DEPARTMENT OF ELECTRONIC ENGINEERING
THE UNIVERSITY OF YORK

Voice Production & Synthesis

PureData Source-filter Speech Synthesiser

Y3761870

CONTENTS

May 2017

1 Principles of voice synthesis

The source-filter model of speech describes voice production as: a power source, which excites a sound source (the vocal folds), which emit sound, which passes through a series of filters (the vocal tract) e.g. the nasal cavity, the mouth, tongue, teeth, and finally radiation from the lips [1].

Voiced speech, such as vowels and some consonants, have a pitch (and therefore a fundamental frequency). This results from the vibration of the focal folds in the larynx [1]. *Unvoiced* speech such as consonants such as /s/ and a /tʃ/ is created from turbulent air travelling through a constriction in the vocal tract [1–3].

When voiced sound is produced, the sound will be modified as it travels along the vocal tract: up the pharynx, throw the nasal and mouth cavities, around the tongue and alveolar ridge, the teeth, through the nostrils and radiating from the lips. Various parts of the tract will have filtering and resonance effects which produces non-harmonic peaks in the sound signal as it is produced [3]. These peaks are called *formants*. The effects of the vocal tract can be summarised as the vocal tract transfer function (VTTF).

The relative positions of these formants are what enables us to perceive different vowels even if they are of the same pitch (whilst people may produce certain vowels at higher pitches compared to other vowels, the pitch itself is not crucial, as can be demonstrated by singing vowels at different pitches and observing that the vowel can still be determined).

For a simple but intelligible synthesiser we must produce a wave rich in harmonics, then pass it through an array of filters modelling the vocal tract and creating formants. The reason it is necessary to have a wave rich in harmonics is so there is enough higher frequency content for VTTF to produce formants from the sound source.

Generally three filters and a sawtooth wave as a sound source would be sufficient for intelligibility if not naturalness.

The vocal tract filters can be modelled in parallel or in series (referred to as cascade synthesis in the literature) and there are pros and cons for either approach. Liljencrants [4] notes parallel makes it easier to preserve correct formant amplitudes and this is the approach I took, as the feedback loop of analysing the output of the synthesiser and making adjustments needed to be as short as possible given the time constraints of the project.

For consonants, additional sounds must be produced. I have focused on fricatives which can be modelled loosely with white noise, although in reality it is not uniformly distributed; it starts to fall off at about 1kHz and falls roughly linearly to zero at 10kHz [3]. The consonants that are closer to vowels can be produced using formant synthesis but with a different spectral tilt to the sound source.

1.1 Naturalness

To develop natural sounding synthesis it is be useful to introduce high frequency *aspiration noise* in the vowels, amplitude-modulated by the voice source [5]. Boosting the relative strength of the fundamental is also helpful for male voices in particular [5].

In order to do this it's worth revisiting the sound source that creates the spectral properties needed to be exhibited by the filters. Fant et al proposed [6] a four parameter model of differentiated glottal flow called the LF-model which has formed the basis of much work in this area. Alternative models do exist such as those that try to create a physical model of the glottis [4].

The LF-model exploits the (assumed) commutative relationship between the voice source, vocal tract, and lip radiation, to combine the effects of the voice source and lip radiation into one model [7].

Li et al [8] define a version of the simplified LF-model expressed for a discrete implementation, replacing the time parameters with ratios and samples (rearranged into equation 1, below). This is the version I implemented in PD.

$$E(k) = -E_e \cdot \begin{cases} \dfrac{e^{\frac{\alpha k}{N}} \sin(\frac{\pi k}{T_p N})}{e^{\alpha T_e} \sin(\frac{\pi T_e}{T_p})}, & 0 \leq k \leq T_e N \\ \dfrac{e^{-\epsilon(\frac{k}{N} - T_e)} - e^{-\epsilon(1-T_e)}}{\epsilon T_a}, & T_e N < k \leq N \end{cases} \tag{1}$$

In the above equation N and k represent the total number of samples and the current sample, respectively. T_p, T_e, and T_a define the time to the maximum glottal flow, the time until the open phase, and the time until the return phase. E_e is the maximum magnitude of glottal closure excitation. ϵ and α control the shape of the curves.

