IDIAP RESEARCH REPORT

A SIMPLE CONTINUOUS EXCITATION MODEL FOR PARAMETRIC VOCODING

Philip N. Garner Milos Cernak Blaise Potard

Idiap-RR-03-2015

Version of JANUARY 16, 2015

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

A simple continuous excitation model for parametric vocoding

Philip N. Garner, Milos Cernak, Blaise Potard [*]

January 16, 2015

Abstract

We describe a continuous-pitch parametric vocoder suitable for speech coding and statistical text to speech synthesis. The spectral model is based on linear prediction. We show that glottal modelling techniques from recent literature can be cherry-picked to produce an excitation signal with properties known to be useful in the above application areas. We further show that the continuous pitch paradigm can be extended to glottal modelling. The resulting vocoder yields synthetic speech that is generally better than without glottal modelling; it has been used in a parametric speech coding application, and is freely available.

Keywords: Speech coding, statistical text to speech synthesis, complex cepstrum, glottal analysis

1 Introduction

Parametric vocoding is of interest for both statistical text to speech (TTS) synthesis [1], and for (very) low bit rate (VLBR) speech coding [2]. Previous work at Idiap [3, 4] has informally identified the vocoder to be the bottleneck in the creation of innovative TTS synthesis applications such as cross-lingual adaptation. The goal of the present work is to create a "reasonable" vocoder to support research in statistical TTS and VLBR speech coding. Reasonable is taken to mean good quality, if not quite state of the art; robust, and freely available.

Speech coding for the purpose of, say, telephone communication, requires as close to exact reproduction of a specific speech signal as possible. It typically achieves this by encoding sample speech frames directly. By contrast, parametric vocoders assume a model for all aspects of a speech signal, and encode the parameters of the model. Of course, if the model is somehow good, the reproduced speech signal will be very close to the original. However, in general, the parametric vocoder is only able to produce a signal with the same statistical properties as the input signal. In TTS and VLBR coding applications, this caveat about statistical properties is not a restriction.

The source-filter model is an intuitive and common parametric model of speech that can be used in a vocoder. Such a model distinguishes a source or *excitation*, and a filter. The excitation is representative of the vibration of the vocal folds and other non-harmonic sounds; the filter then represents the system defined by the vocal tract. Typical models for the vocal tract include the linear prediction (LP) approach of, e.g., Atal and Hanauer [5] and the STRAIGHT method of Kawahara et al. [6].

Generally speaking, vocal tract models are well understood; models for excitation, however, are less mature. A typical excitation model is to use pitch and (binary) voicing estimates to switch between impulses spaced at the pitch period, and white noise. Such an approach has a characteristic "buzzy" quality, normally attributed to over-harmonisation and/or errors in voicing decisions, and possible mismatches between the glottal impulses of speaker and model. Various techniques have been introduced to mitigate the buzzy quality. These typically involve reducing the periodicity at higher frequencies, often by randomising the phase [7].

The pitch extraction required for this and other types of excitation model has to deal with the occasional lack of voicing. Typically, a voicing estimate is produced, allowing the pitch to only be estimated during voiced segments. Continuous pitch estimates utilising interpolation are also possible [8, 9, 10]. Garner et al. [9] in particular showed that such an estimate can lead to a source-filter vocoder with no

[*] This work was supported by armasuisse (the Competence Center for Procurement, Technology, Real Estate and Geodata within the Federal Department of Defence, Civil Protection and Sport) under the RECOD project, and by the Swiss National Science Foundation under SP2: the SCOPES project on Speech Prosody

voicing estimate, relying instead upon a continuous harmonic to noise ratio (HNR) available from the pitch estimation.

