Harmonicity in Speech and Emotion

Isaiah McElvain

29 April 2020

1 Introduction

The goal is to better understand the relationship between emotion and how these emotions are commonly expressed in speech. Understanding this relationship can play an important role in the recreation of these emotions by artificial intelligence or similar software that attempts to replicate human speech.

2 Explanation of the Database

The approach for research was to collect and analyze meaningful data using a publicly available database of speech. The database contained over 1000 audio files. Each audio file is somewhere between 2-5 seconds containing a short phrase. These phrases were spoken by professional voice actors. All of the audio files are lexically identical. Each audio file also given a value between 01 and 08 with each number attributing the file to a specific emotion. For example, 01=neutral, 02=calm, 03=happy, etc. These emotions were classified by untrained adult research participants [1]. These files were also rated on intensity and genuineness, but I was not concerned with these characteristics.

3 Harmonicity

I used the Parselmouth Python library to extract numerical data from each audio file. This extracted data contained values such as average decibels, intensity, RMS, and many others. From all of the variables, I decided to focus on harmonicity.

Harmonicity is the ratio of the harmonic-to-noise content of a file. Speech involves a combination of harmonics and noise and focusing on this aspect lets us analyze how the ratio of this combination evolves and changes depending on the spoken emotion context.

The mean of the harmonicity values was 10.02 db with a standard deviation of 2.49 db and a histogram of the data showed a mostly standard distribution.

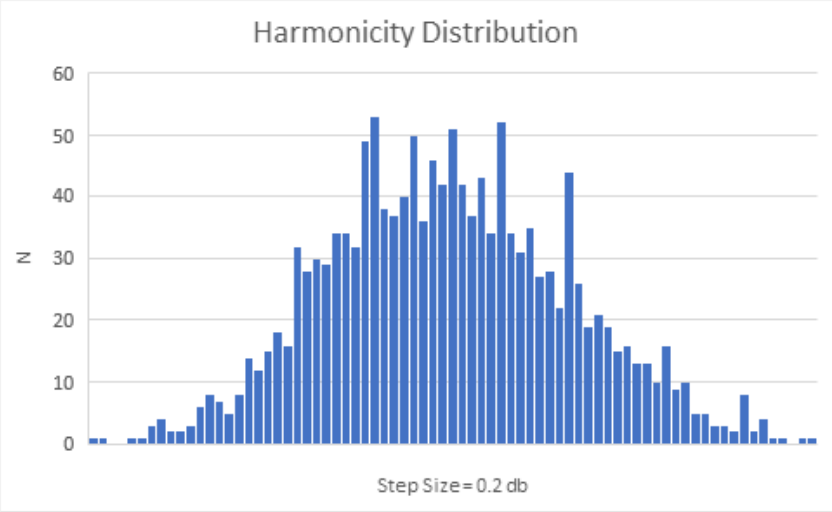

Figure 1: Histogram of the harmonicity data shows a normal distribution.

4 Hypothesis Testing with ANOVA

The null hypothesis is that are no significant differences between the emotions relating to harmonicity in speech. The alternative hypothesis is that differences do exist between the groups of emotions based on harmonicity.

Using the harmonicity data and using the emotional classification of each file I used varies statistical tests to examine the relationship between these two variables.

I decided to perform an ANOVA on the data, with harmonicity being the values and the emotion classification of each audio file establishing the groups. I started with performing an ANOVA with the recording of one voice actor. For one actor, statistically significant differences were found between the groups with an $\alpha < 0.01$.

I wanted to see if this result would generalize for a larger data set, so a second ANOVA was then performed on all of the harmonicity data. For all of data, I found statistically significant differences in the groups with an $\alpha < 0.01$.

The results of both of these tests are sufficient evidence to reject the null hypothesis and accept the alternative hypothesis. This also supports performing deeper analysis of data, and specifically looking at the significance of individual emotions.

5 Hypothesis Testing with T-Tests

With evidence to support looking further into the dataset, I decided to perform a single-sample two-tailed T-test for each emotion group, testing the group means

against the mean of the entire dataset. Six of the eight emotions were found to be significantly different from the mean with a p-value < 0.01.

The emotions that were found to be significantly different from the population mean were: neutral, calm, happy, angry, fearful, disgust, and surprised. The emotions that were not significantly different from the population mean were: sad and fearful. Table 1 shows the average harmonicity and standard deviation of each emotion.

6 Conclusion

These results show that harmonicity of a spoken phrase and the emotion being conveyed are correlated. It is important to note that harmonicity is one variable for a deeply complicated set of parameters that dictate emotion in speech, however this does establish that differences do exist between emotions. This might be an obvious conclusion but it also means that we can start attributing meaningful descriptions to each emotion.

Now that we have established that differences do exist we can start looking at other variables other than harmonicity. These variables could loudness, energy, average pitch, and many others. More advanced analysis techniques should be used such as principle component analysis to find dimensions and features that explain the most variance in the population. Once we find the features that contribute the most, more explicit analysis techniques such as ANOVA and T-Tests should be used.

Hopefully these finding can help guide speech synthesis software or provide an interesting starting-off point for other research projects.

Table 1: Table containing emotions and corresponding p-values of each individual t-test.

Emotion	**Neutral**	**Calm**	**Happy**	**Sad**	**Angry**	**Fearful**	**Disgust**	**Surprised**
Avg. Hamonicity	10.75570107	11.23913492	10.38400165	10.58576859	9.020624764	10.07360077	8.512577095	8.527019368
Std. Deviation	2.041140197	2.100536202	1.724420391	2.846050266	1.959647104	2.032526237	2.205037672	1.503774158
T-value	4.168514976	6.305622571	2.822181789	2.404571733	-4.33327151	0.898061084	-6.108518395	-8.863038582
P-Value	6.7655E-05	9.0045E-09	0.005809326	0.018130218	3.64824E-05	0.371423433	2.1961E-08	4.44224E-14
Significant (.01)	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE

References

[1] Russo FA Livingstone SR. A dynamic, multimodal set of facial and vocal expressions in north american english. 2018.