David Kleinberg
April 9, 2017
Signals & Systems

Project 1

Feature Analysis

- Below are plots of the Zero-Crossings and the Variation of Zero-Crossings analysis for one music file and one speech file, respectively.

Figure 1:

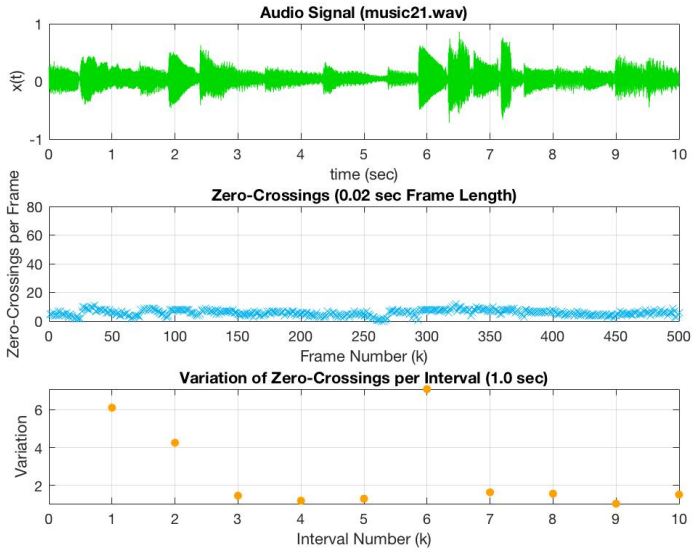
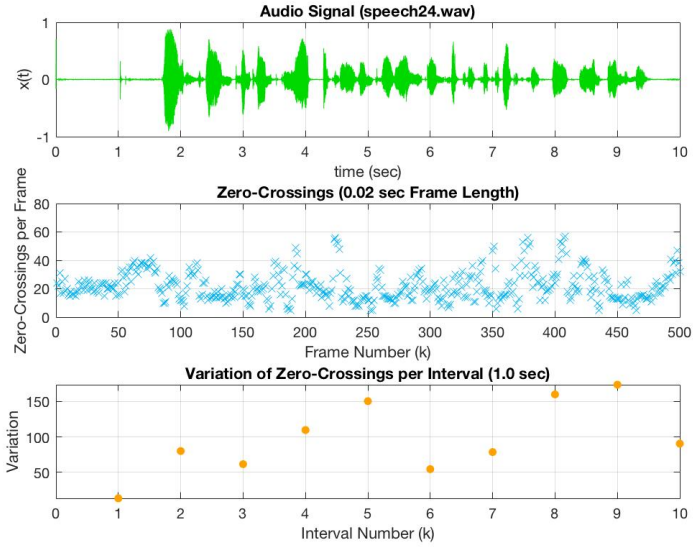


Figure 2:



Zero-Crossings (per Frame) Analysis:

Figure 1, above, shows data from an audio recording that is characteristic of an instrumental signal. The number of Zero-Crossings, per frame, remains relatively consistent throughout the sample. The Variation of Zero-Crossings remains very low, compared to all other samples.

Figure 2, above, shows an audio recording that is characteristic of a speech signal. The number of Zero-Crossings per frame is much higher than the number of Zero-Crossings per frame resulting from the music sample in Figure 1. In fact, almost every frame of the speech sample has a higher number of Zero-Crossings than the frame with the highest number of crossings in the music sample.

While this observation is valid amongst the two samples chosen, the number of Zero-Crossings in a sample is not necessarily telling as to which type of audio signal the values resulted from.

