Synthesis by Rule of Disordered Voices

Jean Schoentgen[1], Jorge C. Lucero[2]

[1]Department of Image and Signal Processing, Université Libre de Bruxelles, Faculty of Applied Sciences, 50, Av. F.-D. Roosevelt, B-1050, Brussels, Belgium
jschoent@ulb.ac.be
[2]Department of Computer Science, University of Brasilia, Brasilia DF, 70910-900 Brazil
lucero@unb.br

Abstract. The synthesis of disordered voices designates the use of numerical methods to simulate the vocal timbre of speakers suffering from laryngeal pathologies or dysfunctions to investigate the link between perceived timbre and speech signal properties. The simulation is based on a mapping of the amplitude of a narrow-band input signal onto the amplitude of a desired output signal, while the cycle lengths of the input and output are identical. The proposed amplitude-to-amplitude mapping, also known as waveshaping, makes possible simulating a wide range of timbres by fixing the control parameters of a cascade of elementary waveshapers. These enable evolving sample by sample the open quotient, pulse onset and offset rounding, speed quotient and formant ripple of the glottal airflow rate. Preliminary perceptual tests show that the perceived naturalness of the synthetic timbres is comparable to or better than the perceived naturalness of timbres generated via template-based waveshaping.

Keywords: speech synthesis, glottal source modeling, voice quality, voice disorders

1 Introduction

Disordered voices refer to voices that are perceived as abnormal with regard to pitch, loudness or timbre. They are often the consequence of laryngeal pathologies or laryngeal, pulmonary or, occasionally, digestive dysfunctions (e.g. reflux). Synthesis here designates the use of models to numerically simulate speech sounds the timbre of which mimics the quality of disordered voices.

Perceptual evaluation plays a central role in the clinical assessment of speech and voice. The relevance of auditory assessment follows from the communicative function of oral speech that by default relies on the auditory channel. In running speech, voice, that is, the sound produced at the glottis via a pulsatile airflow, has a central place because approximately half the speech sounds are voiced. In addition, voice plays a major role in prosody (e.g. intonation) as well as paralinguistic communication (e.g. speaker attitude) and extralinguistic information (e.g. speaker identity). Speakers deprived of voice may therefore be

considered to be unable to communicate orally. One major unresolved issue in the assessment of speech and voice is the ill-understood link between perceived timbre and data that may be obtained instrumentally.

Synthetic speech stimuli have in the past played a major role in the study of the perception of the phonetic identity of speech sounds. Synthetic stimuli have played, however, a minor role only in the investigation of voice timbre. One reason is that the glottal source is usually modeled via the piecewise concatenations of curves that may request observed voice source signals to which they are fitted. Sample-to-sample updating of the source parameters is therefore not possible and continuity and smoothness constraints at the curve junctures are difficult to implement. Also, curve models do not enable source-tract interaction to be taken into account easily.

One approach that circumvents several of these problems is waveshaping, which maps the amplitude of an input signal onto the amplitude of an output signal. The input signal usually is a narrow-band signal the instantaneous frequency of which can be meaningfully interpreted in terms of its Fourier frequency. The cycle lengths of the input signal are identical to the cycle lengths of the output signal, which is the desired glottal area or flow rate or the derivative of the latter with regard to phase [2][6].

Existing applications of waveshaping to the modeling of the glottal area or airflow rate rest on a template cycle, which is turned via its Fourier series into a polynomial waveshaper [3]. The cycle lengths of the output are fixed via the instantaneous frequency of a driving harmonic. When the amplitude of the driving harmonic is changed from 0 to +1, polynomial waveshapers output cycle shapes that continuously evolve from a constant over a quasi-harmonic to the default template shape.

In this article, we discuss template-free waveshaping of different voice qualities, which is based on a cascade of waveshapers the input-to-output maps of which are so simple that they can be directly formulated mathematically on the base of the desired input-output relations.

2 Methods

2.1 Morphological Features of Glottal Area and Volume Velocity

Morphological features of the glottal area and flow rate waveforms that are relevant to timbre have been investigated by Fant and Titze [1][8]. Auditorily-relevant cues of the glottal area are the open quotient, onset and offset rounding as well as formant bandwidth modulation.

