Voice source analysis using biomechanical modeling and glottal inverse filtering

Alan P. Pinheiro[1], *Tuomo Raitio*[2], *Danyane S. Gomes*[3], *Paavo Alku*[2]

[1]Department of Electrical Engineering, Federal University of São João del Rei, Brazil
[2]Department of Signal Processing and Acoustics, Aalto University, Finland
[3]University Center of Patos de Minas, Brazil

alan@ufsj.edu.br, tuomo.raitio@aalto.fi, danyane@unipam.edu.br, paavo.alku@aalto.fi

Abstract

This paper studies the use of glottal inverse filtering together with a biomechanical model of the vocal folds to simulate the glottal flow waveform. The glottal flow waveform is first estimated by inverse filtering the acoustic speech pressure signal of natural speech. The estimated glottal flow is used as a template in an optimization process which searches for a set of parameters for a deterministic vocal fold model such that the model output reproduces the estimated glottal flow. The results indicate that the method can reproduce the main deterministic components of the glottal flow signal with good accuracy.

Index Terms: vocal folds, glottal flow, biomechanical simulation, glottal inverse filtering.

1. Introduction

The voice production process comprehends a set of biological systems that involves the lungs, trachea, larynx, as well as oral and nasal cavities. This process starts within larynx, where the vocal folds generate the excitation signal for the vocal tract through vibratory motions that modulate the air from the lungs. This excitation is known as the glottal flow (U_g). Since the vibration of the vocal folds is of utmost importance in production of voiced speech, several techniques have been developed during the past decades in order to get quantitative information from the vocal folds and the glottal flow. These methodologies involve, for example, videolaryngoscopy [1], electroglottography (EGG) [2], and glottal inverse filtering of speech pressure signals [3].

Videolaryngoscopy uses an endoscopy linked to a high-speed video camera which is inserted in the subject's mouth to directly evaluate movements of the vocal folds. However, the endoscope interferes with the natural voice production. EGG uses a pair of electrodes fixed on the neck skin, next to the larynx, sensitive to the vibrational activity present in this region. EGG provides non-invasive analysis of the vocal fold movements and it benefits especially from accurate extraction of glottal closure instants. EGG might, however, suffer from attenuation of some signal components by the tissues (where the electrodes are fixed) hence resulting in distortion of the EGG information. Finally, glottal inverse filtering is based on non-invasive recording of the acoustical speech signal by a microphone outside the lips. The glottal flow estimate is obtained by first forming a digital model for the vocal tract which is then cancelled from the recorded speech signal.

This paper proposes an approach to simulate the glottal flow using deterministic vocal folds modeling combined with glottal inverse filtering of natural speech. Glottal flow pulseforms are estimated with a glottal inverse filtering method from speech pressure signals of ten speakers. Parameters of a biomedical model are optimized in order to reproduce, with a good precision, the glottal flows estimated from real speech.

2. Material and methods

2.1. Speech data

Speech data were collected from 10 adult Finnish speakers (6 females, 4 males) with no history of speech disorders. Acoustic speech pressure signals were recorded with a condenser microphone (Brüel & Kjær 4188) in an anechoic chamber. Each subject produced a sustained [a] vowel using normal phonation. The data were stored into a computer and downsampled with F_s = 8 kHz.

2.2. Glottal inverse filtering

The glottal flow was estimated from the recorded speech pressure signal using a semi-automatic inverse filtering method described in [4]. Inverse filtering analysis utilized all-pole modeling computed by linear prediction (LP) for the vocal tract. The order of LP was varied between 8 and 12 by the experimenter and he selected the filter order that yielded a flow estimate with least formant ripple. The lip radiation coefficient was set to 0.98 (for further details, see [4, 5]). The analysis was computed using a 100 ms frame which was placed in the middle of the signal. The estimated waveforms were parameterized with two glottal flow parameters which can be extracted automatically. As the frequency domain parameter, we used H1-H2 [6] which measures the dB difference between the levels of the first and second harmonic. Time domain features of the glottal flows were quantified with the Normalized Amplitude Quotient (NAQ) [7], which measures the relative length of the glottal closing phase from two amplitude domain values.

2.3. Vocal fold modeling and parameter estimation

2.3.1 Biomechanical modeling

Physical modeling of the vocal folds was conducted with a two-mass model (2M) based on the study by Ishizaka and Flanagan [8] (see Figure 1). This approach is the most widely-used technique to model the vocal fold vibration with two degrees-of-freedom. Albeit there are more sophisticated models [9], some authors [10] have pointed out that the vibration of the vocal folds is largely dominated by the first two modes (i.e., two degrees-of freedom).

The model assumes that both vocal folds can be represented by a pair of coupled oscillators which vibrate due to the aerodynamic interaction between the model and the airflow from the trachea. When the air (of the lungs) is expelled at sufficient velocity through the glottal area (orifice between vocal folds), the folds start to vibrate producing the glottal flow. The characteristics of the speech signal produced depends greatly on the properties of the vocal fold vibration and the glottal flow waveform.