Figure 3:

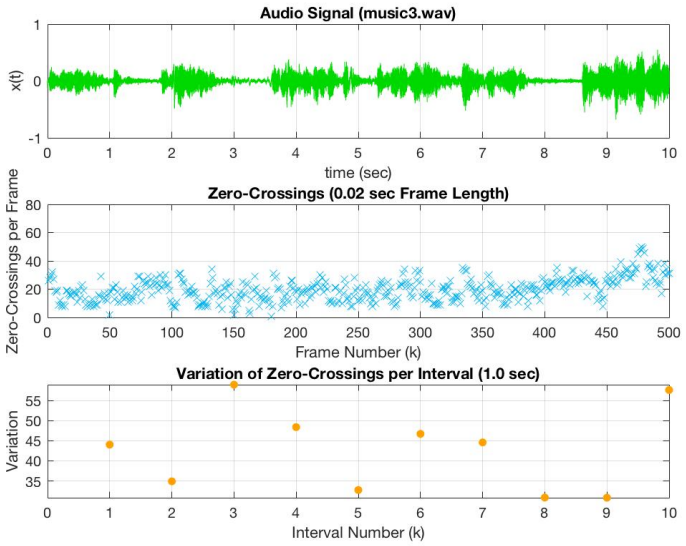
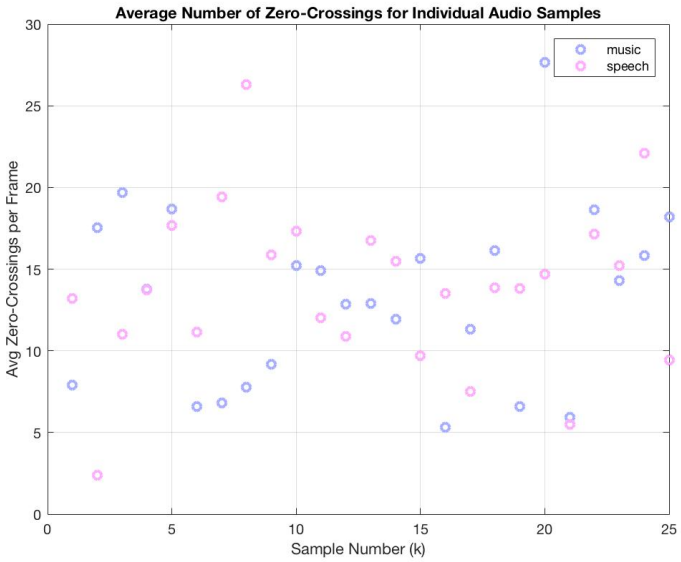


Figure 3, above, shows a different music sample (music3.wav), which contains a signal with 3-4 times as many Zero-Crossings per frame as there were in the music21.wav sample. The audio sample in the figure above is still distinguishable from a speech recording, as its variation of Zero-Crossings is still low compared to that of the speech signal of Figure 2.

Referring back to the second subplot of the characteristic music sample shown in Figure 1: Looking just at the Zero-Crossings, it would be unlikely for that audio file to be classified as a speech recording, because the numbers of crossings per frame are too low for there to be enough variation that deems it speech. It can therefore be concluded that Zero-Crossing values are enough to rule out speech signals, only when the highest number of crossings is below a certain threshold. This property does not however work in the contrary. High numbers of Zero-Crossings does not rule out music signals.

Figure 4:



To further this argument, I plotted the average number of Zero-Crossings per frame in each sample. The results for music are shown with the purple markings and overlaid onto the plot of the speech samples. (In sample 4, the marks overlap completely.) It is clear, from Figure 4, that the number of Zero-Crossings per frame of a sample is not really indicative of the sample's source (instrumental vs. vocal).

Estimated Variation of Zero-Crossings Analysis:

The **variation** in the number of Zero-Crossings per frame was more telling of the source of the signals. Speech files yielded a higher degree of variation than music files. These results were consistent enough to use the ZCRV as a feature with which to classify the two types of signals. while music samples could have many Zero-Crossings per frame, but the number would often remain consistent such that there was little variation.

The variation is far higher, on average, than is the case in the music sample shown above.

