INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

GFM-Voc: A real-time voice quality modification system

Olivier Perrotin[1], *Ian McLoughlin*[2]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France
[2]School of Computing, University of Kent, Medway, UK

olivier.perrotin@gipsa-lab.grenoble-inp.fr, I.V.McLoughlin@kent.ac.uk

Abstract

This article introduces *GFM-Voc*, a new system that allows high-quality and real-time voice modification, including both vocalic formants shifting, and voice quality manipulation. In particular, the system is based on the implementation of a newly developed source-filter decomposition method, called GFM-IAIF, that allows the extraction of both vocal tract and glottis spectral envelopes as a compact set of filter parameters. The latter are then controllable through a GUI, before re-synthesis of the speech with the modified parameters. The system requires no training, and operates on any voice, male or female, without tuning. Given the close link between spectral parameters and speech perception, this system provides an intuitive way to independently manipulate the vocalic formants and the spectral shape of the glottal flow that is responsible for voice quality perception. Additionally, rules have been implemented to link the glottis parameters to high-level voice quality parameters such as vocal force and tenseness. Examples of applications for this system include expressive speech synthesis, by adding the system at the end of a speech synthesiser pipeline, auditory feedback perturbation to study a speaker's response to modified speech, and speech therapy.

Index Terms: speech modification, vocal tract perturbation, voice quality, glottal flow model, real-time processing

1. Introduction

Speech synthesis has now reached a point of high naturalness, prompting current research to shift towards expressive speech synthesis [1]. While intonation, stress, and rhythm have been the main features of interest for expressive speech generation, it has been shown that timbre also conveys substantial expressive information, mainly encoded in the glottal signal spectral envelope [2, 3]. Efforts to include a parametrised glottis signal in text-to-speech (TTS) vocoders [4] have shown improvements in the expressivity of the synthesised speech [5]. However, with the advance of end-to-end speech synthesis and the use of neural vocoders, the parametrisation of vocal signals has disappeared, leading to a loss of control in many recent synthesis systems. To re-introduce full timbral control into speech synthesis, we propose a real-time vocal modification system called *GFM-Voc*, that can modify both vocal tract (VT) and glottis parameters of a speech signal without loss of audio quality. While this paper demonstrates the modification of a speaker's voice in real-time, *GFM-Voc* can also post-process speech from any TTS synthesiser, to adjust the voice quality of the synthesised speech.

The extraction of both VT and glottis components is achieved by a source-filter decomposition method called Glottal Flow Model - Iterative Adaptive Inverse Filtering (GFM-IAIF) [6], detailed below. Source-filter decomposition is common practice in speech analysis [7]. Yet, it is not performed in current real-time voice modification systems such as *Audapter* [8], which modifies the VT by processing the full spectral envelope of the signal that also includes glottis information. By contrast, *GFM-Voc* is the first system that allows independent extraction of both VT and glottis components in real-time, and we claim that the disentanglement of VT parameters allows more subtle modification. Moreover, it also enables real-time glottis modification, which to our knowledge has not been proposed before.

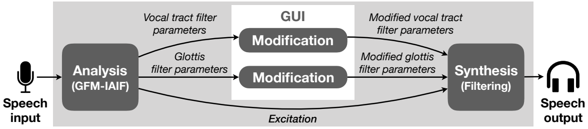
Vocal tract filter parameters
GUI
Modified vocal tract filter parameters
Modification
Analysis (GFM-IAIF)
Glottis filter parameters
Modification
Modified glottis filter parameters
Synthesis (Filtering)
Speech input
Speech output
Excitation

Figure 1: *Schematic diagram of the system framework.*

2. The vocoder architecture

The system architecture is depicted in Fig. 1, showing speech captured from a microphone, vocoded, and then output in real time. The vocoder begins with decomposition using GFM-IAIF, detailed below, to yield the speech excitation signal, and two filters whose frequency responses correspond to the glottis and VT spectral envelopes, respectively. Then, the filters parameters are modified as controlled via a Graphical User Interface (GUI). Speech is finally re-composed by filtering the excitation with the new glottis and VT filters, respectively. The system is coded on Matlab using the Audio Toolbox™ and sampled at 44.1 kHz to ensure high quality audio. It has a latency of 46 ms on a 2012 model MacbookPro with Intel i7 processor. Section 2.1 describes the analysis-synthesis framework further, and section 2.2 details the controllable filter parameters.

