CS578 Speech Processing
Laboratory 0: Part 1
Time Domain Speech Processing

George Manos
csd4333@csd.uoc.gr

Alexandros Angelakis
csd4334@csd.uoc.gr

29 October 2022

1 Implementation

1.1 Details

Using Matlab we created a VUS discriminator that can classify speech segment in either voiced/unvoiced/silence classes with a pretty good rate. Even though VUS is a real time system, for the purpose of this assignment we will consider a non-real time approach.

The discriminator results are represented like the example in Figure 1, where the blue line will correspond to our speech segment classification (1 is for voiced segments, 0.5 for unvoiced and 0 for silence respectively).

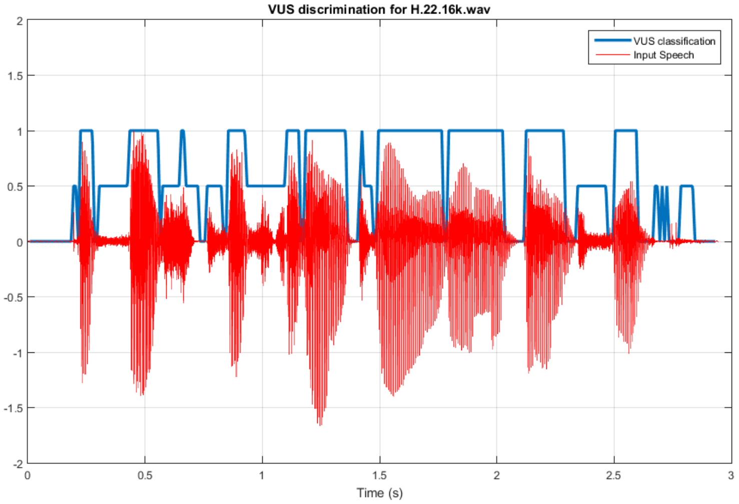


Figure 1: 'The fish twisted and turned on the bent hook'

For our VUS discriminator, we also have to define the frame length, frame shift and the number of frames. Frame length for $t_l = 30$ ms will be equal to $L = f_s \cdot t_l$, frame shift respectively for $t_s = 10$ ms will be equal to $U = f_s \cdot t_s$. For those 2 values, to calculate the total number of frames we can think of the frame as a convolution window of size L that will be applied over a signal of size D with a stride = U, so the formula to calculate it will be $Nfr = \frac{D-L}{U} + 1$.

For each one of the frames, we extract the sub-section of our signal within the bounds $[(i-1)U+1,(i-1)U+L]$, where i corresponds to the index of the $i-th$ window and then multiply that with our Hamming window. Then for that frame result, we calculate energy and zero-crossings using the formulas provided.

Finally, given scalar threshold values for energy and zero crossings e_t, zc_t respectively, our VUS discriminator works as following:

$$VUS[frame] = \begin{cases} 1, & \text{if frame energy} > e_t \\ 0.5, & \text{if frame energy} < e_t \text{ and frame zero crossing} > zc_t \\ 0, & \text{else} \end{cases}$$

On Figure 1, $e_t = \frac{mean(energies)}{2} = 7.64\text{e}{-}04$ and $zc_t = (\frac{3}{2}mean(ZCr) - 0.3std(ZCr)) = 87.9767$ where ZCr is the calculated ZCr value for each frame.

1.2 Experiments

1.2.1 Silence

In order to determine if a speech frame is a silence frame, we can use the thresholds mentioned above, e_t and zc_t. If a frame has less energy than the energy's threshold e_t and less zero crossings than the zero crossings threshold zc_t, then we can easily say that this is a silence frame. This works because we know that in silence frames, there are no big variations of the signal, thus the energy should be less than both voiced and unvoiced frame and the zero-crossings should be also less than the other two cases.

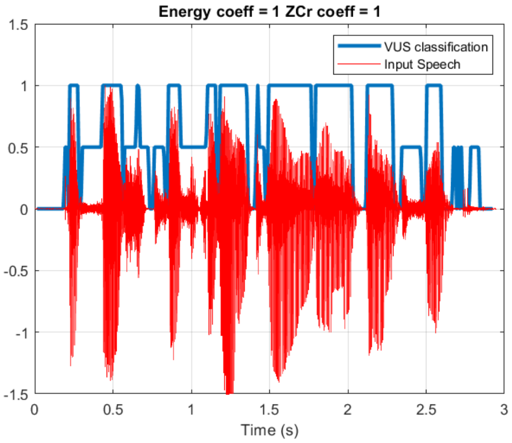
1.2.2 Thresholds

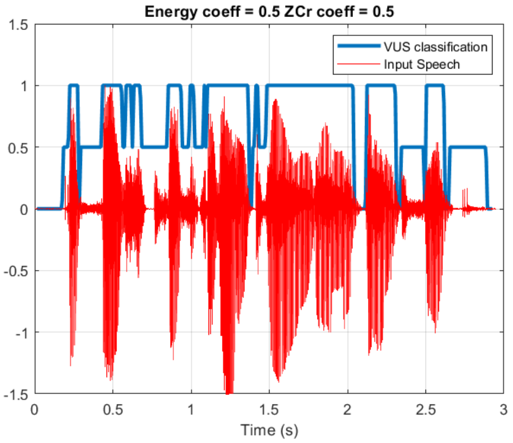
On the example speech signal, we multiplied the preset energy and zero crossings thresholds with a coefficient to test how the results are affected. The way we set the experiment up, we would expect that increasing the Energy threshold will decrease the discriminator's sensitivity to voiced signals and vice versa, while increasing the zero crossings threshold will also decrease the unvoiced sensitivity.