- The open quotient reports the percentage of the glottal cycle length over which the glottis is open. When the ligamental glottis does not close, the open quotient is equal to 100%. But, sound may be produced as long as the folds vibrate and modulate the flow rate.
- The glottis does not open or close abruptly. The area pulse shape has therefore a smooth onset and offset, which has a considerable impact on timbre.

– The bandwidths of the resonances of the vocal tract are modulated by the opening and closing of the glottis that connects or disconnects additional acoustic losses that occur in the trachea and lungs.

The opening and closing of the glottis turns the airflow from the lungs into a pulsatile flow rate that creates sound. Glottal area and flow rate share the open quotient as well as onset/offset rounding, but not cycle shape. Cycle shape skewing and formant ripples as well as the previously mentioned resonance bandwidth modulation are consequences of source-tract interaction.

– The flow rate cycle summit is delayed with regard to the area cycle summit. The delay is a consequence of the inertia of the air in the glottis and pharynx, which does not move instantaneously at glottal opening and which therefore skews the flow rate waveform to the right.
– Ripples at the frequency of the first formant are superimposed on the increasing flow rate at glottal opening. Rippling is a consequence of the vocal source emitting into the vocal tract in which sound waves propagate that disturb the pressure drop experienced by the airflow through the glottis.

2.2 Elementary Waveshapers

A waveshaper comprises one or several memory-less, continuous and smooth maps $y = f(x)$ the input x and output y of which share the same interval so that several waveshapers can be cascaded. The interval $(-1 \leq x, y \leq +1)$ is used the most often, assuming $x(t)$ to be sinusoidal or quasi-sinusoidal.

In waveshapers that comprise two maps, the switch occurs according to whether $\frac{dx}{dt} > 0$ or < 0 to enable increasing or decreasing input signal fragments to be shaped differently. Bi-maps $y = f_{\pm}(x)$ intersect at the interval boundaries, i.e. $f_+(x = \pm 1) = f_-(x = \pm 1)$, and share a common left-hand derivative at $x = +1$ and right-hand derivative at $x = -1$ to guarantee continuity and smoothness of output y.

Hereafter, four elementary waveshapers are discussed that fix (a) the open quotient, (b) pulse onset and offset rounding, (c) the speed quotient, as well as (d) formant ripples. The speed or skewing quotient is the ratio of the time intervals during which the flow rate decreases and increases.

Vertical offsetting. Offsetting consists in shifting sinusoidal input x along the vertical axis to change its duty cycle to fix the open quotient of the glottal area or flow rate waveform.

$$y = (1 - \mathit{shift})x + \mathit{shift}, -1 < \mathit{shift} < +1. \tag{1}$$

Truncating and rounding. Assuming zero leakage, truncating combined with rounding consists in the following.

$y = 0$ if $x \leq x_0$,

$$
\begin{aligned}
&y = x \text{ if } x \geq x_1,\\
&y = interpol(x_0 = x_1 - d_r, y_0 = 0, [\tfrac{dy}{dx}]_{x=x_0} = 0, x_1 = y_1, y_1 = \tfrac{d_r}{\gamma},\\
&[\tfrac{dy}{dx}]_{x=x_1} = 1, x) \text{ if } x_0 < x < x_1.
\end{aligned}
\tag{2}
$$

The interpolator is a cubic Hermite spline, $d_r > 0$ is fixed by the experimenter and $2 < \gamma < 3$ is a parameter that guarantees convex rounding.

Skewing. Formally $y = f_{\pm}(x)$ is written as follows, with $0 < s_k < 1$ fixed by the experimenter.

$$
y = \frac{x}{\sqrt{1 + s_k^2 \pm 2s_k\sqrt{1 - x^2}}} \tag{3}
$$

Waveshaper (3) is a bi-map inspired by [5]. It is easily interpreted and applied only when input $x(t)$ is a sinusoid because then $\pm\sqrt{1 - x^2}$ is the time derivative.

