Jack Cheng Ding Han (150159519, jcdhan1@sheffield.ac.uk)
Afolabi Nathan-Marsh (150159678, ARNathan-Marsh1@sheffield.ac.uk)

Report

Task Allocation

#	Task	Proportion
1	Formant Frequency Research (inc. Conclusions)	Jack
2	Formant Frequency Estimator Implementation	Jack
3	Formant Frequency Estimation (Live)	50% Jack, 50% Afolabi
4	Formant Frequency Estimation (Recorded)	Jack
5	Voiced/Unvoiced Segment Duration Research (inc. Conclusions)	Afolabi
6	Voiced/Unvoiced Segment Duration Implementation	Afolabi
7	Voiced/Unvoiced Segment Duration Estimation (Live)	50% Jack, 50% Afolabi
8	Voiced/Unvoiced Segment Duration Estimation (Recorded)	50% Jack, 50% Afolabi
9	Statistical Code	Jack
10	Graphical User Interface	Jack
11	Report: Estimation Principles	Afolabi
12	Report: Implementation	Jack
13	Report: Recording Results	50% Jack, 50% Afolabi
14	Report: Conclusions	50% Jack, 50% Afolabi

Jack: 17/28 ≈ 60.71%

Afolabi: 11/28 8≈39.29%

Principles Used for Estimating Different Parameters

First and Second Formants

The input signal is filtered between two ranges:

300 to 700Hz for F1, 800 to 2400Hz for F2.

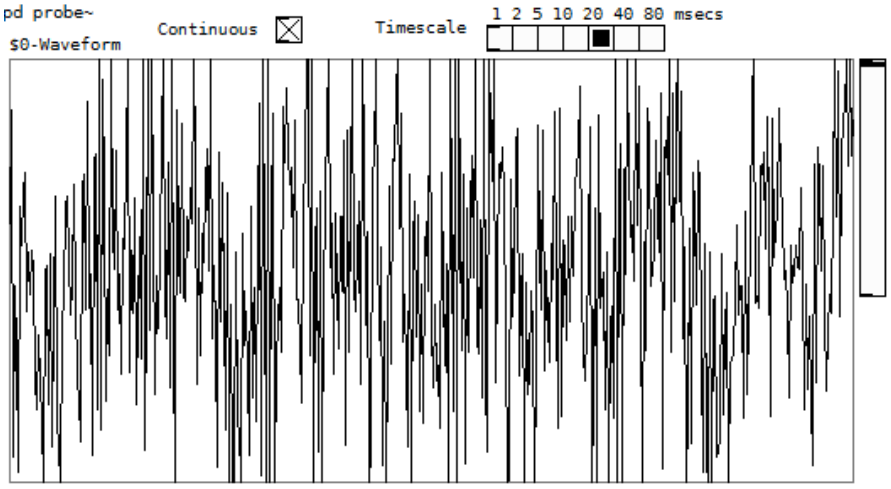
For each formant frequency, the ratio between the maximum amplitudes of frequencies that haven't been attenuated is calculated and the result is passed through a linear equation to retrieve the formant frequency.

Voiced and Unvoiced Segment Durations

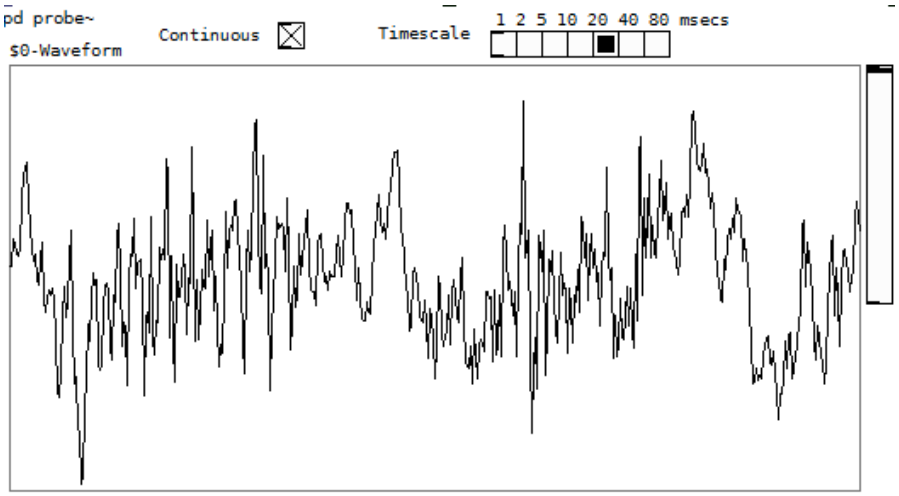
To detect when a segment starts and ends, we check to see if the volume is crossing a certain threshold (which we empirically found to be -12dB) with a positive gradient (start) or a negative gradient (end). To determine whether a segment is voiced or unvoiced, we considered two approaches; measuring the energy of the signal or measuring the number of zero-crossings in a certain interval, we ultimately found the latter to be easier to implement.

