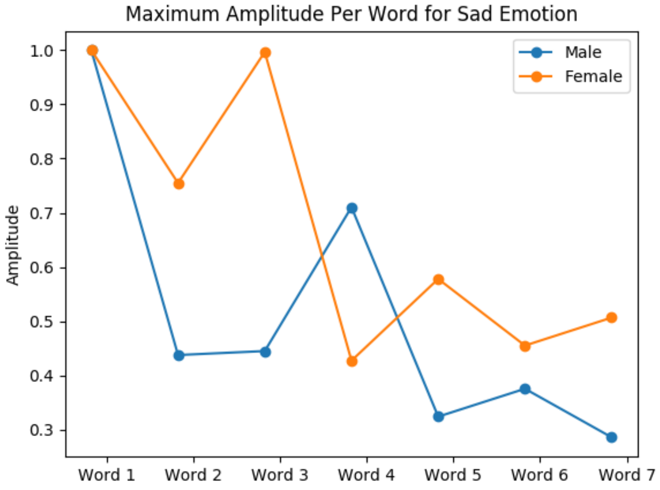

Figure 2.11: Maximum amplitude for male and female sad speech

For the "sad" emotion, illustrated in Figure 2.11, the results are almost completely flipped from that of the "happy" emotion. Quantitative analysis here shows that the female case has higher recordings of maximum amplitude. In words 2, 3, 5, 6, and 7, the female case has higher recordings. The male case has a higher recording only for word 4 and for word 1, both cases being tied at the maximum value of 1.0. Conversely from Figure 2.10, the male and female speech record with roughly the same amplitude between 0.3-0.7, a fairly wide range. Additionally, as referenced by Figures 2.5 and 2.6, the male and female cases registered at roughly the same maximum amplitude before normalization. This evidence suggests that sad female speech is generally louder, and especially so throughout the entire speech utterance. The main takeaway from this experiment can be summarized as follows.

Remark on Sad speech for the Max feature: Female is louder than male throughout most of sentence (5/7 words).

The experiments above show that maximum amplitude as a feature can show the highest registers of volume for each word, and therefore display which gender case is louder within its emotional speech. Through making direct comparisons with the speech's waveform before normalization was performed, maximum amplitude analysis gives a new perspective.

2.4.2 Average Amplitude Feature

The average amplitude for a wave form is taken from averaging each individual amplitude recording over the entire length of the wave form. At the word-level, the average amplitude feature is calculated solely from the amplitude readings within a single constructed frame. Whereas maximum amplitude showed only the highest volume readings, average amplitude gives a clearer picture for the general topographical volume of each word, and therefore each emotional speech instance. The average amplitude for each word is calculated using the Python library NumPy's `mean()` function.

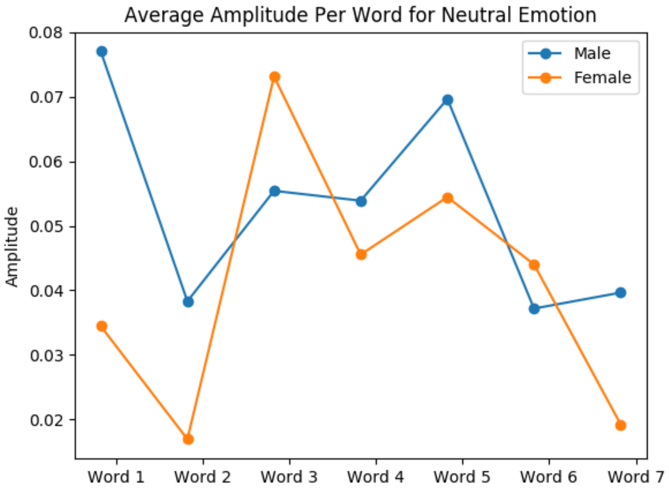

Figure 2.12: Average amplitude for male and female neutral speech

For the "neutral" emotion in Figure 2.12, or baseline, the results are not as even as in Figure 2.9. The average amplitude is higher for the male case in words 1, 2, 4, 5, and 7, whereas the average amplitude is higher for female speech in words 3 and 6. Additionally, by viewing the actual amplitude readings themselves, it can be seen that average amplitude is much lower than maximum amplitude. This should be expected due to the nature of a *maximum* amplitude, but this is more so as a result of amplitude variance within the words themselves. So then how can we explain these readings in more detail? Going back to Figures 2.1 and 2.2, we can see that the peaks in the male wave form are thicker than those in the female wave form. This will be explored further when summed amplitude is discussed. The main takeaway from this experiment can be summarized as follows.

Remark on Neutral speech for the Average feature: Male is louder than female throughout most of sentence, with varying intensity (5/7 words).

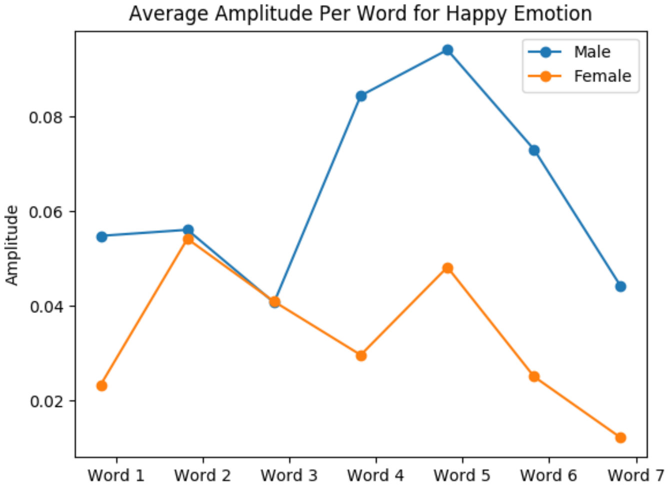

Figure 2.13: Average amplitude for male and female happy speech

For the "happy" emotion, higher average amplitude is registered across the board for the male case (see Figure 2.13). One exception is that for word 3, they registered an equal average amplitude of 0.04. This suggests that happy male speech is generally louder for a longer amount of time for each word. Looking at Figures 2.3 and 2.4, this posit is corroborated due to the thickness of each of the peaks. Happy male speech is clearly louder for the overall duration of the word itself, as opposed to happy female speech which registers extremely high amplitude spikes, but not for very long. Additionally, looking at words 4, 5, 6, and 7 we can see that the happy male speech is louder and with a longer duration towards the latter half of the utterance. This is reflected in Figure 2.10 which shows similar results for the male case. It is interesting to see how low the readings are for happy female speech, especially so when the maximum amplitude readings in the same figure show louder speech for the female case. The main takeaway from this experiment can be summarized as follows.

Remark on Happy speech for the Average feature: Male is louder than females at the beginning and end of the sentence (5/7 words).

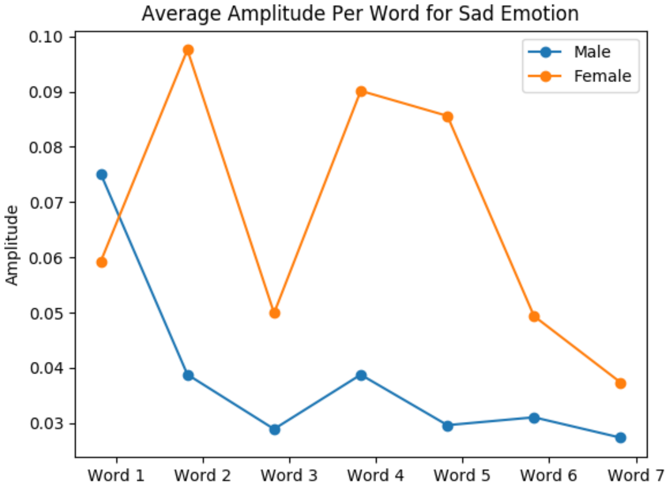

Figure 2.14: Average amplitude for male and female sad speech

For the "sad" emotion (Figure 2.14), higher average amplitude is registered almost universally by the female case. For words 2-7, the average amplitude records higher for female speech. Only word 1 has a higher recording for male speech. Even more drastic are the amplitude readings themselves, which show the female case displacing the male case by upwards of 0.05 (0.09-0.04 for word 2). While this is almost a reflection of Figure 2.11 which shows higher maximum amplitude for female speech, it is clear that sad female speech is louder for longer periods of time as well as overall than sad male speech. What is even more curious is that when looking back to Figures 2.5 and 2.6, the female wave form does not have *much* thicker peaks than the male wave form, except for word 7 which has a lower reading for average amplitude than other words. The main takeaway from this experiment can be summarized as follows.

Remark on Sad speech for the Average feature: Female is louder than male throughout most of sentence (6/7 words).

In this subsection, it was shown that average amplitude can reveal the sustained volume for a word. Specifically in regards to the wave form, it shows the relationship between the thickness of peaks and the word's volume. Average amplitude analysis expands upon maximum amplitude analysis, showing more of the volume contour of the utterance.

2.4.3 Summed Amplitude Feature

The summed amplitude is calculated by summing each individual amplitude recording. At the word-level, the summed amplitude is taken from the amplitude readings across each individual word frame. The information derived from the summed amplitude is a mixture of what is gotten from the maximum and average amplitudes. The overall *loudness* of the word itself is shown in the summed amplitude, as well as the loudness across the duration of the word. As such, this score displays the sustained volume even more so than before. The summed amplitude for each word is calculated using the Python library NumPy's `sum()` function.