Indeed, through our experiments we can see that doubling both thresholds results into more speech parts classified as silenced as the boundaries to classify either voiced or unvoiced are more strict. On the other hand, reducing both thresholds by half results into less parts being classified as silenced, while even more parts previously classified as unvoiced are now considered voiced.

Reducing only the energy coefficient by half with steady zero-crossings coefficient, we can see that while the silenced parts are exactly the same, many parts previously classified as unvoiced are now considered voiced. On the other hand, reducing the zero-crossings coefficient by half with the same energy coefficient, the parts classified as voiced are the same while more parts previously classified as silenced are now considered unvoiced.

The experiment for doubling 1 threshold without affecting the other would have inverse results as described above, but we did not include the plots into this report as they were too many (rule of thumb ☺)





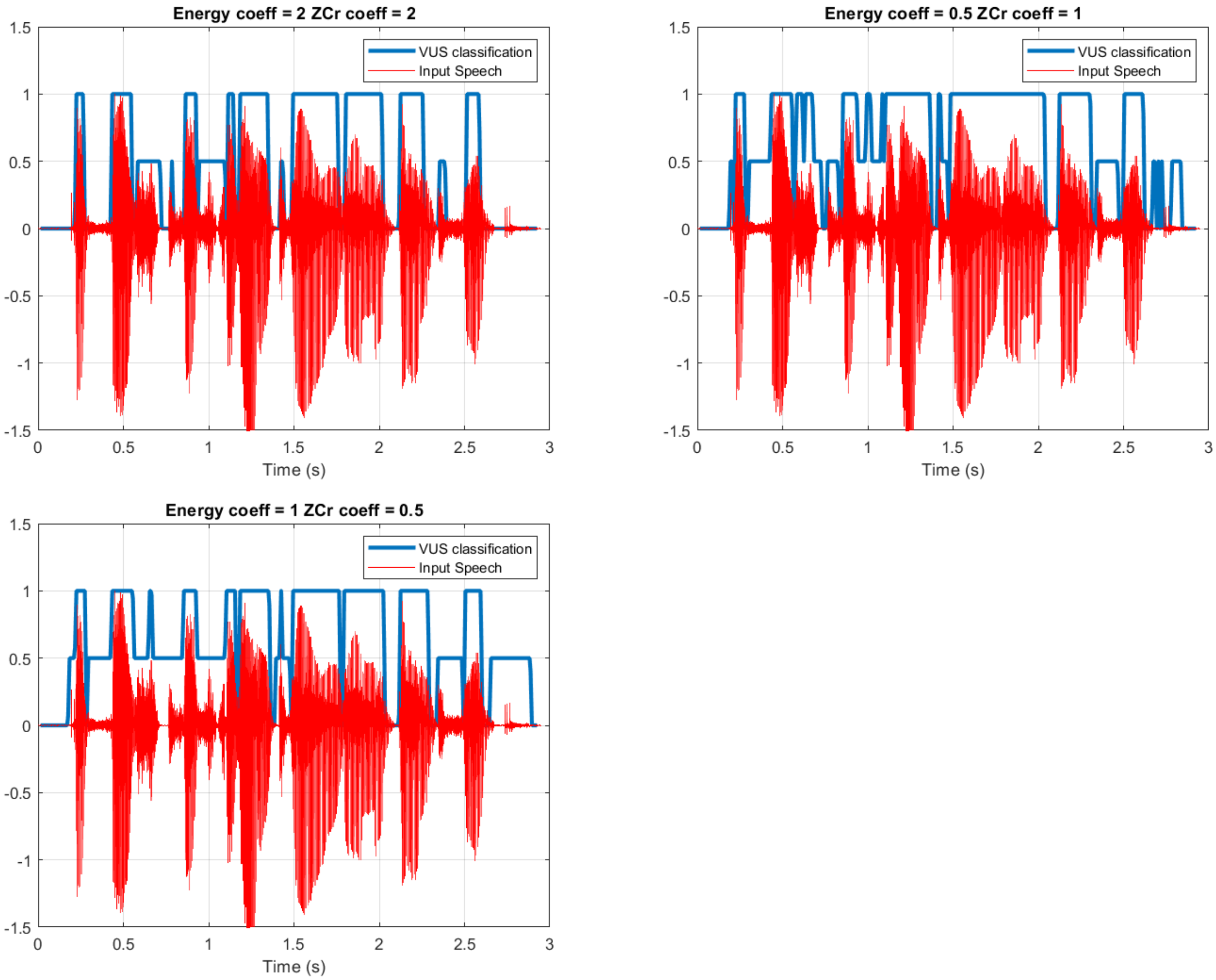


Figure 2: Experimenting with energy and zero crossings thresholds for the same signal

1.2.3 Analysis Window and Frame Rate

For the purpose of this assignment, we will refer to stride as the distance between each frame. This means that as stride value increases, the actual VUS frame rate decreases.