For a random variable vector, *A*, made up of *N* scalar observations, the variance is defined as

$$V = \frac{1}{N-1}\sum_{i=1}^{N}|A_i - \mu|^2$$

where μ is the mean of *A*,

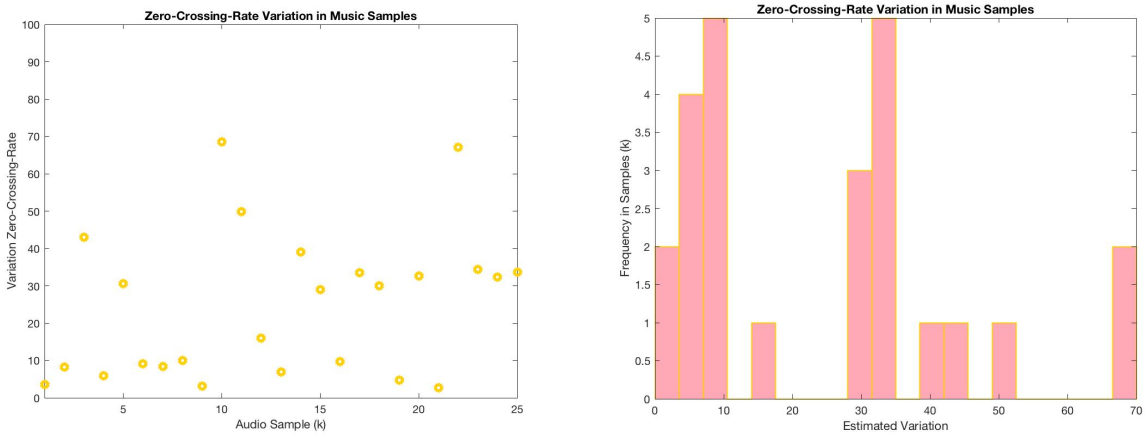
$$\mu = \frac{1}{N}\sum_{i=1}^{N}A_i$$

We therefore refer to these values as *estimated* variations because variation *V* takes into account the mean of the vector, and is thereby dependent on all other values in that vector. This means that adding/removing values from the vector will likely change the variation.

In this case, the length of the vector is finite. In fact, it is only 50 variables in length. Within each of these vectors are 50 values, each representing the count of Zero-Crossings that occurred within a 20msec time frame. Additionally, sampling frequency can always add some level of uncertainty and therefore further imply that these values be called “estimates”.

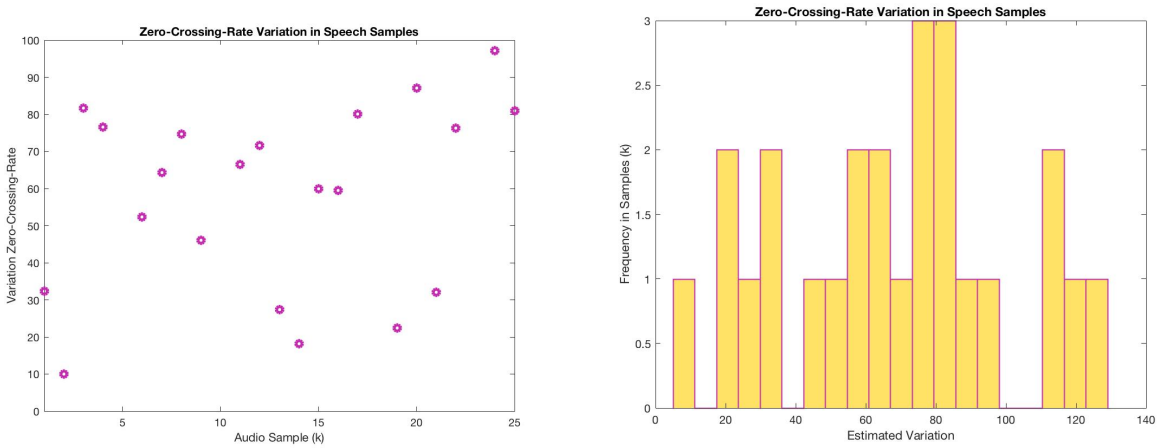
The calculated variation values are likely to change for a different time frame, different levels of overlapping frames, a different sampling frequency, or even a different amount of time the sample is taken over. It is for this reason that it is considered an estimate.

