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 construction considers voice production as essentially a system of a power source exciting a sound source (the vocal folds), which emit sound through a series of filters (the vocal tract) e.g. the nasal cavity, the mouth and its shape being manipulated, the tongue and its position, 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] and can be modelled aeroacoustically as jets.

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 very simple but intelligible synthesiser then, we would have to produce a wave rich in harmonics and then put it through an array of filters to model the formants. The reason it is necessary to have a wave rich in harmonics is so there is enough higher frequency content for the formants to pass through to create intelligible speech.

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

Frication noise 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 formants 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.

1.1 Naturalness

To develop natural sounding synthesis it can 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 helps to revisit 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.

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).

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

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. The limitations of this data are that it is gathered from only three speakers, all Swedish, and all male. Nonetheless 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

A feature of real speech is that the fundamental pitch is not constant, it exhibits some flutter [5]. One approach to model this 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 glossary is available in appendix A.4, page I.

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 below.

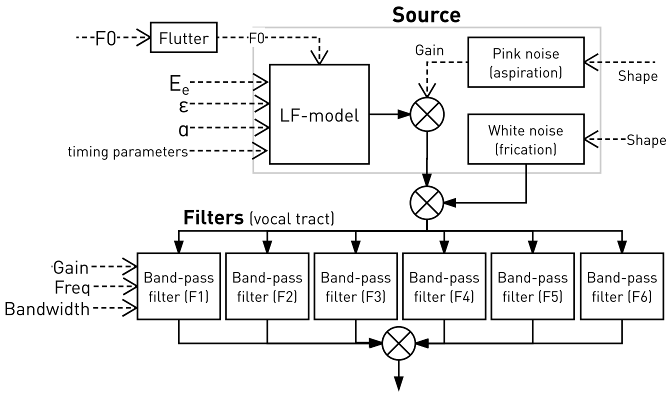

Figure 1: Signal path diagram for the system. Audio signals are represented with black arrows, control signals are represented with dotted arrows.

In the top-level patch 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.

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 cascade 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, although the figure is for cascade rather than parallel formant synthesis so further reading is needed. I reduced my system to four formants and added a fifth to model an additional prominent formant movement in the diphthong of my chosen word.

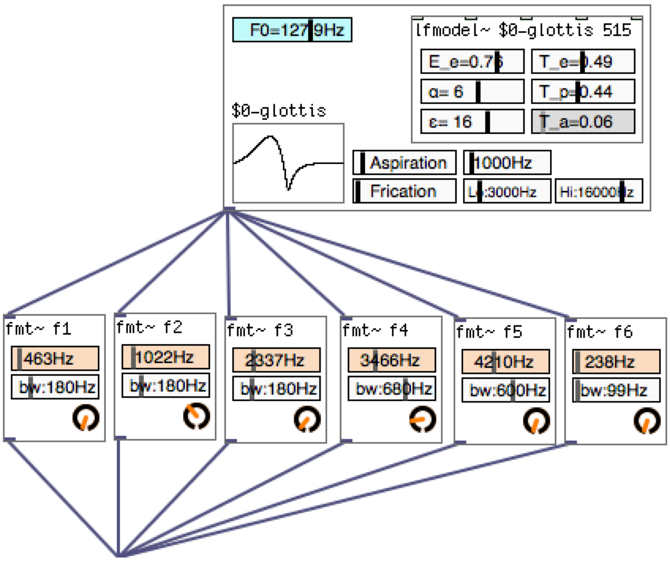

Figure 2: Signal flow in the PD patch: from sound source through the vocal tract

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. This object specifies the bandwidth in octaves so I created some logic to convert to a bandwidth in Hertz to make adjustments quicker. However it does not map the centre frequency precisely because [bandpass] sets the bandwidth logarithmically rather than linearly.

A controller patch (controller.pd, 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.

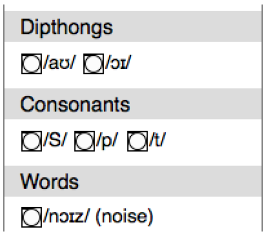

Figure 3: A small section of the [controller] patch showing how presets can be selected.