A revisited LF-model uses a data reduction scheme to reduce the number of control parameters [9]. This can make synthesis simpler by reducing to controls parameters (R_a, R_g, R_k) although it was only partially implemented in my final model. A useful property is given as

$$F_a = \frac{1}{2\pi T_a} \tag{2}$$

which determines the spectral tilt of the waveform [9].

An analysis done by Gobl [10] produces some useful indications for typical male speakers (e.g. $F_a =$ 700Hz). E_e tends to be stronger for vowels and weaker for consonants, though the dataset is limited. It is a useful starting point in lieu of further analysis. Gobl also notes the impact of speech prosody (stress, intonation) on voice source parameters but that, though interesting, is beyond the scope of this system.

A feature of real speech is that the fundamental pitch is not constant, but exhibits flutter [5]. Synthesised speech with a constant F_0 does not sound natural. One solution might be to vary the timing parameters in the LF-model, but Klatt & Klatt note [5] many efforts at randomising the fundamental produce harshness in the resultant voice, and they propose an alternative pseudorandom contour to F_0 where

$$\begin{aligned}\Delta F_0 = &\frac{\mathrm{FL} \cdot F_0}{5000}\big[\sin(2\pi 12.7t) \\ &+ \sin(2\pi 7.1t) + \sin(2\pi 4.7t)\big]\end{aligned} \tag{3}$$

suggesting an FL value of 25. This is implemented in my synthesiser.

2 How the PD patch works

This section is focused on the audio properties of the PD patch and the control structures which operate them specifically. In addition to this, many different general-purpose control objects were created to streamline development. A full list is available in appendix A.3, page IV.

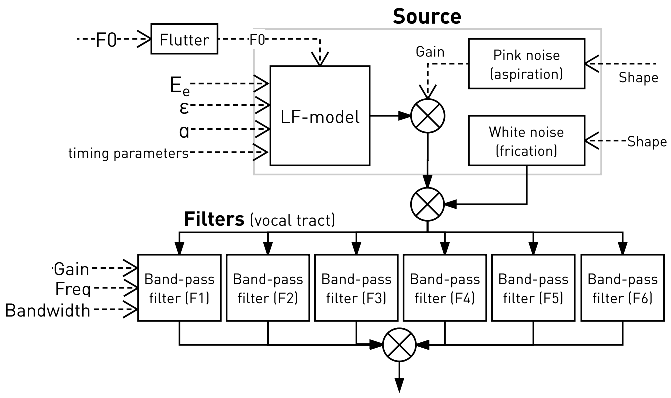


Figure 1: Signal path diagram for the system. Audio signals are represented with black arrows, control signals are represented with dotted arrows.

The synthesiser is configured as a parallel formant synthesiser with an adjustable LF-model of the glottal flow derivative as the sound source. Aspiration and frication noise are provided by two additional noise generators in the voice source. Pseudo-random flutter is implemented using Klatt's algorithm detailed in the previous section. The formants are modelled as band-pass filters with adjustable gain, centre frequency and bandwidth. The signal path is for this system is outlined in the signal flow diagram in Figure 1 above.

In the top-level patch ([synth]) the connecting edges of the objects are reserved only for audio signals to make the data flow clear (see Figure 2 below). Control signals are routed using message receivers.

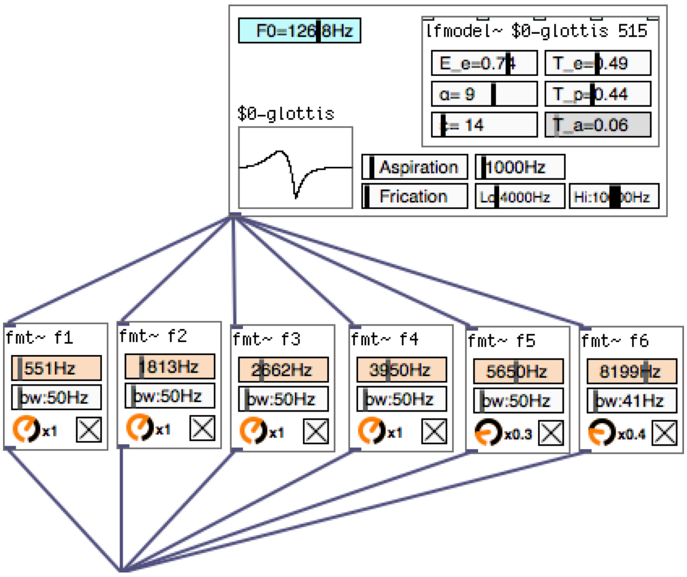


Figure 2: Signal flow in [synth]: from sound source through the vocal tract.

