Acoustic Differences in Languages in Comparison to English

Angelina Zhai and Maisey Perelonia

1 - Abstract

The world is a diverse place, containing thousands of different dialects, thus it is possible to observe different acoustic properties in different languages. This study examines the acoustic differences between various languages and dialects in both timbre and pitch through comparisons with English. We examined nine different languages and found that while there was not a distinct trend present between each language group in our study, differences in fundamental frequencies do exist when the language in question is compared to English.

2 - Introduction

Languages are an interesting acoustic topic for their necessity and diversity. There exist tonal languages such as Mandarin, where the spoken words utilizes four different tones; a wrong tone can indicate a completely different meaning. There also exist atonal languages such as English, where tones are not needed to indicate different meanings. As students who are proficient in more than languages, we are especially interested in the acoustic differences between the languages we speak everyday. Through this project, we examined the differences in fundamental frequencies and timbres of different spoken languages. This report will first discuss the methods data we have collected, then the data pre-processing steps, followed by the statistical analysis and discussion.

3 - Objective

Determine if there exists any acoustic differences between different spoken languages when compared to English..

4 - Study Design and Methods

4.1 - Data Collection

If our resources were unlimited, then the ideal method of our experiment would be to randomly collect recordings from a large sample set of thousands of people from many countries, and then average out the acoustic characteristics within each country. By collecting from such a large number of people, it would produce a sample set that is representative of the entire population and provide a good basis as to how the majority of that country's population speaks.

However, given that we do not have the capacity to conduct a study of such a large scale, we decided to gather nine of our multilingual friends and record them speaking the same sentence in two languages: English (as a control to compare our other languages to) and another language in which they are fluent. The sentence is in the following format:

“My name is _____. I am _____ years old, and I like _____. For a meal, I ate _____, and later, I will be doing _____”.

The nine languages and dialects we have recorded are English, Arabic, Japanese, Mandarin, Shanghainese, Cantonese, Hindi, Tagalog, Indonesian, and French.

4.2 - Materials Used

We used Apple wired headphone microphones to record participants in a quiet room. We used the voice memo app on iPhone for initial recording, and Garageband for data cleaning after the recording process.

4.3 - Data Processing

After all participants were recorded, Garageband was used to clean up the recordings in .m4a format. First, we ensured that the duration of pauses before and after each person’s recording of their sentences were consistent in both recordings, so that our results will not be affected by these discrepancies. Figure 1 below illustrates the amplitude vs. time graph of one of our recordings after cleaning.

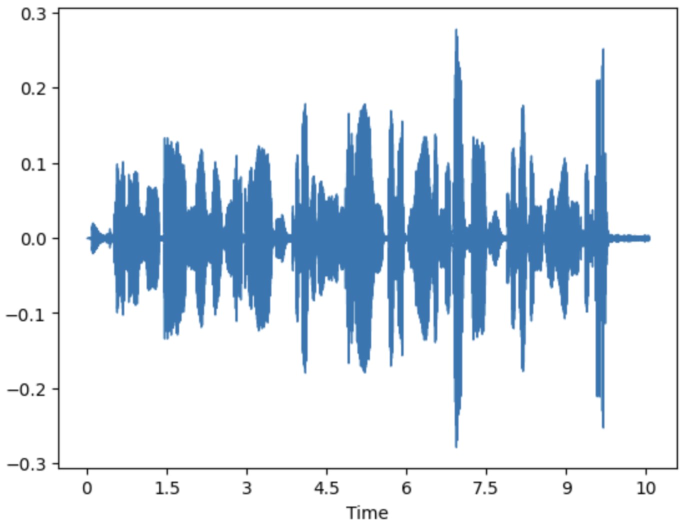

Figure 1: Amplitude vs. Time Graph of one raw audio recording

To examine all frequencies (harmonics included) with respect to time, we used the short-time Fourier Transform (STFT) on each recording. Traditionally, the Fast Fourier Transform (FFT) averages out the frequency components over time in a signal. We decided to use STFT over FFT because it can track the frequencies of a signal over a period of time [1], making STFT a more optimal tool than the FFT, for the purpose of examining how frequently a person varies in pitch.

To determine whether we have cleaned and processed our data properly, we displayed our recordings onto spectrograms, as shown below in Figure 2. We can see that the most prominent frequencies, the fundamental frequencies, emit the brightest and thickest lines on the graphs.