For a list of the control messages see appendix A.3 on page I. Most control messages are terminated with an interpolation time in milliseconds; the receiver will linearly ramp to the new value in the given time. Sequencing chains of these events is handled either in [qlist] objects or using a 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. This behaviour is defined in msgpipe.pd (see appendix A.4 for more details).

The intention to this design was to create a very flexible synthesis system that could be configured rapidly but without a lot of repetition of details. In the pursuit of naturalness it is equally important to enable a rapid turnaround of analysis of synthesis attempts, as without more sophisticated analysis techniques there is a large element of trial and error.

To that end I also created an object 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.

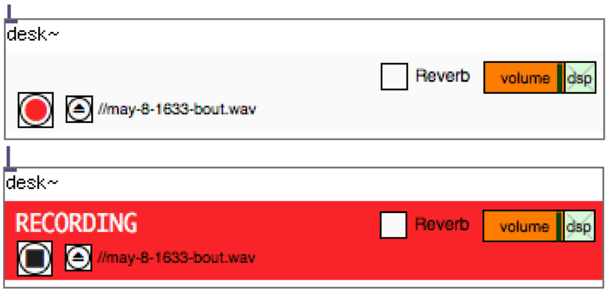

Figure 4: [desk~] object for rapidly saving samples for analysis when stopped (above) and recording (below)

Note that analyser subpatch adapted from forum thread about log-scale graphs.

3 Critical analysis and future improvements

Using 'typical' values for voice source parameters is hard because trans voices differ from typical male or female voices. [11] Particularly 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.

Although the implementation of the LF-model is there I had difficulty utilising it to improve my synthesiser. Determining correct parameters is not trivial as using this 'analysis-by-synthesis' approach to compare the synthesised spectrogram with the reference requires a lot of time in a non-automated system such as this PD implementation.

The additional parameters in the revised model [9] provide further control over the spectrum: increasing R_k raises level of voice fundamental relative to upper parts of the spectrum. Increasing R_g promotes the level of the second harmonic at the expense of the fundamental. This analysis starts to get into the spectral qualities of specific voices, e.g. it is noted that sonorous voices have relatively high F_a of the order of 2000Hz [9].

Fant also discusses a shape parameter, R_d, which predicts the other values, citing a 1994 publication that I was unable to locate a copy of [12]. I tried to implement this as I felt it would make it much more straightforward to find appropriate LF-model parameters for my system, however I had enormous problems matching up the "statistical relations" for the R parameters in the LF-model [9] - particularly due to an initial misunderstanding of the statistical prediction model. The work in progress for this can be seen in the `[lfmodel~]` patch. I would very much like to do more research into this and complete the revisited model.

I found that much of my alterations, though making sense from a theoretical perspective and being audible in contrived examples, made minimal impact when real words were being synthesised. It's probably quite important to do real listening tests.

One issue with my system is that the transitions between parameters, e.g. formant centre frequencies, are linear. This isn't actually true to formants in actual speech which have more gradual ramps between positions. See Figure 5 below for an example in an earlier word. For my final word I compensated for this somewhat by having more interpolation points during the word and ramping between them.

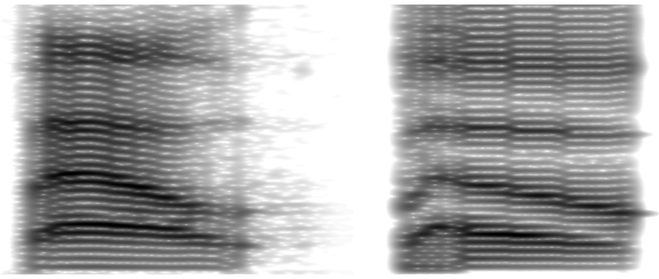

Figure 5: Linear vs. non-linear formant transitions of `/baʊt/` in the reference recording (left) vs synthesised (right).