Jack Cheng Ding Han (150159519, jcdhan1@sheffield.ac.uk)
Afolabi Nathan-Marsh (150159678, ARNathan-Marsh1@sheffield.ac.uk)
Below are screenshots of the amplitude-time graph within Example 3-2 live-voice-analysis, demonstrating the difference between voiced and unvoiced sounds:

Voiceless labiodental fricative /f/



Voiced labiodental fricative /v/



Voiced sounds possess high short-term energy and their signals crosses the time-axis fewer times. On the other hand, voiceless sounds have low short-term energy and their signals cross the time-axis more [1].

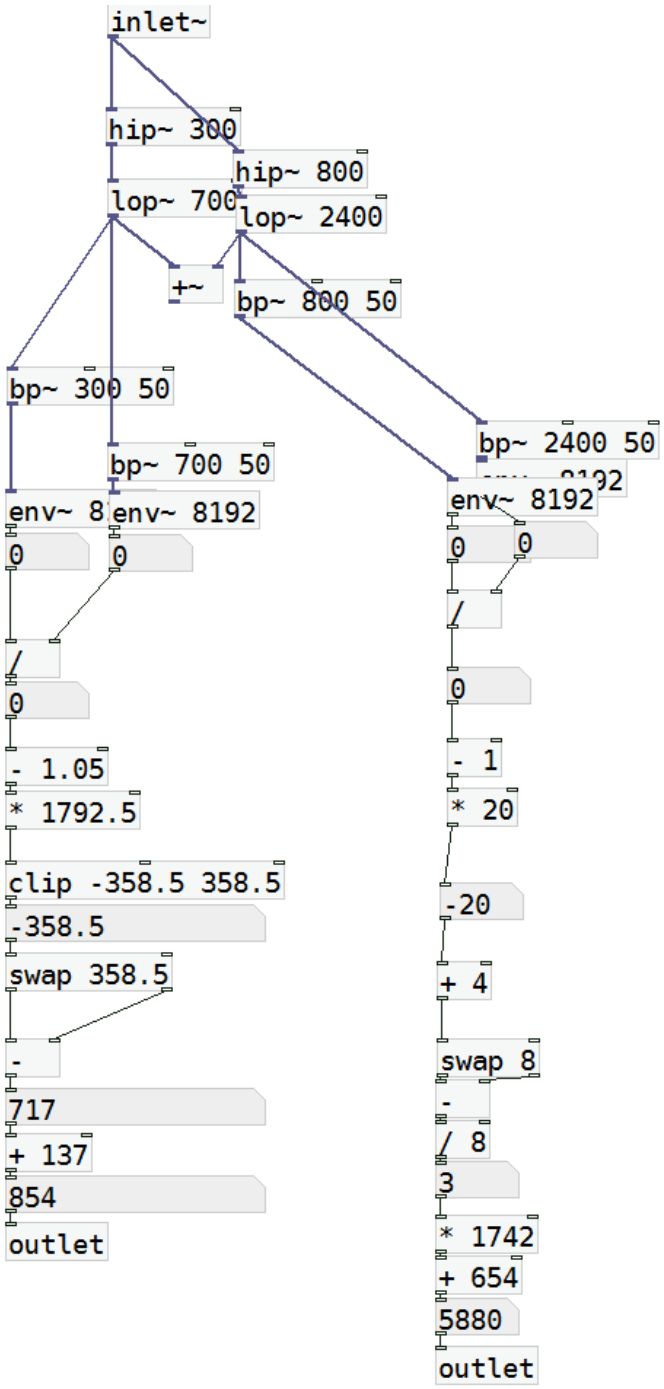
The graphs show that the signal for /f/ crosses the horizontal time axis more often than the signal for /v/.

We determined the threshold for unvoiced segments to be 170 zero-crossings as this was the lowest number of zero-crossings for an unvoiced consonant, which we empirically found to be /sh/.

Jack Cheng Ding Han (150159519, jcdhan1@sheffield.ac.uk)
Afolabi Nathan-Marsh (150159678, ARNathan-Marsh1@sheffield.ac.uk)