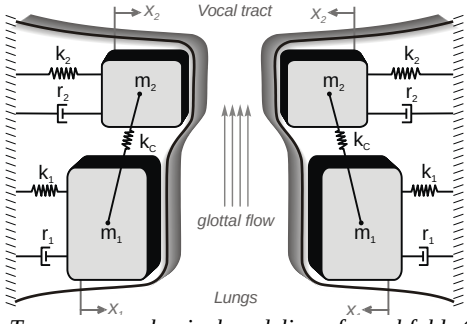

Figure 1: *Two mass mechanical modeling of vocal folds (coronal view). Left and right parts are considered symmetric.*

The two-mass approximation uses the Bernoulli equation to obtain the pressure distribution along the glottal orifice and estimates the displacement x of the masses based on the following equations:

$$\begin{aligned} m_1\ddot{x}_1 + r_1\dot{x}_1 + k_1x_1 + k_C(x_1 - x_2) + I_1(x_1) &= F_1; \\ m_2\ddot{x}_2 + r_2\dot{x}_2 + k_2x_2 + k_C(x_2 - x_1) + I_2(x_2) &= F_2. \end{aligned} \quad (1)$$

In Equation (1), the masses (m_1 and m_2) of the oscillator are driven by mean pressures F_1 and F_2. Terms I_1 and I_2 represent mathematically the (i) collision between the vocal folds in its upper and lower portion (see Figure 1) and (ii) the nonlinear stress-strain curve of vocal folds tissues. Details of the the model can be found in [8].

2.3.2 Optimization procedure for parameter estimation

The most important model factors that define the functioning of the vocal folds are the masses (m_1 and m_2), stiffness coefficients (k_1 and k_2), coupling spring constant (k_c), and the subglottal pressure (P_s). These constants can be mathematically expressed using a parameter vector defined as $\varphi := [m_i, k_i, k_c, P_s \mid i=1, 2]$.

Our goal is to search for a suitable value of φ with which the model can accurately match the glottal flow estimated from natural speech by inverse filtering. In order to reach this goal, an optimization procedure combining genetic algorithms (GAs) and the simplex method was used. This method compares the signal estimated by inverse filtering with the one simulated by the two-mass model using a parameter vector φ searched by the optimization process. Figure 2 describes the major parts of this procedure.

As the model equations might have a non-convex search space, GA is applied in two steps. In the first one, a rough approximation is searched for in order to avoid inappropriate local minima. In the second step, this approximate solution serves as the starting point for a simplex method to refine the approximate solution found by GA. This approach helps in reducing the optimization time.

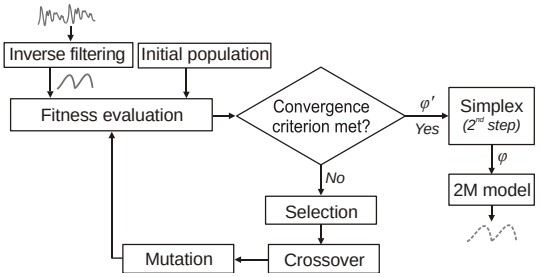

Figure 2: *Flowchart of the optimization procedure.*

GA uses the roulette wheel selection rule [11] and adopts a population of two thousand individuals. The algorithm was programmed to process only one hundred generations to avoid excessive processing time. Each individual in the population represents a potential parameter vector φ for the two-mass model. The search space involved the interval from [1,0.10^{-5}; 5,0.10^{-6}; 10,0; 1,0; 1,0; 600] to [4,0.10^{-4}; 1,0.10^{-4}; 300,0; 80,0; 50,0; 5000] for the six parameters (i.e., chromosomes) of φ. The adopted crossover and mutation probability were 16% and 1%, respectively. Finally, the fitness evaluation was done according to Equation (2) by evaluating the mean squared error between the estimated glottal flow (U_{ge}) obtained by inverse filtering and the simulated glottal signal (U_{gs}) produced by the 2M model using the current parameter set φ.

$$\Psi_1 = \frac{1}{n}\sum_{k=1}^{n}(U_{ge}[k] - U_{gs}[k])^2 \quad (2)$$

In Equation (2), n denotes the sample number of a recorded speech signal. A lower value indicates that the model produces a glottal flow that fits well the signal obtained by inverse filtering. In the second step, the approximate solution found by GA is refined by the simplex method which employs an alternative objective function, defined in Equation (3), based on the ratio between the spectrum of the estimated glottal flow ($E(w)$) and the spectrum of the simulated flow ($\hat{E}(w)$).

$$\Psi_2 = \sqrt{\frac{1}{n}\sum_{w=1}^{n}\left(10\log_{10}\frac{E(w)}{\hat{E}(w)}\right)^2}\ \text{dB} \quad (3)$$