We would expect that decreasing the stride value would decrease the number of VUS points created. As a result of this we should be seeing less classification differences - longer straight straight blue lines. This would also mean that we would be seeing less of those steep drops on the blue line.

On the other hand, increasing just the frame length, we expect to see more smooth patterns with less steep drops on the blue line, as many silence parts would be included to the high energy voiced part, hence being classified as voiced.

Increasing both frame length and stride would result into less VUS centers, hence less sensitivity to those steep drops to the silenced class. Decreasing those 2 parameters would result into the opposite behavior.

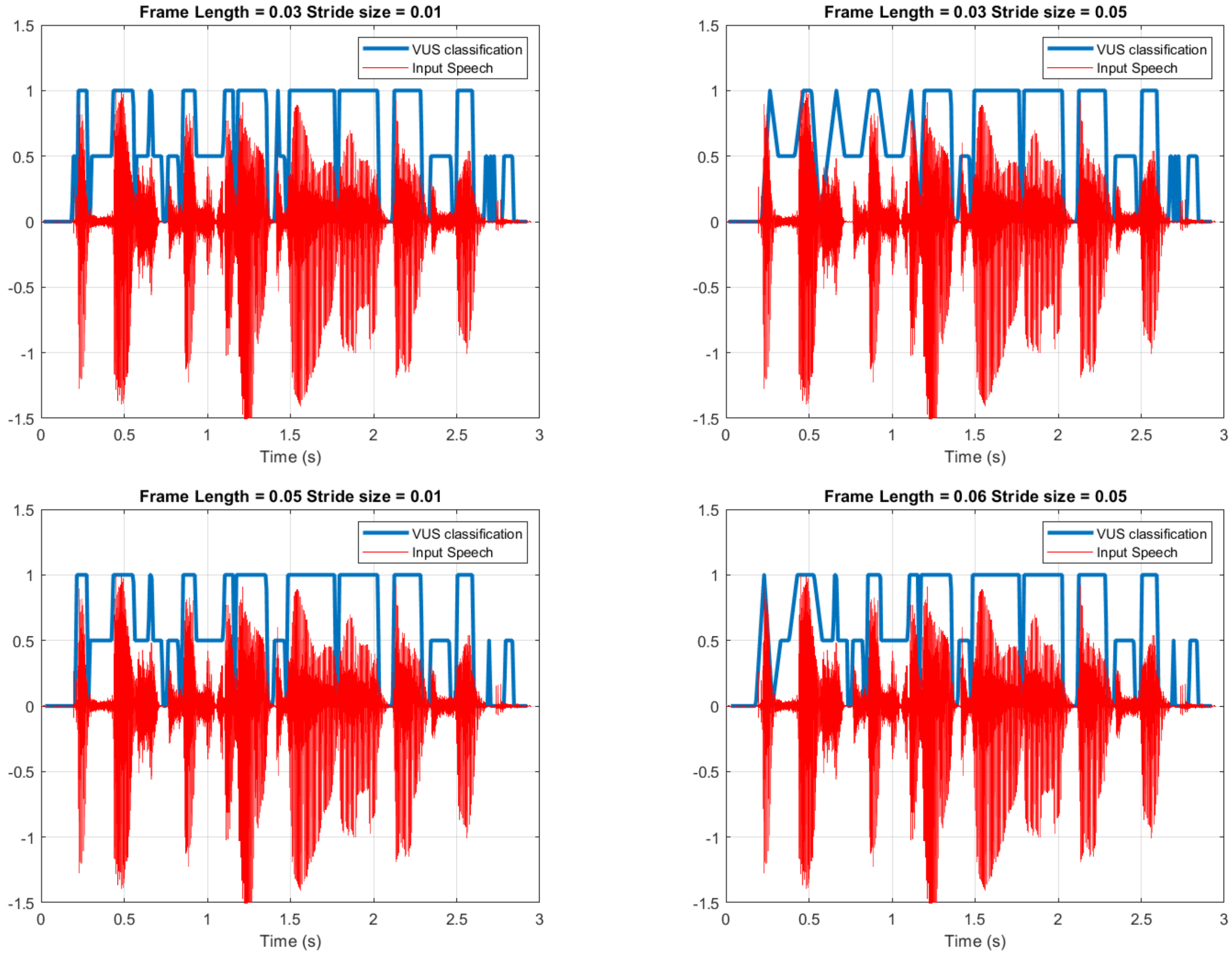


Figure 3: Experimenting with energy and zero crossings thresholds for the same signal

1.2.4 Interpolation

As we can see, by interpolating our VUS discrimator results over the whole speech waveform we can have different results in our final estimates. For all speech signal examples that are spoken sentences, linear and cubic interpolation appear to have the same results. As for the spline interpolation, we can see that the results look like Gibbs' phenomenon as it uses not-a-knot end conditions. The interpolated value at a query point is based on a cubic interpolation of the values at neighboring grid points in each respective dimension. On the other hand, cubic is a shape-preserving piecewise interpolation. The interpolated value at a query point is based on a shape-preserving piecewise cubic interpolation of the values at neighboring grid points. For those examples, we can choose either cubic or linear interpolation as the results appear more clean.

