Physics-based synthesis of disordered voices

Jorge C. Lucero[1], *Jean Schoentgen*[2], *Mara Behlau*[3]

[1]Department of Computer Science, University of Brasilia, Brasilia DF 70910-900, Brazil
[2]Laboratories of Images, Signal and Acoustics, Université Libre de Bruxelles,
Faculty of Applied Sciences, 50, Av. F. D. Roosevelt, B-1050, Brussels, Belgium
[3]Center for Voice Studies, R. Machado Bittencourt 361, Sao Paulo 04044-001, Brazil
lucero@unb.br, jschoent@ulb.ac.be, mbehlau@uol.com.br

Abstract

The purpose of this work is the development of a synthesizer of disordered voices, i.e., a generator of sounds that mimic the vocal quality of speakers that suffer from a voice disorder or laryngeal pathology. A physics-based modeling strategy is followed for physiological fidelity and a direct relation between the physiology and the perception of vocal quality. The synthesizer is based on a smooth and asymmetric version of a lumped mucosal wave model of the vocal folds, coupled to a wave-reflexion analog of the vocal tract. In this paper, we report our progress and present results of synthetic voices with various levels of vocal frequency jitter, additive pulsatile noise at the glottis due to airflow turbulence, and tension imbalance between the left and right vocal folds.

Index Terms: voice, voice synthesis, voice disorders, vocal folds

1. Introduction

Since the pioneering works of Flanagan and Landgraf [1] and Ishizaka and Flanagan [2] on one- and two-mass representations of the vocal folds, respectively, a number of mathematical models have been proposed to characterize the physics of phonation and simulate the production of voice [3]. The models have been widely used in a variety of both normal and abnormal (pathological) configurations. In the pathological cases, particularly, several studies have shown the capability of the models to reproduce complex oscillation patterns and vocal instabilities observed in patients [4, 5, 6]. However, less attention has been given to the actual synthesis of disordered voices, i.e., the generation of sounds that mimic the vocal quality of speakers suffering from a voice disorder or laryngeal pathology [7]. As pointed out by Fraj et al. [7], past studies on disordered voice synthesis have relied on formant synthesizers and on the generation of a glottal excitation signal by curve concatenation techniques. In their work, Fraj et al. adopted a physical description of the vocal tract using a wave-reflection analog [8]. However, the glottal excitation was still generated from a given curve template of the glottal area.

This paper presents an extension of Fraj et al.'s synthesizer which incorporates a model of the oscillating vocal folds. The purpose of the extension is to increase the physiological fidelity of the synthesizer and allow for a direct control of the synthesized sound in terms of laryngeal parameters.

In order to have a simple control of the synthesizer and facilitate practical applications, simplicity of the vocal fold representation is mandatory. Also, the model must produce smooth variations of the glottal flow. Non-smoothness increases the content of higher harmonics, which results in unnatural timbres. Let us note, for example, that the two-mass model and related representations produce a non-smooth variation of the flow at the glottal closure and, in some versions, at the instants when the glottal channel shape changes from divergent to convergent and vice versa. Further, they are prone to numerical instabilities when the glottal area is close to zero. Although some techniques have been proposed to solve those issues [9, 10], a much simpler solution is proposed here, based on a version of a previous mucosal wave model by Titze [11]. Basically, the model is a single mass-damper-spring oscillator; however, it also incorporates the transfer of energy from the airflow to the vocal folds due to the out-of-phase motion of the entry and exit glottal edges. In this way, it captures much of the oscillatory behavior of the popular and more complex two-mass model. Previous studies have applied various versions of the mucosal wave model to analyze normal phonation dynamics [11, 12, 13]. This paper will also test its capability to simulate disordered voices, both in symmetric and asymmetric glottal configurations.

2. Synthesizer

2.1. Vocal tract

The vocal tract is represented as a sequence of concatenated cylindrical tubes, through which propagates a planar acoustical wave with forward an backward components. Both the subglottal and supraglottal tracts are included, and the later also includes the nasal cavities and paranasal sinuses. Further, losses owing to wall vibration, air viscosity and thermal conduction are considered.