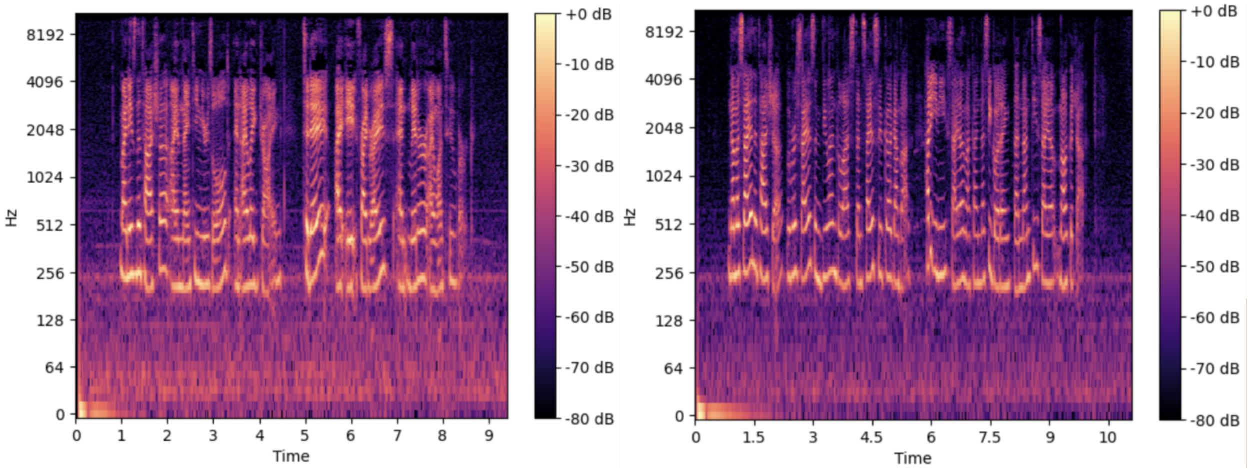

Figure 2: Spectrograms of two recordings, where the x-axis is time in seconds, y-axis is frequency in Hz, and z-axis is loudness in dB

Timbre plays a factor in our analysis because different people will have different timbres due to the variation in vocal properties. More specifically, this is composed of the fundamental as well as the resonant frequencies in the sound that is perceived. We used Yin's algorithm [2] to isolate the fundamental frequencies (F0) at each point in time to eliminate the possibility of timbre to affect our results. An additional benefit of this is that unnecessary background noise can be filtered out. Figure 3 below demonstrates the result of this estimation, and it is quite obvious that only one frequency has been kept at each point. This will further be discussed in the analysis section.

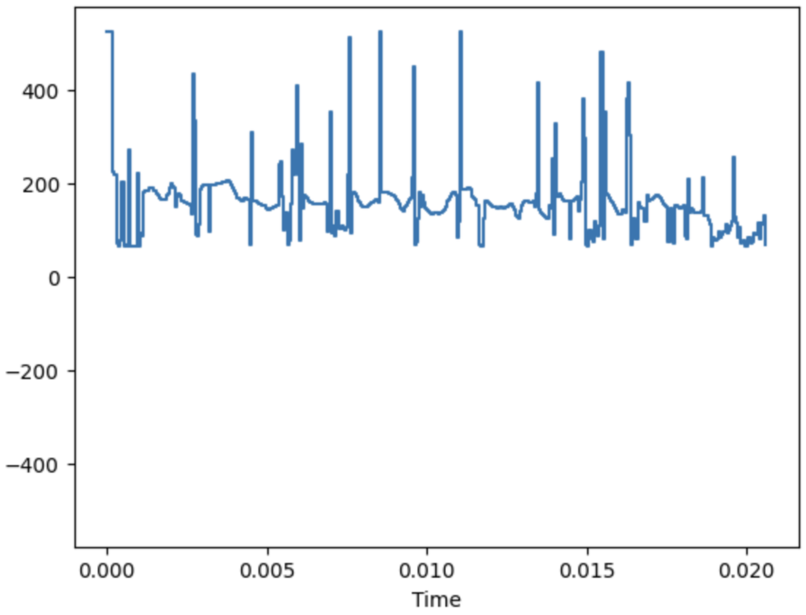

Figure 3: Fundamental frequencies vs. Time representation for one sample audio

5 - Analysis

At the end of the data cleaning process, there remained 9 distinct data pairs in 9 different languages with every recording ranging from 10 to 15 seconds, sampled at a rate of 22050 Hz. Two-sample T-tests were conducted on each data pair to find the corresponding absolute percent difference in mean between the two recordings. The two-sample t-test works under the assumption that the data points in both samples are random, independent from each other, and are normally distributed [3], which are all true in our case, considering the quality of our voice are random, the acoustic features in one language do not affect the features in another language, and given the large amount of data points, we can assume that all our samples are normally distributed. Traditionally, the test is used to test if the means from the two samples are equal, though the test also provides information on how the mean in one sample differs from the mean in the other [4]. The t-test was performed on all data pairs (both fundamental frequency and spectrogram data samples) using the `stats.ttest_ind` function from the SciPy library.

Additionally, we also conducted the two-sample Z-test on all data pairs (both fundamental frequency and spectrogram data samples). This is done using the stats.weightstats.ztest from the statsmodels library. The Z-test serves the same purpose of finding the absolute percent difference in mean, while it is specifically designed for datasets with a very large number of data points [5].

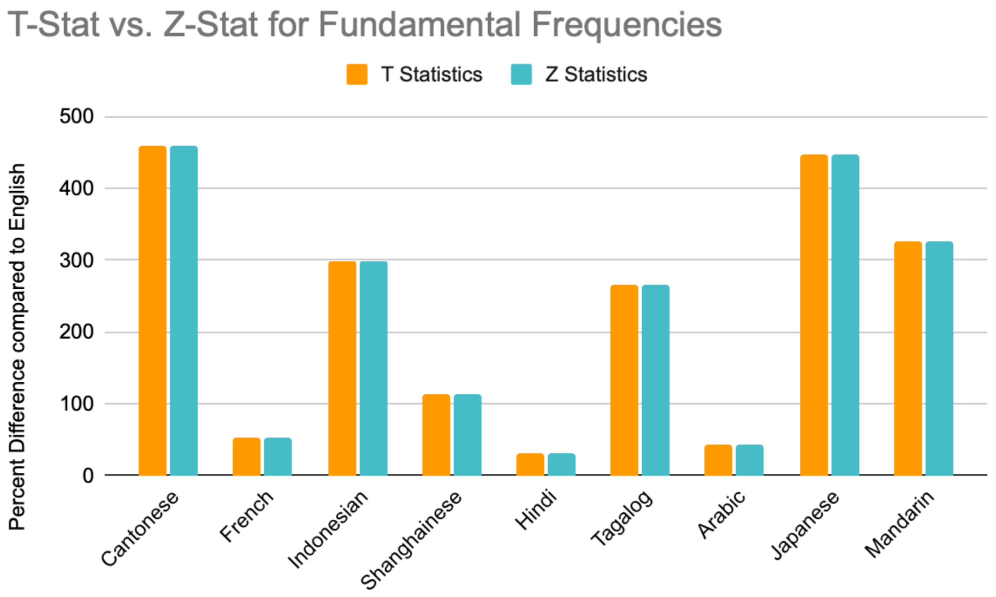

Figure 4: Comparison of Calculations by T-Test vs. Z-Test on Fundamental Frequency Data

When we conducted the two tests on all our fundamental frequency and spectrogram data pairs, it was revealed that there was not a significant difference in result between the two tests. Thus, our discussion on results will be described in terms of the Z-Test from now on.

Additionally, we have examined the graphs of fundamental frequency approximations generated by Yin's algorithm and discovered that the estimations generated are not always correct. In fact, as shown in Figure 5, there seems to be asymptotic behaviour in the approximations whenever the participant begins speaking again after short pauses.

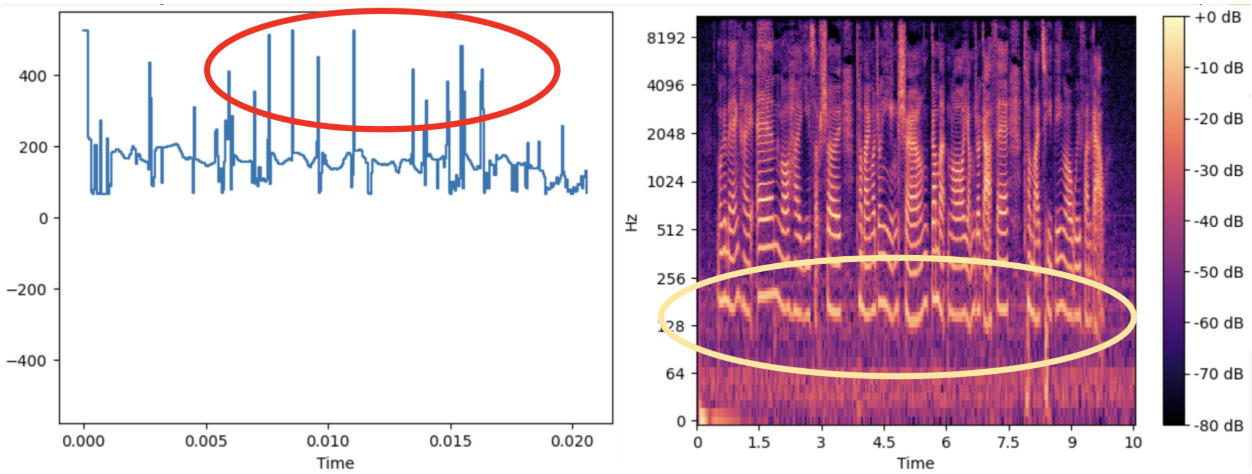

Figure 5: Estimated fundamental frequencies (left) vs. Spectrogram representation of the same audio sample (right). Circled in yellow is the fundamental frequency we are looking for, and circled in red is erroneous fundamental frequency data points generated by Yin's algorithm.

We decided to plot our fundamental frequency data counts as histograms to see if there is a significant amount of outliers, and as shown in Figure 4, there was a noticeable amount of data at the very high and low ends of the spectrum. To ensure that the analysis conducted later will be on fundamental frequency data that is as accurate as possible, the interquartile range (25th percentile to 75th percentile only) of each set of fundamental frequency data is taken by using the `series` function from the Pandas library, which provides lots of useful information such as the dataset's median, mean, and the interquartile range. Taking the interquartile range gets rid of excess information and outliers, and provides a more accurate representation of each dataset. Using the output from the `series` function, we then calculated for the percent differences to find if the language in question is higher or lower in fundamental pitch than English.

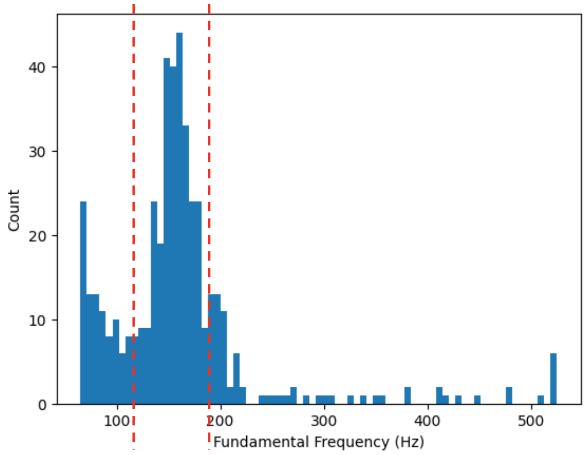

Figure 6: Histogram representation of Yin's Fundamental Frequency. Data points bounded by the red dotted lines are the data points within the interquartile range.

6 - Results and Discussion

Following the analysis methods described in the previous section, we obtain the graphs below, where Figure 7 shows the Z-statistic of audio power between foreign language and English in each data pair, and Figure 8 shows the signed percent difference in the fundamental frequencies between the two languages in each data pair.

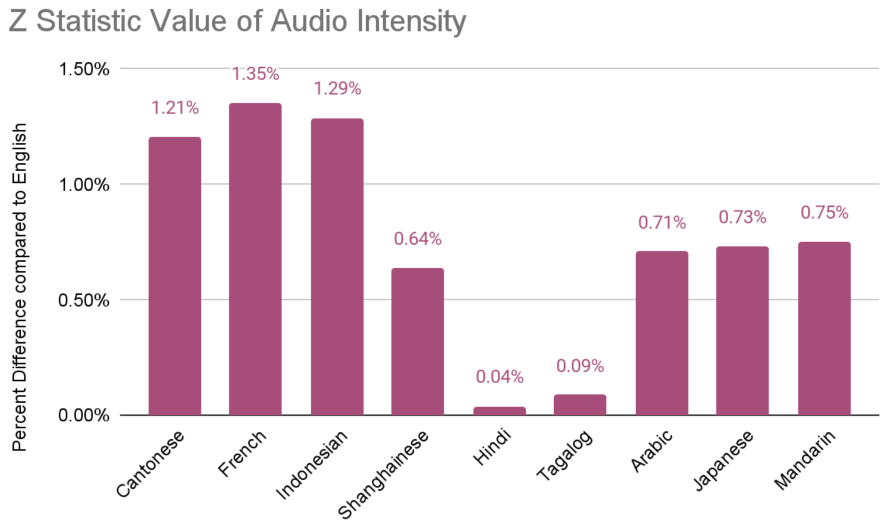

Figure 7: Z-Test results for each language paris

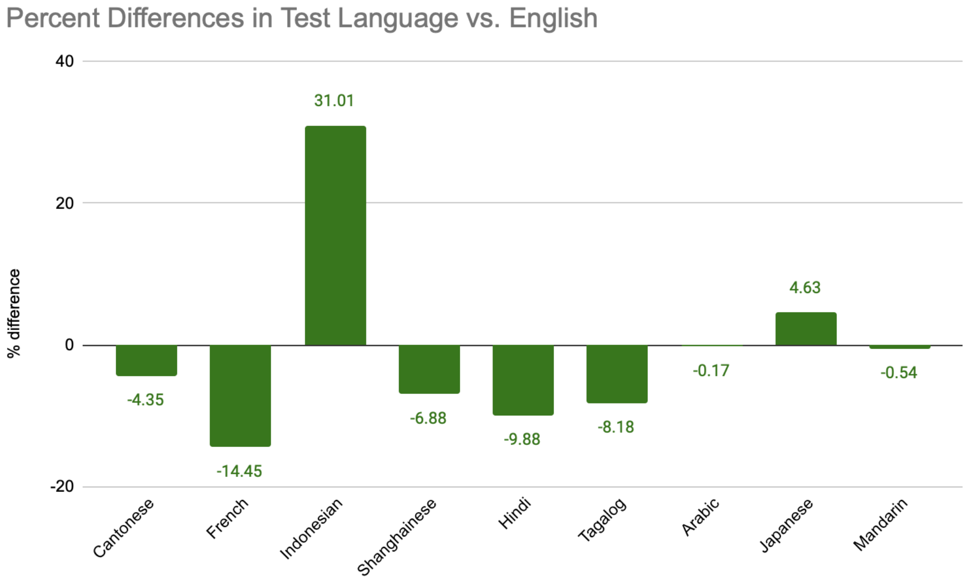

Figure 8: Percent Differences between Fundamental Frequencies in each language pairs

From Figure 7 we can see that the language with the highest difference in audio power is French at a 1.35% absolute difference, followed by Indonesian at a 1.29% absolute difference. The language with the smallest percent difference in audio power is Hindi at a 0.04% absolute difference. As shown, there is not a significant difference between the test and control languages in all data pairs, signifying that all participants spoke at similar volume with similar timbre when speaking the two languages.

In Figure 8, however, more noticeable differences are observed between the fundamental pitches between each speaker's speech samples in foreign language versus in English. The language with the highest percent difference in fundamental frequency when compared to English is Indonesian with a 31% difference in F0 higher than English. The language with the second highest percent difference in magnitude is French, being 14.45% lower in F0 than English. The languages with the lowest percent difference magnitudes are Mandarin and Arabic, at only less than 1% difference.

The results in Figure 8 have indicated that there are indeed significant differences in fundamental frequencies for some foreign languages when compared to the pitch of their speech in English. The precise nature of this difference, such as the difference range for each language examined, the potential causes for this difference (such as participant demographic, language skill level), and behaviours of the difference in tonal versus atonal languages, can be determined by acquiring more participants and language samples.

7 - Conclusion

To conclude, we have determined that while someone's speech timbre and loudness is not likely to change when they are speaking, it is likely to observe differences in pitch of speech in different languages. The specific causes and pattern for this difference can be studied further in depth with a larger participant population and diversity in demographics, which can be studied in the future by recruiting more participants with different skill levels, age, gender identities. It will also be optimal to recruit participants proficient in more languages and conduct analysis according to language groups (such as by language families and tonality). It is fascinating to see how people of different backgrounds can come together and communicate similar words yet sound very different while doing so.

8 - References

[1] “librosa.stft,” *librosa.stft - librosa 0.10.0.dev0 documentation*. [Online]. Available: https://librosa.org/doc/main/generated/librosa.stft.html. [Accessed: 09-Dec-2022].

[2] A. de Cheveigné and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[3] “Two-Sample Independent t-Test,” *Statistics Online Support*. [Online]. Available: https://sites.utexas.edu/sos/guided/inferential/numeric/onecat/2-groups/independent/two-sample-t/. [Accessed: 09-Dec-2022].

[4] J.M. Curran, “The Frequentist Approach to Forensic Evidence Interpretation,” Encyclopedia of Forensic Sciences, pp. 289-291, 2013.

[5] Andrew P.King and Robert J.Eckersley, “Inferential Statistics II: Parametric Hypothesis Testing”, *Statistics for Biomedical Engineers and Scientists*, pp.91-117, 2019.