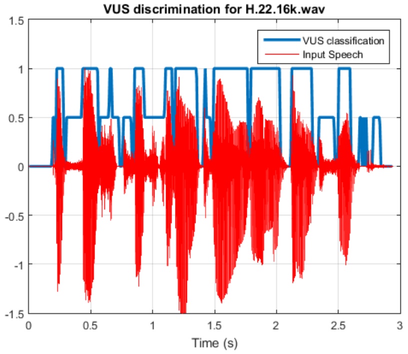


Figure 4: Linear Interpolation

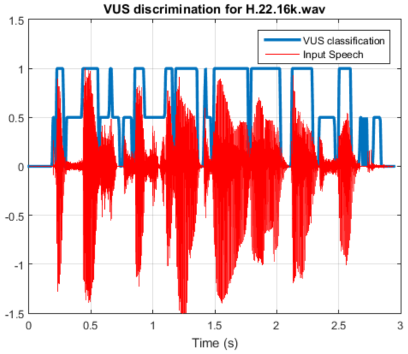


Figure 5: Cubic Interpolation

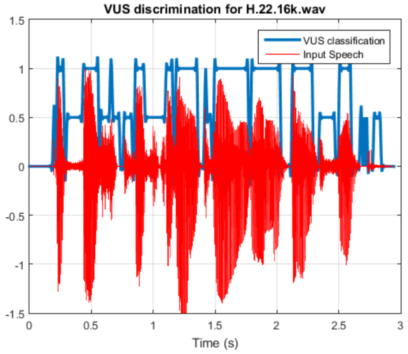


Figure 6: Spline Interpolation

For phoneme speech signals, every interpolation method has different results. Cubic interpolation has a curve before we classify the frame as a voiced one, starting from negative numbers. The same goes to spline interpolation, but there is no curve at the beginning of our discriminator and it starts from very low numbers. Although both cubic and spline interpolation methods work pretty well for the most part of the signal, we will choose the linear one as it keeps the line within the classification boundaries (0, 0.5, 1).

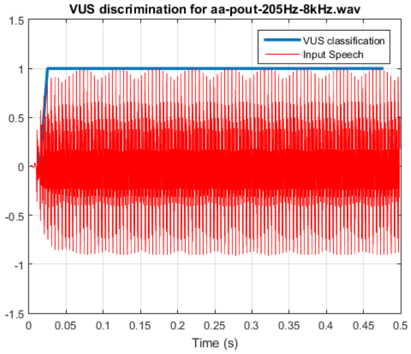


Figure 7: Linear Interpolation

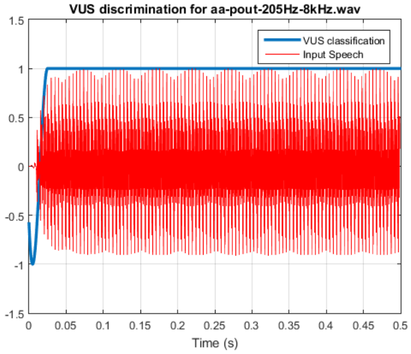


Figure 8: Cubic Interpolation

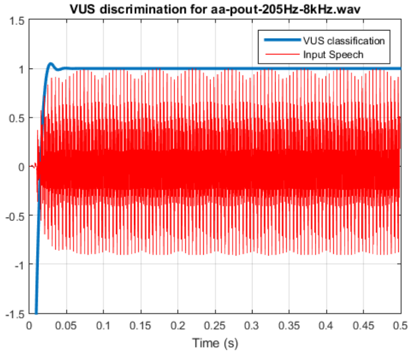


Figure 9: Spline Interpolation

1.2.5 Speaking Styles

On first view, we don't notice any major differences between the signals themselves, apart from the fact that the left, where the speaker is more stressed, seems more dense than the right one.