Implementation

Formant Frequency Estimation

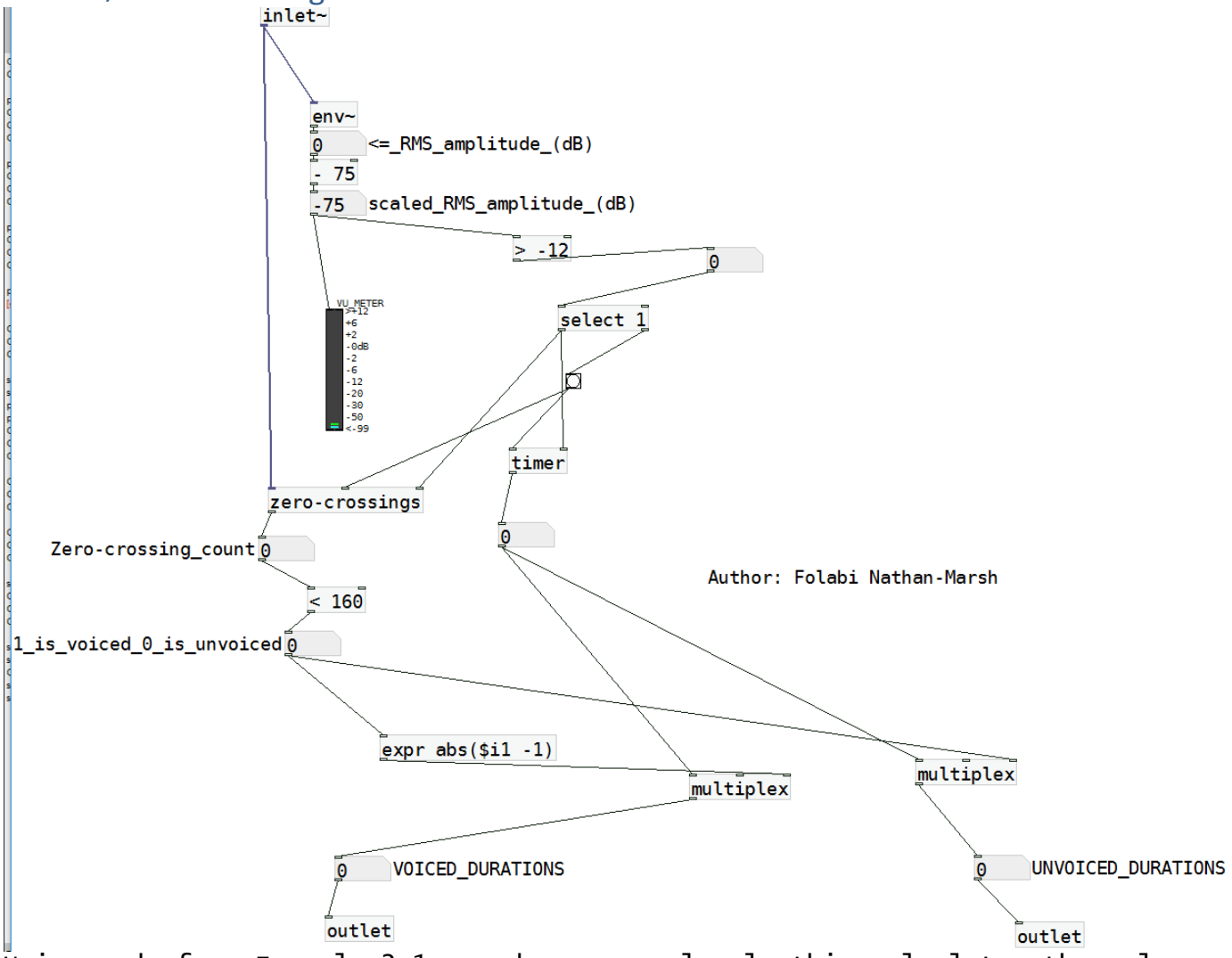


Jack used code from Example 8-2 formant tracker; High-pass, low-pass and band-pass filters are used to get frequencies within the ranges mentioned in the previous section. The maximum amplitudes we need to calculate ratios for are retrieved using the envelope object.

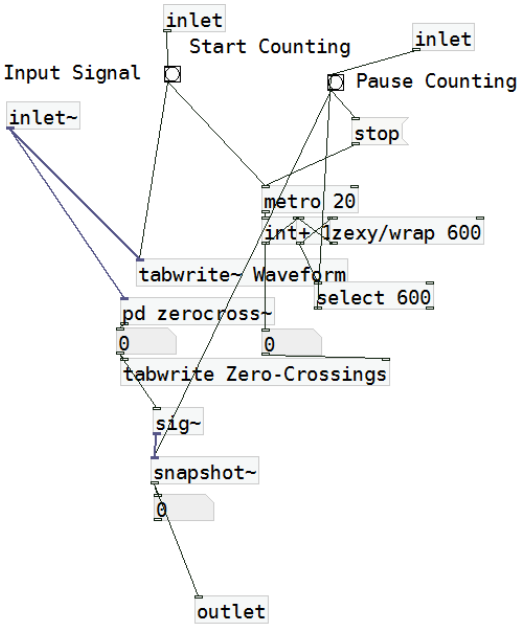
We studied code from Example 8-1 vowel-quadrilateral to find out how to map ratios to frequencies. We implement this mapping from ratios to frequencies using multiplication, addition, subtraction and clipping objects.

Jack Cheng Ding Han (150159519, jcdhan1@sheffield.ac.uk)
Afolabi Nathan-Marsh (150159678, ARNathan-Marsh1@sheffield.ac.uk)

Voiced/Unvoiced Segment Duration Estimation



Using code from Example 2-1 sound pressure level, this calculates the volume of the incoming signal. If it is greater than the threshold -12 dB, a bang is sent to start a timer object, if the signal goes below -12 dB again, the timer stops counting.