Figure 5:



$$\overline{ZCRV_m} = 24.4783$$

Figure 6:



$$\overline{ZCRV_s} = 67.4819$$

In the histogram of ZCRV for music samples, the bins are of size 3.5 (variance). In the histogram of ZCRV for speech samples, the bin size is ~6.5 (variance).

- Below are plots of the Signal Power (per frame) and the Low Power Frames analysis for one music file and one speech file, respectively. The music and speech files are identical to the samples used for Figure 1 and 2, respectively.

Figure 7:

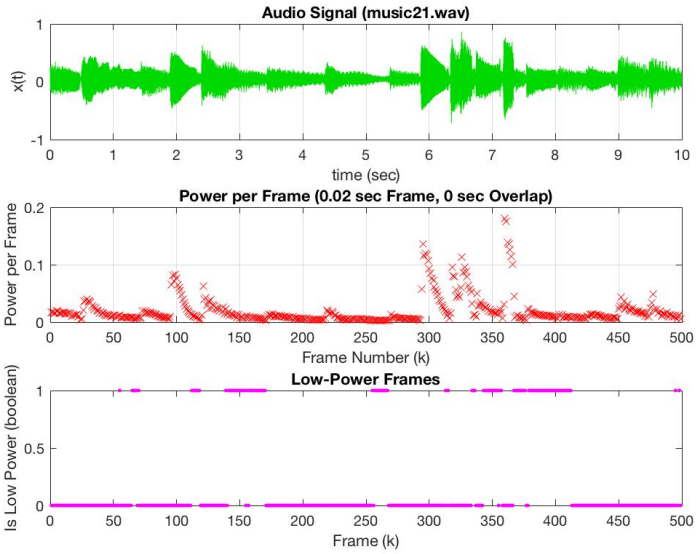
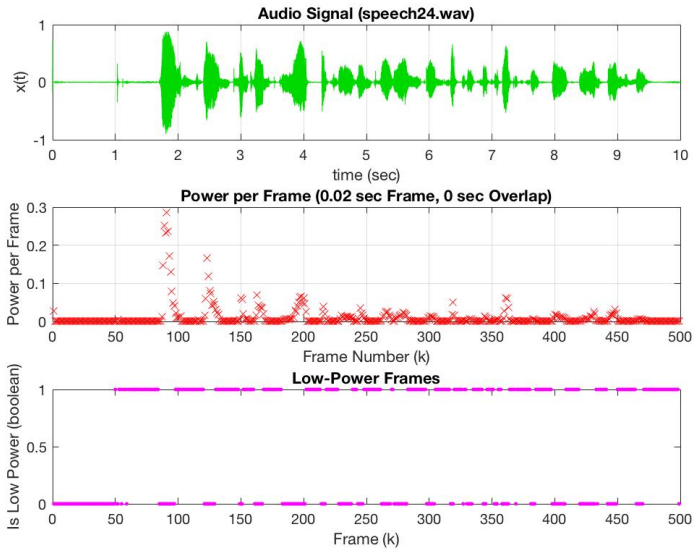


Figure 8:



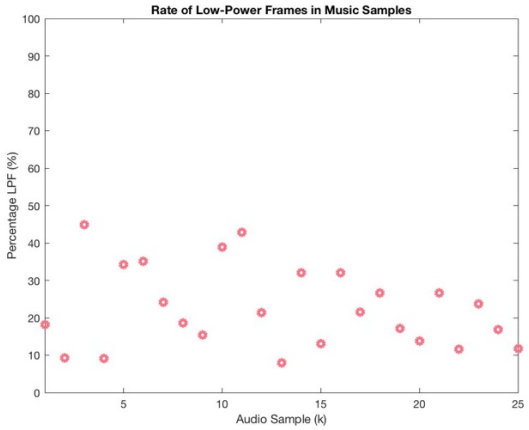
Low Power Frames Analysis:

