University of Sheffield

COM3502-4502-6502
Speech Processing

Main Programming Assignment

Jake Sturgeon

Ruairí Sinkler

Department of Computer Science

December 7, 2018

QUESTION 1 ***(worth up to 5 marks)***

Provide a screenshot of `[wsprobe~]` for a typical voiced sound, and explain the features in the waveform and spectrum that distinguish it from an unvoiced sound. *Hint: use the 'snapshot' feature in* `[wsprobe~]` *to obtain a static display.*

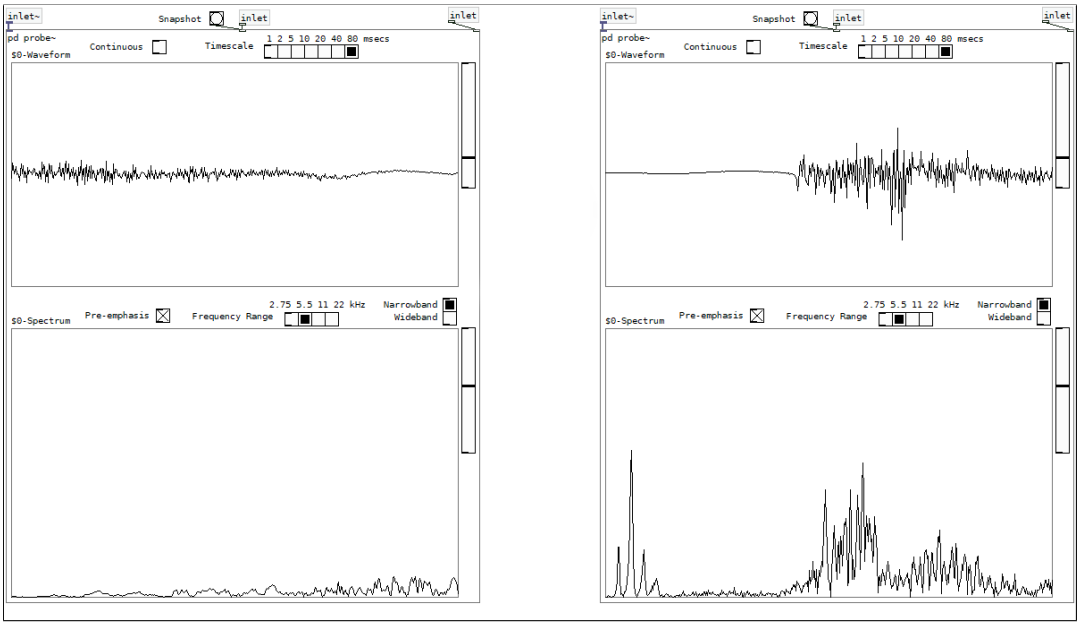

Figure 1: Left: a snapshot of the letter "T" being said. Right: a snapshot of the letter "D" being said.

On the in figure 1 above we have taken a snapshot of the letters T and D. We chose to pick these two letters because they are produced in exactly the same way (alveolar plosives) apart from the fact that D is voiced and T is unvoiced so they are a good pair to compare.

It can be seen that the waveforms are pretty similar for both but the difference between them lies in the Spectrum below. The spectrum show how in the voiced D (right) there is fairly strong low-frequency energy which is produced by the voiced sounds whereas in the voiceless T (left) there is only a low-level white noise across all frequencies.

QUESTION 2 ***(worth up to 5 marks)***

Which sounds are most affected when the low-pass cut-off frequency is set to around 500 Hz - vowels or consonants - and why?

As low pass filter only allows low frequencies past, a lot of the higher energy frequencies are removed. This means a lot of the voiceless fricatives are filtered away as they fall in the range above 500Hz.

Therefore, consonants are affected more than vowels.

QUESTION 3 ***(worth up to 5 marks)***
How is it that the speech is still quite intelligible when the high-pass cut-off frequency is set to 10 kHz?

As the high-pass filter is set very high, at 10kHz, it removes the majority of the low-energy frequencies. This means a large amount of the voiced sounds are reduced or eliminated and so vowels become a little less clear. The consonants remain largely unaffected, however, and since consonants effectively carry more information than vowels, in terms of understanding speech, this means the speech is still intelligible.

This is similar to how we can still understand speech even when it is whispered and no voiced sounds are used.

It is also worth noting that the filters used are not perfect "brickwall" filters but rather approximation of idealised ones. This means there is "roll-off" below the cut-off frequency and so some frequencies below 10kHz are still getting through because of the slope between the pass-band and stop-band as well as the attenuation that occurs within the stop-band itself so even frequencies within the stop-band are getting through in a minimal way.