Author: Folabi Nathan-Marsh

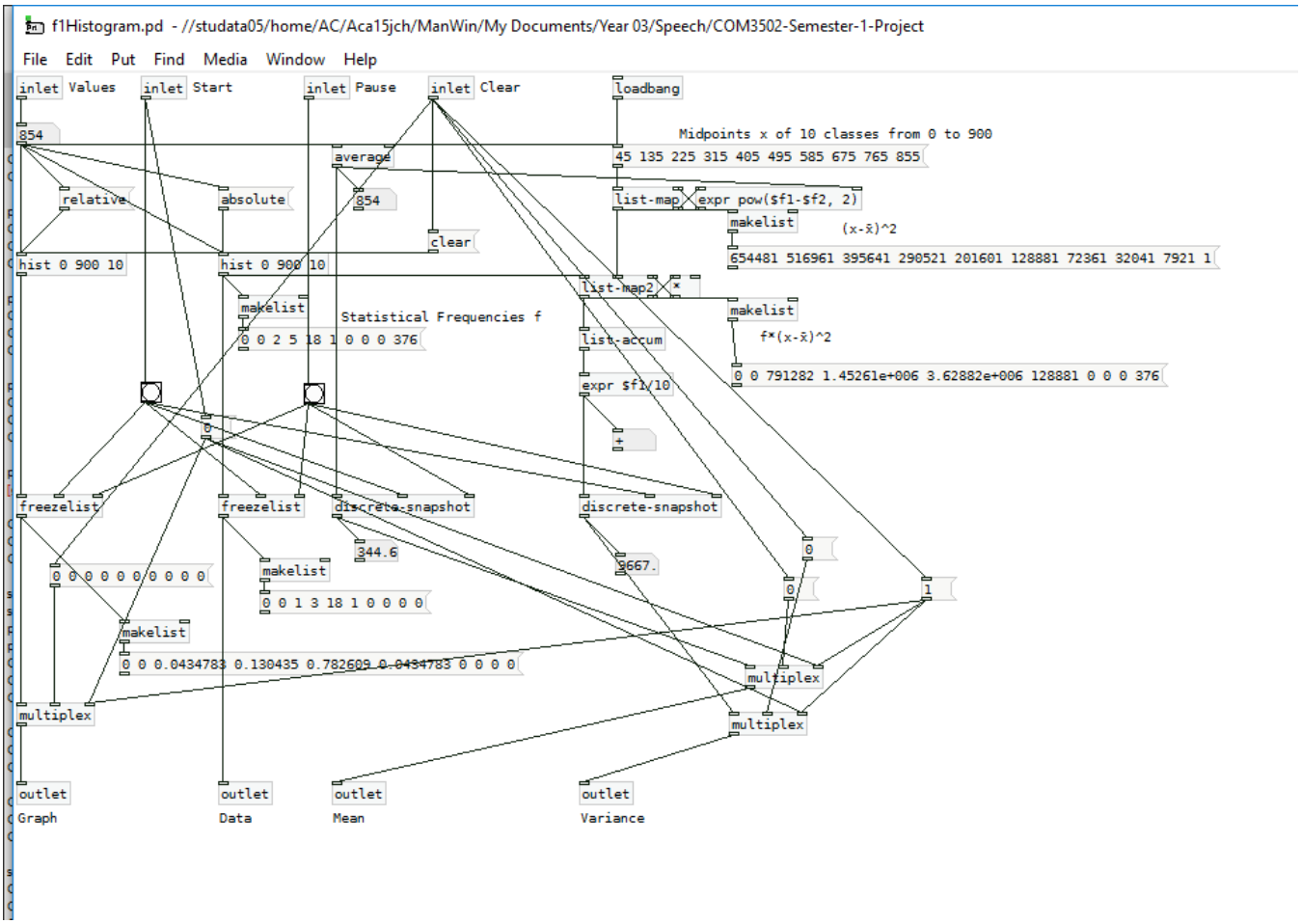
When the timer begins, it sends a bang to the zero-crossings object Folabi wrote to start counting zero crossings, when it ends, the zero-crossings object outputs the number of zero crossings it counted in that time-window.

Folabi modified code from Example 14-3 zero-crossings. His additions to it include the ability to send bangs to start counting and pause counting which is triggered by our code for determining when a segment starts and when a segment ends. It sends out a value of either 0 to 1, if it is 1, the duration from the timer object is sent to the leftmost outlet for voiced durations. If it is 0, the duration is sent to the rightmost outlet for unvoiced durations.

Jack Cheng Ding Han (150159519, jcdhan1@sheffield.ac.uk)
Afolabi Nathan-Marsh (150159678, ARNathan-Marsh1@sheffield.ac.uk)

Statistical Code

Jack wrote code for producing histograms, means and standard deviations for all four parameters. They are all identical, only differing in the range of data and the list of midpoints. Here is the code for First Formant.



The first inlet is a constantly changing number outputted by the estimators. The second inlet starts measuring, the second inlet freezes the outputs, the third inlet freezes the outputs as well as clear them to 0.