2.1. GFM-IAIF for the analysis-synthesis framework

Acoustic speech production theory describes speech as formed by an excitation passing through a first filter *G* describing the glottis signal spectral envelope, and a second filter *VT* that describes the vocal tract spectral envelope [9]. It has been shown that the glottis spectral envelope can be modelled by a 3rd order filter that combines a resonance called the glottal formant (described by its central frequency F_{GF} and bandwidth B_{GF}), and an additional low-pass filter called spectral tilt (described by its cutting frequency F_{ST}) [10]. The *VT* filter is described by a set of resonances, called vocalic formants, described by their centre frequencies F_i and bandwidths B_i.

The GFM-IAIF source-filter decomposition method [6] allows the excitation, *G* filter, and *VT* filter, to be disentangled from the speech signal. The filters are described by sets of LP coefficients of order 3 and 46 for *G* and *VT*, respectively. The parameters described above are extracted from the analogue model of the filters, computed using the bilinear transform [11]. The plots in Fig. 2 display the real-time extraction of the *G* (left)

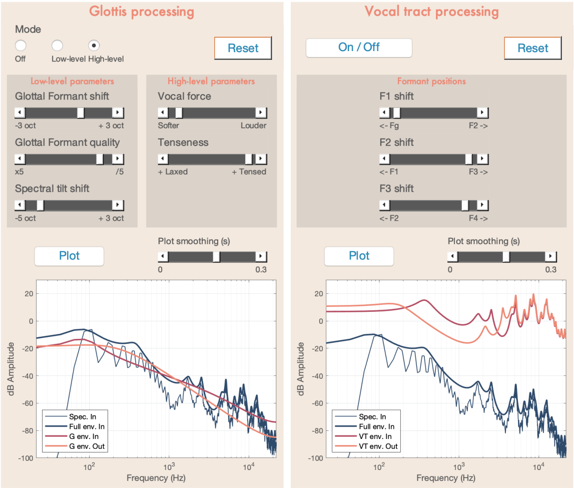

Figure 2: *Screenshot of the GUI which controls the system.*

and *VT* (right) filters in dark pink, superimposed on the speech spectrum (blue) and spectral envelope (thick blue).

2.2. Control of the filters parameters

A GUI, shown in Fig. 2, has been designed to allow real-time filter parameter modification. Basic controls are provided for the sake of demonstration, but could easily be replaced with more sophisticated VT and glottis parameter transformations.

2.2.1. Control of vocal tract parameters

The right panel of Fig. 2 shows the *VT* filter control. It includes the shifting of the three first formants, that are responsible for vowel perception: the first, second and third formants being roughly linked to jaw opening, tongue position (back-front), and lip rounding, respectively. Each formant shift is controlled using a logarithmic scale to account for the log-perception of frequencies, and each formant can be shifted from the previous to the next one. The plot in the lower right shows an example of *VT* modification operating on the input filter (in dark pink). The output filter (in light orange) has the 1st formant shifted downwards and the next two shifted upwards.