A highly regarded vocal model is the harmonic plus noise model (HNM) of Stylianou [11]. HNM involves identifying and removing each harmonic of the signal, leaving a noise component that can be modelled using LP. It has been applied in *concatenative* TTS [12]. Although it is trickier to apply in statistical TTS, owing to the harmonics being modelled individually, Erro et al. [13] report an implementation. In a comparison by Hu et al. [14], the HNM based techniques were more favourably evaluated than the source-filter models. One possible reason for this is the concept of the maximum voiced frequency employed in HNM. In short, harmonics are not modelled above a certain frequency (around 4 kHz); it is these (missing) harmonics that are believed to lead to the buzzy quality.

More recent excitation models have sought to understand the glottis more thoroughly. Much of this work is based on the LF model of Fant, Liljencrants and Lin [15]. The LF model is defined in terms of time domain parameters; Vincent et al. [16] and Cabral et al. [17] have attempted parameter extraction for TTS, and the model is still widely used [18]. Whether or not the LF model is accurate, it brings to light two or three important concepts (at least from the point of view of the present paper): One is the existence of a glottal formant; the other is the suppression of high frequency harmonics (cf. HNM).

Another key observation from the LF model is that the glottal opening phase can be modelled using a negatively damped second order system, i.e., two poles in the unstable part of the s-plane. This is a maximum phase filter. It follows that techniques distinguishing minimum and maximum phase can be used to infer information about the glottis. The LF model was cast as a linear filter model by Doval et al. [19], who coined the term CALM (causal anticausal linear model). Measures of the glottal formant follow directly [20, 21].

Recent, more general, work considers the whole maximum phase response via the *complex* cepstrum of, e.g., Oppenheim and Schafer [22]. Drugman et al. [23] compare the complex cepstrum with the zeros of the z-transform (ZZT) method, finding both give similar results with the complex cepstrum being faster. Maia et al. [24] subsequently use the maximum phase cepstrum to infer the impulse response of a filter that can be used as the glottal excitation.

Given the techniques introduced and built upon in the above literature, we aim to show that a relatively simple excitation model, which is close to the human voice production process, can be constructed based on the following three insights:

1. The complex cepstrum can be used in conjunction with LP analysis to infer parameters of a CALM-based glottal model.

2. CALM model parameters can be interpolated to infer a continuous CALM glottal model.

3. A CALM glottal model combined with a simple noise model via the continuous HNR has properties similar to HNM systems.

The resulting model is demonstrated qualitatively; a quantitative evaluation remains a subject for future research.

2 The glottal model

2.1 Derivation from LF

The LF model (figure 1) is derived for the most part from fitting curves to inverse-filtered waveforms. The curves are observed to fit well, and authors have shown that synthetic speech based on the fit is well received [16, 17]. However, automatic parameter extraction is difficult. For the present model we invert this situation, defining a model where parameter extraction is straight-forward, but making no quantitative claim about the resulting waveform. Rather, qualitatively it results in a waveform that has similar features.

LF models a glottal flow *derivative*; i.e., the signal that is observed by the encoder, and should be reproduced by a decoder. Notice [15] that the opening phase of the LF model corresponds to a maximum phase conjugate pole pair in the *s-plane*; the closing phase corresponds to a single real minimum phase pole. It follows that a similar *digital* model can be defined, where the opening phase corresponds to a maximum phase pair in the *z-plane*, and the closing phase corresponds to a minimum phase real pole. This the basis of the CALM model [19].

The CALM model is essentially a model of the glottal flow. To yield a derivative, a zero can be added at $z = 1$. However, this has the effect of cancelling out the high frequency roll-off of the real pole. To

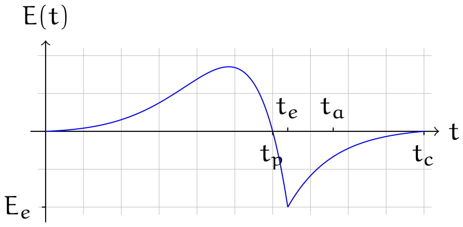

Figure 1: Time domain LF model. Opening phase to the left of t_c, closing phase to the right. t_a is normally an offset from t_c.

counter this effect, we use a double real pole. The minimum phase part of the resulting filter can then be thought of either as single pole plus a high-pass filter, or as a second order critically damped system (double real pole) plus the lip radiation (single zero). The second interpretation is more appealing from a physical point of view. The resulting system is illustrated in figure 2.