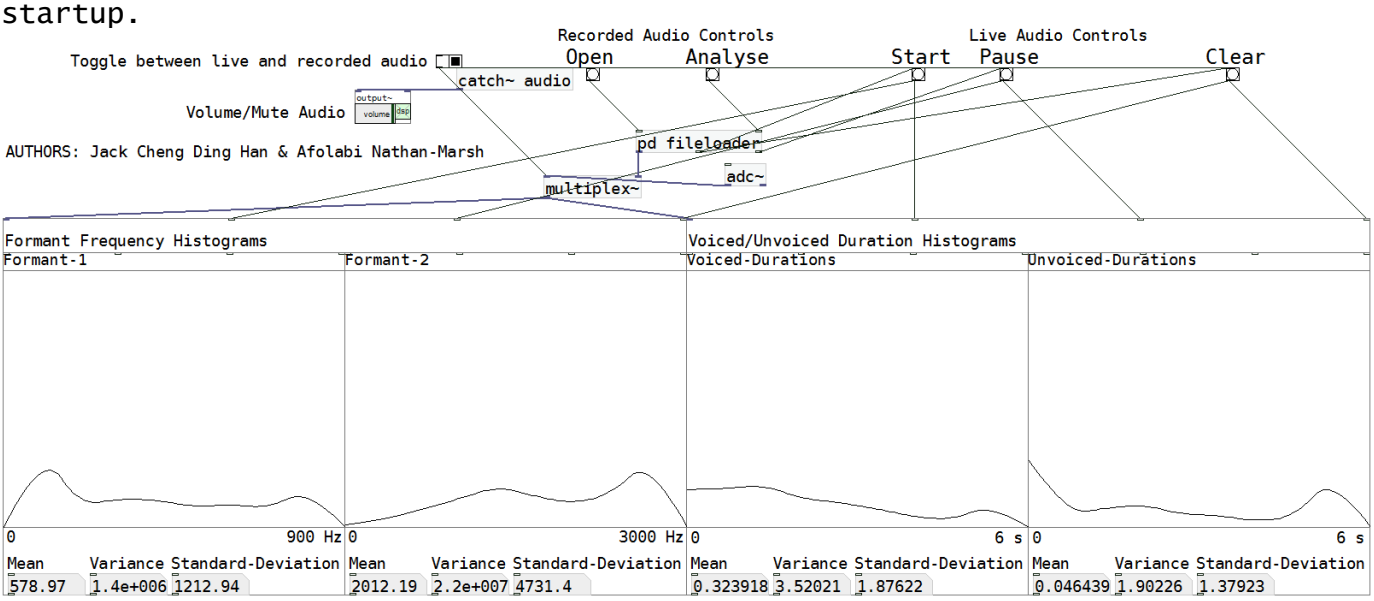
The outlets are: a list that can be displayed in an array in the GUI, the distribution itself for testing purposes, the mean and the variance (later on in the GUI, passed through an expression object to return the standard deviation)

The objects makelist, freezelist and discrete-snapshot were all written by Jack.

Jack Cheng Ding Han (150159519, jcdhan1@sheffield.ac.uk)
Afolabi Nathan-Marsh (150159678, ARNathan-Marsh1@sheffield.ac.uk)