One compromise I made was to have frication noise production independent of the LF model. In the human vocal tract, the jet stream is interrupted when the glottis is closed. This is why voiced consonants have less power than unvoiced. To implement this I would need to integrate the glottal derivative to get the actual glottal flow, and amplitude modulate the frication noise with this signal. I'm not sure to what extent the difference would be perceptible, but it would perhaps make it easier to get the power of the voiced vs unvoiced source correctly distributed for mixed phonemes. Another choice that may have made the model more helpful was if the frication noise bypassed the formants altogether or used a different formant chain. 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, although the vocal tract does modulate the power that this source receives. Filtering this signal through the vocal tract is therefore removing a lot of the signal that shouldn't be filtered. I compensated for this somewhat by setting the bandwidth of the formants widely.

A problem with my LF-model is that when parameters are adjusted it will rewrite the glottal derivative waveform once per DSP cycle (as soon as one parameter is updated it blocks all others until that rewrite is complete).

This would be fine if it could update the glottal waveform in one DSP cycle (approximately 23µs per cycle). When I timed the calculation on my machine it took 1310µs to complete a cycle which is a significant loss of resolution.

A better solution might be to continually update the glottal waveform, in this case interpolation would need to be done between values to prevent introducing a lot of high frequency distortion when the time-domain amplitude changes rapidly based on a completely different glottal derivative (this effect is somewhat masked by the per-cycle update because the glottal waveform tends to gravitate towards zero at the start and end of the cycle).

An alternative to the LF-model has been proposed that is more computationally efficient and demonstrated to be perceptually equivalent in a psychoacoustic experiment [13]. In order to move the system towards real-time manipulation of the glottal derivative, and or to increase the sample rate, this could be useful to implement.

Sometimes there is interpolation noise/distortion which really is unacceptable and should be eliminated although because PD can be somewhat obtuse in the order it processes things and the timings of commands being sent this can be somewhat difficult.

One problem was that the more control parameters were added to the synthesiser the harder it got to experiment to find the right sound (partially down to a bug in OS X PD where the typing caret does not appear in message boxes). At the start it was much easier to pin down a certain sound before more 'realistic' parameters were added and iteration became much slower.

Can show waveforms of my voice vs. what comes out of the synthesiser. Note it would be nice to be able to record the voice things with an extremely flat microphone at consistent distances with measured freq responses in an anechoic chamber for proper comparison (and normalize the levels).

There's a few things I attempted to develop in this project but that I had trouble with that would be potential room for development. I spent a lot of time researching voice source models to try and build a system that could produce characteristic sound of a human voice. Part of this voice source model involves the voice source parameters changing in response to certain things. For example the acoustic characteristics of the source in one vowel may be different to another, or for consonants, or even due to stress, loudness, tiredness, and endless other factors. Although I did try implementing a system that boiled down, using statistical relations, the voice source parameters for various vowels, I did not manage to get it working due to time and availability constraints. This could form part of a large system that would allow the modulation of the voice in terms of emotion, stress, and prosody. To get an accurate set of voice source parameters I could use inverse filtering to negate the effects of the vocal tract on a voice recording, although it may be difficult to reproduce sound precisely with the LF-model and this approach due to recurrent patterns and randomness in normal human speech [9].

I would like to implement a complete set of IPA phonemes in the synthesiser. For the most part this would be as simple as analysing them in Praat and keying the values into the synthesiser. For some types of consonants this would be more complicated. It seems like an optimised synthesiser would allow automated targeting of speech and would try and fit the parameters to this as best as possible.

The noise for the fricatives just uses PD's `pink` object. Whilst this is more realistic than white noise it doesn't quite share the same spectral properties as the fricative noise actually produced by the body [3]. It would be good to create a more sophisticated model of the source of unvoiced sounds such as there is for voiced sounds.

A more sophisticated model for future implementation could use a physical model which mimics the details of the glottal excitation process [4]. It would be important to determine however that the difference is perceptible.

It has been demonstrated that there's an association between excitation amplitude E_e with F_0 in Swedish. [12]. Assuming this extends to English, it could be a factor for an improved system that can convey some more extra-lingual semantics onto the speech synthesis system, e.g. making the synthesised speech sounding more stressed.

discuss bandwidth of formants not showing much affect on naturalness perception

discuss relatively cumbersome development process in PD vs other processes

limitations of building phoneme by phoneme. success for single words but more involved in full sentences

Additional features such as vocal fry would require more work. [10] states glottal flow for creaky phonation differing substantially.