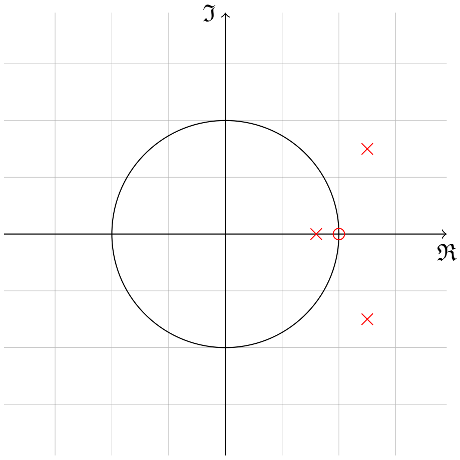

Figure 2: Proposed z-domain glottal model. The pole inside the unit circle is doubled.

2.2 Glottal parameter estimation

Given that the opening phase of the above CALM model is maximum phase, it may be hypothesised that its parameters can be discerned from (complex) cepstral analysis. This is a stronger hypothesis than that of Maia et al. [24] in that it assumes a particular model; nevertheless, it is in the same spirit.

An algorithm follows in the spirit of the work of Drugman et al. [23]. Speech frames are calculated with a period suitable for vocoding (5 to 10 ms), but with a size suitable for glottal period identification (25 ms). For each frame, LP analysis (auto-regression) is performed to determine spectral shape. To identify glottal closures, the (windowed) LP residual is calculated, and the largest peak is chosen as the exemplar glottal closure. Because of the windowing, this peak is normally close to the centre of the frame. The authors note that more advanced glottal closure detectors are available.

The glottal period is estimated using the (continuous) pitch estimation of Garner et al. [9], allowing a frame of two glottal periods around the exemplar to be taken. This shorter frame is windowed using a Nuttall window, which is in the spirit of that recommended [23]. The complex cepstrum is then calculated.

Given the complex cepstrum, the positive (minimum phase) half is set to zero, and the zero'th coefficient divided by two. The inverse DFT then yields the maximum phase spectrum. Squaring (periodogram) and again applying inverse DFT yields the autocorrelation, from which a second order LP analysis can be calculated. The two conjugate poles of this analysis then correspond to minimum phase versions of the maximum phase pair in the above model.

Figure 3 shows a scatter plot of the extracted poles for an example (female) utterance. Notice that two quite broad clusters are evident, illustrating that a glottal formant appears to be consistently extracted. Noisy (complex) and/or failed (real) poles are also present. 180° corresponds to 1000 Hz; the