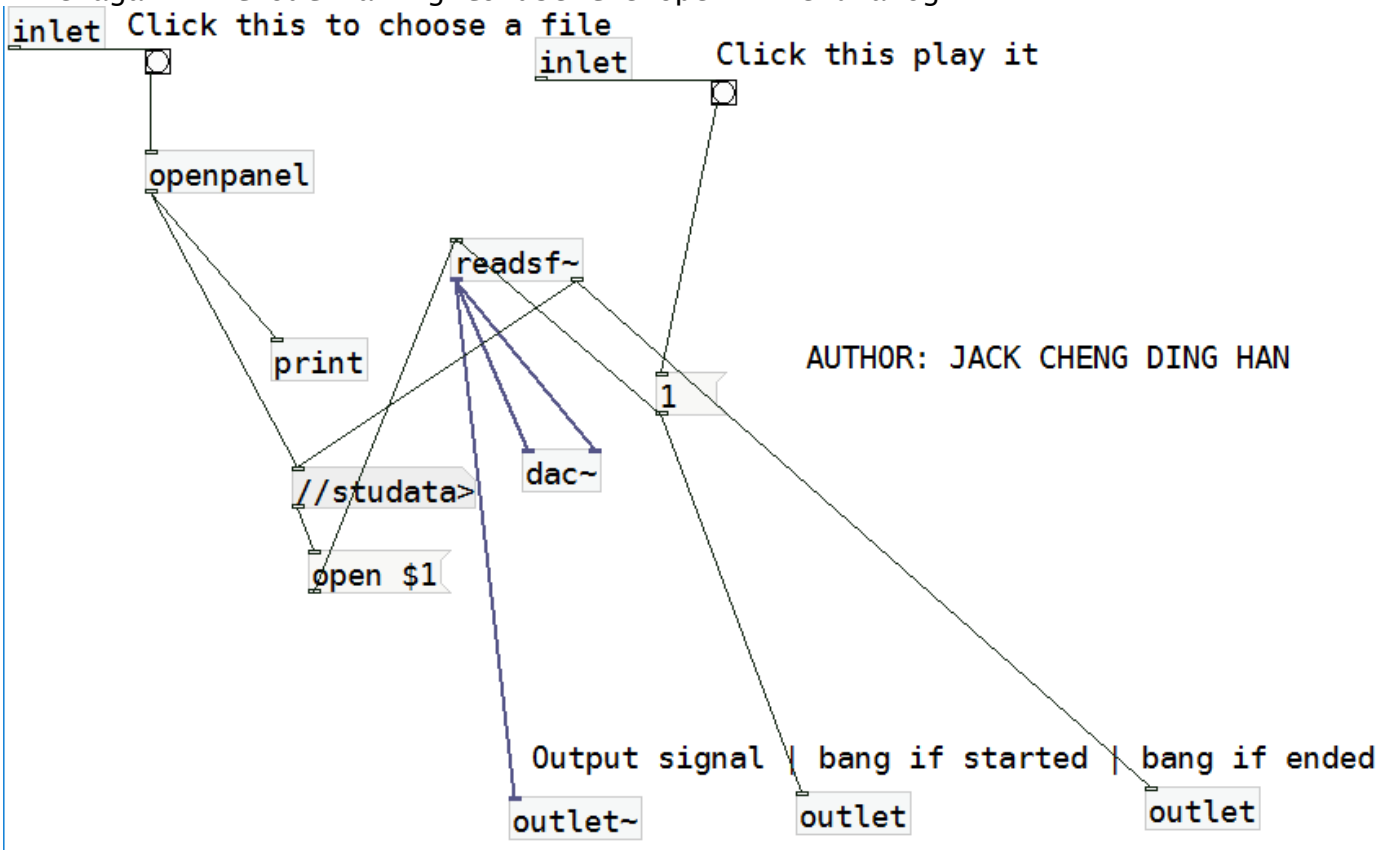
Graphical User Interface

Jack wrote the graphical user interface. This is what the user sees upon startup.



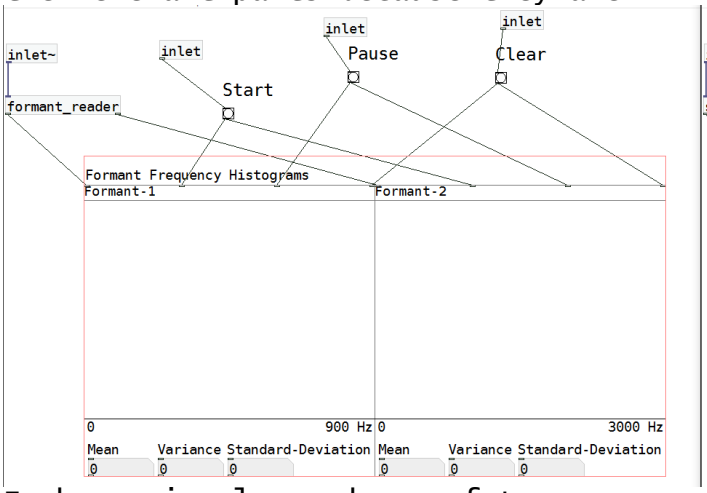
The top half are the controls, there are two radio buttons that lets the user determine if they want to use live audio (the default and leftmost setting) or recorded audio.

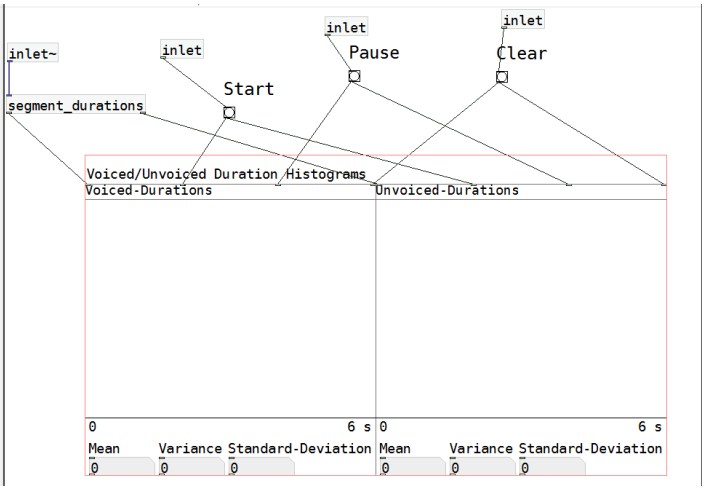
If they use the recorded audio, they first have to click the second radio button. Next, they click the leftmost bang (labelled “Open) to launch an open file dialog to select their desired audio file, then they click the bang labelled “Analyse” to play file, activate the histograms and the return measured quantities. When the file stops playing, the histograms, means, variances, and standard deviations automatically freeze so the user can see the final results. Using the patch “fileloader”, the user can analyse/play the file again without having to use the open file dialog.



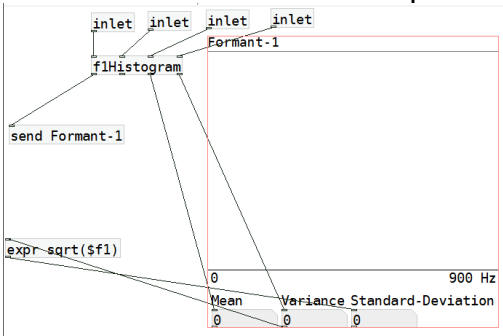
Jack Cheng Ding Han (150159519, jcdhan1@sheffield.ac.uk)
Afolabi Nathan-Marsh (150159678, ARNathan-Marsh1@sheffield.ac.uk)
If the user wants to analyse live audio, they click the first radio button, and use the bangs labelled “Start”, “Pause” and “Clear” directly, these all correspond to the inlets in each of the histogram abstractions.