The acoustical wave propagation is solved numerically at a sampling rate of 88.2 kHz, which fixes the length of the elementary tubes at approximately 0.4 cm. The trachea is modeled with 36 tubes with a constant cross-sectional area of 2.5 cm^2 [14], and the geometry of the supraglottal tract is fixed from published data [15, 16]. Details on the models and the numerical algorithms may by found in Fraj. et al.'s work [7]. Here, we focus on the glottal portion of the synthesizer.

2.2. Vocal fold model

Motion of the right vocal fold is described by the equation

$$M_r\ddot{x_r} + B_r(1+\eta_r x_r^2)\dot{x_r} + K_r x_r = P_g, \qquad (1)$$

where x_r is the tissue displacement at the midpoint of the glottis, M_r, B_r and K_r are the mass, damping and stiffness, respectively, per unit area of the vocal fold medial surface, η_r is a nonlinear damping coefficient and P_g is the mean air pressure

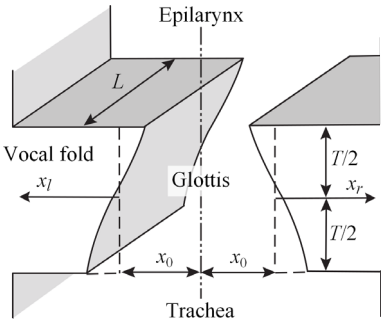

Figure 1: Schematics of the vocal folds.

at the glottis [11, 12, 13]. A similar equation describes motion of the left fold.

When the glottis is open, the mean glottal air pressure is

$$P_g = P_i + \left(\frac{P_s - P_i}{k_t}\right)\frac{a_1 - a_2}{a_1}, \qquad (a_1, a_2 > 0), \quad (2)$$

where P_s is the subglottal pressure, P_i is the supraglottal pressure (at the epilarynx), k_t is a transglottal pressure coefficient, and a_1 and a_2 are the cross-sectional glottal areas at the lower and upper edges of the vocal folds, respectively [11]. The glottal areas are given by

$$a_{1,2}(t) = L[x_0 + x_r(t \pm \tau_r)] + L[x_0 + x_l(t \pm \tau_l)], \quad (3)$$

where L is the vocal fold length, x_0 is half the glottal width when the the vocal folds are at rest, and $\tau_{r,l}$ is the time delay for the surface wave to travel half the glottal height T. The positive and negative signs correspond to a_1 and a_2, respectively.

To simplify the equations, the glottal area factor in Eq. (2) is approximated by the linearization

$$\frac{a_1 - a_2}{a_1} \approx \frac{1}{x_0}(\tau_r \dot{x_r} + \tau_l \dot{x_l}). \quad (4)$$

This linearization eliminates the danger of division by a small number, which may occur in Eq. (2), at the expense of a less accurate representation. Also, the glottal area is evaluated at the midpoint of the glottis, as

$$a(t) = \begin{cases} L[2x_0 + x_r(t) + x_l(t)], \\ \qquad\qquad \text{if } x_r(t) + x_l(t) > -2x_0, \\ 0, \quad \text{otherwise} \end{cases} \quad (5)$$

The glottal airflow is computed using $a(t)$, as explained below, regardless of the values of $a_1(t)$ and $a_2(t)$.

Glottal closure occurs when $a(t) = 0$, and the glottal pressure is set to $P_g = (P_s - P_i)/2$. Contact forces for the colliding vocal folds are modeled through the addition of a spring with stiffness $K_{rc} = 3K_r$.

Values of the various parameters in the equations above were adopted from the cited references, for an adult male configuration, and are presented in Table 1. All units are in cgs system.

Further, a standard value of $P_L = 6000$ dyne/cm^2 was adopted for the lung pressure.