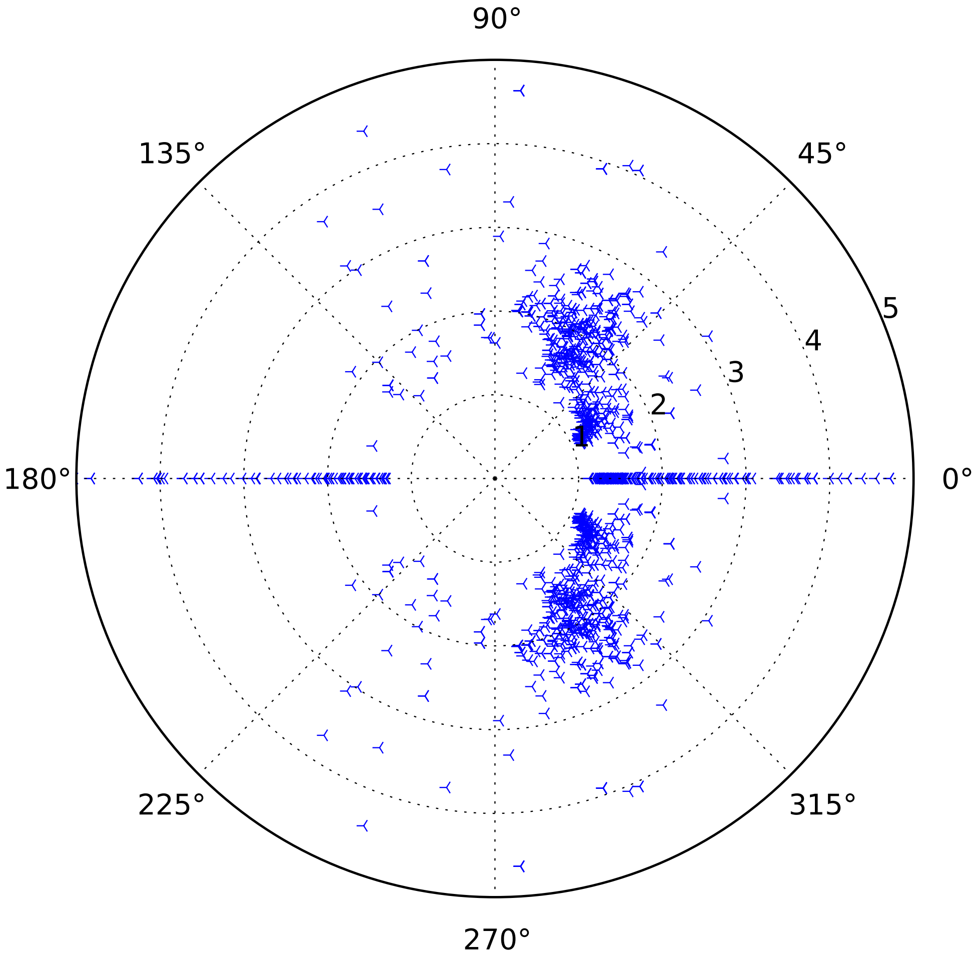

Figure 3: Extraction of glottal formant. The unit circle is the innermost one; all poles are shown as maximum phase. 180° corresponds to 1000 Hz.

periodogram was truncated to 1000 Hz before calculating the autocorrelation. This is to maximise the chance of calculating poles based on voiced segments.

2.3 On phase

The LP analysis always yields the minimum phase solution. In principle, however, the two resulting poles should be maximum phase. In CALM [19], the maximum phase is simulated by applying the filter time reversed. In the present implementation, the filter is applied forwards in a minimum phase sense. We note that in preliminary experiments we were unable to perceive any difference between the two implementations.

2.4 Interpolation

In pitch estimation, the segments that are unvoiced must be handled appropriately. In the case of continuous pitch, these segments are somehow interpolated. Where the vocoder is required to be continuous, the same caveat applies to the estimate of the maximum phase poles. The algorithm above will readily produce estimates for these poles; the difficulty is that below some HNR they will be very noisy; this is visible in figure 3. Although for very low HNR this should not matter since the noise is dominant in the reconstruction, it is *intuitively* of concern for the transition periods between low and high HNR; it also relies on robust HNR estimation.

A solution, the one used in the present implementation, is to use a Kalman smoothed estimate of the maximum phase poles. This is the same algorithm used to smooth and interpolate the pitch in the estimator of Garner et al. [9]. In summary, a function of the HNR is taken as a frame-dependent variance on a harmonic measurement; a Kalman smoother is then able to give reasonable estimates of the