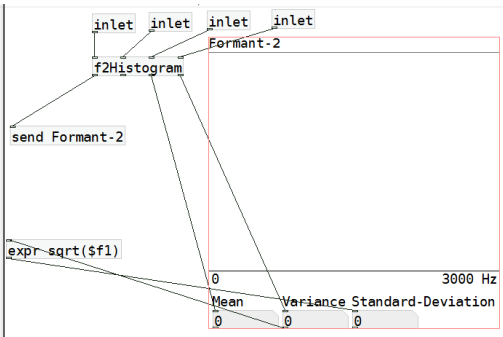
The results are displayed in the bottom half of the interface and only show the relevant parts because they are in fact two canvases:

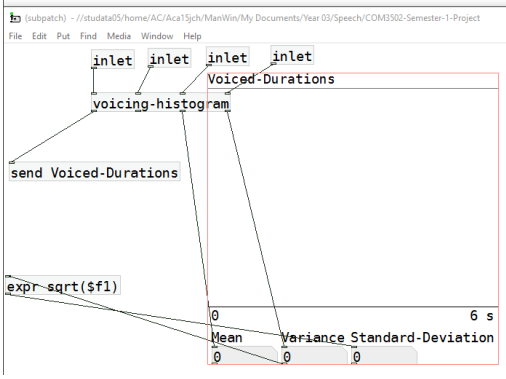


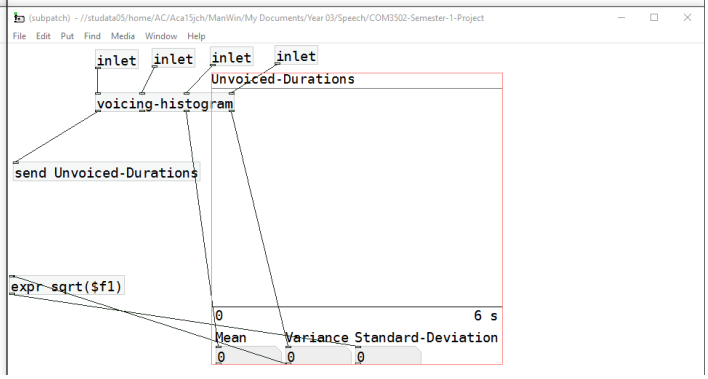


Each one is also made up of two canvases:









Jack Cheng Ding Han (150159519, jcdhan1@sheffield.ac.uk)
Afolabi Nathan-Marsh (150159678, ARNathan-Marsh1@sheffield.ac.uk)

Evaluation

Formant Frequency Estimates

Q: *What are the average F1 and F2 values for the four 'cardinal' vowels [i], [a], [ɑ] and [u]?*

Live

Jack produced the closed cardinal vowels /i/ and /u/, Afolabi produced the open cardinal vowels /ɑ/ and /a/.

Frequency (Hz)	F1 Means				F2 Means			
Trial	1	2	3	Average	1	2	3	Average
i	161.6	347.3	325.4	278.1	2170	1958	2207	2111.67
u	261.5	342.1	293.8	299.13	838.5	920.6	1108	955.7
ɑ	691.9	849.8	653.6	731.77	569.6	656.6	959.6	728.6
a	774.2	771.4	848.8	798.13	1585	1670	1675	1643.3

Q: *How much speech (in seconds) do you need in order to obtain stable estimates of the parameters?*

The following table describes how much speech (in seconds) we needed to obtain stable estimates for the formant frequencies:

Stabilisation times (s)	F1	F2
i	7	4
u	2	2
ɑ	2	3
a	3	4
Average	***3.5***	***3.2***

Recorded

Jack recorded files for the four cardinal vowels.

Frequency (Hz)	F1 Means				F2 Means			
Trial	1	2	3	Average	1	2	3	Average
i	345.0	344.6	347.2	345.6	2012.2	2024.2	2034.4	2023.6
u	475.0	470.2	468.8	471.33	651.2	644.0	645.6	646.93
ɑ	787.8	787.2	788.7	787.9	958.8	969.0	960.5	962.77
a	837.0	839.6	835.3	837.3	1267.5	1275.4	1261.5	1268.13

Conclusions

According to Catford [2], the formant frequencies for the four cardinal vowels are as follows:

Vowel	First Formant (Hz)	Second Formant (Hz)
i	240	2400
u	250	595
ɑ	750	940
a	850	1610

Jack Cheng Ding Han (150159519, jcdhan1@sheffield.ac.uk)
Afolabi Nathan-Marsh (150159678, ARNathan-Marsh1@sheffield.ac.uk)
For the vowels we analysed live, the first formants appeared very close to these whereas the second formant were more accurate with the front vowels /i/ and /a/.

In contrast, Jack's recorded files for /i/ and /a/, gave slightly lower second formant values than predicted.