I initially used a large number of formants (10). This did create added realism but put a lot of burden when trying to make adjustments. Fant [9] cites a formula

$$m = F_s(l_e/c) \tag{4}$$

which defines the minimum number of VTTF poles required to maintain correct spectral distribution for formant synthesis. Assuming an average male vocal tract length of $l_e = 17.655$cm and a sampling rate of $F_s = 44100$Hz this indicates a need for 22 formants. This seems well beyond what is likely to be practical to implement in PD. To determine the resonant frequency of a particular formant Johnson [3] gives the formula

$$F_n = \frac{(2n-1)c}{4l_e} \tag{5}$$

which suggests that the tenth formant comes in at approximately 9.4kHz. As human hearing drops off after 10kHz I see no particular need to implement these higher formants in any case. For simplicity I reduced my system to five formants and added a sixth to model an additional prominent formant movement in the diphthong of my chosen word. I encourage the reader to experiment with selecting the bypass switch on the formants on my patch to hear what difference the formants above F_3 make.

In order to recreate the spectrum more precisely I had to use more configurable filters than the standard array of [bp~], [lop~] and [hip~]. I found for the high and low-pass filters especially that the rolloff was too gradual to be of much use. Instead I used the built-in [biquad~] filter using [bandpass] to calculate the coefficients for the formants. One problem with this is that if the control parameters are not interpolated smoothly enough it can create artefacts in the synthesis. However, reducing the time-grain for [line] can introduce a lot of processor overhead. I tried to strike a balance. [bandpass] specifies the bandwidth in octaves so I created some logic to convert to a bandwidth in Hertz to make adjustments more quickly. However, it does not map the centre frequency precisely because [bandpass] sets the bandwidth logarithmically rather than linearly.

A [controller] patch—Figure 3 below—has any array of buttons for different phonemes and words, and when they are pressed sends control messages to the vocal tract and voice source patches to configure them.

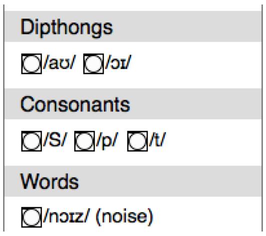


Figure 3: A small section of the [controller] patch showing how presets can be selected.

Most control messages are terminated with a linear ramp time in ms. Sequencing chains of these events is handled with [qlist] and a message delay pipe, see [msgpipe] in appendix A.3 for more details.

I aimed to create a flexible synthesis system that could be configured rapidly. In the pursuit of naturalness it is equally important to have rapid turnaround of analysis of synthesis attempts, as there is a large element of trial and error without using more sophisticated techniques. To aid with this I created [desk] to allow recording directly from the synthesiser as easily as possible (see Figure 4 below). Once a location has been selected with the 'eject' button, pressing record will record to that file until stop is pressed.

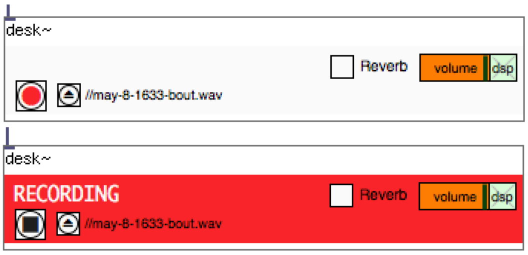


Figure 4: [desk~] object for rapidly saving samples for analysis when stopped (above) and recording (below).

3 Critical analysis and future improvements

Listening back to the final synthesiser it is clear that my approach was most effective for synthesising vowels rather than a whole word. If you use the controller to hold a particular vowel and then manipulate the parameters of the LF model it gives a window into the flexibility this approach can have over the spectral qualities of the voice. I think this would make it highly suited for use a singing synthesiser.

For words I think the system falls short. There is a small selection of words in the system: /nɔɪz/, /baʊt/, and /sip/, and each one does give a different window into the synthesiser as the qualities of the voice change substantially for each. For my chosen word, /nɔɪz/, the synthesiser produces quite a buzzy quality and attempts to minimise this with the LF-model tended to make it sound rather flat. I think ideally I need to determine the voice source parameters from my own reference recording, using inverse filtering.