Table 1: *Glottal parameters of the synthesizer.*

Parameter	Value
M_r	0.5
B_r	50
η_r	800
K_r	170 000
τ_r	0.0008
x_0	0.05
L	1.4
k_t	1.1

2.3. Glottal area smoother

The above equations produce a non-smooth variation of the glottal area at the boundaries between the open and closed periods of the glottis. To obtain a smooth variation, interpolation with a cubic Hermite spline $p_3(x)$ is applied. The Hermite spline is defined with endpoints at $x_0 = r(1-\gamma)/\gamma$, $x_1 = r/\gamma$, and $p_3(x_0) = 0$, $p_3'(x_0) = 0$, $p_3(x_1) = x1$, $p_3'(x_1) = 1$. Parameter r controls the level of smoothing and was fixed at $r = 0.2$, $2 < \gamma < 3$ guarantees monotonicity and was fixed at $\gamma = 2.5$. Outside the interval $[x_0, x_1]$, $p_3(x)$ is defined with values $p_3(x) = 0$, for $x < x_0$, and $p_3(x) = 1$, for $x > x_1$.

Thus, a smooth glottal area $a_m(t)$ is computed by $a_m = p_3(a)$. This transformation models the fact that glottal closure is not produced abruptly by flat collision of the opposite vocal folds, but rather that a more or less smooth transition from a fully open to a fully close glottis occurs.

2.4. Glottal flow and coupling to the vocal tract

Calculation of the glottal flow is done following Titze's [17] model

$$U_g = \pm\left(\frac{a_m c}{k_t}\right)\left\{\frac{-a_m}{A^*} + \left[\left(\frac{a_m}{A^*}\right)^2 \pm \left(\frac{4k_t}{c^2\rho}\right)\left(P_s^+ - P_i^-\right)\right]^{1/2}\right\} \quad (6)$$

where A^* is the effective glottal area computed from the input areas A_i and A_s to the supraglottal and subglottal tracts, respectively, as $1/A^* = 1/A_i + 1/A_s$, $\rho = 0.00114$ g/cm^3 is the air density, $c = 35000$ cm/s, is the speed of sound, and P_s^+ and P_i^- are the incident pressures at the glottal entry and exit, respectively. The plus sign corresponds to $P_s^+ \geq P_i^-$ and the minus sign to $P_s^+ < P_i^-$. Once the flow is known, the reflected pressures P_s^- and P_i^+ are computed with

$$P_s^- = P_s^+ - (\rho c/A_s)U_g \quad (7)$$

$$P_i^+ = P_i^- + (\rho c/A_i)U_g \quad (8)$$

2.5. Vocal frequency jitter

Jitter is created by introducing a stochastic perturbation to the stiffness coefficients, which simulates an irregularity of muscle tension. The following stochastic term is added to the left side of Eq. (1):

$$a_r \varepsilon_r K_r x_r \quad (9)$$

where a_r is a scaling coefficient and ε_r is a noise factor. This factor is computed by using, first,

$$\varepsilon_r = \begin{cases} +1, & p = 0.5 \\ -1, & p = 0.5 \end{cases} \quad (10)$$

where p designates de probability, and next band-pass filtering using a second order resonator with a peak frequency at 40Hz and a bandwidth of 150Hz.

2.6. Additive noise

Turbulence noise produced by the airflow is simulated with an additive pulsatile stochastic term proportional to the flow rate,

$$b\varepsilon U_g \tag{11}$$

where b is a scaling coefficient and ε is a noise factor computed as in Eq. (10) and next low-pass filtered with a bandwidth of 1000 Hz.

2.7. Vocal fold asymmetries

A vocal tension imbalance may be simulated by considering an asymmetry in the stiffness coefficient. An imbalance coefficient $Q = K_l/K_r \geq 0$ is introduced, while the left-right values of all other parameters are kept equal. Note that only the interval $0 \leq Q \leq 1$m must be considered. This range of values for Q means that the left vocal fold is less stiff than the right, as in, e.g., a left side paralysis. The smaller Q, the more severe the pathology. If $Q > 1$, then $K_l > K_r$, and a simple change of time scale will rewrite the equations with an equivalent asymmetry parameter $0 \leq Q \leq 1$ applied to the right oscillator instead of the left one.

During the periods of glottal closure, the asymmetric contact dynamics is modeled following Sommer et al. [18]

3. Simulation results

3.1. Normal voices

The above equations were numerically solved using a standard Euler-Maruyama's algorithm [19] implemented in Python.

Figs. 2 and 3 shows results for a normal male configuration during the production of vowel /a/, without any jitter or noise, and with symmetric vocal folds. The glottal area shows some skewing to the left, which is consequence of the nonlinear factor in Eq. (1). Nevertheless, the glottal flow shows the required skewing to the right, as well as an added formant ripple, both consequence of the acoustical interaction with vocal tract.