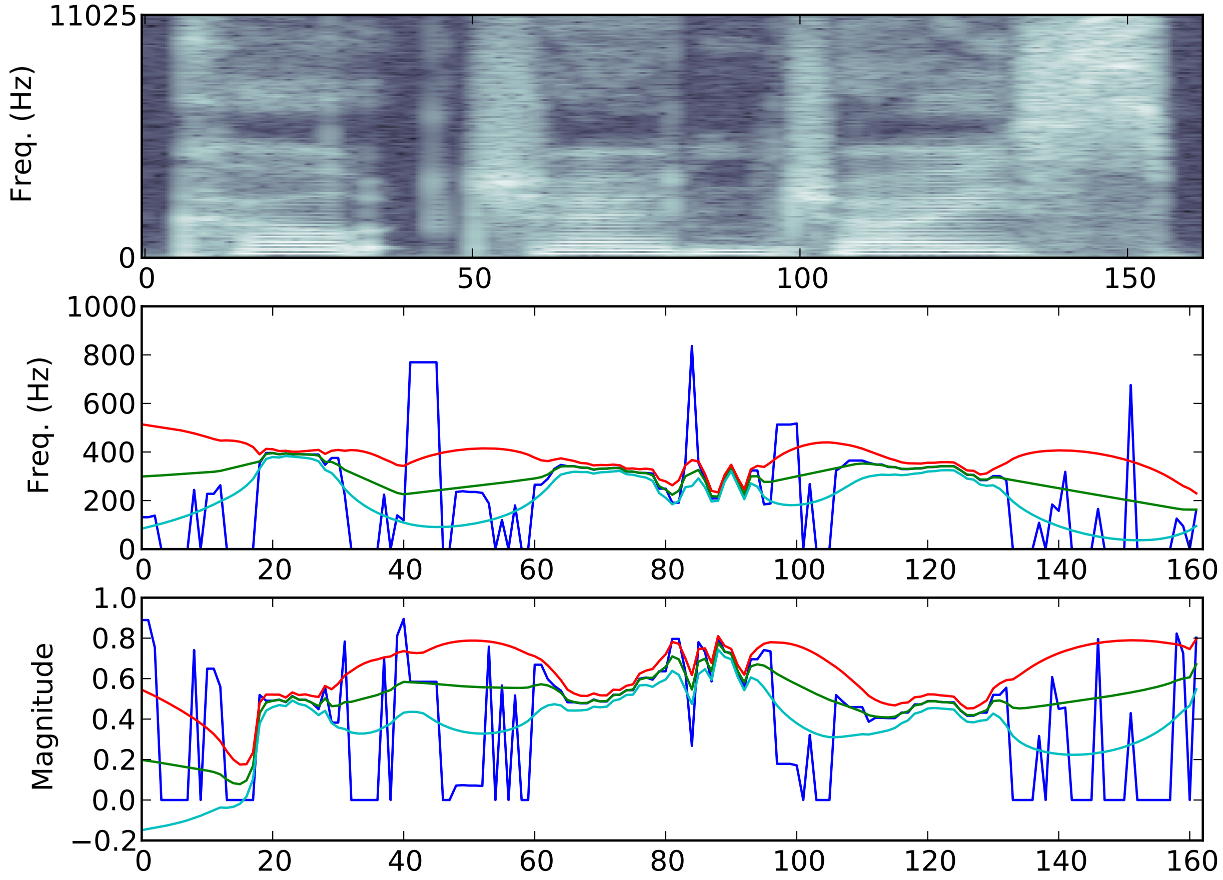

Figure 4: Interpolation of glottal formant. The plots are, respectively, spectrogram, angle and (minimum phase) magnitude. Blue is the raw measurement from LP; the Kalman smoothed, plus and minus one standard deviation are, respectively, red, green and cyan.

measurements where the variance is small (high HNR) and interpolate where the variance is large (low HNR). The pole with a positive argument from the LP is smoothed in polar coordinates. The dynamical system for an observation p_t is defined as

$$p\left(\rho_t \mid \rho_{t-1}\right) \sim \mathcal{N}(\rho_{t-1}, \phi^2), \tag{1}$$

$$p\left(p_t \mid \rho_t\right) \sim \mathcal{N}(\rho_t, \sigma_t^2), \tag{2}$$

where ρ is the state, "$\sim$" is taken to mean "is distributed as" and $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.