Comparing the spectrogram (see Figure 5) the formant transitions appear similar. In the reference recording there are fewer higher frequency harmonics or aspiration noise at the onset of the /n/ consonant – I think it would be possible to achieve a more similar spectral distribution by manipulating the parameters of the LF-model.

The spectral distribution of the fricative /z/ is very different in the synthesised output however I found that when trying to duplicate the reference precisely it sounded much less intelligible. In the end I used figures for the spectral distribution of fricatives from Johnson [3], who notes the difficulty in measuring fricatives due to several spectral peaks of shifting amplitude. It is hard to tell what properties of the reference spectrogram are from the voice vs. the recording environment. I would like to record the reference recordings in an anechoic chamber with a completely flat high quality microphone.

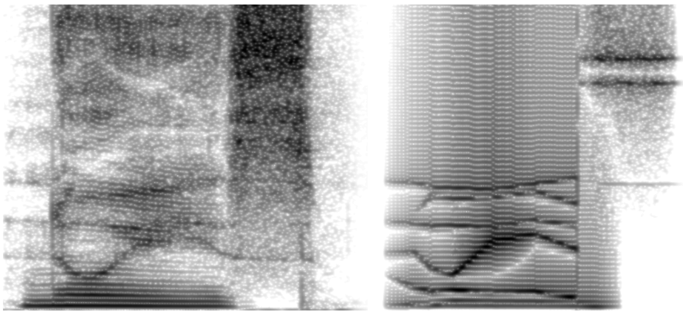


Figure 5: Comparison of spectrograms taken from Praat for /nɔɪz/. Reference (left) vs synthesised (right). 0-10kHz, 0.02 window size.

Whilst writing this report I realised that I could have viewed the spectral slice in Praat which would have made direct analysis much easier, and fitting much more straightforward (see Figure 6 below).

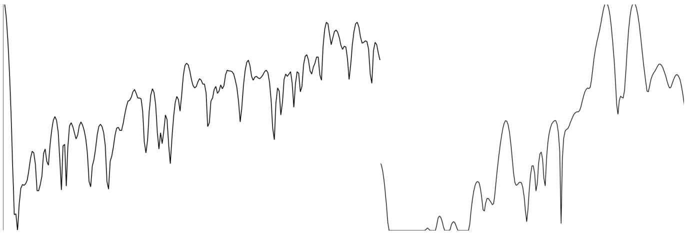


Figure 6: Comparison of of the spectral slice of /z/ from 0-10kHz linear. There are more prominent peaks on the synthesised (right).

Determining LF-model parameters requires a lot of time. The additional parameters in the revised model [9] provide further control over the spectrum: increasing R_k raises relative level of voice fundamental. Increasing R_g promotes the level of the second harmonic at the expense of the fundamental. This analysis starts to explore spectral qualities of specific voices, e.g. sonorous voices have relatively high F_a of the order of 2000Hz [9]. Fant discusses a shape parameter, R_d, which predicts the other values, citing a 1994 publication that I was unable to locate a copy of [11]. I tried to implement this as I felt it would make fitting parameters easier, however I had enormous problems matching up the "statistical relations". The work in progress for this can be seen in the [lfmodel~] patch. I would like to complete this.

Using 'typical' values for voice source parameters is difficult with my reference voice as I am a trans woman and my voice differs from typical male or female voices; for trans women there is a tendency for F_3 to be raised compared to male voices, with slower speech at reduced loudness and a higher F_0 [12]. This raises into question the 'typical' LF parameters suggested for males or females in the literature.

The transitions between formant centre frequencies in my system are linear. This isn't true to formants in human speech, which have more gradual ramps between positions. See Figure 7 below for an example. For my final word I compensated for this by having more interpolation points during the word and ramping between them.

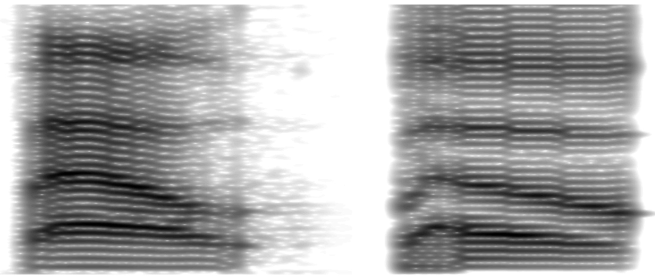


Figure 7: Non-linearity of F_n transitions in Praat for /baʊt/ reference (left) vs synthesised (right).