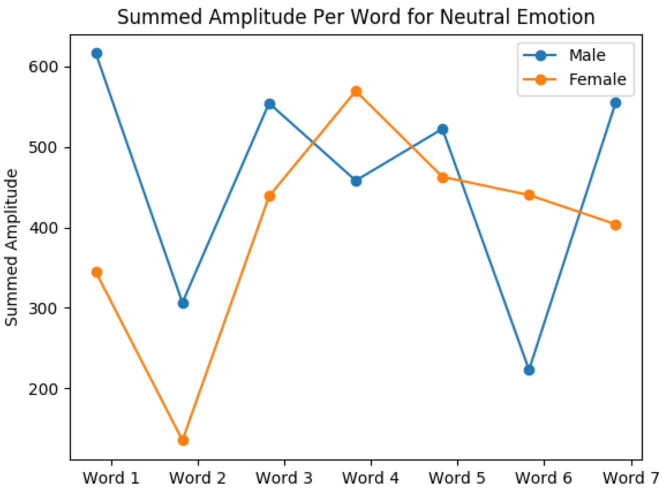

Figure 2.15: Summed amplitude for male and female neutral speech

For the "neutral" emotion, or baseline, the summed amplitude in Figure 2.15 shows very similar results to the average amplitude from Figure 2.12. Here, the summed amplitude trends higher for male speech, as shown in words 1, 2, 3, 5, and 7. Only in words 4 and 6 is the summed amplitude recorded higher for female speech. What is different between these two analyses is the magnitude of the readings. Summed amplitude will be higher than average amplitude due to the nature of calculating an average, but still in this result the summed amplitude for each word is roughly within the same margin. In Figure 2.12, the average amplitude showed higher readings at the beginning of the speech utterance. While the male case did register higher for summed amplitude, these results are compelling due to the lower margin of difference across the word frames. This harkens to the ideal of neutral speech as a baseline model for emotional speech. The main takeaway from this experiment can be summarized as follows.

Remark on Neutral speech for the Sum feature: Male is louder than female throughout most of sentence (5/7 words).

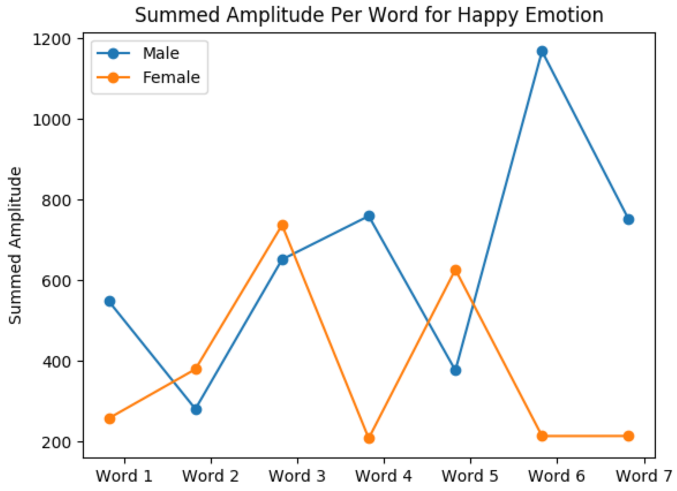

Figure 2.16: Summed amplitude for male and female happy speech

For the "happy" emotion shown in Figure 2.16, the summed amplitude is fairly split equally between the male and female cases, though the male case is higher in summed amplitude for one more word. Specifically, the summed amplitude is higher for male speech in words 1, 4, 6, and 7, whereas it is higher for female speech in words 2, 3, and 5. What is immediately noticeable is the displacement between the male and female readings when the male case has a higher summed amplitude. The most dramatic instance is in word 6, wherein the male case has a positive displacement reading of 1000 (1200-200). This suggests that happy male speech is louder overall, and its volume is higher for sustained periods of time word to word. In the case of words 6 and 7, the male case tends to have a higher, sustained volume towards the latter half of its speech utterance. The main takeaway from this experiment can be summarized as follows.

Remark on Happy speech for the Sum feature: Male is louder than female towards the end of the sentence, but is similar in volume throuhgout (3/7 words).

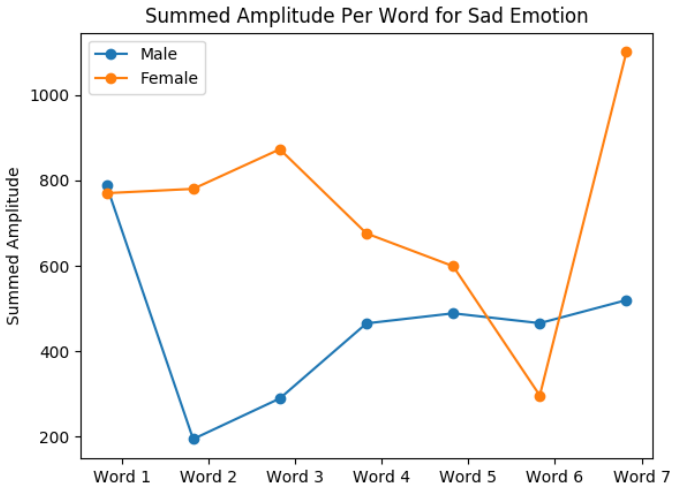

Figure 2.17: Summed amplitude for male and female sad speech

For the "sad" emotion, the summed amplitude is once again higher overall for female speech, as shown in Figure 2.17. It is almost the same in terms of higher-volume-gender-breakdown per word as in Figure 2.14, except for word 6. In fact, there are higher summed amplitude readings for female speech in words 2, 3, 4, 5, and 7. Words 1 and 6 have higher summed amplitude readings for male speech. However, in this case, there is less displacement between the male and female readings than in Figure 2.14. This is perhaps due to the maximum amplitude values having overall less displacement accordingly, as seen in Figure 2.11. Topographically, this suggests that sad male and female speech are relatively similar in the overall volume relative to the overall sustained volume for each word. However, the summed amplitude analysis for sad speech shows that sad *female* speech has higher overall volume and higher overall sustained volume than sad male speech. The main takeaway from this experiment can be summarized as follows.

Remark on Sad speech for the Sum feature: Female is louder than male throughout most of sentence (5/7 words).

The summed amplitude feature suggests that there is valuable information not only in the height of the peaks within the wave form, but in the length and thickness of the peaks as well. The former shows the intensity of the volume for each word, and the latter displays the magnitude of the overall sustained volume of words in emotional speech utterances. Quantitative analysis builds off of the information derived from the feature extraction performed on the original speech signal in the qualitative analysis phase, but provides discrete evidence to build these conclusions. Here, using three different features (max, average, and summed amplitude) we characterized three emotions in a male and a female. But these initial results are not independent of each other. Their information can be referenced against each other to inform a cumulative hypothesis. This is illustrated in Table 2.1 which summarizes the six remarks about our analyses.

	Max Feature	Average Feature	Summed Amplitude Feature
Happy	Male louder, varying intensity (5/7 words)	Male louder at beginning/end (5/7 words)	Male louder at end (3/7 words)
Neutral	Male/Female similar, varying intesnity (5/7 words)	Male louder, varying intensity (5/7 words)	Male louder (5/7 words)
Sad	Female louder (5/7 words)	Female louder (6/7 words)	Female louder (5/7 words)

Table 2.1: Remarks on the maximum, average, and summed amplitude features

The information in 2.1 shows consistency across all three features. One hypothesis generated by all three features is that males are louder when happy while females are louder when sad.

CHAPTER 3

Classification Using the Summed Amplitude Feature

Problems within the machine learning domain can be divided into two categories: classification and regression. Classification problems serve to identify or group data into various pre-specified classes, or tag the data using custom labels. A classifier is a machine learning model that can predict the categorical class labels for new data. The analysis performed by classification is important in providing a better understanding of the data at large due to the organizational techniques deployed [6]. Instead of calculating a pre-supplied label for the data, regression problems instead produce a real or continuous value from its analysis. These values are a direct prediction formed from learning the relationship between features within the data. Regression can be a simple process in the case of linear regression, where a line of best fit is calculated. Multivariate regression, however, becomes much more complex due to the consideration of multiple variables or features from the data. It is not a question of which machine learning method is *better*, but rather to decide which method is best for the problem *at hand*.

For the problem presented in this study, classification is the better method. There are multiple categories that the data can be classified into. When considering emotion, the data can be grouped into different classes for *neutral*, *happy*, or *sad* for example. This is the original problem presented: can we derive emotion purely from audio waveforms? Through the processes in Chapter 2, however, another group of labels presents itself for classification: gender. The RAVDESS dataset provides both emotion and gender labels for each datapoint (audio recording). This offers up a new problem: are emotions expressed differently between males and females? Can we derive the gender expression of the emotion purely from the audio waveform? This is surely not a regression problem, because we use discrete features to characterize the original data. The original data itself (amplitude from the waveform) is continuous, but is transformed into discrete features before any analysis is performed. With the knowledge that this problem is one of classification, the best machine learning model can then be chosen.

3.1 Clustering Algorithms for Classification

Basic classification involves using pre-supplied class labels to form predictions for data. The model is trained using these class labels to predict *new* data, not to describe the original dataset. The data in the RAVDESS dataset is provided with class labels for emotion and gender, which does allow for the prediction and classification of new audio samples. However, the purpose of this study is to analyze the audio waveforms *for* classification. The process of deriving emotion and gender from the data samples presupposes that the class labels are unknown. They need to be discovered. This is where clustering comes in.

Clustering is a form of unsupervised learning, as it does not use any external information such as class labels [10]. Clustering is the process of grouping sets of data into multiple groupings or clusters, where data within a single cluster has a high level of similarity, and is not similar to data in other clusters. Simply, it is the act of partitioning a set of data into smaller subsets [6]. Therefore, data within a cluster is distinct in the similarity of one or more attributes, and so a virtual class is formed for the cluster and its members. As such, we will be performing clustering to form a better understanding of the audio data itself, in an attempt to analyze and describe the latent information in the clustered groups [10].