[5] states transition from a voiceless consonant to a vowel often includes a short interval of breathy voicing in which the first-harmonic amplitude is increased. This could indicate the need to add aspiration noise as well as adjust the voice source parameters to increase the first-harmonic.

[5] notes that breathiness increases for unstressed and final syllables, and at the margins of voiceless consonants.

[5] also discusses scenarios where different formant models, cascade or parallel, are applicable, and amplitude modulation of aspiration and frication noise to simulate the effect of vocal-fold vibration:

". There is a cascade formant model of the vocal-tract transfer function for laryngeal sound sources, and a parallel formant model with formant amplitude controls for frication excitation. A third vocal-tract model in which the vocal-tract transfer function for laryn- geal sound sources is approximated by formants configured in parallel is useful for some specialized synthesis applica- tions, but is normally not used. As was the case in the origi- nal formant synthesizer, the aspiration and frication noise sources are amplitude modulated, to simulate the effect of vocal-fold vibration, if AV is nonzero."

Would probably reimplement in something like Super collider but if had to continue using PD would work more on control primitives so that prototyping is much faster.

As it was not implemented fully in my final design I have removed the details from this report

The 'shape parameter' $R_d = (U_0/E_e)(F_0/110)$. [9] discusses some statistical relations, cited from a 1994 publication that I was unable to locate a copy of. [12]. These are the following predicated values, as they relate to R_d, and an estimation of R_d from the geometrical constraints of the LF model:

$$R_a \approx (-1 + 4.8R_d)/100 \tag{5}$$
$$R_k \approx (22.4 + 11.8R_d)/100 \tag{6}$$
$$R_d \approx (1/0.11)(0.5 + 1.2R_k)(R_k/4R_g + R_a) \tag{7}$$

The main range of variation is $0.3 < R_d < 2.7$, and the upper range is intended for transitions towards complete abduction as in prepause voice terminations.

Here's an example of the variation from [9]: " A pronounced vocal-tract narrowing, as in the [i:] and [y:] and the maximally rounded [u:] and [ʉ:], causes a loss of transglottal pressure which modifies the glottal flow pulse towards a greater R_d value and a somewhat lower E_e."

Higher R_d is typical of female vs male phonations, but also found in voiced consoants and aspirated vowels vs regular vowels. [9]

Klatt [5] discusses a diplophonic double pulsing in whcih pairs of glottal pulses migrate toward one another and the first of the pair is usually attenuated in amplitude. Tends to occur when the fundamental is low (voicing is unstable). (possible further improvement).

Found consonans hadest. [3] notes that it is harder to measure acoustic characteristics of fricatives because there may be several spectral peaks which change relative amplitude from one utterance to another

A Appendix

A.1 IPA transcription and VPM description

/nɔɪz/

A.2 Acoustic characteristics of the vowel and consonants

A.3 Control messages for the synthesis system

A.3.1 Voice source

A.3.2 Formants

A.3.3 Control

A.4 Glossary of custom PD objects

	Arguments	Inlets	Outlets
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.

	Arguments	Inlets	Outlets
defaultarg			

	Arguments	Inlets	Outlets
fmt~			

	Arguments	Inlets	Outlets
formatlabel			

	Arguments	Inlets	Outlets
interface			

	Arguments	Inlets	Outlets
lfmodel~			

	Arguments	Inlets	Outlets
master~			

	Arguments	Inlets	Outlets
`msgpipe`			

	Arguments	Inlets	Outlets
`quasirandomdrift`			

	Arguments	Inlets	Outlets
`rdefault`			

	Arguments	Inlets	Outlets
`recall`			

	Arguments	Inlets	Outlets
`rls`			

	Arguments	Inlets	Outlets
`round`			

	Arguments	Inlets	Outlets
`rqlist`			

	Arguments	Inlets	Outlets
`snapchange~`			

	Arguments	Inlets	Outlets
`source~`			

	Arguments	Inlets	Outlets
`synth`			

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

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

[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.