In the real vocal tract the jet stream of frication noise is interrupted when the glottis is closed. To implement this I would need to integrate the glottal derivative and amplitude modulate the frication noise with this signal (the glottal flow). It may have been better if the frication noise bypassed the VTTF for vowels altogether, or used a different VTTF. This is because the 'sound source' for fricatives is not the glottis but rather the area where the jet stream hits a wall, e.g. the teeth. Filtering this signal through the vocal tract is removes a lot of the signal too early.

My LF-model has four parameters changing every 1ms. A DSP cycle is 23µs at 44100kHz; on my machine this calculation takes 1310µs, which is a loss of resolution and exceeds the 1ms time grain. A computationally efficient alternative has been proposed and demonstrated to be perceptually equivalent in a listening tests [13]. This could be used.

As a final note, I had extensive problems implementing a flexible system in PD which has very limited metaprogramming abilities. Additionally, PD-extended had its last update four years ago and can be somewhat buggy on OS X. Rarely, the [biquad~] filters will become locked in an unstable state and the patch will need to be restarted (please do this if you hear an aggressive siren or chopping sound). I would love to implement in another system such as SuperCollider or a higher level programming language to really explore the possibilities.

A Appendix

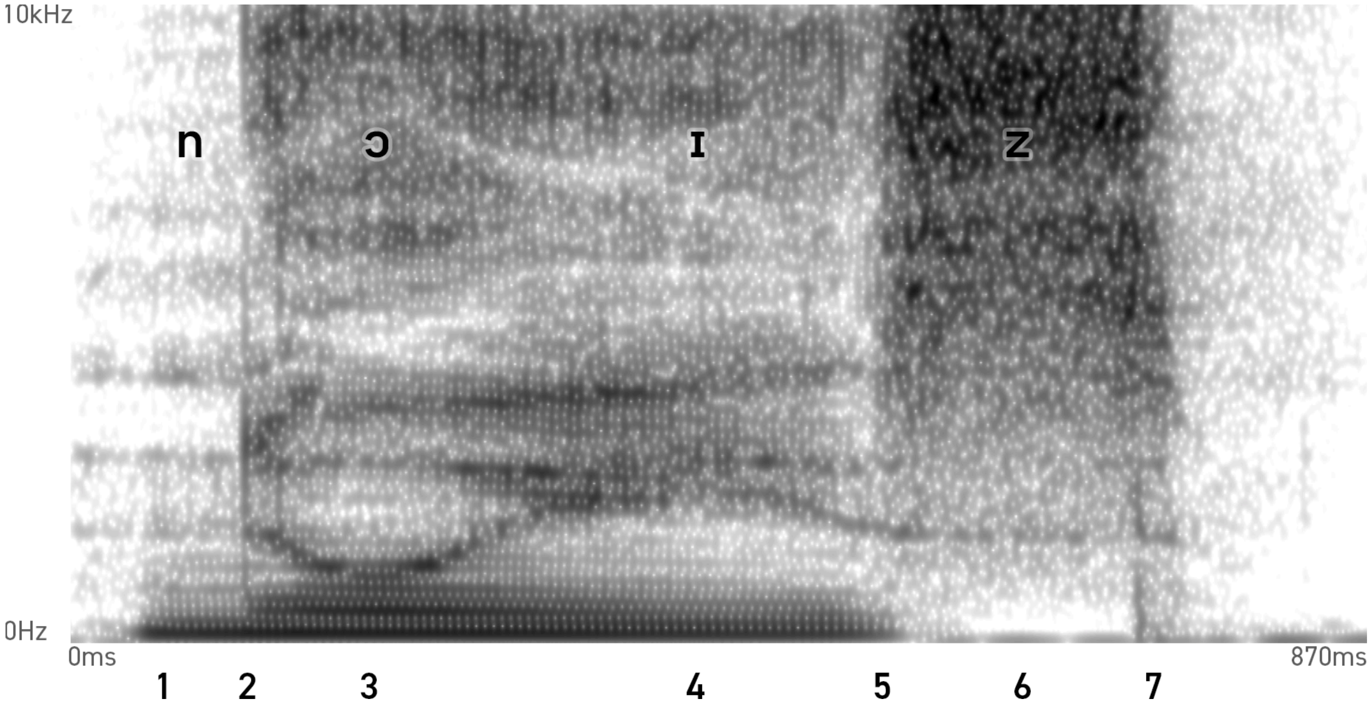
A.1 IPA transcription and VPM description