3.1.1 K-Means Clustering

K-means clustering is one of the most widely used clustering algorithms [10]. It is also one of the oldest clustering algorithms at over 50 years old. The "*k*" in *k*-means is an integer value specified by the user which will dictate how many clusters to form during the algorithmic process. These *k* clusters are additionally non-overlapping within the data [10]. Clusters each contain a centroid, which is calculated as the mean value of the points residing in the cluster. The clustering process initially selects *k* centroids at random (or as previously specified). Then, all of the data points are assigned to the cluster centroid closest to them, forming a cluster of points. This is calculated using a distance function.

For this study, the Scikit-learn library's *k*-means function from the Scikit-learn clustering package is used. The distance function used in this *k*-means function is the Euclidean distance function, described here:

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

where the distance between the n-dimensional points p and q is calculated. Then, the centroid of each cluster is re-calculated based on the data points assigned to its cluster. This process is repeated until cluster membership stabilizes, and no points change clusters between rounds [10]. Furthermore, *k*-means tends to produce clusters of uniform size, regardless of the size of the "true clusters" [11].

K-means as a clustering algorithm has lasted so long in relevancy because of its robustness and effectiveness on numerous data types. *K*-means is also often used as a benchmark test for other methods of clustering and unsupervised classification [10]. This algorithm is not without its disadvantages, however. It is highly sensitive to outliers due to the continuous re-calculating of centroids, which would take into account the large distance in relation to outliers. Also, the algorithm will not always converge with "true clusters" in a global optimum. More often, it will converge with a local optimum for cluster membership [6]. As such, when selecting initial centroids at random, it is useful to run the algorithm multiple times, as different results may occur with local convergence. K-means clustering has a linear runtime relative to the data sample size and number of clusters: $O(nk)$ for n samples and k clusters [5].

3.1.2 Agglomerative Clustering

Agglomerative clustering is a type of hierarchical clustering algorithm. Hierarchical clustering algorithms work by organizing the data points into a hierarchy of clusters, such that the data is repeatedly clustered into smaller subgroups [6]. Agglomerative clustering works from a bottom-up approach. Each individual data point starts in its own cluster, then clusters are merged together iteratively to form larger clusters. Eventually, a hierarchy or tree of clusters can be seen (in a time-series format, where time steps are represented as the merge points). Typically this process iterates until there is one cluster formed, or alternatively a pre-specified termination point is reached.

For each merge point, two clusters that are closest together (using a pre-specified distance function) are merged. Due to this, agglomerative clustering algorithms require at most n iterations for n data points [6]. This results in clusters of high quality, or closer to "true cluster" status, but there is an exorbitant time cost. The runtime for agglomerative hierarchical clustering is at least $O(n^2)$ and at worst $O(n^2 \log n)$ [5]. However, more recent versions and adaptations to this clustering method are cognizant of that fault, and focus on efficiency while maintaining quality of clusters.

Additionally, different linkages may be used. Through a *single* linkage, the distance between clusters is calculated as the distance between the closest pair of data points. The *complete* linkage calculates based off of the distance between the farthest pair of data points. Finally, the *average* linkage calculates the average distance between a pair of data points [4].

3.2 Clustering Results

3.2.1 Clustering Summed Amplitude

Once a clustering algorithm has been selected (for this study, both *k*-means and agglomerative clustering is considered), then clustering can be performed on a set of features. Out of the various quantitative features examined in Chapter 2, it was determined that the summed amplitude feature was the most informationally rich due to it encapsulating the relationship between the volume and sustained length of the utterance. Figure 3.1 displays the summed amplitude per word, for each emotion sample in consideration. Each word's summed amplitude is plotted on the same vertical plane so that direct comparisons may be made between each sample.

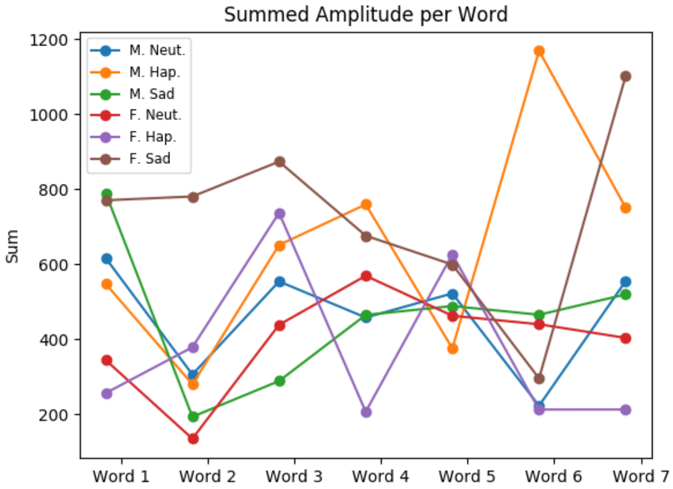

Figure 3.1: Summed amplitude per word for male and female neutral, happy, and sad speech

Trends can immediately be spotted in this plot, with words 2, 6 and 7 having considerable distribution, or variation of range, between samples. The *male happy* and *female sad* samples have the highest summed amplitude throughout the sentence, while the *female neutral* has the lowest summed amplitude throughout the sentence. It can be seen, however, that the *male neutral*, *male sad*, *female neutral*, and *female happy* samples are roughly similar in terms of their overall topology throughout the sentence. This will be explored further during the clustering process.

The goal of the clustering process is to either confirm qualitative observations of the data through the algorithmic processes, or to create new observations and theories for the dataset through the groupings that are constructed. A feature matrix of the summed amplitude per word for each of the six emotion samples (male neutral, male happy, male sad, female neutral, female happy, female sad) is then created. This results in a 6×7 matrix of features as input to the clustering algorithms. Each sample (row) is fed into the clustering algorithm one by one from the matrix. The two clustering algorithms are each run using the same feature matrix, with each algorithm forming 2, 3, and 4 clusters successively.

Number of Clusters	Method	MN	MH	MS	FN	FH	FS
2 Clusters	*k*-Means	1	1	1	1	1	0
	Agglomerative	1	0	1	1	1	0
3 Clusters	*k*-Means	0	1	0	0	0	2
	Agglomerative	0	0	0	0	2	1
4 Clusters	*k*-Means	1	3	1	1	0	2
	Agglomerative	1	3	1	1	0	2

Table 3.1: Results of *k*-Means and Agglomerative clustering using 2, 3, and 4 clusters for the summed amplitude of the male neutral (MN), male happy (MH), male sad (MS), female neutral (FN), female happy (FH), and female sad (FS) emotions

For each test, the cluster membership of an audio sample (the right side of the table) is displayed by a value between $[0,1]$ for two clusters, $[0,2]$ for three clusters, and $[0,3]$ for four clusters. Audio samples that share a value (for example, MH and FS for Agglomerative clustering for 2 clusters) thus share a cluster and are therefore similar. Now, each clustering method per number of clusters will be discussed in turn for: (a) two clusters; (b) three clusters; and (c) four clusters.

(a) For two clusters, k-Means clustering found the *female sad* sample to be in one cluster, and the rest of the samples in another cluster. This follows from Figure 3.1, as the *female sad* sample lies mostly above the others for total summed amplitude per word. This suggests that sad female speech is distinguishable from the other forms of emotional speech based on its magnitude of volume combined with the length of each word utterance. Agglomerative clustering grouped together the *female sad* sample with the *male happy* sample, and grouped the remaining samples in their own cluster. Looking back to Figure 3.1, we can see that these two samples have consiberable distribution towards the end of the sentence, with the *male happy* sample having a large summed amplitude on word 6, and the *female sad* sample having a large summed amplitude on word 7. This grouping suggests that happy male speech and sad female speech is similar in volume and length of utterance, and that these two emotions can be recognized reciprocally.

(b) For three clusters, k-Means clustering found that the *male happy* and *female sad* samples to be in separate clusters of their own, while the rest of the samples were in the remaining cluster. This is an accurate representation of the plot in Figure 3.1, because while the *male happy* and *female sad* samples have higher distribution and a larger summed amplitude towards the end of the sentence, they have a high distribution between themselves. Note how the *female sad* sample starts with a higher summed amplitude at the beginning of the sentence, and then dips on word 6, while the *male happy* sample does the opposite, starting with a lower summed amplitude and peaking on word 6. This information is lost when clustering with two clusters, as these two samples are similar only at the end of the sentence. This clustering suggests that these two samples are unique from the rest, and that they may be easily distinguishable, based on which part of the sentence is analyzed. Agglomerative clustering found, conversely, that the *female happy* sample was in its own grouping instead of the *male happy* sample. This is curious, because this sample does not have very high distribution from the rest of the samples that are grouped together (*male neutral*, *male happy*, *male sad*, *female neutral*), except on word 4, where it is significantly lower in summed amplitude than the rest of the samples.

(c) For four clusters, *k*-Means clustering found that the *male happy*, *female happy*, and *female sad* samples were in their own clusters, and the *male neutral*, *male sad*, and *female neutral* samples are in one single cluster. This result is a combination of the results for three clusters wherein *k*-Means found *male happy* and *female sad* to be in their own clusters, and Agglomerative found *female happy* and *female sad* to be in their own clusters. This results from the *male happy* sample having high distribution on word 6, the *female happy* sample having high distribution on word 4, and the *female sad* sample having high distribution on word 7. This suggests that happy male and sad female speech can be distinguished by a high magnitude of volume at the end of the full utterance, and that happy female speech can be distinguished by a drop in volume in the middle of the utterance. Agglomerative clustering found the same cluster groupings.