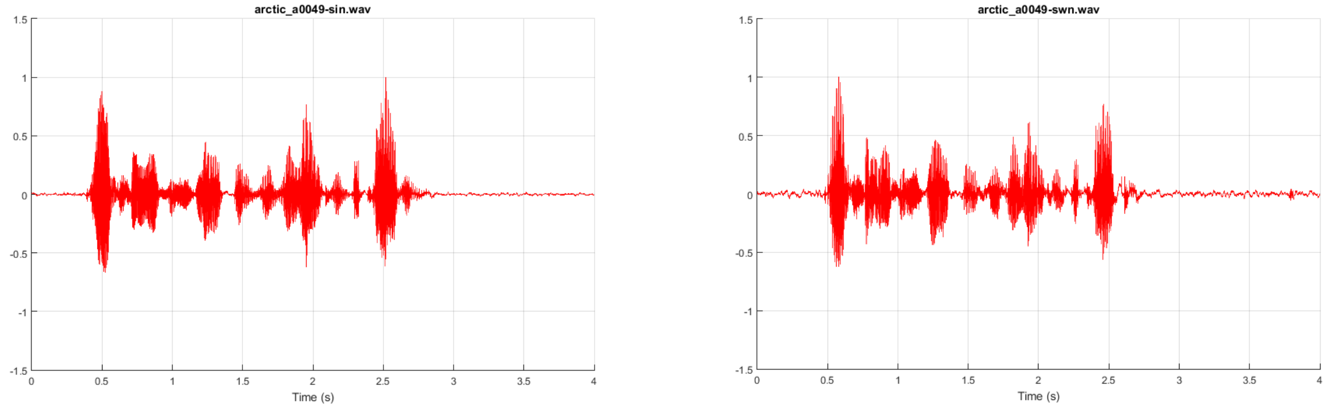


Figure 10: Speech signals with the same content, but the left has a more stressed speaking style and the right is speaking to a quiet place.

However, the differences appear when you plot the energy and zero-crossings distributions. As one can see, the energy distribution for the signal, although the patterns seem almost identical, the scale on y-axis is 10 times lower, meaning that the short-time energy in the stressed speaker speech signal is almost 10 times more than the short time energy for the not-stressed speaker speech signal.

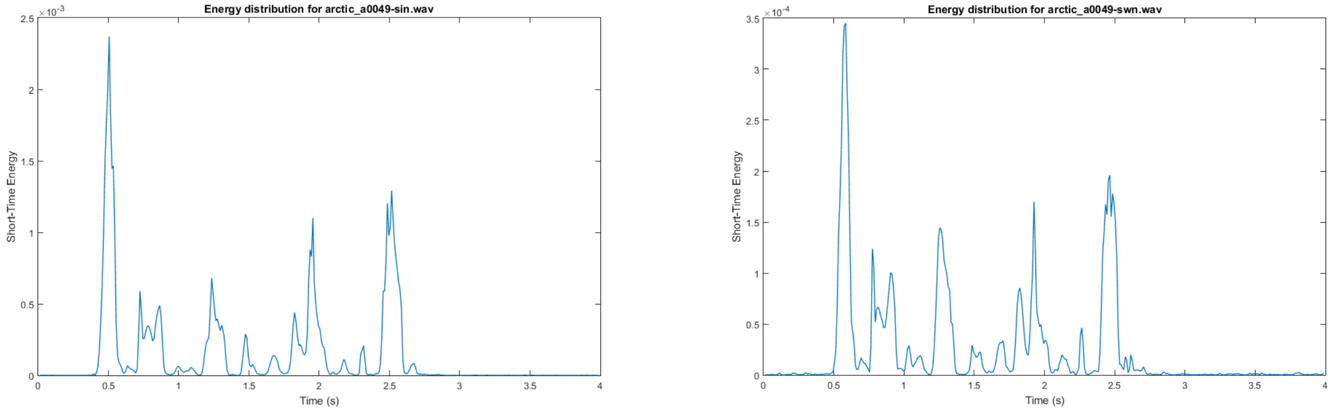


Figure 11: Energy distributions for both stressed and calm speaker.

On the Zero-crossings plots, the plots are again pretty similar but one can observe more zero crossings on the stressed speaker signal as the scale again is different. The major peaks are also following different patterns around the 1st second of the time signal. This is exactly the part where the speaker says ”Gregson was asleep”. Through those plots, we could probably say that Lombard speech not only increases the energy and the zero crossings on speech, but it also changes the zero crossings pattern on fricative letters like ”s” (in ”Gregson”, ”gs” is pronounced more like a ”z”, same for the ”s” in ”was”, but this pattern also happens in ”asleep”).

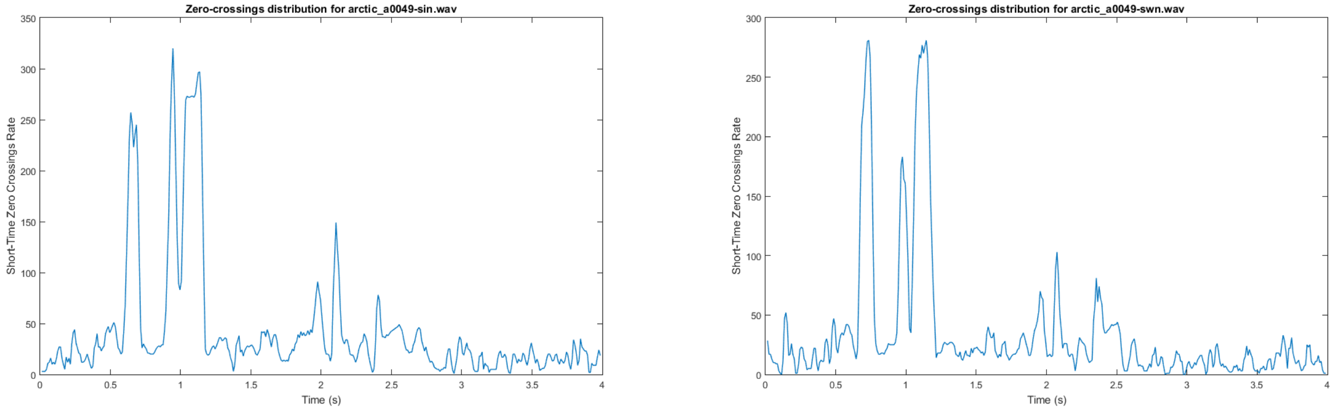


Figure 12: Zero-crossings distribution for both stressed and calm speaker.