QUESTION 4 ***(worth up to 5 marks)***
COM3502-4502-6502: The `[GraphicEqualiser~]` object uses an FFT internally; what does FFT stand for and what does an FFT do?
COM4502-6502 ONLY: What is a DFT and how is it different from an FFT?

FFT stands for Fast Fourier Transform and is an efficient algorithm used when processing signals. It uses Fourier Analysis to resynthesise an input signal, which is in the time-domain, into a spectrum of the signal, in the frequency domain.

It is useful because FFT generates the magnitude and phase, which is used for inverse to generate the original waveform.

QUESTION 5 ***(worth up to 10 marks)***
With `speed = 50` and `depth = 0.5`, what are the minimum and maximum amplitudes of your LFO output, and how do they vary with changes in these two settings? Also, please provide two screenshots: (a) your `[LFO~-help]` object and (b) the internal structure of your `[LFO~]` object.

Minimum Amplitude: -0.5 Maximum Amplitude: 0.5

These values directly correspond to -depth and +depth. Speed has no effect.

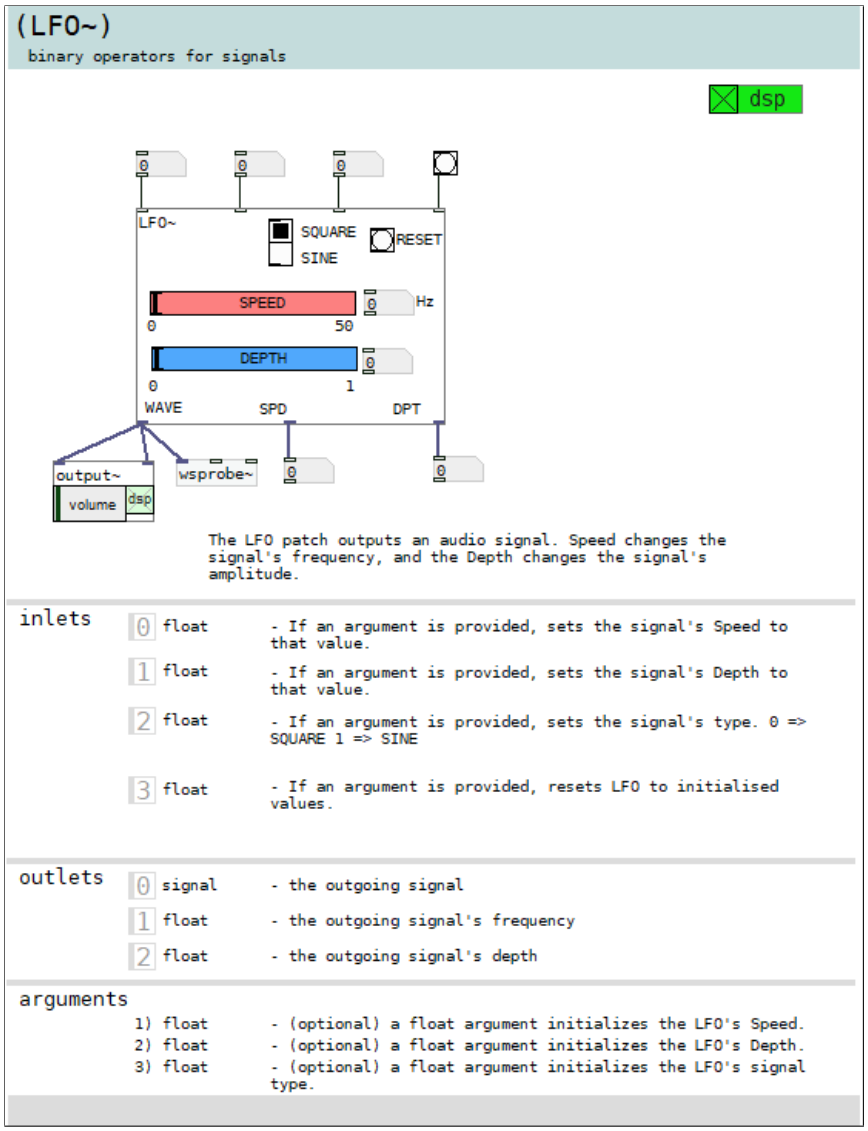

Figure 2: Our LFO-help.