While *k*-Means clustering worked quite well for grouping the samples and making distinctions between samples for both three and four clusters, the *female happy* sample does not have nearly the same level of distribution overall as the *male happy* and *female sad* samples. Thus, *k*-Means clustering with three clusters has the best and most significant findings for this limited study of one male and one female.

3.2.2 Clustering New Summed Amplitude

Something worth noting is the summed amplitude's dependence on length as a feature. Looking back at Figures 2.3 and 2.6 (the audio signals of the happy male and sad female speech), we can see that at the end of the sentence, where summed amplitude was higher for these samples, the word length is longer than in other samples. While length may be an important factor in classifying emotion at a machine level, human beings typically do not process length of utterances as easily. Thus, we can divide the summed amplitude by its length to eliminate this factor, and focus purely on the volume intensity of the utterance. This is performed by dividing the summed amplitude value obtained for each word frame by the length of the frame itself. The new summed amplitude values are plotted in Figure 3.2.

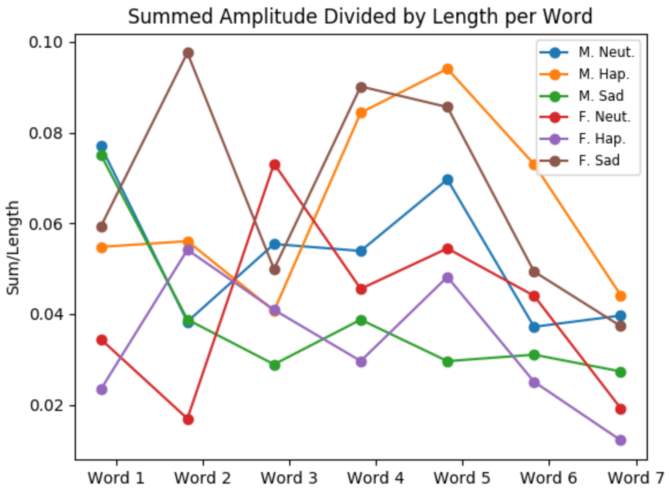

Figure 3.2: Summed amplitude divided by length per word for male and female neutral, happy, and sad speech

Comparing Figure 3.2 with Figure 3.1, we can see considerable differences. The overall topology of the plot has more distribution and a wider range of values than the original summed amplitude, which saw lower values in the beginning of the sentence and higher values towards the end of the sentence. To that point, the values for sum over length towards the end of the sentence are considerably lower with an equal distribution than the original plot of summed amplitude. This shows just how much length skews the original data. Due to this new plot having a large amount of uniform distribution overall, it is not very clear how the various samples will be grouped via clustering, and thus which samples are easily distinguished. However, it can be seen that the *male happy* and *female sad* samples once again lie higher on the topology of the plot, and thus they are higher in summed amplitude overall than the other samples.

Now, a new feature matrix of the summed amplitude divided by length per word for each of the six emotion samples is created. This results in a 6 × 7 matrix of features, akin to the original feature matrix, to serve as input to the clustering algorithms. The algorithms are each run attempting to form 2, 3, and 4 clusters successively.

Number of Clusters	Method	MN	MH	MS	FN	FH	FS
2 Clusters	*k*-Means	0	1	0	0	0	1
	Agglomerative	0	1	0	0	0	1
3 Clusters	*k*-Means	2	1	2	0	0	1
	Agglomerative	2	1	2	0	0	1
4 Clusters	*k*-Means	0	1	0	3	2	1
	Agglomerative	0	1	0	3	2	1

Table 3.2: Results of *k*-Means and Agglomerative clustering using 2, 3, and 4 clusters for the summed amplitude divided by length of the male neutral (MN), male happy (MH), male sad (MS), female neutral (FN), female happy (FH), and female sad (FS) emotions

Similar to Table 3.1, cluster membership is displayed by a value between [0,1] for two clusters, [0,2] for three clusters, and [0,3] for four clusters. Speech samples that share a value thus share a cluster and are therefore similar. Now, each clustering method per number of clusters will be discussed in turn for: (a) two clusters; (b) three clusters; and (c) four clusters.

(a) For two clusters, *k*-Means and Agglomerative clustering had the same results. They both show the *male happy* and *female sad* samples in one cluster, with the rest of the samples in the other cluster. These results are identical to the original Agglomerative clustering result for two clusters. Looking at Figure 3.2, we can see how the *male happy* and *female sad* samples may be distinguishable from the rest, as they have a higher overall summed amplitude for words 4-7, and thus the majority of the sentence. This suggests that these two samples are distinguishable from the other samples purely from their higher volume for the latter half of the sentence.

(b) For three clusters, *k*-Means and Agglomerative clustering had the same results. They both found three clusters with two samples each: the *male happy* and *female sad* samples in one cluster (similar to the first round with two clusters), the *male neutral* and *male sad* samples in another cluster, and the *female neutral* and *female happy* samples in the final cluster. Looking at Figure 3.2, we see that the *male neutral* and *male sad* samples have identical summed amplitudes for words 1 and 2, and are fairly close at words 4, 6 and 7. This could explain their resulting grouping. The *female neutral* and *female happy* samples are similarly close in summed amplitude (though not identical) for words 3, 5, and 7. These clustering results are motivational in a possible classification for gender among the emotional speech samples. Disregarding the *male happy* and *female sad* samples, which have a higher overall summed amplitude and are thus easily distinguishable, the rest of the samples are clustered by gender. While the distribution and placement of the summed amplitude of the values on the plot do not allow for a generalized characterization of volume for the two gendered clusters of samples, it does suggest an overall similarity in volume for (neutral, sad) male and (neutral, happy) female speech.

(c) For four clusters, *k*-Means and Agglomerative clustering had the same results. They also showed similar results to the clustering for three clusters, except the *female neutral* and *female happy* samples are split into their own separate clusters, while the *male neutral* and *male sad* samples remain in the same cluster. This follows from Figure 3.2, as the *male neutral* and *male sad* samples are highly similar at the start and end of the sentence, whereas the *female neutral* and *female happy* samples are not as similar, and thus are distinguishable from each other. However, this does not result in these samples being distinguishable from the other samples. This merely creates groupings without a further enrichment of the data.

From these results, a few hypotheses can be made:

(1) As seen in both Tables 3.1 and 3.2, the *male happy* and *female sad* samples were the most distinguishable from the rest of the samples, and according to Figure 3.2 are the strongest emotions in volume, making them the most intense overall. This suggests that males are very intense in volume when they are happy, and female are very intense in volume when they are sad.

(2) Males are also seen to be less distinguishable in the other emotion samples (*male neutral*, *male sad*) and thus are neutrally intense in volume.

(3) Finally, females convey a wider range of intensity of volume for other emotion samples (*female neutral*, *female happy*), as seen in the varying results of clustering for these samples.

Notably, these clustering results are in agreement with the remarks in Table 2.1 from Chapter 2, which were derived by simply visualizing the data.

CHAPTER 4

Further Analysis for Emotional Speech Modeling

Thus far, baseline metrics for various features such as maximum, average, and summed amplitude have been extracted from two whole data samples (a single sample meaning one gender, multiple emotion samples), and classification has been performed on these samples as well. While this is highly informative, it is not necessarily representative for the dataset as a whole, which has many more of these samples, or representative of emotion and gender at large. To build a more accurate environment for characterizing the data, more samples need to be analyzed and compared to the previous work. This collection of samples and collective data analysis lays out the groundwork for a model that standardizes the relationships between the data's properties, and helps to further characterize future data against the model.

Now, two more male and two more female samples are brought in from the dataset. To create an actual data model, many more samples may need to be considered to paint an accurate picture of the dataset and standardize the relationships. However, due to the manual process of frame construction (see: Figure 2.8), only four more samples (two male, two female) are considered and therefore analyzed. In the case of an automated implementation of frame construction, considering a large scale of samples is not an issue. These four samples follow the same process as before: they must be read from the .wav file and converted into a usable data form, then the data is normalized, then frames are constructed manually, and finally features are extracted from the samples' frames.

4.1 Qualitative Analysis of Happy and Sad Emotions Using the Summed Amplitude Feature

In section 2.2 the summed amplitude feature revealed more information within the process of clustering, specifically regarding the *male happy* and *female sad* samples. They were shown to be the most intense emotions (when considering volume) and most distinguishable from the other samples. Therefore, in considering new samples these two emotions can be compared sample to sample within gendered constraints to observe the potential similarity/standardization of the model. Specifically, the *male happy* and *female sad* emotion samples from three different male samples and three different female samples are compared against each other. They are plotted together in Figure 4.1.