Selecting + or − in front of $\sqrt{1 - x^2}$ defines a map that is a slanted smooth hemi-"8" with $\frac{dy}{dx} \to \pm\infty$ when $x \to \pm 1$. The sign of the limit of $\frac{dy}{dx}$ depends on the sign of $\frac{dx}{dt}$. A thin "8" ($s_k \approx 0$) outputs a quasi-harmonic $y(t)$ the increase and decrease of which are similar in length. A fat "8" ($s_k \approx 1$) outputs an $y(t)$ the extrema of which are delayed with regard to the extrema of $x(t)$ and the cycle shape decrease of which occurs faster than the increase.

Rippling. The waveshaper simulates formant ripple during flowrate cycle increase by adding an exponentially decaying cosine to x when $\frac{dx}{dt} > 0$. No ripple is added when $\frac{dx}{dt} < 0$ because the formants are assumed to be dampened when the glottis is open. Ripple and first formant have opposite phases because positive acoustic pressure in the vocal tract decreases the pressure drop across the glottis and vice versa. Ripple size a_{ripple} is fixed by the user.

The decaying cosine is multiplied by sigmoidal Gompertz curves *gptz* that assign to (4) an envelope with a slope $\frac{dy}{dx}$ equal to +1 when $x = \pm 1$ to guarantee smoothness when the waveshaper switches maps.

Formant frequency F_1 and bandwidth B_1 are normalized with regard to vocal frequency f_0 because $t = 1$ corresponds to one glottal cycle, irrespective of f_0.

$$
\begin{aligned}
&t = 0.5x + 0.5\\
&y = x + gptz(t, a, b, c)(-a_{ripple})e^{-\pi\frac{B_1}{f_0}t}\cos(2\pi\tfrac{F_1}{f_0}t) \text{ if } \tfrac{dx}{dt} > 0 \text{ and } t < 0.5\\
&y = x + gptz(1 - t, a, b, c)(-a_{ripple})e^{-\pi\frac{B_1}{f_0}t}\cos(2\pi\tfrac{F_1}{f_0}t) \text{ if } \tfrac{dx}{dt} > 0 \text{ and } t \geq 0.5\\
&y = x \text{ if } \tfrac{dx}{dt} \leq 0, \text{ with } a = 1,\ b = -6,\ c = -50.
\end{aligned}
\tag{4}
$$

2.3 Modulation Noise and Additive Noise

Other features involved in the genesis of vocal timbre that have been implemented are frequency modulation noise (vocal frequency jitter and vocal frequency tremor) as well as aspiration noise and pulsatile noise owing to turbulent

airflow in the glottis. The size of pulsatile noise evolves with the clean flow rate, whereas aspiration noise is stationary. The added noise is Gaussian white noise low-pass filtered with the cut-off frequency in the interval $600Hz - 800Hz$, depending on perceived strain.

Jitter in natural voices is reported as a single quantity in % [7]. Expression (5) is therefore a quasi-definitional model of vocal frequency jitter depending on a single control parameter a_{jit} and white noise dW that perturbs intonation frequency f_0. Together they determine instantaneous frequency $\frac{d\theta}{dt}$ of the driving sinusoid of the waveshapers that output the glottal area or airflow rate.

$$d\theta = 2\pi f_0 dt + a_{jit} dW \tag{5}$$

However, voices simulated with (5) are perceived as hoarser than expected given the measured cycle perturbations at the glottis. The degree of perceived hoarseness can be brought into line with cycle length perturbations observed in natural voices by linear second-order low-pass filtering perturbations dW in the vicinity of $600Hz - 800Hz$ so that glottal cycle length jitter of 1% causes voices to be perceived as borderline hoarse. A second beneficial consequence of filtering is that the constraint $2\pi f_0 dt > a_{jit} dW_{lowpass}$ is satisfied for modal voices as well as feebly and moderately hoarse voices, which agrees with the usual view that dW in (5) is perturbative. Voices that are perceived as severely hoarse can be simulated by means of (5) only when the previous constraint is not satisfied by all samples of $dW_{lowpass}$.

Vocal frequency tremor is simulated by adding to model (5) Gaussian white noise that is filtered by a linear second-order filter [3]. The gain, center frequency and bandwidth of the filter respectively fix tremor size, tremor frequency and tremor irregularity. Experimenting with tremor and jitter confirms that perceived hoarseness depends not only on perturbation size but also on perturbation bandwidth. Data with regard to the latter are however not reported in the literature, neither for jitter nor tremor.