/nɔɪz/:

- /n/ voiced alveolar nasal (sonorant)
- /ɔɪ/ transitions loosely between three states with the latter two vowels taking up the majority of the transition /ʌɔi/:
    - /ʌ/ starts back open-mid, moves further open, then closes and moves forward
    - /ɔ/ moves further open
    - /i/ closes and moves forward
- /z/ voiced alveolar fricative (obstruent) although, in this case, the voicing is low level and decays before the end of the fricative, which takes on an /s/ quality

A.2 Acoustic characteristics of the vowel and consonants

Below is an account of the spectral behaviour of the word. I have moved most of the discussion of formant frequencies below this figure and slightly lower down (section A.2.1). The first figure is taken from Praat from 0-10kHz with a window length of 0.015.



1. Onset of /n/, 69ms. Strong fundamental relative to harmonics, formants are present but not prominent. Intensity increasing gradually until 2.

2. Rapid change to vowel, about 12ms during which time the peak amplitude increases and there are much more prominent higher frequency noise and harmonics. The formants are more prominent and characteristic of /ʌ/.

3. Over the course of about 85ms the vowel transitions to /ɔ/. F_2 lowers distinctly to 1kHz. F_1 lifts slightly to meet it. The prominence of the fundamental increases relative to the harmonics.

4. Transition to /i/ over 200ms. F_2 raises to 2300Hz. F_1 lowers slightly. The prominence of the fundamental decreases but only slightly.

5. For another 100ms F_2 drops slightly and then there's a small window of about 30ms where the level of the harmonics drops sharply (the spectral tilt is much more weighted towards the fundamental) and the level of the fundamental starts decreasing. Then noise is introduced as the /z/ begins.

6. The noise is well represented up to about 17kHz before it starts to drop off (not shown above). The distribution is indicated in Figure 6, page 4, with spectral peaks at 9300kHz and 8500kHz, although this is not constant. The noise lasts for 160ms. For the first 40ms the fundamental is still prominent but it begins to drop off and the sound is more characteristic of /s/.

7. Noise decays relatively uniformly over 33ms as the word ends.

A.2.1 Formants

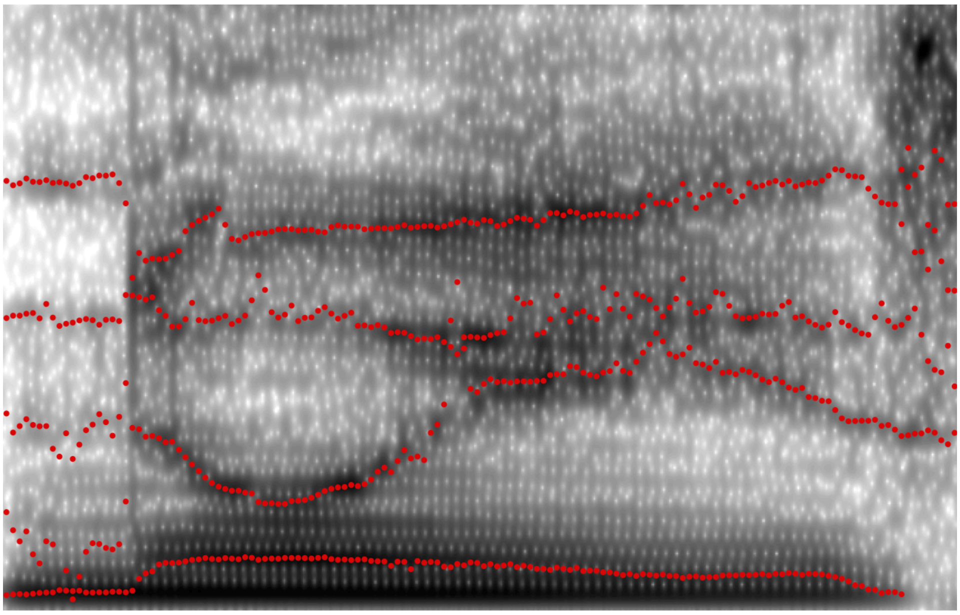


Figure 8: Movement of the first four formants according to Praat. Max formant limited to 4800Hz, window length 0.015. Spectrogram 0-6kHz, window length 0.01, dynamic range 70dB.