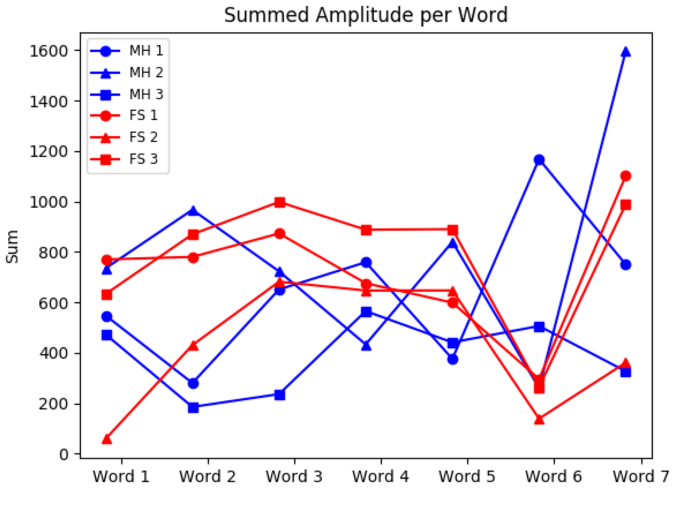

Figure 4.1: Summed amplitude per word for three samples of male happy speech and three samples of sad female speech

In this graph, the three *male happy* samples are plotted in blue while the three *female sad* samples are plotted in red, with separate markers identifying each distinct sample. The (red) *female sad* samples here are fairly similar in their general topography and value of summed amplitude, especially for words 3-7. However, the second *female sad* sample is not as similar for words 1-3. This is motivational in establishing a standard characterization of sad female speech across the dataset. The (blue) *male happy* samples on the other hand are not as uniform. The first and third *male happy* samples are fairly similar for words 1-4, but otherwise the three samples are not very similar, with a high range variation of summed amplitude across the sentence utterance.

Also worth exploring is the variation of the summed amplitude feature from Section 3.2.2, wherein the summed amplitude per word is divided by the length of the word frame. This was useful in further characterizing the two most intense emotional samples, *male happy* and *female sad* and also eliminated the length of the word as a factor that could possibly skew the data. This exemplifies the pure volume or intensity of each audio sample. In Figure 4.2 this variation of the summed amplitude feature is plotted for the three distinct *male happy* and *female sad* samples.

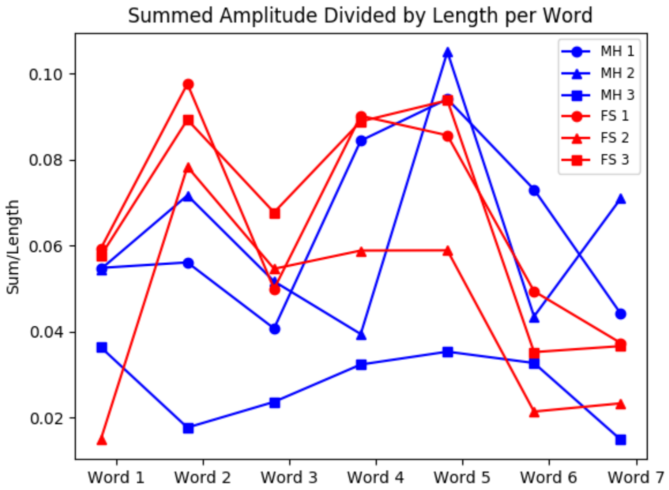

Figure 4.2: Summed amplitude divided by length per word for three samples of male happy speech and three samples of sad female speech

In this plot, the similarities and dissimilarities between the gendered samples become more clear. For the three (red) female samples, the overall topography of the summed amplitude divided by length becomes fairly consistent, with a small amount of range variation between the second *female sad* sample and the other two. Despite this distribution, the samples each follow the same pattern of crests and troughs (increasing and decreasing zones in the graph). This shows a very high level of similarity for all three *female sad* samples, and is compelling in establishing a data model for characterizing sad female speech.

The three *male happy* samples on the other hand had differing, less-motivational results from Figure 4.1. In Figure 4.2 the first and second *male happy* samples are more similar in topography and summed amplitude divided by length value than the third sample. However, while they are closer in value, the crests and troughs of the plot do not match up in the same manner as the female samples, therefore showing the dissimilarity prevalent between them. The third *male happy* sample is not similar at all in topography or summed amplitude divided by length value. Previously in Figure 4.1 the second *male happy* sample was the least similar within the male samples, while in Figure 4.2 the third sample is least similar. This high level of variation across the two summed amplitude features is not compelling for establishing a data model for happy male speech.

4.2 Quantitative Analysis of Neutral, Happy, and Sad Emotions Using k-Means Clustering

With the two new gendered samples (each with neutral, happy, and sad samples), clustering is revisited. Considering each of the six emotion samples from all six gendered samples (three males and three females), this increases the dimensions of the feature matrix to be [18 samples × 7 features] for the summed amplitude per word feature. Since *k*-means led to more compelling results in previous iterations, it is solely used as a clustering algorithm in this particular analysis. It is then run with the feature matrix forming 2,3,4,5, and 6 clusters successively.

Clusters	MN1	MN2	MN3	MH1	MH2	MH3	MS1	MS2	MS3	FN1	FN2	FN3	FH1	FH2	FH3	FS1	FS2	FS3
2	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	0
3	0	0	0	1	2	1	1	0	0	1	1	0	0	1	0	2	0	2
4	0	0	2	3	1	2	2	2	0	2	2	0	0	2	0	1	0	1
5	3	3	1	4	2	1	1	0	3	1	0	3	3	0	3	2	3	2
6	2	4	2	5	1	2	2	3	3	2	3	4	0	3	4	1	0	1

Table 4.1: Results of *k*-means clustering using 2, 3, 4, 5, and 6 clusters for the summed amplitude of the male neutral (MN), male happy (MH), male sad (MS), female neutral (FN), female happy (FH), and female sad (FS) emotions for three male and three female samples

For each test, the cluster membership of the audio sample is displayed by a value between $[0,1]$ for two clusters, $[0,2]$ for three clusters, $[0,3]$ for four clusters, $[0,4]$ for five clusters, and $[0,5]$ for six clusters. Audio samples that share a value are members of the same cluster, and are therefore similar in regard to the feature's qualities. Now, the results for each incrementing cluster number will be discussed in turn for: (a) two clusters; (b) three clusters; (c) four clusters; (d) five clusters; and (e) six clusters. In consideration to the magnitude of the samples being clustered, only significant results are discussed.

(a) For two clusters, the second *male happy* sample, and the first and third *female sad* samples are clustered together, with the rest of the samples grouped in the other cluster. The *male happy* and *female sad* samples were clustered together in previous iterations, and therefore significantly distinguishable from other samples, so this clustering is motivational in maintaining this trend. Worth noting is that instead of the first *male happy* sample, which was clustered previously with the first *female sad* sample, the second *male happy* sample is within this grouping. This clustering once again shows the similarities between happy male and sad female speech, particularly its intensity of volume in relation to its length.

(b) For three clusters, the largest cluster representation is the first cluster (marked with value 0). This includes all three *male neutral* samples, as well as the second and third *male sad* samples. This would suggest that sad male speech is quite similar to neutral male speech, the baseline, and thus is not very distinguishable from other male samples. Looking at the female samples at large, it is worth noting that each emotion variant almost entirely shares a cluster. The first and third *female happy* samples share one, the first and second *female neutral* share one, and the first and third *female sad* share one. This suggests that the female emotion samples are more easily distinguishable from each other via volume relative to length. While these samples do share clusters with male samples, it is still motivational in establishing a model for classifying female emotions from the utterance.

(c) For four clusters, the results for the female samples persist. The first and third *female happy* samples share a cluster, the first and second *female neutral* share another, and the first and third *female sad* also share another. That these results persist over different clustering iterations is highly motivational in establishing a model for female emotion samples. This is not the case for the male samples. The first and second *male neutral* samples are clustered together, but they also share a closter with five other samples, four of which are female. Also, the first and second *male sad* samples are clustered together, but they share too share a cluster with five other samples, three of which are female. When considering only the first and second male samples we can establish a distinguishable cluster representation that could classify the emotion of the (male) utterances, but the third male samples breaks this down.

(d) For five clusters, our previous results begin to break down. As more clusters are formed, the separation between samples is given more weight in determining groupings. However, this may present the most accurate results for possible classification. Within the female samples, only the first and third *female happy* samples share a cluster and the first and third *female sad* samples share a luster. The female neutral samples all fall in different clusters. This suggests that really the happy and sad female speech is truly distinguishable and classifiable, whereas the neutral female speech is not as easily distinguishable. For the male samples, only the first and second *male neutral* samples share a cluster. All other male samples (across emotions) are in separate clusters. This could help in establishing a *lack* of emotion in the utterance, considering the neutral emotion is most distinguishable.

(e) For six clusters, the previous results for the female samples break down even further. Only the first and third *female sad* samples share a cluster. This follows from our previous plotting and clustering results, which showed that sad female speech is the most intense in volume of the female emotional utterances. With six clusters having been formed, these results are highly motivational in establishing a possible model for sad female speech. Additionally, the second and third *male sad* samples share a cluster. While they also share a cluster with two female samples, they do not share one with any other male samples. This suggests that sad male speech may be distinguishable from other male emotional utterances. Previously the *male sad* samples were clustered with the *male neutral* samples. The formation of more clusters means a finer grained analysis for distinguishing between the male emotional utterances.

This round of clustering gave more insight into the possible classification of gendered emotional speech, particularly in regard to female emotional speech samples. On average, k-means clustering established one cluster for each female emotional constraint. Additionally, the highest volume of clusters ($k = 6$) found the sad female speech to be most distinguishable.

Now, a new feature matrix of the summed amplitude divided by length per word for each of the 18 emotion samples is created. In Section 2.2.2, this version of the summed amplitude feature provided similar, but more compelling results. Similar to the previous feature matrix used, it is also dimensionally [18 samples × 7 features]. Only the k-means algorithm is run, forming 2, 3, 4, 5, and 6 clusters successively.