A male voice generated by the synthesizer is given in the supplementary audio file male_normal.wav. It was generated while decreasing the stiffness coefficient by 10% from the initial values in Table 1, and with jitter and pulsatile noise levels set to zero.

3.2. Frequency jitter and additive noise

Besides the normal voice, eight additional examples of synthetic male voices with symmetric vocal folds and various levels of stiffness perturbation and additive pulsatile noise are included in supplementary files. Their file names, measured jitter and noise SNR are listed in Table 2.

3.3. Asymmetry

Fig. 4 shows results when varying the tension imbalance. In the interval $0.4 \leq Q \leq 1$, the right and left vocal folds oscillate in locked regimes at the same frequency. Within this region, the oscillation frequency decreases as Q decreases, i.e., as the left vocal fold becomes more flaccid and its natural frequency lowers (see Fig. 5). For smaller values of Q, complex patterns appear with intervals of toroidal oscillations (two frequencies in irrational relation), frequency locking in subharmonic regimes,

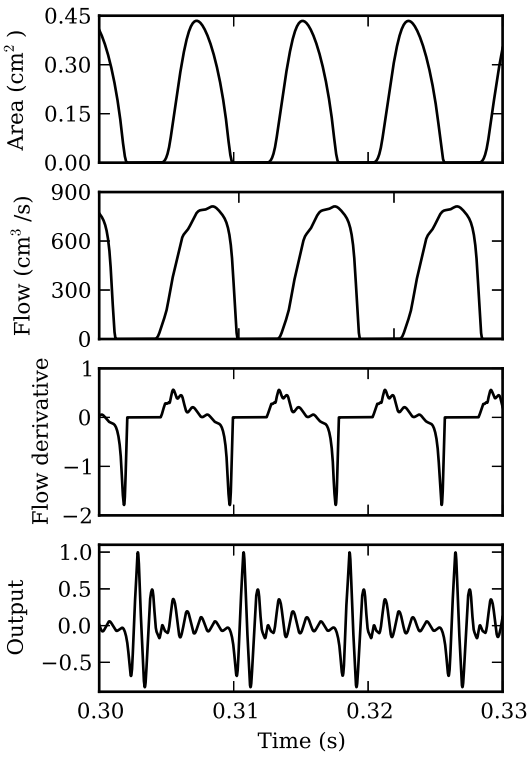

Figure 2: Simulation of a normal male voice during the production of a sustained /a/. The panels, from top to bottom, show the glottal area, glottal airflow, glottal airflow derivative, and radiated acoustic pressure. The flow derivative and radiated pressure are given in arbitrary units.

Table 2: *Simulated voices with symmetric vocal folds.*

File	Jitter (%)	SNR (dB)
male_normal	0.0	220
male_1.wav	0.1	43
male_2.wav	0.1	37
male_3.wav	0.8	220
male_4.wav	0.9	42
male_5.wav	0.7	37
male_6.wav	1.2	220
male_7.wav	1.3	43
male_8.wav	1.2	37

and irregular oscillations. The lower plot shows a detail of the region for $0.15 \leq Q \leq 0.22$.

Three synthetic male voices generated with asymmetric vocal folds, as listed in Table 3, are given in supplementary audio files. They were produced by varying only the Q coefficient, while keeping all other parameters at their normal values in Table 1 (as in voice male_normal.wav). For $Q = 0.5$, the oscillation is frequency-locked at 1:1 regime. The voice pitch is low, compared to the normal synthetic voice, due to the lower natural frequency of the left fold. For $Q = 0.2$, the vocal folds oscillate in a 1:2 frequency-locked regime. The glottal area pattern, in Fig. 6 (top), shows two maxima per glottal cycle. The perceived pitch is much lower, due to the presence of the subharmonic component at half the fundamental frequency. When

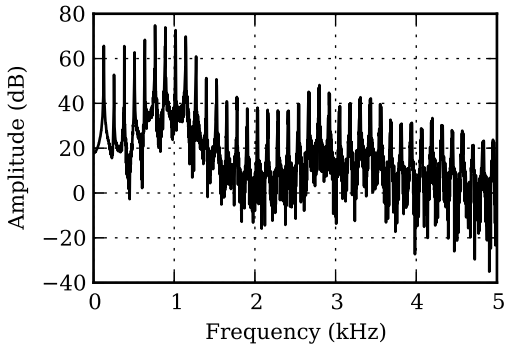

Figure 3: Spectrum of the radiated pressure shown in Fig. 2.