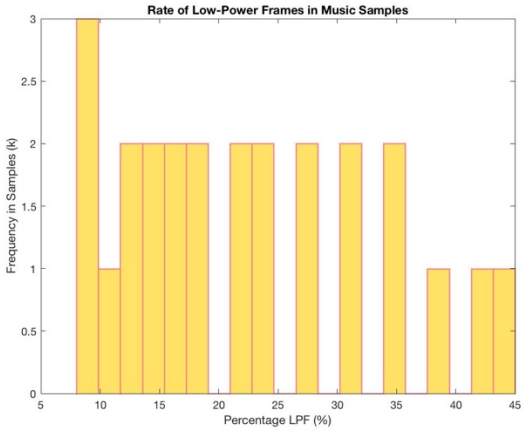
Similar to the case with the number of Zero-Crossings per frame of an audio signal, the power per frame of an audio signal is not very indicative of its nature. The two plots show very similar

quantities even though one signal fell far along the music side of the spectrum and the other was easily classified as speech.

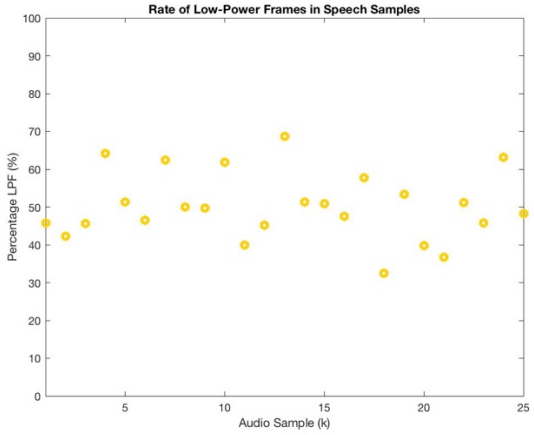
Figure 7 shows data from a music recording. The number of Low-Power frames is significantly lower than the number of Low-Power frames resulting from the speech signal in Figure 8. This feature (LPF) is very indicative of the nature of the signal.

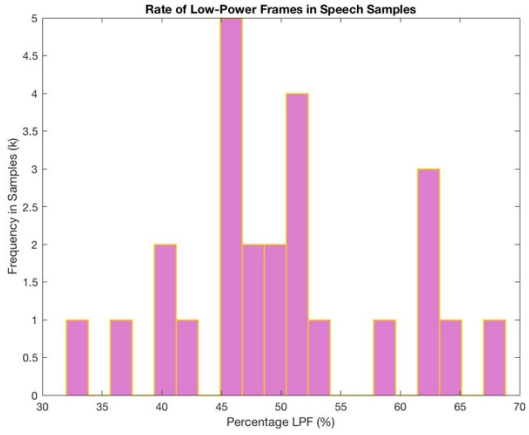
Figure 9:





$$\overline{LPF_m} = 22.6844$$





$$\overline{LPF_s} = 50.0533$$

In the histogram of LPF for music samples, the bins are of size ~1.85 (%LPF). In the histogram of LPF for speech samples, the bin size is ~1.8 (%LPF).

As shown in the plots above, the percentage of Low-Power Frames in music signals is generally lower than those of speech.

Signal Classification

As concluded from the analyses above, the two values that are most indicative of the nature of the audio signal are the ZCRV and the LPF. To classify an unknown signal, I will start to develop the thresholds of each type with respect to these two features.