Clusters	MN1	MN2	MN3	MH1	MH2	MH3	MS1	MS2	MS3	FN1	FN2	FN3	FH1	FH2	FH3	FS1	FS2	FS3
2	0	0	0	1	1	0	0	0	1	0	1	0	0	1	0	1	1	1
3	2	2	0	1	1	0	0	0	1	2	1	2	0	1	2	1	1	1
4	0	0	2	3	1	2	2	2	1	0	1	0	2	1	0	3	1	3
5	2	2	1	4	3	1	1	1	0	2	0	2	1	0	2	4	0	4
6	1	1	3	4	5	3	3	3	4	1	0	1	3	0	1	2	0	2

Table 4.2: Results of k-means clustering using 2, 3, 4, 5, and 6 clusters for the summed amplitude divided by length of the male neutral (MN), male happy (MH), male sad (MS), female neutral (FN), female happy (FH), and female sad (FS) emotions for three male and three female samples

Similar to Table 4.1, cluster membership of the audio sample is displayed by a value between $[0, 1]$ for two clusters, $[0, 2]$ for three clusters, $[0, 3]$ for four clusters, $[0, 4]$ for five clusters, and $[0, 5]$ for six clusters. Speech samples that share a value are members of the same cluster, and are therefore similar. Now, each incrementing cluster number is discussed in turn for: (a) two clusters; (b) three clusters; (c) four clusters; (d) five clusters; and (e) six clusters.

(a) For two clusters, more samples are grouped in the smaller cluster than in the previous round of clustering. All three *female sad* samples and the first and second *male happy* samples are grouped in this cluster. This follows from previous analysis where happy male and sad female speech were found to be the most intense in purely volume. This serves to illustrate that claim. Worth noting is that the second female sample is clustered all together across the emotional constraint. This suggests that this sample is not distinguishable in terms of emotion.

(b) For three clusters, samples across the gender constraint share cluster groupings. When looking at just the male samples, the first and second male samples each are clustered together for each emotion. The first and second *male happy* samples are clustered together, the first and second *male neutral* samples are clustered together, and the first and second *male sad* samples are clustered together. This suggests that emotion can easily be classified from male speech from the summed amplitude divided by length. This result is motivational in creating a model for emotional male speech.

(c) For four clusters, the previous results already start to break down. The happy emotion overall (for both genders) has the highest level of diversity, with its samples being grouped into all four clusters. This suggests that male and female speech cannot be easily distinguishable for happy emotional utterances. For the male samples, now only the first and second *male neutral* samples and the first and second *male sad* samples are in their own clusters. Thus, these emotion samples are uniformly distinguishable for male speech. Similarly, the first and third *female neutral* samples and the first and third *female sad* samples are in their own clusters. Similar logic follows for the distinguishability of these samples. Also worth noting is that the first and second *male neutral* samples and the first and third *female neutral* samples are clustered together, suggesting that neutral speech overall between both genders is similar. This is motivational in establishing a baseline metric for modeled emotional speech.

(d) For five clusters, similar results follow. The neutral and sad male samples are clustered similarly as the clustering round with four clusters, and the neutral and sad female samples as well. The second female sample is clustered all together across the emotion constraint, showing a lack of distribution for that sample. Finally, just the first *male happy* sample is clustered together with the first and third *female sad* sample, suggesting that the similar relationship of intensity of volume is not as strong in distinguishing these emotion samples as previously indicated.

(e) For six clusters, some previous findings persist. The neutral samples for both genders are mostly clustered together, maintaining the assertion that truly neutral speech is similar regardless of gender. The first and third *male neutral* and *male sad* samples are clustered separately by emotion, maintaining the assertion that these emotions (or lack thereof, for the neutral emotion) are easily distinguishable for male speech. The highest variation occurs for the happy emotion across both genders, in which five out of the six clusters are represented. This suggests that happy speech as a whole has a wide range of intensity of volume

The clustering results for the summed amplitude divided by length feature were somewhat illuminating in establishing a similarity for neutral speech, and a dissimilarity for happy speech. Many of the successive clustering rounds grouped together either most of the neutral samples regardless of gender, or most of the neutral samples within each gender in individual clusters. This shows a high level of similarity for neutral speech, and establishes a baseline model for neutral speech as a whole, which itself serves as a baseline for emotional speech. Finally, the high level diversity of cluster membership for the happy samples (in which diversity increases with the number of clusters formed) shows great variation in the intensity of volume for happy speech, regardless of gender. This is significant, though not motivational in modeling such emotional speech.

This full second round of clustering took a closer, finer-grained approach to quantitative analysis. With a larger amount of clusters being formed ($k = 6$), criterion for cluster membership quantitatively gets more narrow due to the euclidean distance metric. Therefore, cluster groupings of multiple samples that are found, especially with a higher number of clusters, displays more discrete and nuanced similarities between samples. Some results from this more nuanced clustering process were motivational: the groupings of neutral samples from the summed amplitude divided by length feature show a dependence on pure intensity of volume for classifying baseline neutral speech, and the groupings of female samples for each emotional variant for the summed amplitude proper feature allow for the modeling of female emotional speech in respect to the utterance's volume relative to the utterance length.

A quantitative approach, specifically in regard to classification, was hypothesized to be effective in the classification of emotion from human speech and possibly between gender. However, through its exploration we have seen it to be not completely effective in discretely grouping separate emotional speech or gendered speech between each other. Thus, while some results are *motivational* for creating a model for gendered, emotional speech, they are *incomplete* in their findings and their ability to accurately cluster based off of purely gender and/or emotion. In the final chapter, qualitative analysis will conclusively establish beginning prototypical models for gendered, emotional speech.

CHAPTER 5

Building Prototypes for Emotional Speech

The results of the quantitative analysis in Chapter 4 were not immediately successful in obtaining desired emotion or gender groupings of samples, though they did provide relative information of the auditory relationship between different types of gendered, emotional speech.

With the inclusion of more samples, we revisit qualitative analysis to try to identify more visual relationships and trends between the samples for each emotional variant. This information is equally as valuable as the clustering performed previously in its ability to show the nature of how people convey emotion in their speech, in regard to intensity of volume. These trends can be encapsulated into a prototypical model for each emotional variant. The prototype can be compared against future speech data to try to distinguish the emotion conveyed within the utterance.

5.1 Neutral Speech Prototype

In an attempt to view how the speech is uttered for each emotional variant, regardless of gender, all six samples (three male, three female) are plotted. This is shown for the neutral emotion in Figure 5.1.

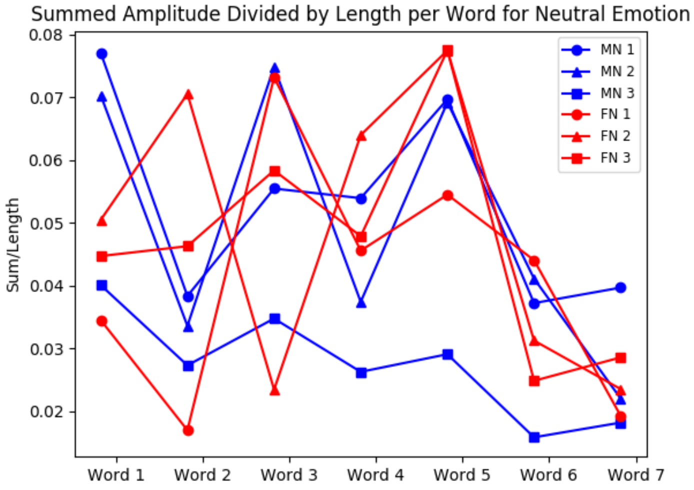

Figure 5.1: Summed amplitude divided by length per word between three male samples and three female samples for the neutral emotion

Some visual trends can be spotted among the topography of each sample, but this is still not completely clear. To alleviate the confusing nature of plotting three different samples for each gender, the summed amplitude divided by length is averaged between the three samples. This results in one plotted instance for each gender, and is more clear in establishing trends. This is shown for the neutral emotion in Figure 5.2.

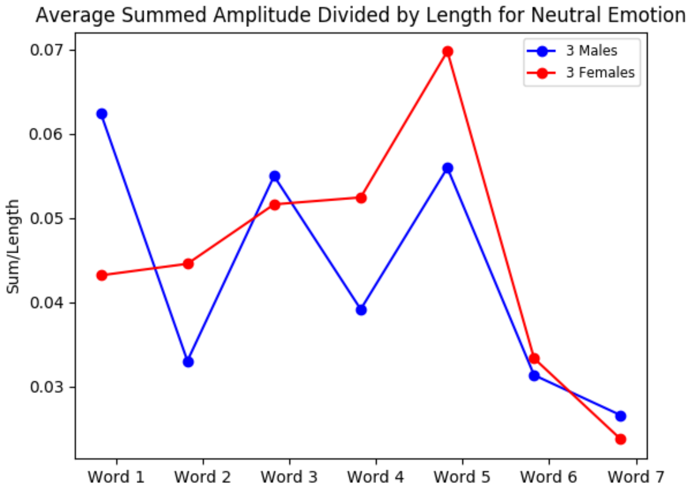

Figure 5.2: Average summed amplitude divided by length per word between three male samples and two female samples for the neutral emotion

This plot is more clear and easier to differentiate between each gendered sample(s). However, it can be seen that neutral speech is not similar between male samples and female samples looking at the general topography of both lines (spotting increasing and decreasing zones). This contradicts the results of the last round of clustering, which showed that neutral speech between males and females is *highly* similar. So, where is the discrepancy? Viewing Figure 5.1, we can see that the second female neutral sample (denoted as "FN 2") has a high degree of variation from the other two female neutral samples. Additionally, its topography is extremely different, if not opposite, from these samples. If we leave out the second female neutral sample as an exception, then the results may be less skewed. This is shown in Figure 5.3.