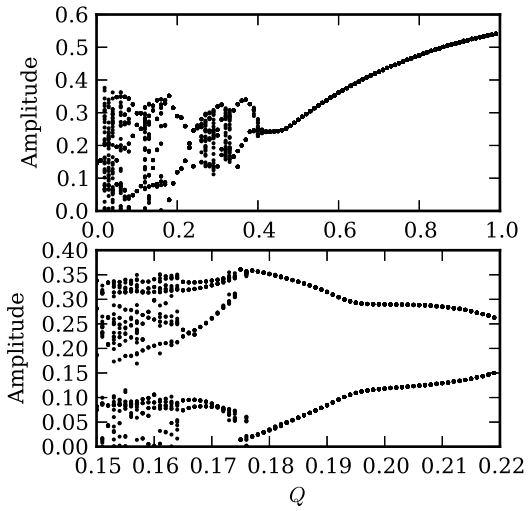

Figure 4: Maximum of glottal area vs. vocal fold tension imbalance Q.

$Q = 0.157$, the oscillation has an irregular pattern, shown in Fig. 6 (bottom).

Table 3: *Simulated voices with asymmetric vocal folds.*

File	Q	Regime
male_asym_1.wav	0.5	1:1
male_asym_2.wav	0.2	1:2
male_asym_3.wav	0.157	Irregular

3.4. Assessment of the synthetic voices

A set of 20 synthesized samples with various levels of jitter, pulsatile noise and tension imbalance was analyzed by three speech-language pathologists, trained clinicians, in order to identify the naturalness of each sample, by consensus. Vocal jitter was the parameter which yielded the most natural samples; followed by reduced levels of pulsatile noise samples and tension imbalance. Breathy voices produced by high levels of

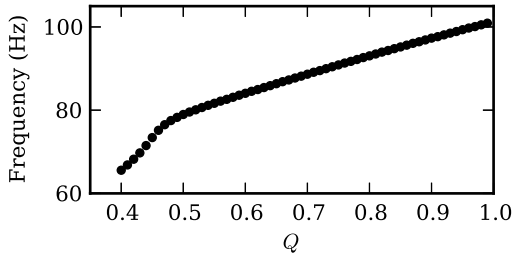

Figure 5: Oscillation frequency vs. vocal fold tension imbalance Q, within the region of 1:1 frequency locking.

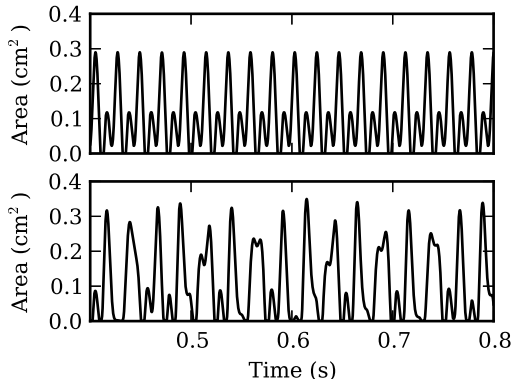

Figure 6: Glottal area patterns in asymmetric vocal folds, for $Q = 0.2$ (top) and 0.157 (bottom).

pulsatile noise appeared to be less natural than all the other samples.

4. Conclusions

The synthesizer is capable of simulating disordered voice sounds with good degree of naturalness and at controlled levels of vocal frequency jitter, additive pulsatile noise, and right-left tension asymmetry. Following similar modeling strategies, other factors may be easily incorporated, such as vocal microtremor, additive aspiration noise, perturbations to the lung pressure and to the glottal adduction, and asymmetries in the mass and mucosa properties of the vocal folds.