In the case of the pole magnitude, $\rho_{m,t}$ (overloading the notation slightly with subscript m being the magnitude and subscript ω the angle),

$$\sigma_{m,t}^2 = \left(\frac{1 - r'(\tau_{\max,t})}{r'(\tau_{\max,t})}\right)^2. \tag{3}$$

The other parameters are $\rho_{m,0} = 0.5$, $\sigma_{m,0}^2 = 1.0$ and $\phi_m^2 = 0.1$. $r'(\tau_{\max,t})$ is the autocorrelation from the pitch tracker [9]. For the pole angle, $\rho_{\omega,t}$, these become

$$\sigma_{\omega,t}^2 = \left(\frac{1 - r'(\tau_{\max,t})}{r'(\tau_{\max,t})} \times (\omega_{\theta hi} - \omega_{\theta lo})\right)^2, \tag{4}$$

with $\omega_{\theta hi} = 500$Hz (converted to radians), $\omega_{\theta lo} = 0$, $\sigma_{\omega,0}^2 = \omega_{\theta hi}^2$ and $\phi_\omega^2 = (\rho_{\omega,0}/4)^2$. $\rho_{\omega,0}$ is set to the mean pitch[1]. Although these parameters are grounded in intuition (how much one might expect a frequency or magnitude to vary per frame in different HNR conditions), there is inevitably a heuristic

[1] We note that the value actually tends to be around twice the pitch.

	Male	Female
Original	EM1_ENG_0001_0	EF2_ENG_0001_0
Naive	EM1_ENG_impulse	EF2_ENG_impulse
Glottal	EM1_ENG_cepgm	EF2_ENG_cepgm

Table 1: Vocoded examples from the EMIME database. Sampling rate is 22 kHz with 24th order LP. Frame rate 12 ms.

element. Note that only one pass is required as the doubling and halving errors associated with pitch estimation do not appear to happen in pole smoothing.

The smoother operation is illustrated in figure 4; notice that the behaviour remains intuitive, with a narrow distribution for voiced segments and a wider but smoothed one for unvoiced (noisy) segments.

The above algorithm addresses only the maximum phase pole pair. We find by experimentation that placing the minimum phase double pole arbitrarily close to the unit circle ($z = 0.99$) is necessary to suppress high frequency harmonics. However, it is still subjective and remains a matter for research; see the discussion below.

2.5 Harmonic plus noise

In the vocoder of Garner et al. [9] (upon which we build), the excitation is a mixture of impulses and noise, added in the ratio implied by the HNR from the pitch estimation. In the present implementation, the impulses are modified by the interpretation of CALM described above. The question arises of how to modify the noise. Certainly a zero to model lip radiation should be added. However, at the time of writing, we have no means to discern other noise shape for excitation; we simply leave it to the LP vocal tract model.

Algorithmically, the CALM and noise signals are simply added according to the HNR. A further LP analysis is done on this signal, and it is inverse filtered. This yields a signal with a flat spectrum suitable to excite the original LP filter modelling spectral shape.

2.6 Discussion

The fact that the excitation is actually a mixture of harmonics and noise raises some rather general questions about the analysis procedure. Where the signal is purely noise, e.g., during fricatives, the model may be expected to work. During purely harmonic segments, it should also work. However, it is less clear whether, for instance, the high frequency roll-off will already have been modelled by the LP. This is a generic issue for glottal analysis by inverse filtering in general.

There is a third case, where the harmonics and noise are mixed more or less equally. In this case, much of the high frequency roll-off of the glottal filter is masked by the noise; it cannot be expected that the LP analysis model it, as the LP cannot distinguish the two components. Rather, in this case, the glottal roll-off can be expected to define a frequency at which the noise is (perceptually) higher than the harmonics. This is entirely analogous to the maximum voiced frequency of the HNM model [11]. It follows that the minimum phase double pole in the CALM inspired model should serve to define a perceptually maximum voiced frequency. Although for the present paper we do not suggest anything other than a heuristic to define the value of the pair, the maximum voiced frequency is one way to proceed [25].

