Report

# Principles Used for Estimating Different Parameters

## First and Second Formants

The input signal is filtered between two ranges:

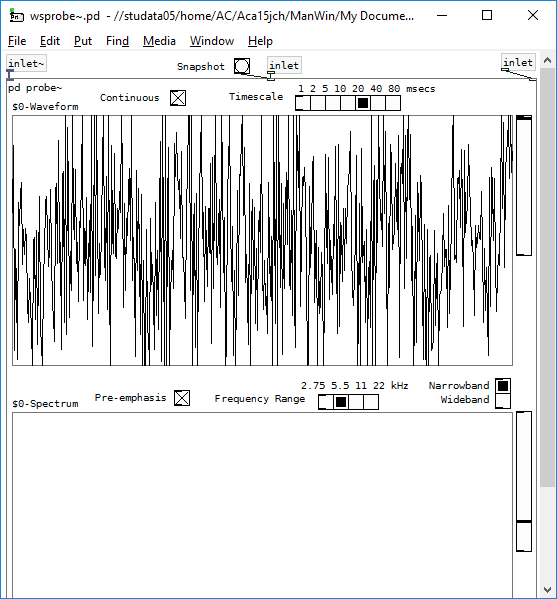
300 to 700Hz for F1, 800 to 2400Hz for F2.

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.

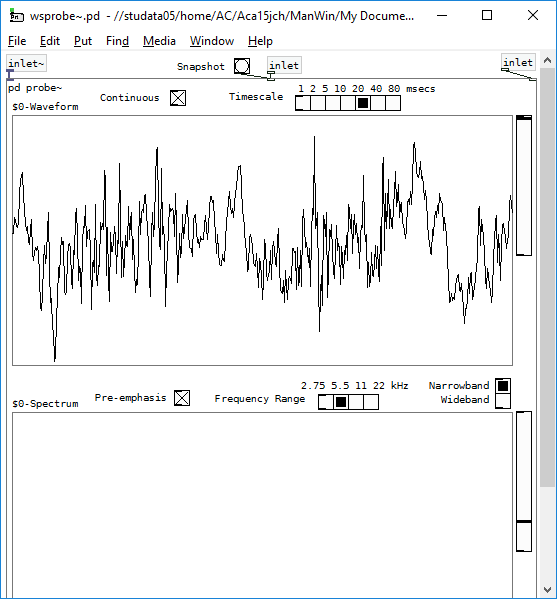
## Voiced and Unvoiced Segment Durations

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. 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/



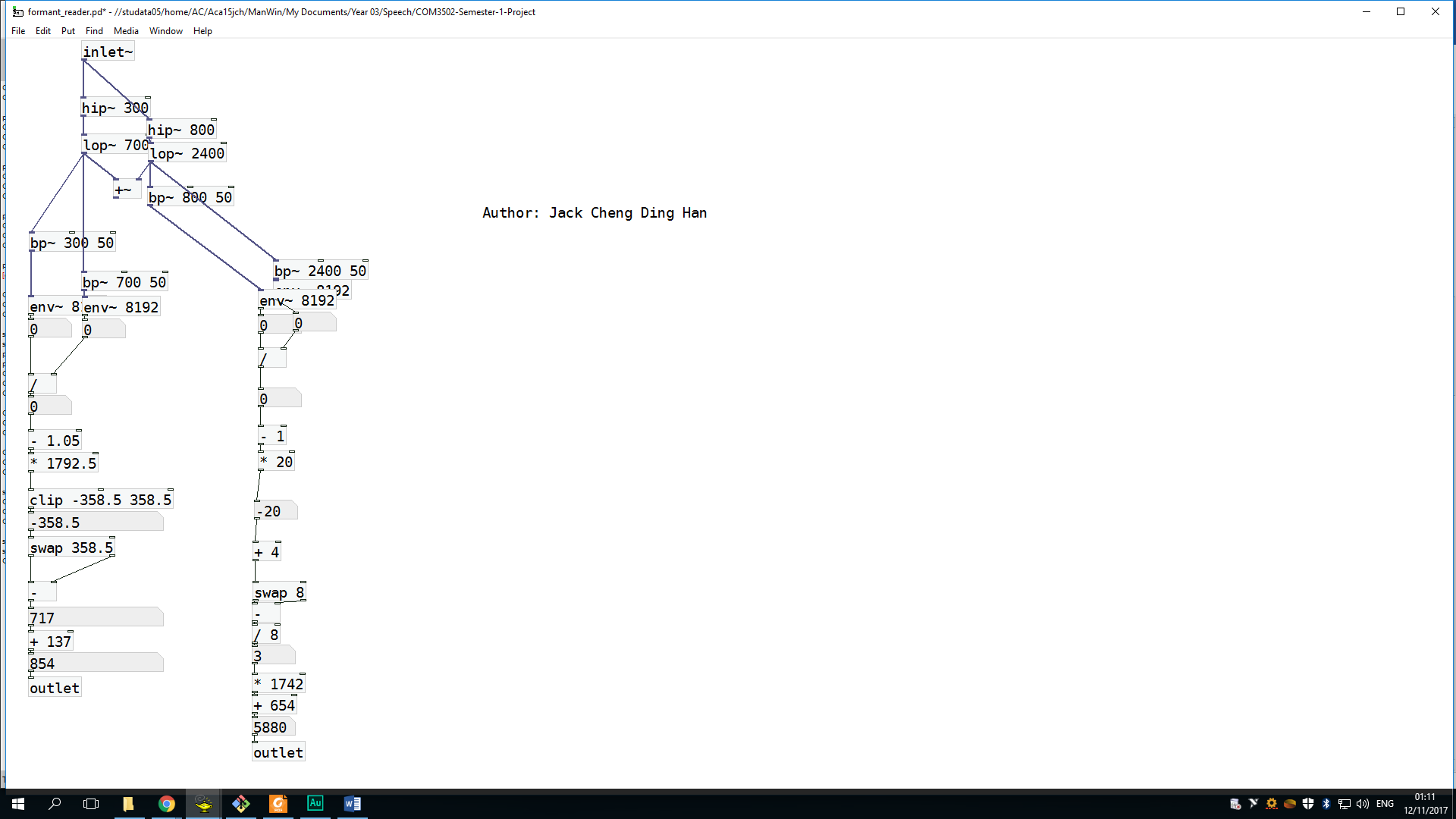
Voiced sounds possess high short-term energy and their signals crosses the time-axis fewer times. On the other hand, voiceless sounds have low short-term energy and their signals cross the time-axis more [1].

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/.

# Implementation

## Formant Frequency Estimation

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.

## Voiced/Unvoiced Segment Duration Estimation

# Evaluation

## Formant Frequency Estimates

### 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 |

The following table describes how much speech (in second) 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 |

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

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 |

### Conclusions

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

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

The consonant-vowel clusters are:

CV CVC CV CVC CVC VC ССVCV 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. Our results match this prediction.

# 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. |