2 VUS Discriminator Results

Visually, we would say that our VUS has achieved some pretty great results in classification. We can see that it correctly classifies most voiced samples, while the main issues it faces is with unvoiced and silence. For the short time high sampled speech signals we can see that it classifies almost all samples as voiced, which is the case, where we notice almost totally straight lines. In those signals, there is silence in the first 0.02 seconds so our discriminator classifies these frames as silence, and then it jumps up to 1, classifying the signal as a voiced one till the end of the classification. There is also one voiced sample (/eh/) where the discriminator is 1 from the beginning to the end.

However, this behavior changes as we experiment with the aforementioned hyperparameters and thresholds, not necessarily for the better.

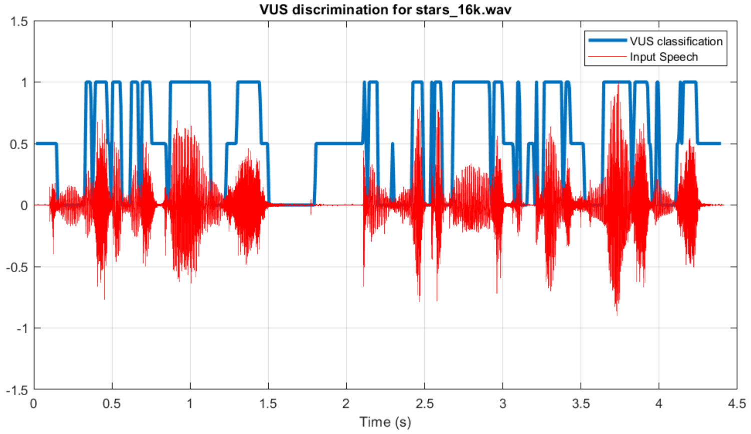


Figure 13: 'Can you see the stars? Can you recognize the constellations?' Recorded by George Manos

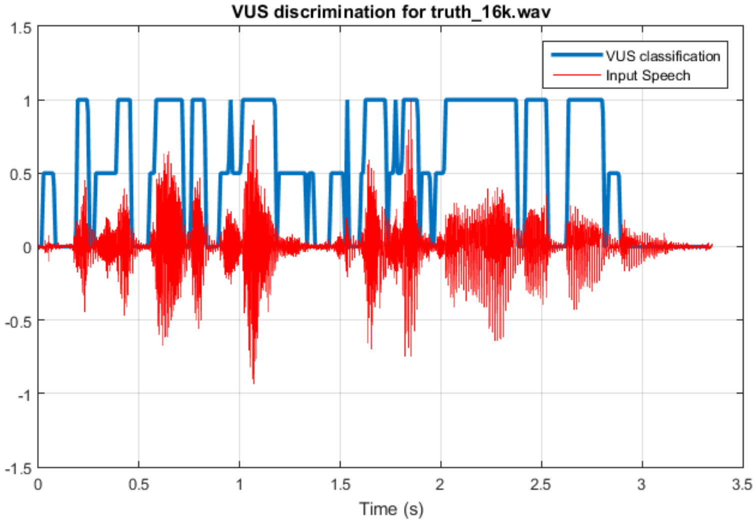


Figure 14: 'If you tell the truth, you don't have to remember anything.' Recorded by Alexandros Angelakis