Time	F_1 Hz	F_2 Hz	F_3 Hz	F_4 Hz
0.00	187.13	735.41	1840.91	2923.55
0.02	199.93	640.85	1783.54	2909.98
0.04	208.56	679.34	1947.53	3095.68
0.06	308.63	1731.04	3038.28	3434.92
0.08	468.05	1633.86	2976.16	3543.13
0.10	515.25	1386.34	2920.32	3789.28
0.12	531.65	1206.70	2907.87	3782.18
0.14	535.80	1111.19	2938.47	3730.53
0.16	540.88	1150.62	2909.00	3745.76
0.18	524.20	1227.30	2845.50	3760.99
0.20	486.81	1307.88	2754.43	3766.51
0.22	466.05	1520.45	2707.38	3777.33
0.24	479.58	1997.24	2658.23	3797.71
0.26	475.20	2245.22	2701.83	3820.43
0.28	451.89	2279.80	2838.13	3838.63
0.30	428.37	2306.08	2893.53	3884.25
0.32	410.37	2350.00	2932.86	3894.82
0.34	379.65	2374.36	2993.90	3900.67
0.36	362.91	2472.55	2995.21	3975.63
0.38	350.22	2455.72	2992.44	4055.83
0.40	358.40	2367.09	2938.43	4119.43
0.42	370.27	2281.11	2900.42	4179.20
0.44	371.15	2190.59	2890.91	4203.89
0.46	345.23	2008.37	2814.51	4226.80

A.3 Glossary of custom PD objects

- `[analyser~]` – $input$passthrough Display a time-domain and a frequency-domain representation (log-scale FFT) of the signal put into it. Does not alter the signal in any way.
- `[defaultarg]`
- `[fmt~]`
- `[formatlabel]`
- `[interface]`
- `[lfmodel~]`
- `[master~]`
- `[msgpipe]` Programmable message delay pipe: any global message prefixed with `[wait t(` where *t* is a number of milliseconds will be delayed by that time by that amount of milliseconds. For example the message `[wait 15 f0 100 50(` would wait 15 ms before sending the message `[100 50(` to the receiver `f0`.
- `[quasirandomdrift]`
- `[rdefault]`
- `[recall]`
- `[rls]`
- `[round]`
- `[rqlist]`
- `[snapchange~]`
- `[source~]`
- `[synth]` Master patch - start here

B Bibliography

[1] D. M. Howard and D. T. Murphy, *Voice Science Acoustics and Recording.* Plural Publishing, Inc., 2008.

[2] S. Narayanan and A. Alwan, "Noise source models for fricative consonants," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 328–344, 2000.

[3] K. Johnson, *Acoustic and Auditory Phonetics.* Blackwell Publishing, 2 ed., 2003.

[4] J. Liljencrants, I. Karlsson, G. Fant, and M. Büvegürd, "Analysis by synthesis of glottal airflow," *Speech Maps (Esprit/Br No 6975), Deliverable 27, WP 1.3, RP3*, no. 6975, pp. 1–14, 1995.

[5] D. H. Klatt and L. C. Klatt, *Analysis, synthesis, and perception of voice quality variations among female and male talkers*, vol. 87. 1990.

[6] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Stlqpsr*, vol. 4, pp. 1–13, 1985.

[7] A. Del Pozo and S. Young, "The linear transformation of LF glottal waveforms for voice conversion," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1457–1460, 2008.

[8] H. Li, R. Scaife, and D. O'Brien, "LF model based glottal source parameter estimation by extended Kalman filtering," *Proceedings of the 22nd IET Irish Signals and ...*, 2011.

[9] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech.*, vol. 2-3, pp. 121–156, 1995.

[10] C. Gobl, "Voice Source dynamics in connected speech," *Speech Transmission Laboratory, Quarterly Progress and Status Report*, vol. 29, no. 1, pp. 123–159, 1988.

[11] G. Fant and A. Kruchenberg, "Notes on stress and word accent in Swedish," *Stl-Qpsr*, no. 2-3, pp. 125–144, 1994.

[12] D. Günzburger, "Acoustic and perceptual implications of the transsexual voice," *Archives of Sexual Behavior*, vol. 24, no. 3, pp. 339–348, 1995.

[13] R. Veldhuis, "A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 566–571, 1998.