DEPARTMENT OF ELECTRONIC ENGINEERING
THE UNIVERSITY OF YORK

Voice Production & Synthesis

PureData Source-filter Speech Synthesiser

Y3761870

CONTENTS

1 Principles of voice synthesis 1

2 How the PD patch works 2

3 Critical analysis and future improvements 2

A Appendix I

A.1 IPA transcription and VPM description I

A.2 Acoustic characteristics of the vowel and consonants I

A.3 Control messages for the synthesis system I

A.4 Glossary of custom PD objects I

A.5 Indicative screen-shots of the PD patch III

B Bibliography III

May 2017

1 Principles of voice synthesis

The source-filter theory 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.

This can be modelled relatively simply with some kind of sound source and the vocal tract transfer function (VTTF).

In the case of simple implementation with PD the sound source can be modelled with some kind of wave rich in harmonics, e.g. a sawtooth wave or, as we'll see, a more sculpted waveform, and the VTTF as a parallel set of filters.

- Talk about what formants are - Talk about noise production

Sound can be split into two general categories: voiced sound which involves a repeated waveform with a fundamental and a series of harmonics, and unvoiced sound which can be modelled as noise (discuss splitting of this into frication/aspiration in the model)

Need to discuss how this relates to consonants and vowels

Frication noise in consonants is not white noise, it starts to fall off at about 1kHz and continues to drop roughly linearly until 10kHz when it is approximately zero. [1] Fricatives work like turbulence - the sound of a jet of air hitting an obstacle. [1]

More general stuff can be referenced from [2]

Discuss formants as the resonant freqencies of the cavities in the vocal track [1]

Source-filter theory of speech construction.

used parallel rather than series/cascade formant synthesis. This makes it easier to preserve correct formant amplitudes [3] which makes for easier feedback loop on analysing formants and correcting.

LF-model models differentiated glottal flow rather than real glottal flow.

Revisited LF-model uses a data reduction scheme to adjust parameters using a reduction to a few other control parameters [4]:

Spectral tilt $F_a = 1/(2\pi T_a)$, alternative to $R_a = T_a/T_0$.

$$R_a = \frac{T_a}{T_0} \tag{1}$$

$$R_g = T_0/(2T_p) \tag{2}$$

$$R_k = (T_e - T_p)/T_p \tag{3}$$

$$OQ = T_e/T_0 = (1 + R_k)/(2R_g) \tag{4}$$

$$R_d = (T_d/T_0)(1/110) \tag{5}$$

$$= (U_0/E_e)/(F_0/110) \tag{6}$$

$$\approx (0.5 + 1.2R_k)(R_k/(4R_g) + R_a)/0.11 \tag{7}$$

Use [5] for indications of parameters for typical male speakers (e.g. F_a = 700Hz, R_k = 0.30 R_g = 1.20. Notes E_e tends to be stronger for vowels and weaker for consonants. Voiced consonants weaker than vowels. Some limitations on this data: it's gathered from only 3 speakers, all Swedish, and all male. This also demonstrates the impact of prosody on voice source parameters. Notes that decrease of E_e is generally accompanied by an increase of r_a and r_k.

Increasing R_k raises level of voice fundamental relative to upper parts of the spectrum. Increasing R_a (and thereby decreasing F_a) gives a secondary effect of a relative boost of the fundamental which occurs in breathy phonation. Increasing R_g promotes the level of the second harmonic at the expense of the fundamental. Sonorous voices have relatively high F_a of the order of 2000Hz [4]

The 'shape parameter' $R_d = (U_0/E_e)(F_0/110)$. [4] discusses some statistical relations, cited from a 1994 publication that I was unable to locate a copy of. [6]. 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:

$$R_a \approx (-1 + 4.8R_d)/100 \tag{8}$$

$$R_k \approx (22.4 + 11.8R_d)/100 \tag{9}$$

$$R_d \approx (1/0.11)(0.5 + 1.2R_k)(R_k/4R_g + R_a) \tag{10}$$

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 [4]: " 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. [4]

Using 'analysis-by-synthesis' to determine LF model parameters, comparing recorded sound with that of a synthesiser and then adjusting parameters until the spectrograms are close to each-other, this is the approach used. [4]

Naturalness has been shown to be linked with aspiration noise introduced at higher frequencies in the vowels, and also to the relative strength of the fundamental component (less so in the case of a female voice). [7]

On naturalness of the voice, one approach is a small slowly varying F_0 pseudorandomness [7] (period-to-period flutter) – maybe adjusting the timing parameters of the LF model. The same study 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).