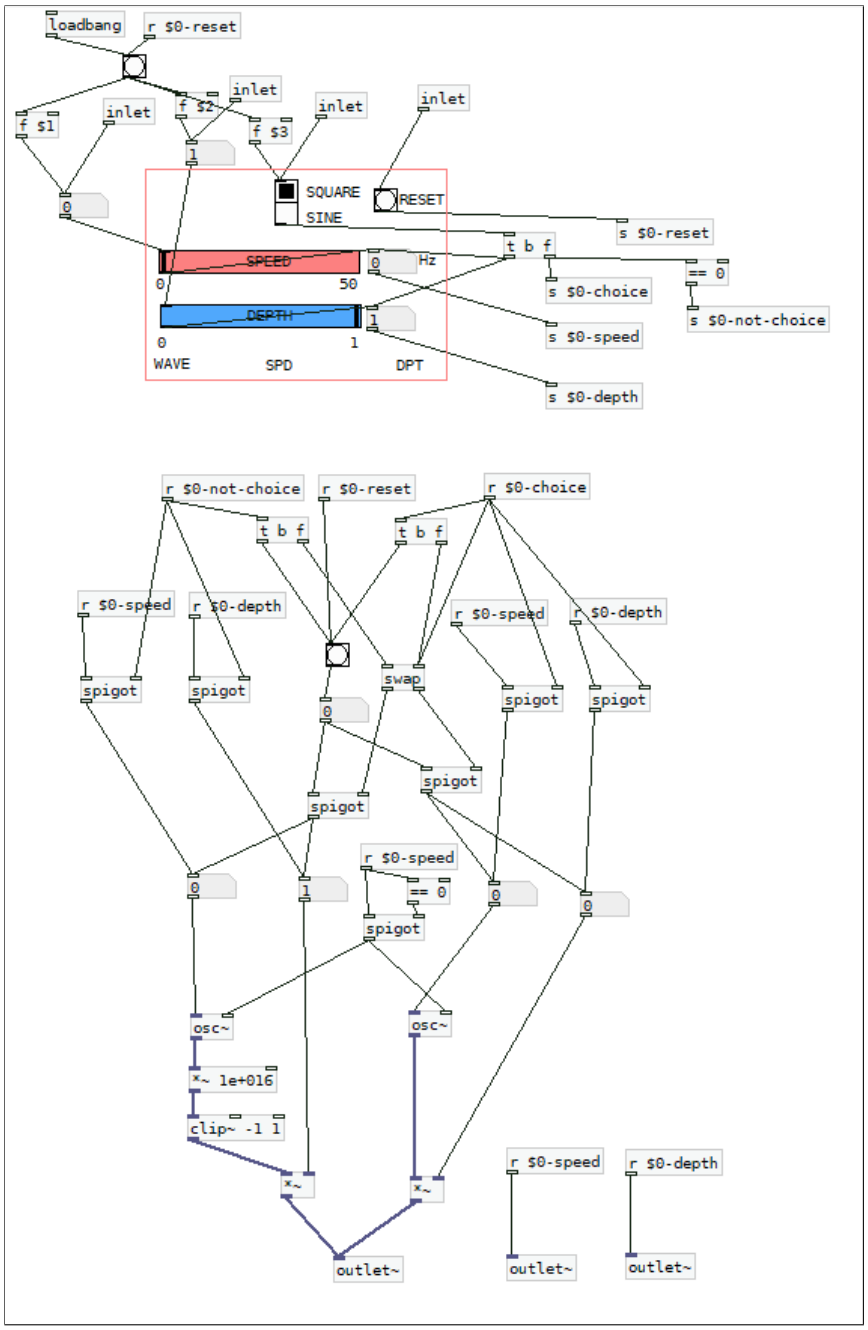

Figure 3: Internal structure of our LFO.

QUESTION 6 ***(worth up to 5 marks)***

In your own words[1], why is this effect known as 'ring modulation'?

Ring Modulation involves multiplying an input signal by a wave (usually sine) to rapidly alter the polarity of the input signal. The name comes from back when it was implemented using electronics, using hardware rather than software, where a set of 4 diodes would be linked in a circle, or ring, to create the same effect.

[1] I.e. do not plagiarise from Wikipedia.

QUESTION 7 ***(worth up to 5 marks)***
Why is SSB commonly used in long-distance radio voice communications?

To boost the amplitude of a signal it is multiplied by another signal (e.g. a sine wave) but in doing so this essentially doubles the data present since the result consists of the sums AND differences of the input signals. By using SSB one of these is removed and so the signal data is then halved again so the bandwidth required to transmit it will remain the same as the bandwidth required to transmit the original signal, but the amplitude has still been boosted.

QUESTION 8 ***(worth up to 5 marks)***
COM3502-4502-6502: Why can the voice be shifted up in frequency much further than it can be shifted down in frequency before it becomes severely distorted? /emphHint: look at `[wsprobe~]`.
COM4502-6502 ONLY: Your frequency shifter changes all the frequencies present in an input signal. How might it be possible to change the pitch of a voice *without* altering the formant frequencies?