Zero-Crossing-Rate Variation:

$STD(ZCRV_m) = 19.4690$
$STD(ZCRV_s) = 32.1884$

- The majority of ZCRV in music samples lie below ~50.
- The speech samples do not have as consistent a cutoff with this signal feature, however the clear one is a ZCRV > ~10.
    - There are still many music samples with ZCRV's that fall below this cutoff, so can still be helpful in the Signal Classification to distinguish between the two types of signals.
- As shown with the standard deviation calculations above, the ZCRV values are not as distinct between the two types of signals. Not only is the standard deviation large, but the ZCRV values can also fall within the range of the mean ± standard deviation.

Percentage Low-Power Frames:

$STD(LPF_m) = 10.8204$
$STD(LPF_s) = 9.0033$

- The standard deviation values show much more promise in the LPF feature when it comes to classifying unknown signals. The deviations are much lower than those of the ZCRV and the window of mean ± standard deviation for the two types of signals do not overlap:

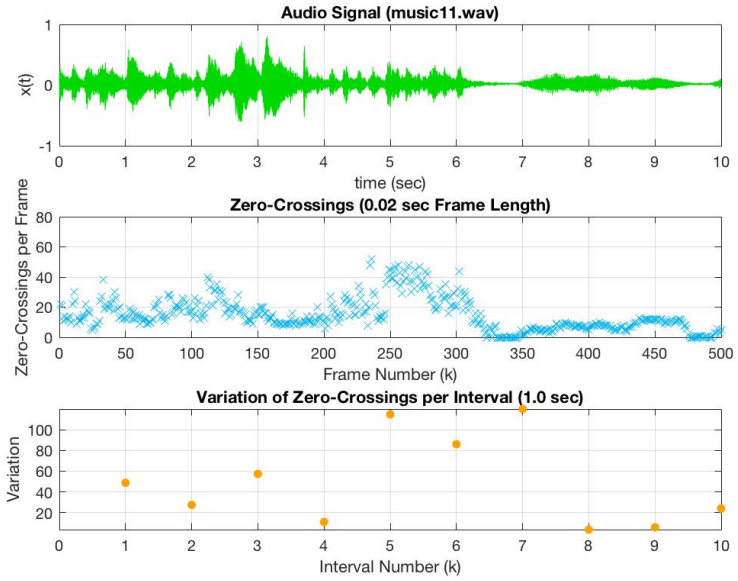
$$\overline{LPF_m} + STD(LPF_m) < \overline{LPF_s} - STD(LPF_s)$$
$$22.6844 + 10.8204 = 33.5048 < 41.0203 = 50.0533 - 9.0033$$

➔ The Signal Classification can use these values to accurately determine the source of the signal

- The histograms also provided good insight as to where to draw the cutoffs. Based on the data derived from the training samples, a signal with a %LPF > 45 is likely to be a speech recording. Similarly, a ZCRV > 70 was also likely to be speech. (The contrary does not hold however.)

Examples of Corner Cases:

speech21.wav made the cutoff %LPF of 36.5, but this low range value could not be adjusted any higher to yield more accurate classifications of music signals, because the %LPF of this recording was only marginally higher. This audio file was examined more closely to determine the cutoff.
speech18.wav had the lowest %LPF, but amongst the highest ZCRV so it did not fall into this problem.
However, music11.wav proved to be very problematic. Its ZCRV data is shown below (very characteristic of a speech recording). And it did not have a favorable %LPF value, as was the case for speech18.wav for the classifier function to characterize it accurately.



All in all, the classifier had 100% accuracy on the training speech audio files and all of the test recordings. It was 88% accurate with the sample music files given the 2 or 3 that had very similar characteristics to speech recordings and fell in the corner cases.

<u>Neither vocal nor instrumental:</u>

While the SoundClassifier.m function will only return a 0 or a 1, some educated guesses as to how the “undefined” signal will be classified can be made.

- The else is a 1. Therefore, there is potentially a larger window for the signal to fall in and be classified as a music recording.
- This, of course, relies on the “undefined” signal not having characteristics of a vocal/instrumental audio signal.
- If the “undefined” signal had properties that fell between those of music and speech, it is more likely for the function to classify the signal as speech because the window was more generous for this type of file.