Note [7] discusses the limitations of adding this pseudorandom flutter to F_0, referencing that other efforts have led to a harsh voice quality (Rozsypal and Millar, 1979). They propose a slow quasirandom drift to F_0 contour thru FL flutter control parameter. The sum of three slowly varying sine waves. Suggested FL value of 25

$$\begin{aligned}\Delta F_0 =&(FL/50)(F_0/100)(\sin(2\pi 12.7t) \quad (11)\\ &+\sin(2\pi 7.1t)+\sin(2\pi 4.7t)\text{Hz} \quad (12)\end{aligned}$$

[7] also discusses naturalness by mixing an impulse train and noise as the source waveform (Kate et al 1967; Holmes 1973) - specifying a cutoff frequency below which the source consists of harmonics, and above which the source if lat-spectrum noise. also see rothenberg et al (1975) and makhoul et al (1978)

The LF model exploits the (assumed) commutative relationship between the voice source, vocal tract, and lip radiation, to combine the ffects of the voice source and lip radiation into one model. [8]

2 How the PD patch works

Discrete version of LF-model from [9]

Frication/aspiration noise amplitude modulated by glottal flow (mentioned in [1]) but also mentioned more specifically re: amplitude modulation in a paper I can't remember.

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. [10] 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.

I had enormous problems matching up the "statistical relations" for the R parameters in the LF-model [4] - particularly due to it containing linear models of non-linear parameters. The work in progress for this can be seen in the `lfmodel~` patch.

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.

bp vs biquad vs bpw2

Non-linear ramping/tweening would be good

Discussion of whether frication noise should be emitted when glottis isn't open - tricky to implement in my model

One problem is that when LF parameters are adjusted it only attempts to rewrite the glottal derivative waveform once per DSP cycle. (And as soon as one parameter is updated it essentially blocks all others until that rewrite is complete).

This would be fine if it could update the glottal waveform in one DSP cycle (bearing in mind at 44K it's approx .000022676 seconds or 23µs per cycle). But on my machine when timing it, it took 1.31ms or 1310µs to complete a cycle. So obviously this indicates a significant loss of resolution. I'm not sure what effect receiving new bangs has on this process/what impact interrupt has (which will surely occur in this situation).

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).

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 're-alistic' 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 comparion (and normalize the levels).

Went up to 9 formants.. but this ended up meaning tweaking and iterating was taking a long time so I stripped it back down to 5. Lost some realism on sustained vowels but gained it on words.

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 [4].

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 [1]. It would be good to create a more sophisticated model of the source of unvoiced sounds such as there is for voiced sounds.

Wanted to get LF-model controlled by a single or two parameters and then determine these on a per phoneme basis, but had trouble getting the maths right. In the end due to running out of time just went for manually dialing in the timing parameters. I would really love to explore a more advanced model.

[3] shows an example of a physical model, which "mimics" the details of the glottal oscillation process, with the intent on providing realism not possible in a simplified model.

It has been demonstrated that there's an association between excitation amplitude E_e with F_0 in Swedish. [6]. 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.

An alternative to the LF-model has been proposed that is more computationally efficient and demonstrated to be perceptually equivalent in a psychoacoustic experiment. [11]. Currently for 512 samples the system seems to perform adequately with the LF-model but perhaps for higher-resolutions this alternative would be better suited.

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. [5] states glottal flow for creaky phonation differing substantially.

[7] 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.

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

[7] 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.

A Appendix

A.1 IPA transcription and VPM description

"Bout" – /ba t/
b – bilabial plosive stop voiced a t - alveolar voiceless plosive stop

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
rls			

	Arguments	Inlets	Outlets
round			

	Arguments	Inlets	Outlets
rqlist			

	Arguments	Inlets	Outlets
`snapchange~`			

	Arguments	Inlets	Outlets
`source~`			

	Arguments	Inlets	Outlets
`synth`			

A.5 Indicative screen-shots of the PD patch

B Bibliography

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

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

[3] 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.

[4] 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.

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

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

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

[8] 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.

[9] 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.

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

[11] 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.