2.4 Asthenic versus Pressed Voice

During speech, open quotient, pulse onset and offset rounding as well as skewing are not expected to evolve independently for a given phonatory setting. One may expect larger open quotients and onset and offset rounding to co-occur with smaller skewing. The latter gives rise to voices that are perceived as asthenic, whereas the reverse causes voices to be perceived as pressed. Upper and lower limit values have therefore been fixed for parameters *shift*, d_r and s_k in (1), (2) and (3). They are evolved linearly between their extreme values by means of one control parameter comprised between -1 (asthenic) and $+1$ (pressed).

2.5 Modeling of the Glottal Area, Airflow Rate and Speech Signal

The glottal area is modeled by inserting a harmonic driving function into waveshaper (1), followed by truncating and rounding (2). If the user wishes to skew

the glottal area waveform, the harmonic is first inserted into skewing waveshaper (3), followed by (1) and (2) in that order.

The output, which is comprised between 0 and +1, is then multiplied by the maximal glottal area $a_{g,max}$ and inserted into Fant's model that simulates the glottal flow rate taking into account lung pressure and tracheal and vocal tract loads [1]. Fant's model involves solving numerically a differential equation the solution of which may become unstable if rapidly evolving equation parameters are included, such as time-variable formant frequencies and bandwidths.

As an alternative, the glottal flow rate may be simulated exclusively by means of waveshapers by inserting the harmonic driving function into the skewing waveshaper, followed by rippling, offsetting, truncating and rounding. Output $0 \leq y \leq +1$ is then turned into the glottal airflow rate u_g by means of the following expression, with ρ = density of air at human body temperature, K = recovery coefficient ≈ 1 and p_L = lung (excess) pressure [1].

$$u_g = a_{g,max} y \sqrt{\frac{2}{\rho K} p_L (1 - \mathit{shift})} \tag{6}$$

Parameter *shift* that is identical to the control parameter of waveshaper (1) enables simulating soft vowel onsets and offsets when the parameter evolves from +1 to its target value $\mathit{shift}_{target} \geq 0$ and back. In speech, shift_{target} is not expected to be smaller than 0, which corresponds to an open quotient of 50%.

When the glottis is wide open (i.e. $\mathit{shift} = 1$) then flow rate $u_g = 0$ because $p_L(1 - \mathit{shift})$ is zero. When the open quotient is at a minimum (i.e. $\mathit{shift} = \mathit{shift}_{target}$) pulmonary (excess) pressure evolves to its maximum $p_L(1 - \mathit{shift}_{target})$.

Finally, flow rate u_g is inserted into a formant synthesizer. The implementation of the vocal tract resonances closely follows the proposal by Klatt [4]. The sampling frequency of the synthetic speech signal equals $50kHz$. The first three formants are controllable by the user, the following two are fixed at default values that are $3500Hz$ and $4500Hz$ and the final 20 formants up to $25kHz$ are fixed to equally spaced default frequencies and large default bandwidths. The purpose of the latter is to keep the transfer function trend flat up to $25kHz$. Sound radiation is simulated by a numerical derivative of the formant synthesizer output and the synthetic stimuli are saved in .wav format.

3 Results and Discussion

Figures 1 to 4 demonstrate the sample by sample evolution of four morphological features of the glottal flow rate. They respectively show the glottal airflow rate with rounding, skewing, open quotient and formant ripple increasing with time.

Informal listening tests show that the perceived naturalness of the vocal timbre of the sounds synthesized via waveshaping by rule is equivalent to or better than the naturalness of the timbre of speech sounds simulated by means of a template-based waveshaper [3].

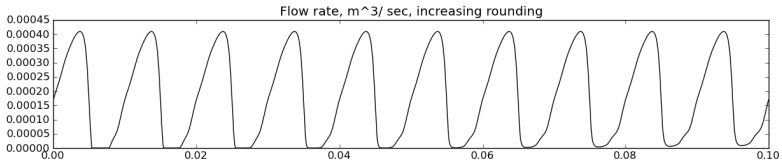

Fig. 1. Glottal airflow rate with increasing pulse onset and offset rounding r_d. The horizontal axis is in s, the vertical axis in $\frac{m^3}{s}$.