Also in his recordings; the first formant values for /u/ were slightly higher than predicted, whereas the second formant values our application calculated for /u/ was closer than the live calculations.

The second formant values for /ɑ/ were more accurate when recorded.

Voiced/Unvoiced Segment Duration Estimates

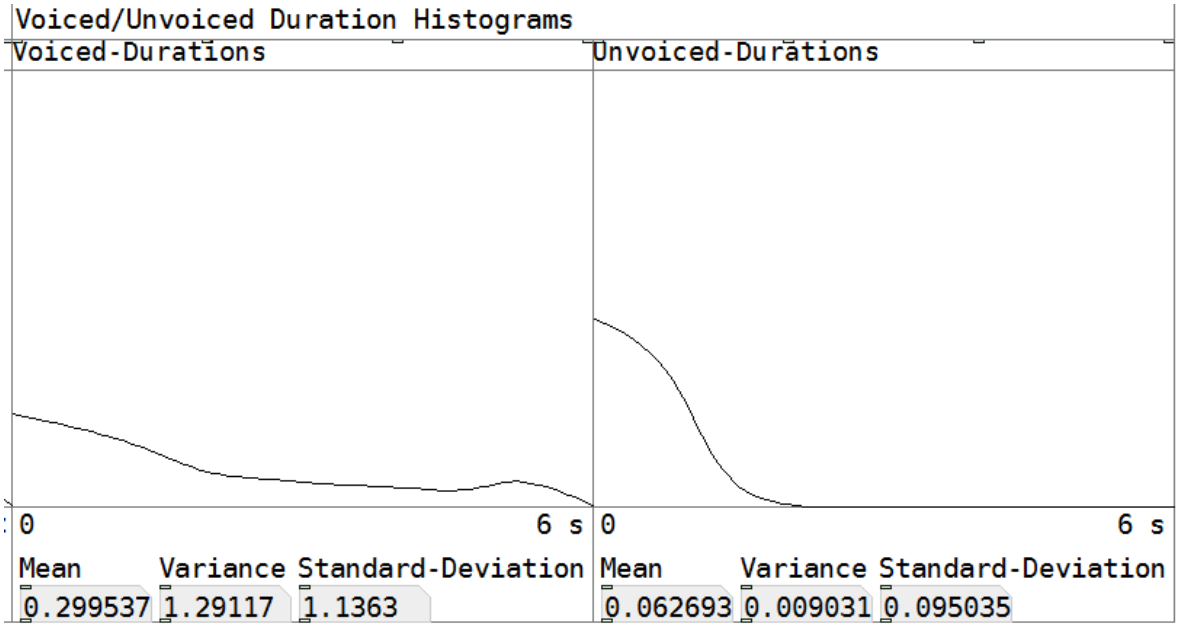
Q: *What are the average durations of voiced (V) and unvoiced (UV) segments in the utterance: "She had your dark suit in greasy wash water all year"?*

We both spoke the sentence "She had your dark suit in greasy wash water all year"

Method	Speaker	Total Duration (s)	Mean Voiced Duration (s)	Mean Unvoiced Duration (s)
Live	Jack	4.0	0.284444	0.204336
Recorded	Jack	3.8	0.226395	0.193887
Live	Afolabi	3.33	0.493424	0.101007
Recorded	Afolabi	2.8	0.493424	0.186921
	Average	3.4825	***0.374422***	***0.171538***

Q: *How much speech (in seconds) do you need in order to obtain stable estimates of the parameters?*

As elaborated on further in the section below, we expected the mean voiced duration to be higher than mean unvoiced duration, while our tests for under 4 seconds matched this prediction, the estimates for voiced durations were lower than expected, especially when Jack spoke the sentence.
However, when Jack stretched the sentence out the sentence to ***10 seconds***, we obtained the mean of 0.299537 seconds for voiced durations. While this is low, the histogram showed two peaks and the standard deviation was relatively high. In contrast, the unvoiced durations had a low mean and low standard deviation.



Jack Cheng Ding Han (150159519, jcdhan1@sheffield.ac.uk)
Afolabi Nathan-Marsh (150159678, ARNathan-Marsh1@sheffield.ac.uk)

Conclusions

“She had your dark suit in greasy wash water all year” in the Renounced Pronunciation dialect is:

[ʃi həd jə dɑːk suːt ɪn ˈgriːsi wɒʃ ˈwɔːtər ɔːl ˈjiə]

The consonant-vowel clusters are:
CV CVC CV CVC CVC VC CCVCV CVC CVCV VC CV

Let us mark voiced consonants with a G:
CVCVGGVGVCCVCVGGGVCVGVCGVCVVGGV

There are 8 unvoiced sounds out of 31, so there are more voiced sounds being produced in the entirety of the sentence. Therefore, the mean voiced duration should be greater than the mean unvoiced duration.

References

[1] M. Greenwood and A. Kinghorn, “Automatic Silence/Unvoiced/Voiced Classification of Speech,” The University of Sheffield, Sheffield, 1999.

[2] J. C. Catford, in *A Practical Introduction to Phonetics*, Oxford, Oxford University Press, 1988, p. 161.