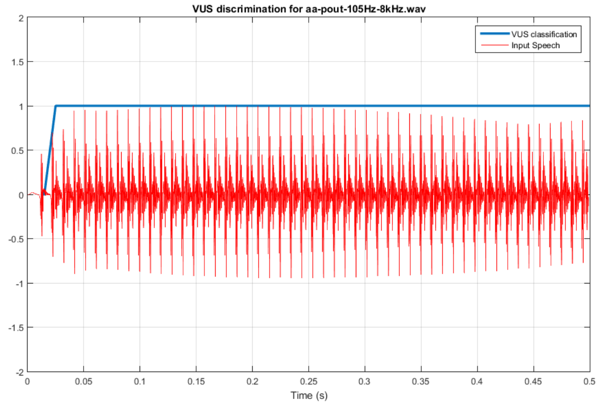


Figure 15: /aa/ speech signal

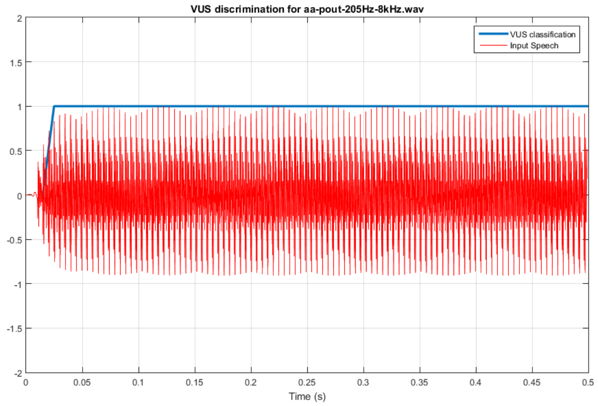


Figure 16: /aa/ speech signal

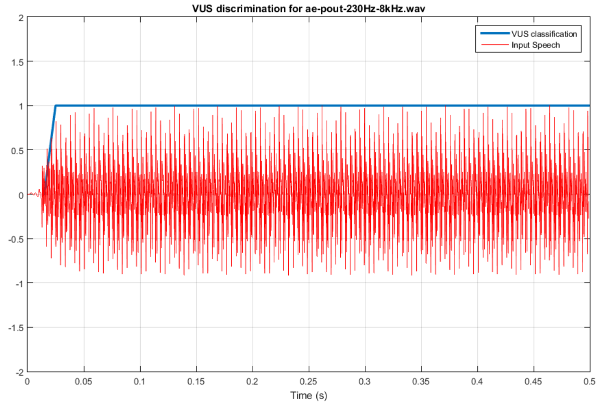


Figure 17: /ae/ speech signal

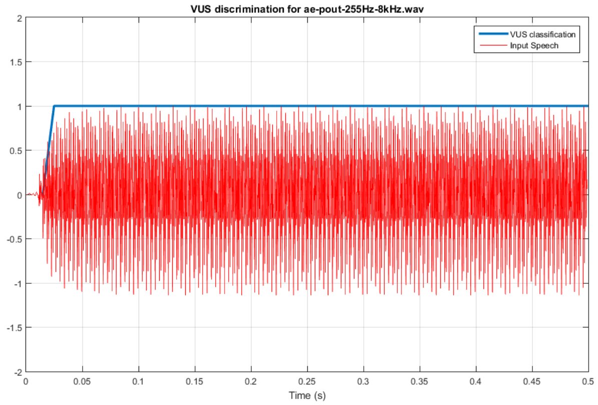


Figure 18: /ae/ speech signal

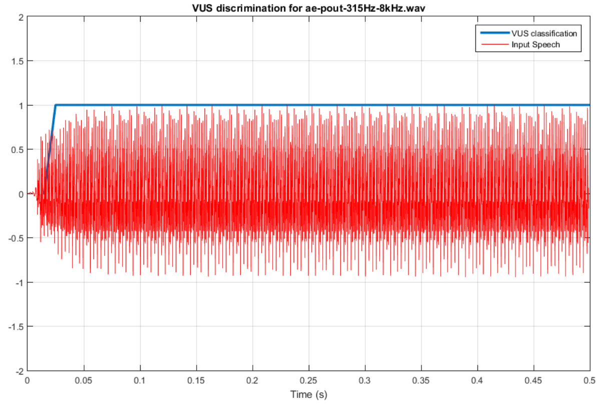


Figure 19: /ae/ speech signal

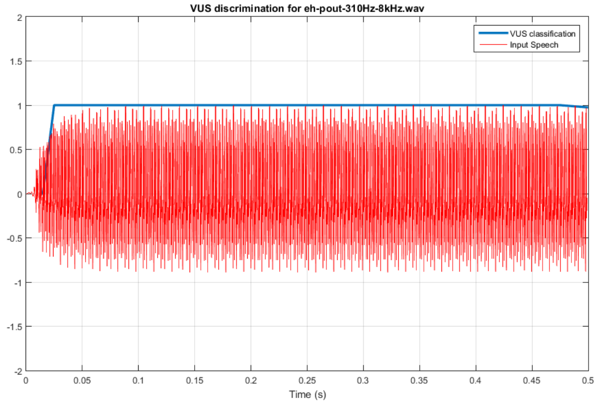


Figure 20: /eh/ speech signal

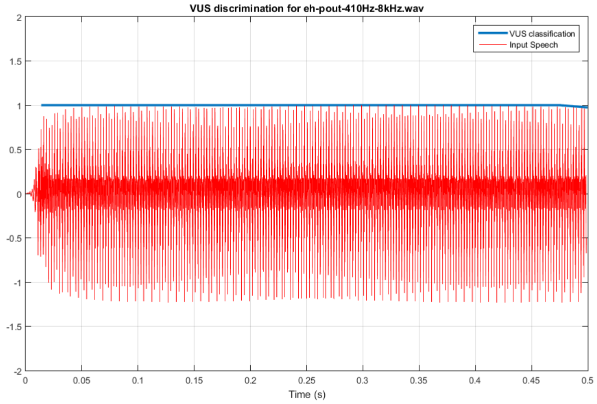


Figure 21: /eh/ speech signal

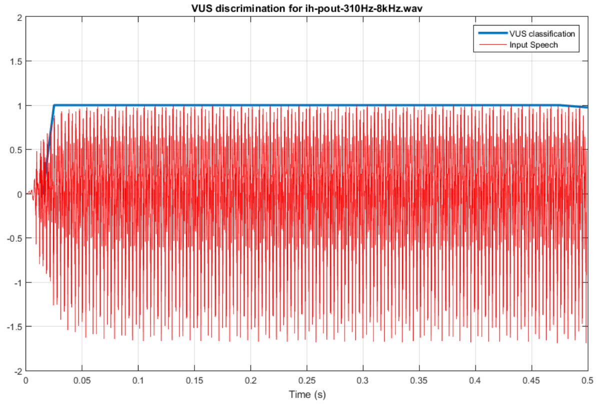