Finally, note that whilst the noise is taken to be flat in the present model, that is unlikely to be true in practise. The minimum phase pair is then taken to be compensating for this lack of noise modelling. In this sense, its perceptually optimal value may not correspond to its actual value insofar as the glottal model is correct.

3 Evaluation

In evaluating the vocoder at this stage in the research, we do not mean to claim superiority or otherwise over other techniques. Rather, the purpose of the present paper is to demonstrate that the insights described in the previous sections do indeed lead to a functioning vocoder. To this end, we present some vocoded examples[2], comparing with the naive impulse plus noise excitation.

[2] See http://www.idiap.ch/paper/2955/ for samples.

	Female
Original	`luke2-77-ryle`
Naive	`luke2-impulse`
Glottal	`luke2-cepgm`

Table 2: Vocoded examples from an audio-book. Sampling rate is 48 kHz, again 24th order LP. Frame rate 8 ms.

	Female
Original	`hisson_005_00097`
Naive	`hisson_005_impulse`
Glottal	`hisson_005_cepgm`

Table 3: Vocoded examples from an audio-book. Sampling rate is 16 kHz, again 24th order LP. Frame rate 8 ms.

Table 1 lists examples from the EMIME database [26]; the same ones used before [9]. In each case, the glottal modelled version sounds more acceptable than the naive version. However, the female voice is noticeably better than the male.

Tables 2 and 3 list examples from audio books, and show the extremes of the current capability. The first of these (luke2) represents the best quality that we have been able to achieve. The second (hisson) represents one that does not work well as well as the naive excitation. This latter case emphasises that the naive excitation is not necessarily too bad in the context of continuous pitch since the excitation is always mixed.

In short, the examples demonstrate that the glottal model has the basic properties that would be expected. However, the performance is dependent upon parameters such as sample rate and frame rate, not to mention speaker. In this sense, it is still work in progress.

Finally, we note that the vocoder has been used in a VLBR coding application [27]. As well as producing acceptable voice quality, this application also confirms that the parameters of the vocoder are suitable (in this case) as inputs and outputs of (deep) neural networks.

4 Conclusions and future directions

We have shown that a continuous pitch vocoder can be improved by continuous glottal excitation modelling. In particular, the combination of complex cepstrum, LP, CALM and continuous pitch can reduce the buzzy quality associated with more naive models. The glottal modelling and parameter extraction are intuitive extensions of known techniques; the novelty is in the combination of technologies rather than the individual technologies.

The work has been validated qualitatively, by demonstration. Good quantitative results will most likely require analysis of the cases for which the techniques work less well. In particular, a better understanding of the relationship between maximum voiced frequency and the minimum-phase double pole should lead to an analytic means to set that parameter. Nevertheless, the framework has been used to improve a VLBR coder [27].

Sophisticated glottal modelling techniques continue to appear. For instance, recent work by Raitio et al. [28] suggests use of neural networks for excitation modelling. Given the interest in neural techniques for TTS in general, such approaches may turn out to be state of the art. However, the present excitation model should lend itself to applications where the parameters need to be adapted or warped.

An implementation of the vocoder is freely available in the SSP (Speech Signal Processing) package on GitHub[3].

References

[1] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1154, November 2009.

[3] https://github.com/idiap/ssp

[2] Milos Cernak, Xingyu Na, and Philip N. Garner, "Syllable-based pitch encoding for low bit rate speech coding with recognition/synthesis architecture," in *Proceedings of Interspeech*, Lyon, France, August 2013.

[3] Hui Liang, *Data-Driven Enhancement of State Mapping-Based Cross-Lingual Speaker Adaptation*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, October 2012.

[4] Lakshmi Saheer, John Dines, and Philip N. Garner, "Vocal tract length normalization for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 7, pp. 2134–2148, September 2012.

[5] B. S. Atal and Suzanne L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, August 1971.

[6] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, April 1999.

[7] Alan V. McCree and Thomas P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, July 1995.