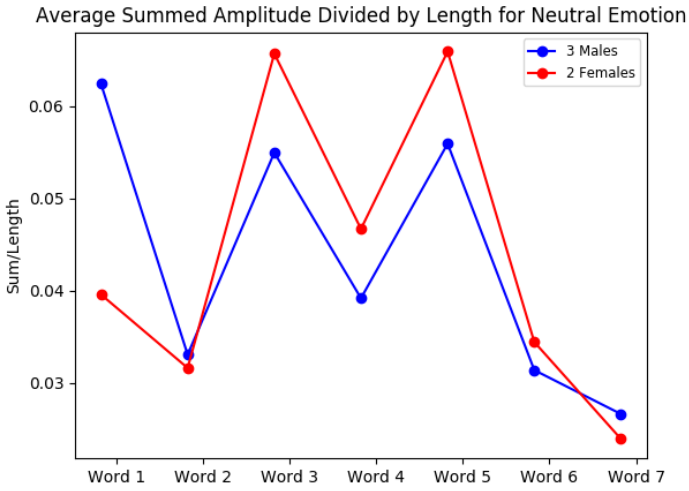

Figure 5.3: (**Exception: Female 2**) Average summed amplitude divided by length per word between three male samples and two female samples for the neutral emotion

The results of leaving out the second female neutral sample as an exception is immediately noticeable. The topography between the two plotted groupings is highly similar. Not only are the values for summed amplitude divided by length close for each word, but the intervals between each word follow the same trajectory: decreasing between words 1 and 2, increasing between words 2 and 3, decreasing between words 3 and 4, increasing between words 4 and 5, decreasing 5 and 6, and finally decreasing again between words 6 and 7. Using this trajectory as a prototype, we can try to spot a similar trend in future samples to identify it as possible neutral speech.

5.2 Happy Speech Prototype

For the happy emotion, the six samples (three male, three female) are plotted in Figure 5.4.

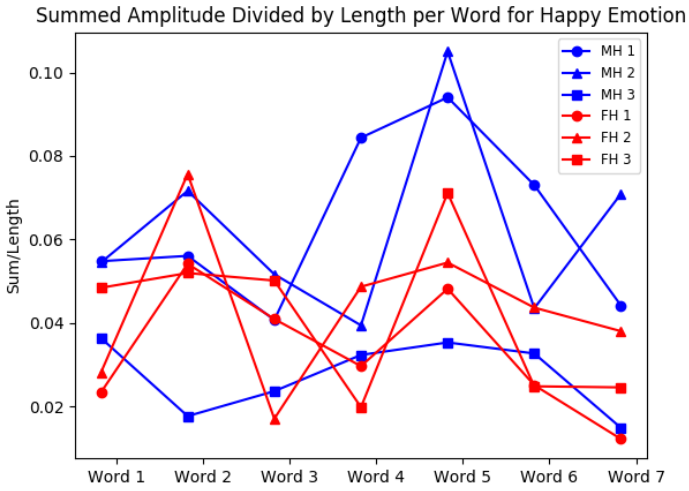

Figure 5.4: Summed amplitude divided by length per word between three male samples and three female samples for the happy emotion

We then average the three samples for each gender to obtain a more clear view of the topography of each gendered emotional sample. Before doing so, we can see that the third male happy sample (denoted as "MH 3") is dissimilar to the other two male happy samples and has a high level of variance. Therefore, following the same procedure as in Figure 5.3, we leave out the third male happy sample as an exception. This is shown in Figure 5.5.

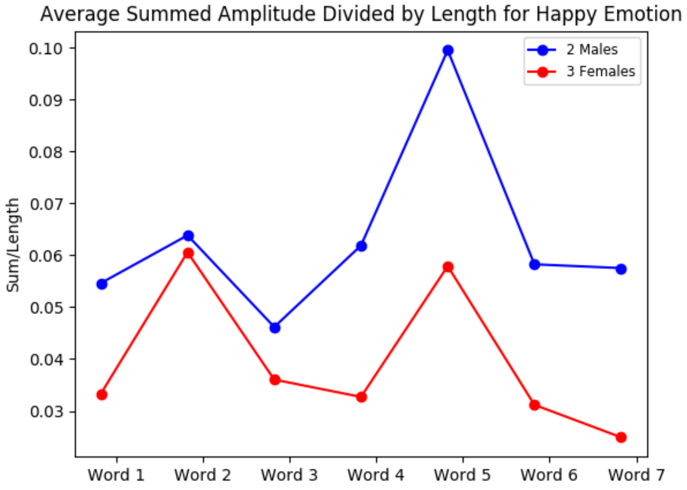

Figure 5.5: (**Exception: Male 3**) Average summed amplitude divided by length per word between two male samples and three female samples for the happy emotion

The results of averaging the happy male and happy female samples (barring the third male sample) presents a motivating trend in the topography for happy speech overall. The values of summed amplitude divided by length are less close than in Figure 5.3, but have a similar trajectory of intervals. They are as follows: increasing between words 1 and 2, decreasing between words 2 and three, increasing between words 4 and 5, decreasing between words 5 and 6, and finally decreasing further between words 6 and 7. There is a discrepancy, however, between words 3 and 4. The averaged male samples present an increasing interval whereas the female samples show a decreasing interval. While these results for the happy emotion are not as streamlined as the results for the neutral emotion, they still present a possibly prototype to spot similar trends in trajectory and identify happy speech in future samples.

5.3 Sad Speech Prototype

Finally, the six samples (three male, three female) representing the sad emotion are plotted similarly to Figures 5.1 and 5.4. This is show in Figure 5.6.

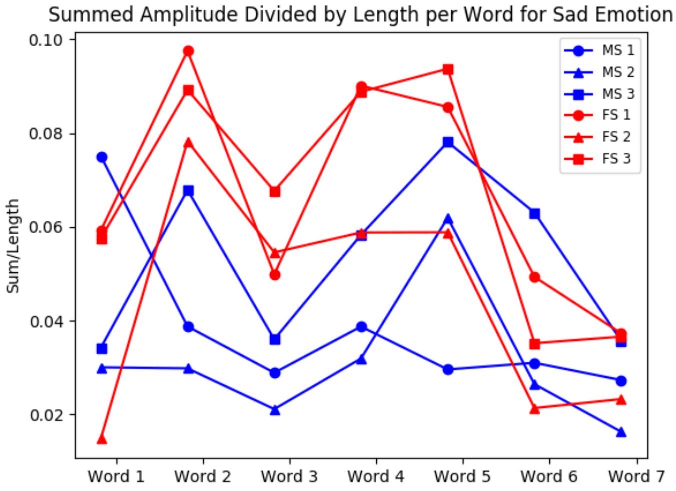

Figure 5.6: Summed amplitude divided by length per word between three male samples and three female samples for the sad emotion

Now, we average the three samples for each gender similar to Figures 5.3 and 5.5. Additionally, we can see that the first male sample (denoted as "MS 1") has an opposing trajectory and topography from the other two male samples. We move to leave out the first male sad sample as an exception, as done in previous examples. The results of this averaging (and exception) are shown in Figure 5.7.

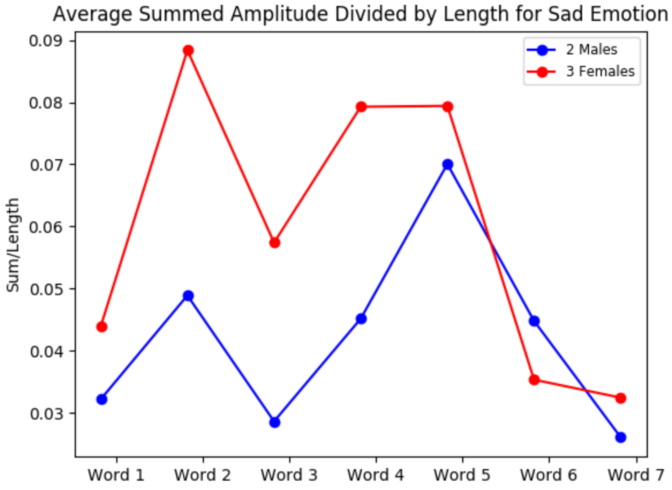

Figure 5.7: (**Exception: Male 1**) Average summed amplitude divided by length per word between three male samples and two female samples for the sad emotion

For the majority of the sentence, up to word 5, the averaged male and female samples are not very close in value for summed amplitude divided by length. However, the topography appears to be very close in terms of both trajectory, and magnitude of interval (the amount of increase/decrease between word). The trajectory is as follows: increasing between words 1 and 2, decreasing between words 2 and 3, increasing between words 3 and 4, increasing further (though at different magnitudes of interval) between words 4 and 5, decreasing between words 5 and 6, and finally decreasing further (at different magnitudes of interval) between words 6 and 7. The high distribution between the male and female samples, however, are motivational in the classification between gender. The similar topography and trajectory is motivational in modeling happy emotional speech, but the somewhat transformation-like distance of value between the male and female samples has the potential to model gendered speech as well.

These plots offer up prototypes for how males and females convey neutral, happy, and sad speech. For additional samples, their topographies (for summed amplitude divided by length per word) can be compared to these prototypes to identify the emotion (from these three emotional variants). Figure 5.8 plots the averaged results (with exceptions) to summarize our findings and illustrates the possible prototypes for neutral, happy, and sad speech.