Figure 22: /ih/ speech signal

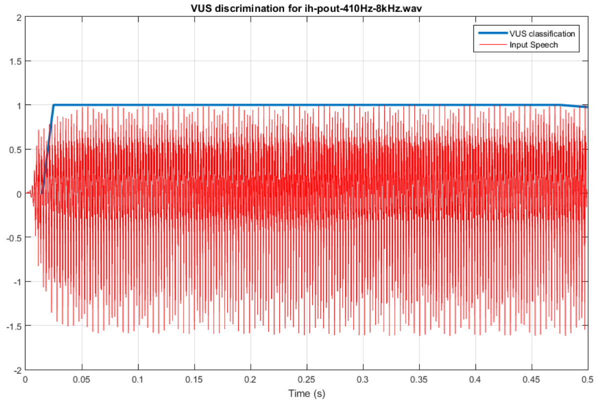


Figure 23: /ih/ speech signal

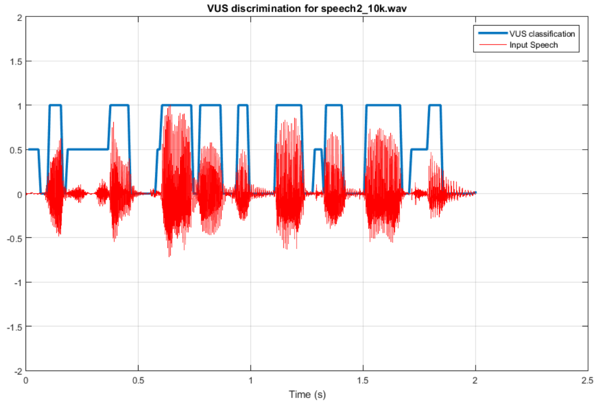


Figure 24: ’Which tea party did Baker go to?’

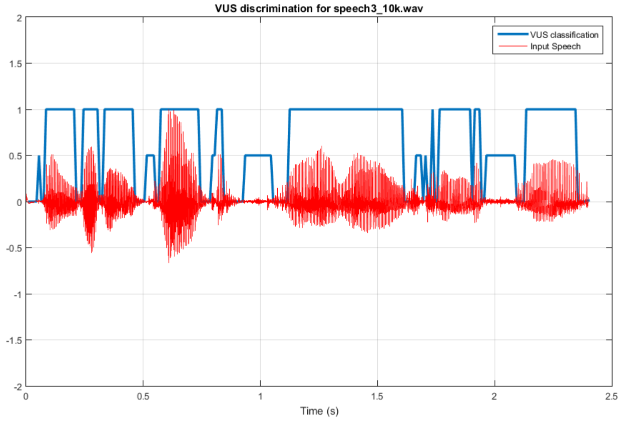


Figure 25: ’A little blanket laid around on the floor’

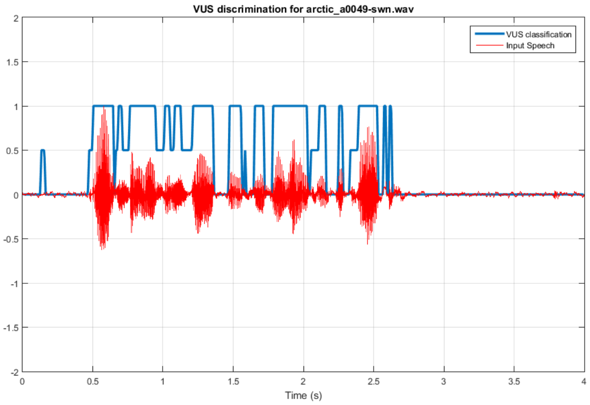


Figure 26: 'Gregson was asleep when he re-entered the cabin (quiet place)'

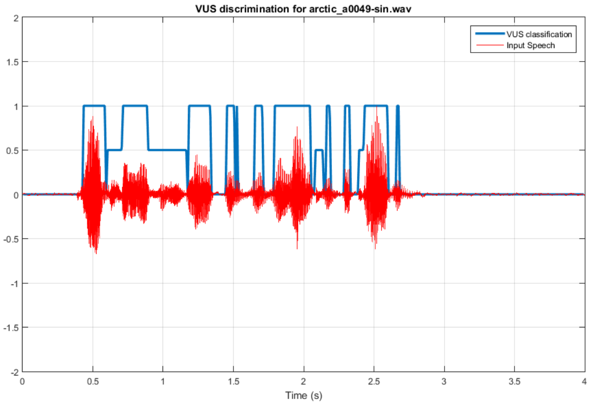


Figure 27: 'Gregson was asleep when he re-entered the cabin (noisy place)'

3 Appendix

The part we have all been waiting for: