Challenge 2 Report

Juan Reyes
University of California, Santa Barbara

Marco Chavez
University of California, Santa Barbara

***Abstract*—The focus of this report is on the implementation of three different methods to discriminate between speech and music. It was implemented using MATLAB. The three methods created are using *Percentage of Upper and Lower Amplitude Samples*, which from our test samples achieve accuracy of about 95%, *Low Energy Frames*, which had 90% accuracy, and *Zero Crossing Rate*, with accuracy of about 88%.**

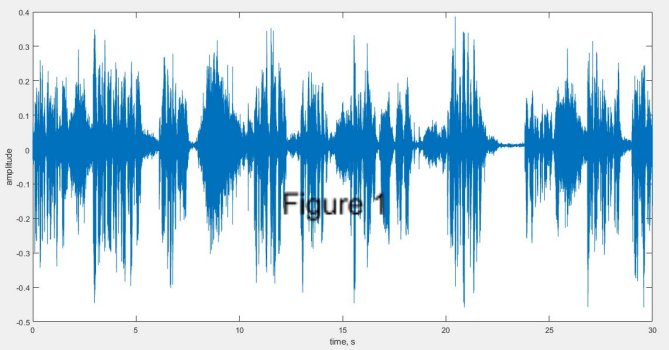
I. INTRODUCTION

Signal processing is used for various useful applications, which usually involves audio or images. With audio, things like convolutions with another signal can create interesting sounds, such as recreating the sound of an instrument being played in different building. The audio can also be processed for identification, by detecting whether it is music or something else completely. In this paper, audio is processed to determine the nature of it, whether it is music or speech. This can be challenging as the line between speech and music can be blurry. Therefore, three different methods were used to accomplish this discrimination of speech and music. The first method is *Percentage of Upper and Lower Amplitude Samples*, this method is simple and the analysis is done in the time domain of the signal. The next method, *Low Energy Frames*, is analyzed in the frequency domain. The last method, *Zero Crossing Rate,* is analyzed in the time domain. The last two method analyze the audio data using bins, which allows for flexibility in audio.

II. METHODS

A. *Percentage of Upper and Lower Amplitude Samples*

The implementation for this method is rather simple, it analyzes the signal in the time domain and takes advantage of the general waveform of speech. Figure 1 demonstrates this, it shows what a general recording of speech would look like. The x-axis represents time and y-axis is the amplitude of the sample.. It has many pauses or areas that have amplitudes of zero or almost zero. This is so because when someone speaks, there are pauses between their words. An analysis of the music waveform shows that it has a consistent width, there are little to no



samples where the amplitude is zero. Figure 2 illustrates this really well, similarly to Figure 1, the x-axis is time and y-axis is amplitude. Using this observation a percentage of how many samples are below a certain number and the percentage of how many are over a certain number was used to distinguish whether a signal was speech or music

B. *Low Energy Frames*

This method analyzes energy densities as follows. Before analyzing low energy frames the data is filter to remove any outliers, that can potentially skew the classification of audio signal. Each audio file is partitioned into bins, where each bin hold the same number of samples. Then each bin is applied the fast fourier transform to convert the data from time-domain to frequency domain. Once in the frequency-domain multiplying each sample with its complex conjugate yield the energy of each sample. The threshold value, E_L, to distinguish low energy bin is set, by calculating the normalized energy of bin. This threshold is combine with a threshold value, S_L, for the number of acceptable samples to be below E_L. If the total number of samples below E_L is not acceptable, the bin is considered a low energy bin. The total number of low energy bins per audio file is used discriminate between music and speech using a third threshold.

C. *Zero Crossing Rate*

This method discriminates music from speech by measuring how frequent the audio signal amplitude goes from negative to positive. To avoid false zero crossings due

to noise, values close to zero are zeroed out. The audio signal is then put into bins of equal size. To discriminate speech from music, two threshold values are set. The first threshold is the number of zero crossing per bin, that dictates whether a bin has high number of zero crossings. The second threshold is the number of high zero crossing bins allowed to be consider speech.

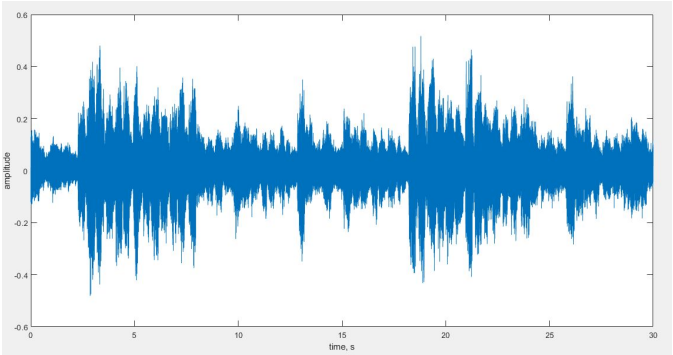


Figure 2

III. RESULTS

A. *Percentage of Upper and Lower Amplitude Samples*

Many different values were tested for these two bounds, the best options for the lower bound and upper bound were 0.055 * (max value of sample) and 0.75 * (max value of sample), respectively. A ratio of each of these samples and the entire number of samples was calculated to determine if it was music or speech. The calculation was if the percentage of samples lower than 0.055 * Max was less than 0.15% and the percentage of samples greater than 0.75 * Max was greater than 80% then it was labeled as music, otherwise it falls into the speech category. Using this approach and with the thirty second audio samples provided this method achieved approximately 95% accuracy, with better results detecting music than speech.

B. *Low Energy Frames*

From intuition it make sense for speech to have more low energy frames, but the results are contradicting. The reason for this is the way that the threshold value E_L was calculate. Music has more stable energy pattern than speech. This causes the music audio to have a higher threshold, E_L, value where the majority of the bins are considered low frames. Speech on the other hand has periodic low energy frames follow by high energy frames, lowers the E_L threshold for speech signals. For speech the only the truly low energy frames are counted as low energy frames.

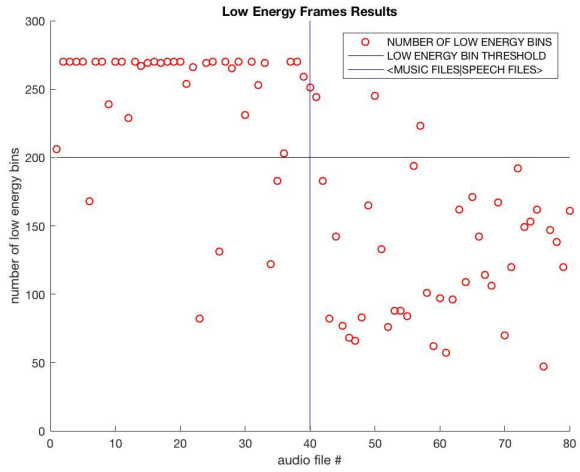


Figure 3: Low Energy Frames results of analyzed audio files with the discriminating threshold marked horizontally. Everything the left of the vertical line are music audio test and to the right of the line are speech audio test.

From empirical analysis, this algorithm had its highest success when the thresholds were set as explained next. The energy threshold, E_L, was set to 0.503*normEnergyOfBin. The threshold for the number of low energy samples per bin was set to 0.4 * binSize. These threshold values makes sense given the explanation above. The third threshold the number of bins pero audio file to discriminate between speech and music was set to 0.74*numOfBins.

C. *Zero Crossing Rate*

After running various music and speech files through this algorithm, speech audio tended to have less bins with high zero crossing and music audio tend to have high zero crossing for most bins. This insight was a key factor in deciding ideal values for thresholds. Given this insight the threshold for the number of bins required to have high zero crossing rate was fairly high, 0.98*numOfBins.

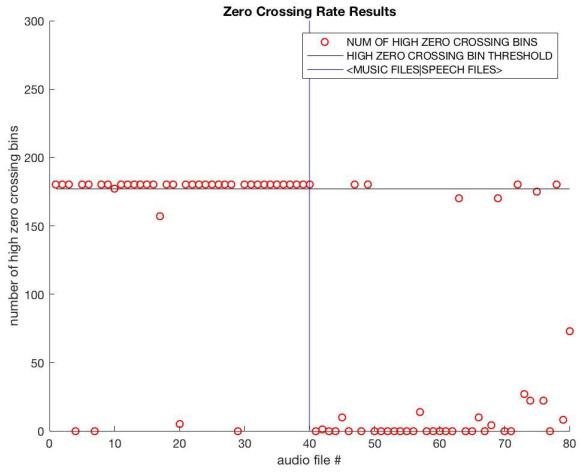


Figure 4: Zero Crossing Rate results of analyzed audio files with the discriminating threshold marked horizontally. Everything the left of the vertical line are music audio test and to the right of the line are speech audio test.

IV. DISCUSSION

A. *Percentage of Upper and Lower Amplitude Samples*

In comparison to the other methods, this one seemed to work the best with relatively high accuracy. The testing set of audio files was a bit small so the results may not be extremely reliable for many more cases. Also the audio samples were thirty seconds long and the music samples could have been filled with music content, while some music may have pauses every once in a while. So in theory it may not work as well for longer samples. In regards to it performing better with music than speech, may be an indication that the speech samples were crossing over to seem more like music. This can happen if the audio recording has loud background noise so the waveform is more continuous closer to a music waveform.

B. *Low Energy Frames*

After revising the audio files that did not pass the test it is understandable why. In some music files the audio had sequences of intense music followed by quieter periods. When analyzing the energy of such a recording it would seem that it was created from someone speaking. Similarly for speech audio that were misinterpreted either the person was speaking really fast, which reduced the amount of silence. Another speech audio that failed was a stand up comedy recording. Either the comedian was talking fast or the audience was laugh. Both the example above seem to have stabilized energy, therefore being misinterpreted as music.

C. *Zero Crossing Rate*

This method was noticeably slower than the two other methods. It is harder to explain why an audio test failed, because there is no intuitive understanding on why music or speech would have high zero crossing rate.

V. CONCLUSION

In conclusion, all the methods seemed to perform well with method A being the most accurate. Method A and method B ran with linear time complexity which good for scalability of audio quality and test size. It would probably take a more complex method that measures or analyzes more of the signal than just the amplitudes, in order to get closer to 100% accuracy. But this might not be possible as the distinguishing factors of music and speech can be vague and blurry.

REFERENCES

- *Automatic speech/music discrimination in audio files* by Lars Ericsson: http://www.speech.kth.se/prod/publications/files/3437.pdf
- *Speech/Music Discrimination via Energy Density Analysis* by Stanislaw Kacprzak and Mariusz Zi´olko https://link.springer.com/content/pdf/10.1007%2F978-3-642-39593-2_12.pdf

Juan Reyes

- Implemented method A
- Helped debug code
- Helped write code to read in files

Marco Chavez

- Implemented method B and C
- Helped debug code