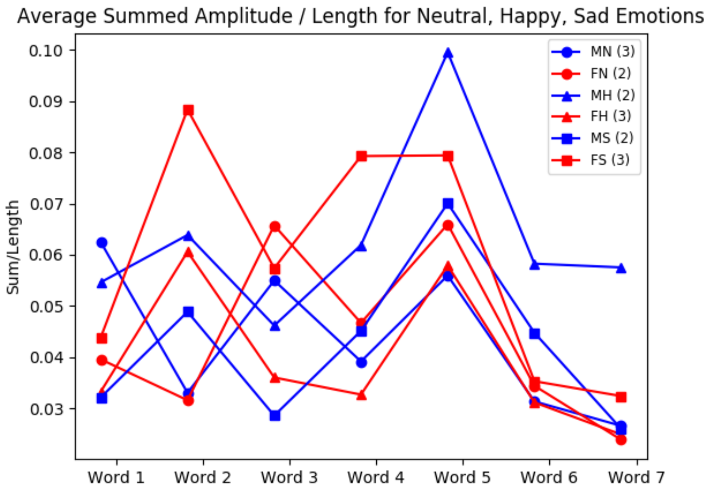

Figure 5.8: (**Exceptions: Neutral Female 2, Happy Male 3, Sad Male 1**) Average summed amplitude divided by length per word for the neutral, happy, and sad emotions

From Figure 5.8, the following observations can be made:

Observation 1: Each of the three emotions has a distinct trajectory pattern for both males and females.

Observation 2: For the neutral emotion, intensity of volume has similar magnitude for both males and females.

Observation 3: For the happy emotion, males are generally louder, or have a larger magnitude of intensity of volume.

Observation 4: For the sad emotion, females are generally louder, or have a larger magnitude of intensity of volume.

Observation 5: The results for the neutral emotion are distinctly different than the trends found for happy and sad speech.

Observation 1 supports the claim that each emotion has a distinct prototypical model, regardless of gender. Observations 2 and 5 support previous hypotheses of the neutral emotion being a baseline model for emotional speech, and is transcendent of gender. For Observations 3 and 4, previous experiments support the hypotheses described. For each case of averaging the male and female samples, it is quite compelling that the topographies were so similar. This qualitative feature was lost during the quantitative analysis process, particularly during clustering. The various gendered, emotional plots were so close in value that the Euclidean distance function could not properly cluster them and thus could not spot these trends and similarities. This simple, qualitative approach to model and prototype creation is more accessible for the bounds of this study as well as the amount of data samples used.

CHAPTER 6

Conclusions

Throughout this study, qualitative and quantitative analysis has been employed hand in hand to make judgements and predictions of the auditory features of emotional speech. Qualitative analysis was utilized as a natural first step in viewing the waveforms themselves. From here, trends were spotted in the waveforms to provide hypothetical baselines for the volume and length of the gendered, emotional utterances. Soon after, we used quantitative analysis measures to attempt to gain *objective*, mathematical findings for the auditory features for the speech. Then, the quantitative analysis attempted to automatically group the speech into groups of gender and/or emotion utilizing classification techniques. Finally, prototypical models for each of the three emotions analyzed (neutral, happy, sad) were constructed, which provided observational hypotheses for the auditory nature of neutral, happy, and sad speech.

Thus far we have been investigating how well our features perform at modeling or clustering emotional speech. Initial results presented here show potential in this direction. With this in mind, we propose the algorithm framework outlined in Figure 6.1 for building an automated emotional speech classifier.

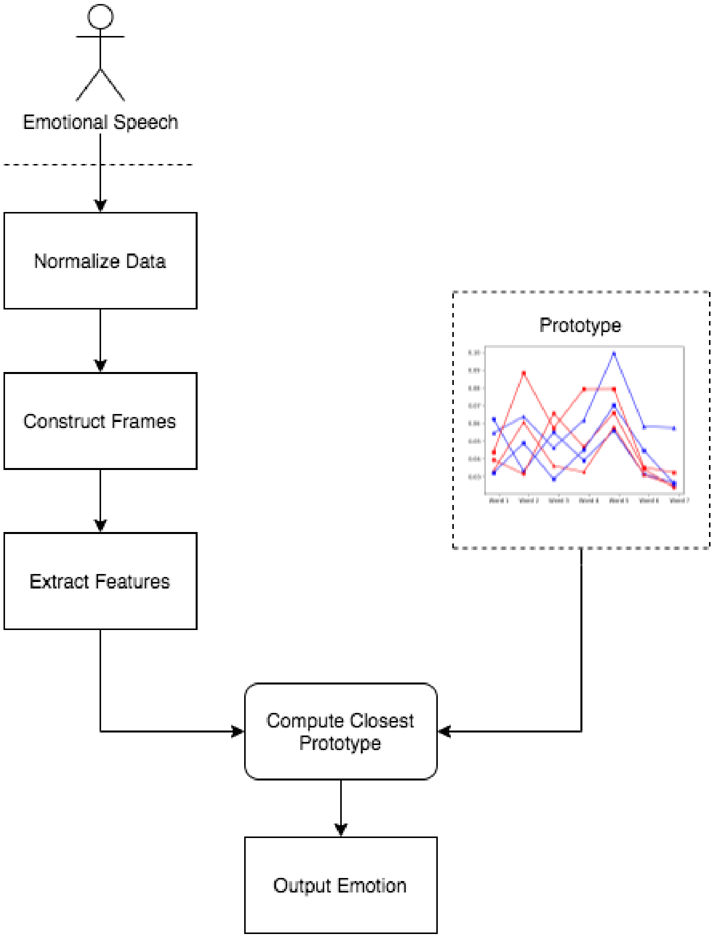

Figure 6.1: Framework for automated emotional speech classificaiton using a previously constructed prototype

This framework (shown in Figure 6.1) works as follows. A new speech sample is imported to the module where the data will be read and preprocessed as described in Sections 2.1 and 2.3, including the normalization phase. Then, sound frames representing the words are constructed so that features may be extracted per word. Features are then extracted, namely the summed amplitude and summed amplitude divided by length features. These features are then compared directly to the prototypes built during the training phase (encapsulated by the dotted line). Finally, our model predicts the emotion conveyed by the speech sample through the previous comparisons.

6.1 Future Work

To further extend the capacities of the qualitative analysis performed in Chapter 5, it may be compelling to look further into the interval between words when plotted as (average) summed amplitude divided by length. The magnitude of the interval itself or the angle between the point (word) and line could serve as possible quantitative features to represent the interval. This is important in further delineating between various emotion prototypes which may have similar patterns for trajectories. For example, in Figures 5.5 and 5.7, we can see that the trajectories are the same (increasing → decreasing → increasing → increasing → decreasing → decreasing), but the magnitude of the interval between words are different. The inclusion of this feature will create a more robust prototype for the emotion models to better classify emotional speech.

Another limiting factor of the analysis performed in this study is the quantity of data samples used. In total, six data samples (three male, three female) were included, with three separate emotions (neutral, happy, sad) analyzed. The RAVDESS dataset as a whole contains 24 data samples (12 male, 12 female) each with 8 different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised). This means that only 18 audio files out of 192 were analyzed (9.4 %). This is quite small of a sample size, and a possible reasoning for less compelling quantitative analysis and clustering results. A larger sample size means more data to be compared against each other, resulting in a more robust model and more emotion prototypes. A limiting factor for the inclusion of more data samples in this study is the manual process of frame construction. Automating the frame construction process is a natural next step, as this would optimize the total time of computation (through eliminating a human-factor), and maximize the batch-size of samples being processed. The time of constructing frames took multiple hours with the manual process, but that time would be shaved down to mere seconds with an automated process.

For a future extension of this study, computational and feature extraction processes need to be automated so that a mass amount of data samples can be analyzed and included into the model. Then, other classification techniques can be employed on the larger data set to more effectively classify gendered, emotional speech. With a robust enough model, emotional speech samples *outside* of the data set can be compared and therefore classified, even without pre-specified class labels. The burgeoning fields of data science and machine learning present a myriad of possibilities for the classification of emotional speech. As presented by this study, the choices of which machine learning technique to employ and which data samples to include are paramount for the effectiveness of this process. Thereafter, the research questions, "Can we computationally distinguish emotion from the human voice?" and "How do males and females convey emotion in their speech?" may be answerable.

References

1. Matplotlib Library. URL https://matplotlib.org/. 7

2. NumPy Library. URL https://numpy.org/. 7

3. About pandas. URL https://pandas.pydata.org/about/index.html. 6

4. Marcel R. Ackermann, Johannes Blömer, Daniel Kuntze, and Christian Sohler. Analysis of agglomerative clustering. *Algorithmica*, 69:184–215, 2014. 32

5. Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, and Andy Song. Efficient agglomerative hierarchical clustering. *Expert Systems with Applications*, 42:2785–2797, 2015. 31

6. Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann (Elsevier), 3rd edition, 2012. 28, 29, 31

7. Daniel T. Larose and Chantal D. Larose. *Data Preprocessing. In: Discovering Knowledge in Data: An Introduction to Data Mining*, chapter 2, pages 16–50. John Wiley Sons, Inc., 2nd edition, 2014. 12, 13

8. Steven R. Livingstone. RAVDESS emotional speech audio. URL https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio. 2, 5

9. Warren Weckesser. wavio 0.0.4. URL https://pypi.org/project/wavio/. 8

10. Junjie Wu. *Cluster Analysis and K-means Clustering: An Introduction. In: Advances in K-means Clustering.*, chapter 1, pages 1–16. Springer, Berlin, Heidelberg, 2012. 29, 30

11. Junjie Wu. *The Uniform Effect of K-means Clustering. In: Advances in K-means Clustering.*, chapter 2, pages 17–35. Springer, Berlin, Heidelberg, 2012. 30