The synthesizer is in development stage and much work is still required to improve the output sound. Nevertheless, relevant applications may be foreseen, e.g., in the investigation of the signal properties and laryngeal conditions that cause a particular abnormal vocal quality, for providing controllable synthetic stimuli in perceptual assessment, and as a computational tool for training of clinicians, clinical diagnosis and treatment.

5. Acknowledgments

This work has been supported by a Cooperation Program between CNPq (Brazil) and F.R.S./F.N.R.S. (French-speaking Community of Belgium), and by CAPES (Brazil).

6. References

[1] Flanagan, J. L. and Landgraf, L. L.,"Self-oscillating source for vocal-tract synthesis", IEEE Tran. Audio Eletroacoust., AU-16:57–64, 1968.

[2] Ishizaka, K. and Flanagan,J. L., "Synthesis of voiced sounds from a two-mass model of the vocal folds", Bell Syst. Tech. J., 51:1233–1268, 1972.

[3] Alipour, F., Brucker, C., Cook, D. D., Gommel, A., Kaltenbacher, M., Mattheus, W., Mongeau, L., Nauman, E., Schwarze, R., Tokuda, I. and Zorner, S., "Mathematical models and numerical schemes for the simulation of human phonation", Curr. Bioinform., 6:323–343, 2011.

[4] Ishizaka, K. and Isshiki, N., "Computer simulation of pathological vocal-cord vibration", J. Acoust. Soc. Am., 60:1193–1198, 1976.

[5] Steinecke, I. and Herzel, H., "Bifurcations in an asymmetric vocal-fold model", J. Acoust. Soc. Am., 97:1874–1884, 1995.

[6] Tokuda, I. and Herzel, H., "Detecting synchronizations in an asymmetric vocal fold model from time series data", Chaos, 15:13702, 2005.

[7] Fraj, S., Schoentgen, J. and Grenez, F., "Development and perceptual assessment of a synthesizer of disordered voices", J. Acoust. Soc. Am., 132:2603–2615, 2012.

[8] Titze, I. R., Myoelastic Aerodynamic theory of Phonation, National Center for Voice and Speech, 2006.

[9] Birkholz, P., Krger, B. J. and Neuscheafer-Rube, C., "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis", Proc. Interspeech 2011, 2681–2684, 2011.

[10] Pelorson, X., Hirschberg, A., van Hassel, R. R., Wijnands, A. P. J. and Auregan, Y., "Theoretical and experimental study of quasisteady?flow separation within the glottis during phonation. Application to a modified two?mass model", J. Acoust. Soc. Am., 96:3416–3431, 1994.

[11] Titze, I. R., "The physics of small-amplitude oscillation of the vocal folds", J. Acoust. Soc. Am., 83:1536–1552, 1988.

[12] Laje, R., Gardner, T. and Mindlin, G. B., "Continuous model for vocal fold oscillations to study the effect of feedback", Phys. Rev. E, 64:056201, 2001.

[13] Lucero, J. C., Koenig, L. L., Loureno, K. G., Ruty, N. and Pelorson, X., "A lumped mucosal wave model of the vocal folds revisited: recent extensions and oscillation hysteresis", J. Acoust. Soc. Am., 129:1568–1579, 2011.

[14] Fant, G., Acoustic Theory of Speech Production, Mouton & Co., 1960.

[15] Story, B. H., Titze, I. R. and Hoffman, E. A., "Vocal tract area functions from magnetic resonance imaging", J. Acoust. Soc. Am., 100:537–554, 1996.

[16] Story, B. H., Titze, I. R. and Hoffman, E. A., "Vocal tract area functions for an adult female speaker based on volumetric imaging", J. Acoust. Soc. Am., 104:471–487, 1998.

[17] Titze, I. R., "Parametrization of the glottal area, glottal flow, and vocal fold contact area", J. Acoust. Soc. Am., 75:570–580, 1984.

[18] Sommer, D. E., Erath, B. D., Zañartu, M. and Peterson, S. D., "Corrected contact dynamics for the Steinecke and Kerzel asymmetric two-mass model of the vocal folds", J. Acoust. Soc. Am., 132:EL271–6, 2012.

[19] Kloeden, P. E. and Platen, E., Numerical Solution of Stochastic Differential Equations, Springer, 1995.