2.2.2. Control of glottis parameters

The left panel of Fig. 2 shows the *G* filter control. Two sets of parameters are available. First, a low-level control allows modification of the glottal formant frequency F_{GF}, its quality factor $Q_{GF} = F_{GF}/B_{GF}$, and the spectral tilt cutting frequency F_{ST}. Again, frequencies are shifted on a logarithmic scale. It has been shown that a joint increase in F_{GF} and F_{ST} leads to the perception of higher vocal force [10]. Moreover, an increase in F_{GF} with Q_{GF}, yielding a flatter glottal formant, increases the perception of tenseness in the voice. These relations have also been implemented, from the Cantor Digitalis real-time singing synthesiser rules [12], to provide a high-level control of the glottis parameters. They are summarised in Fig. 3. While the high-level parameters enable the control of physiologically-like parameters, the low-level parameters allow direct control of the glottis to explore the capacity of the model for unconstrained voice modification. The bottom left plot in Fig. 2 shows an example of voice quality modification (in light orange) of the input glottal filter (dark pink) to lower vocal force and increased tenseness. This leads to a flatter and higher glottal formant, and a lower spectral tilt cutting frequency.

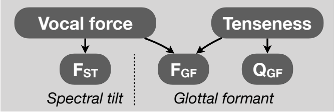

Figure 3: *Mapping of voice quality to glottis parameters.*

3. Conclusions

GFM-Voc is the first framework allowing high-quality and real-time controlled modification of both vocalic formants and voice quality. It can benefit the study of responses to auditory feedback perturbation [13], and has great potential as a post-processor for speech output from expressive speech applications, including TTS. Future developments include excitation signal modification, and more sophisticated high-level rules for filters parameter control.

4. References

[1] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *Proc. of Machine Learning Research*, Stockholm, Sweden, July 10-15 2018, pp. 5180–5189.

[2] D. G. Childers and C. K. Lee, “Vocal quality factors: Analysis, synthesis and perception,” *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[3] C. Gobl and A. Ní Chasaide, “The role of voice quality in communicating emotion, mood and attitude,” *Speech Communication*, vol. 40, no. 1, pp. 189–212, April 2003.

[4] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, “Glottdnn — a full-band glottal vocoder for statistical parametric speech synthesis,” in *Proc. of Interspeech*, San Francisco, CA, USA, September 8-12 2016, pp. 2473–2477.

[5] J. Lorenzo-Trueba, R. Barra-Chicote, T. Raitio, N. Obin, P. Alku, J. Yamagishi, and J. M. Montero, “Towards glottal source controllability in expressive speech synthesis,” in *Proc. of Interspeech*, Portland, OR, USA, September 9-13 2012, pp. 1620–1623.

[6] O. Perrotin and I. V. McLoughlin, “A spectral glottal flow model for source-filter separation of speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, May 12-17 2019, pp. 7160–7164.

[7] P. Alku, T. Murtola, J. Malinen, J. Kuortti, B. Story, M. Airaksinen, M. Salmi, E. Vilkman, and A. Geneid, “Openglot – an open environment for the evaluation of glottal inverse filtering,” *Speech Communication*, vol. 107, pp. 38–47, 2019.

[8] J. A. Tourville, S. Cai, and F. Guenther, “Exploring auditory-motor interactions in normal and disordered speech,” in *Proc. of Meetings on Acoustics*, Montreal, Canada, June 2-7 2013, pp. 1–8.

[9] G. Fant, *Acoustic Theory of Speech Production*. Mouton, 1970.

[10] B. Doval, C. d’Alessandro, and N. Henrich, “The spectrum of glottal flow models,” *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[11] R. Bristow-Johnson. (2001, March) Audio-eq-cookbook. [Online]. Available: https://music.columbia.edu/pipermail/music-dsp/2001-March/041752.html

[12] L. Feugère, C. d’Alessandro, B. Doval, and O. Perrotin, “Cantor digitalis: Chironomic parametric synthesis of singing,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2, 2017.

[13] S. Cai, S. S. Ghosh, F. H. Guenther, and J. S. Perkell, “Adaptive auditory feedback control of the production of formant trajectories in the mandarin triphthong /iau/ and its pattern of generalization,” *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2033–2048, 2010.