[8] Thomas Ewender, Sarah Hoffmann, and Beat Pfister, "Nearly perfect detection of continuous F_0 contour and frame classification for TTS synthesis," in *Proceedings of Interspeech*, Brighton, UK, September 2009, pp. 100–103.

[9] Philip N. Garner, Milos Cernak, and Petr Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102–105, January 2013.

[10] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014, pp. 2513–2517.

[11] Yannis Stylianou, *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*, Ph.d. thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, January 1996.

[12] Yannis Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, January 2001.

[13] Daniel Erro, Iñaki Sainz, Eva Navas, and Inma Hernáez, "Improved HNM-based vocoder for statistical synthesizers," in *Proceedings of Interspeech*, Florence, Italy, August 2011, pp. 1809–1812.

[14] Qiong Hu, Korin Richmond, Junichi Yamagishi, and Javier Latorre, "An experimental comparison of multiple vocoder types," in *Proceedings of the 8th ISCA Speech Synthesis Workshop*, Barcelona, Spain, September 2013, pp. 135–140.

[15] Gunnar Fant, Johan Liljencrants, and Qi-Guang Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 001–013, 1985, Paper presented at the French-Swedish Symposium, Grenoble, April 22–24, 1985.

[16] Damien Vincent, Olivier Rosec, and Thierry Chonavel, "Estimation of LF glottal source parameters based on an ARX model," in *Proceedings of Interspeech*, Lisbon, Portugal, September 2005, pp. 333–336.

[17] João P. Cabral, Steve Renals, Korin Richmond, and Junichi Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *Proceedings of the 6th ISCA Speech Synthesis Workshop*, Bonn, Germany, August 2007, pp. 113–118.

[18] Fernando Villavicencio, "A strategy for LF-based glottal-source & vocal-tract estimation on stationary modal singing," in *Proceedings of the European Signal Processing Conference*, Lisbon, Portugal, September 2014.

[19] Boris Doval, Christophe d'Alessandro, and Nathalie Henrich, "The voice source as a causal/anticausal linear filter," in *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)*, Geneva, Switzerland, August 2003, pp. 15–20.

[20] Baris Bozkurt, Boris Doval, Christophe d'Alessandro, and Thierry Dutoit, "A method for glottal formant frequency estimation," in *Proceedings of the International Conference on Spoken Language Processing*, Jeju Island, Korea, October 2004, pp. 2417–2420.

[21] Thomas Hézard, Thomas Hélie, and Boris Doval, "A source-filter separation algorithm for voiced sounds based on an exact anticausal/causal pole decomposition for the class of periodic signals," in *Proceedings of Interspeech*, Lyon, France, August 2013, pp. 54–58.

[22] Alan V. Oppenheim and Ronald W. Schafer, *Discrete Time Signal Processing*, Signal Processing Series. Prentice-Hall Inc., Englewood Cliffs, New Jersey, 07632, 1989.

[23] Thomas Drugman, Baris Bozkurt, and Thierry Dutoit, "A comparative study of glottal source estimation techniques," *Speech Communication*, vol. 26, pp. 20–34, 2012.

[24] Ranniery Maia, Masami Akamine, and Mark J. F. Gales, "Complex cepstrum for statistical parametric speech synthesis," *Speech Communication*, vol. 55, pp. 606–618, 2013.

[25] Thomas Drugman and Yannis Stylianou, "Maximum voiced frequency estimation: Exploiting amplitude and phase spectra," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1230–1234, October 2014.

[26] Mirjam Wester, "The EMIME bilingual database," Report EDI-INF-RR-1388, The University of Edinburgh, September 2010.

[27] Milos Cernak, Blaise Potard, and Philip N. Garner, "Phonological vocoding using artificial neural networks'," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, Australia, April 2015, To appear.

[28] Tuomo Raitio, Heng Lu, John Kane, Antti Suni, Martti Vainio, Simon King, and Paavo Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proceedings of the European Signal Processing Conference*, Lisbon, Portugal, September 2014.