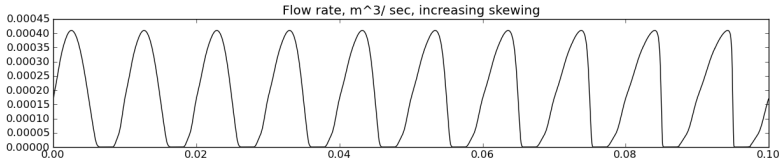

Fig. 2. Glottal airflow rate with skewing s_k evolving from 0 to 0.9. The horizontal axis is in s, the vertical axis in $\frac{m^3}{s}$.

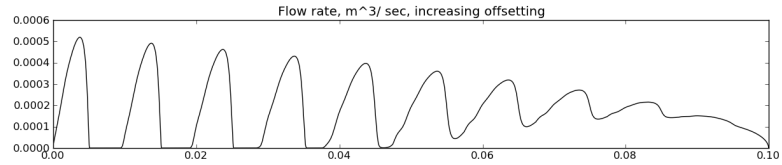

Fig. 3. Glottal airflow rate with offsetting *shift* evolving from 0 to 1. The horizontal axis is in s, the vertical axis in $\frac{m^3}{s}$.

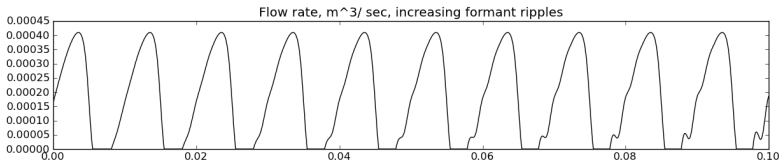

Fig. 4. Glottal airflow rate with formant ripple size a_{ripple} evolving from 0 to 0.4 Formant frequency and bandwidth are equal to $300Hz$ and $100Hz$ respectively. The horizontal axis is in s, the vertical axis in $\frac{m^3}{s}$.

The difference between waveshaping by rule and by template is that for the former the user has a finer control over the morphological and timing features of the glottal area and flow rate, with a concomitant increased risk of selecting combinations of features that are unlikely. Also, waveshaping by rule does not contain an in-built protection against accidental aliasing.

In waveshaping by template, open quotient, rounding and skewing co-evolve plausibly with the amplitude of the driving harmonic. Also, waveshaping by template enables fixing the number of harmonics, making accidental aliasing un-

likely when vocal frequency f_0 evolves slowly. Finally, waveshaping by template enables synthesis to be based on observed glottal areas or flow rates. Polynomial waveshapers are indeed linked to the Fourier series coefficients of an observed template cycle via a linear constant transform [6].

Acknowledgments

Part of the research reported here has been supported by CNPq (Brazil) and F.R.S-F.N.R.S. (French-speaking Community of Belgium).

References

1. T. V. Ananthapadmanaba, G. Fant, Calculation of true glottal flow and its components, Speech Comm. 1, 167-184, 1982.
2. G. Fant, J. Liljencrants, Q. G. Lin, A four-parameter model of glottal flow, STL-QPSR 4, KTH, Stockholm, 1985.
3. S. Fraj, J. Schoentgen, F. Grenez, Development and perceptual assessment of a synthesizer of disordered voices, J. Acoust. Soc. Am. 132, pp. 2603-2615, 2012.
4. D. Klatt, Software for a cascade/parallel formant synthesizer, J. Acoust. Soc. Am. 67, pp. 971-995, 1980.
5. B. T. Peters, J. M. Haddada, B. C. Heiderscheit, R. E.A. Van Emmerik, J. Hamill, Limitations in the use and interpretation of continuous relative phase, J. Biomechanics 36, pp. 271-274, 2003.
6. J. Schoentgen, Shaping function models of the phonatory excitation signal, J. Acoust. Soc. Am. 114, pp. 2906-2912, 2003.
7. J. Schoentgen, Vocal cues of disordered voices, Acta Acustica 92, pp. 667-682, 2006
8. I. Titze, The myoelastic aerodynamic theory of phonation. Denver CO, Iowa City IA: National Center for Voice and Speech, 2006.