As the voice is shifted up in frequency, the voice's fundamental frequency and harmonics are also increased, which gives the effect of a restricted vocal tract. This means all of the energy shifts upwards in frequency with no information loss.

However, as the voice is shifted down in frequency, the voice's fundamental frequency and harmonics are decreased. Furthermore, when the frequency shift is lower than 0 the fundamental frequency becomes negative which causes the waveform to reflect on the x and y-axis. This creates interference as the reflected portion of the wave interacts with the regular portion and so information is lost. Therefore, as the frequency shift tends to -infinity, more of the waveform is interfering with itself. Thus, becoming increasingly distorted.

QUESTION 9 ***(worth up to 5 marks)***
In a practical system, why is it important to keep the feedback gain less than 1?

If the feedback is ever more than 1 then the delayed and returned signal explodes exponentially by adding larger and larger versions of itself to the input signal, rather than decaying off.

QUESTION 10 ***(worth up to 50 marks[2])***
Please provide a short[3] description of the operation of your `[VoiceChanger]` application, together with a screenshot of your final GUI.

[2] 25 for functionality, 15 for design/layout, 5 for `Pd` features, 5 for innovations
[3] no more than 200 words

Our final `[VoiceChanger]` application makes use of all of the various signal processors implemented throughout this assignment. We have organised them into various panels with the different kinds of processors separated out:

- Time-Filters - A set of time-domain based filters
- LFO-Based - A set of LFO-based modulators
- FrequencyShifters - A set of modulators that shift the frequency of the input signal
- Time-Delay - A set of modulators based on altering a time delay on the signal

Our `[VoiceChanger]` loads with all of the signal processors set to parameters that will not affect the input signal. Each individual processor can be reset with the RESET button in the top-right of its panel. In turn each set (including the entire set) can be reset with RESET buttons in the top-right of their respective panels (RESET_ALL for the top-level set).

It can take live voice input or a sound file, controlled by the toggle button in the top left. When the toggle is switched to "SOUND FILE INPUT" an "open" panel pops up to select a sound file which then loops until "LIVE VOICE INPUT" is chosen again.

Finally we have included some preset buttons that insert some interesting voice changing parameters.

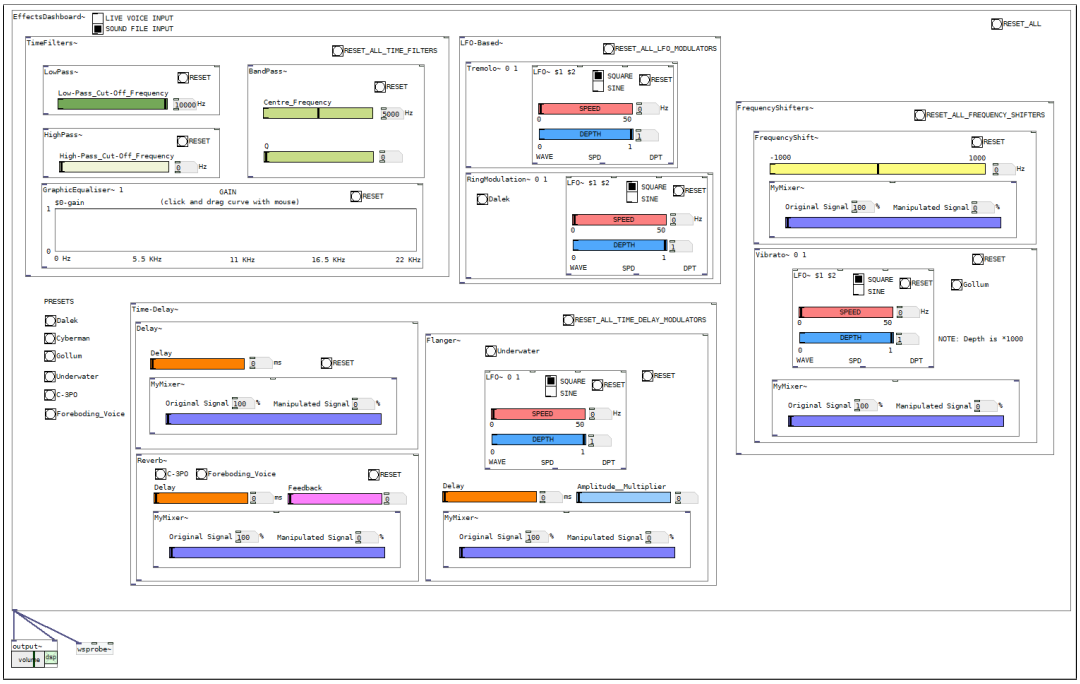

Figure 4: Screenshot of our final `[VoiceChanger]` GUI.