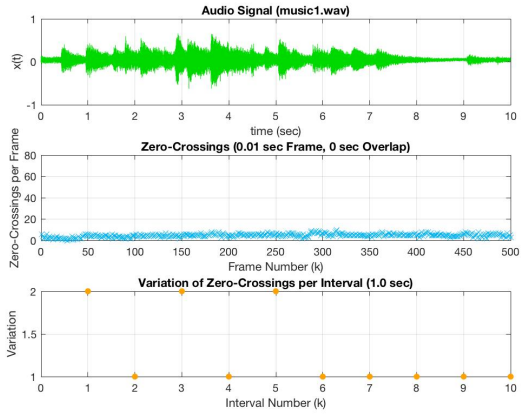
A good visual representation of this is as follows:

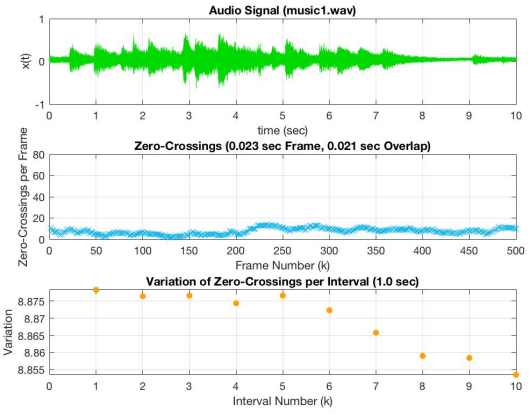
The spectrum above demonstrates the classification of music, speech, and other files. Music identification is represented by the yellow portion, speech by the green, and "other" by the blue. If the audio file has characteristics that fall in the yellow segment, it will be ranked a "1" and therefore classified as a music sample. If the audio file has any shared characteristics to music and speech (but is still neither), it will be classified as speech, because this window is much more forgiving/generous in the Matlab code. Anything outside of this falls in the blue section, which will be given a "1" and therefore be considered a music recording.

The colors form a gradient because there are of course samples that have characteristics which fall in both categories, so the boundaries are not black and white.

Bonus Section

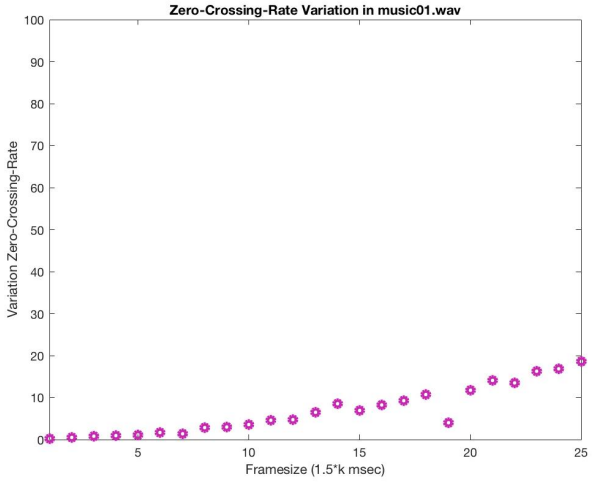
I rewrote some of the Matlab code to adjust the graphs properly for different frame sizes. This required another function "OverlapFunction.m", which determines how much overlap should be granted to each frame such that the audio sample can be divided into even parts. It was very painstaking, but the results allow for a better analysis of the effects frame lengths and overlap percentages on the signal features.





I found that the larger the frame size was, the more variation of Zero-Crossings there would be. Below is a plot of the ZCRV values for just one music sample, with varying frame size. It appears

that the amount of overlap does not affect the variation as much as the size of the frame itself.



Below is the plot of LPF values for varying framesize of the same audio sample. In this case, there is not a clear trend, which leads me to believe that the amount of overlap might be the cause of the variations.