3. Results

To evaluate the accuracy of the method, a validation procedure was developed. Fifty synthetic glottal flow waveforms were generated by the 2M model with predetermined and well known parameters. The estimated parameters set (φ_e) was compared to the true one (φ_s) according to Equation (4). Similarly, the estimated glottal flow (U_{ge}) was compared to the original synthetic glottal flow (U_{gs}) by using Equation (5).

$$E_1 = \frac{1}{m}\sum_{k=1}^{m}\frac{|\varphi_s(k) - \varphi_e(k)|}{\varphi_s(k)}.100\% \quad (4)$$

$$E_2 = \frac{1}{n}\sum_{k=1}^{n}\frac{|U_{gs}(k) - U_{ge}(k)|}{U_{gsp}}.100\% \quad (5)$$

In Equation (5), U_{gsp} denotes the peak value of the synthetic glottal flow waveform, m the number of parameters within φ.

Again, *n* indicates the sample number of a recorded speech signal.

The results of the validation are described in Figure 3. Error terms E_1 and E_2 are shown in the upper and lower panel, respectively, for all the fifty synthetic sounds. The mean and standard deviation of E_1 was 6.9% and 3.1%, respectively. The mean and standard deviation of E_2 was 1.9% and 1.2%, respectively.

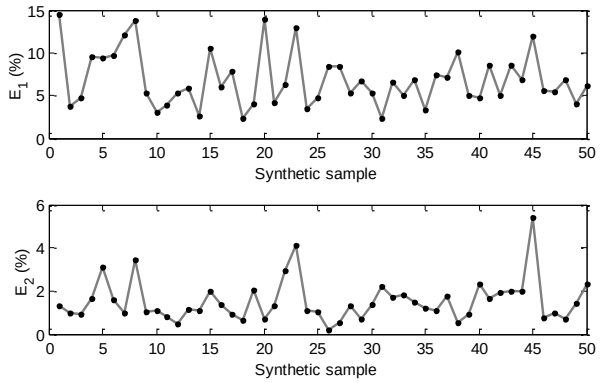

Figure 3: *Validation results. Upper panel: relative error, defined in Equation (4), between the true and estimated parameter set. Lower panel: relative error, defined in Equation (5), between the true and estimated glottal flow.*

By analyzing Figure 3 graphically some observations can be made about the optimization procedure. In sample no. 20, for example, a relatively large value of E_1 is observed. However, the glottal flow waveform produced by this parameter set shows a good match. Probably, a local convergence may explain this result. Nevertheless, the error between simulated and synthetic parameters for this sample is low enough to produce a good approximation to the global solution.

Table 1 shows the parameter values computed from the glottal flows that were obtained by inverse filtering the vowels produced by the ten subjects. Values of error measures computed by Equations (3) and (5) are also given in the table. Glottal flow waveforms, their time-derivatives and spectra are shown for three selected speakers in Figure 4.

Table 1. *Parameters of the 2M model estimated from the speech samples produced by the ten subjects.*

N°	Parameters (m_1; m_2; k_1; k_2; k_c; P_s)[1]	Error	
		Ψ_2 (dB)	E_2 (%)
1	$1.71.10^{-4}$; $5.22.10^{-5}$; 83.9; 4.8; 11.0; 3461	3.4	5.0
2	$7.84.10^{-5}$; $4.16.10^{-5}$; 72.9; 13.0; 19,4; 3656	3.6	5.1
3	$5.97.10^{-5}$; $1.79.10^{-5}$; 85.1; 7.6; 12.3; 3475	3.8	4.3
4	$1.76.10^{-5}$; $2.90.10^{-5}$; 14.7; 18.1; 27.2; 2396	4.1	3.6
5	$1.03.10^{-4}$; $8.51.10^{-6}$; 108.7; 3.4; 6.4; 2851	2.9	4.4
6	$5.88.10^{-5}$; $1.69.10^{-5}$; 90.9; 13.4; 11.8; 4038	4.0	7.1
7	$1.35.10^{-4}$; $3.94.10^{-5}$; 61.2; 18.1; 17.6; 3382	3.4	3.3
8	$1.28.10^{-4}$; $5.32.10^{-5}$; 63.1; 26.6; 26.0; 3841	2.7	2.6
9	$3.50.10^{-4}$; $8.86.10^{-6}$; 134.7; 0.6; 1.5; 3271	3.2	3.1
10	$4.76.10^{-5}$; $5.30.10^{-6}$; 54.4; 4.3; 3.9; 2229	3.1	4.6

[1] Units given in international system of units (Kg; N/m; Pa).

For Subject 2 (Figure 4, left column), the 2M model yields a good match for the glottal flow estimated by inverse filtering. In addition, the flow derivatives show an accurate fit during the glottal closing phase. However, there is a considerable difference between the signals during the closed phase and the beginning of the opening phase. In the frequency domain, the mismatch is most prominent at higher frequencies and is explained mainly by the presence of a stochastic noise component in the flow signal obtained by inverse filtering that cannot be reproduced by the model.

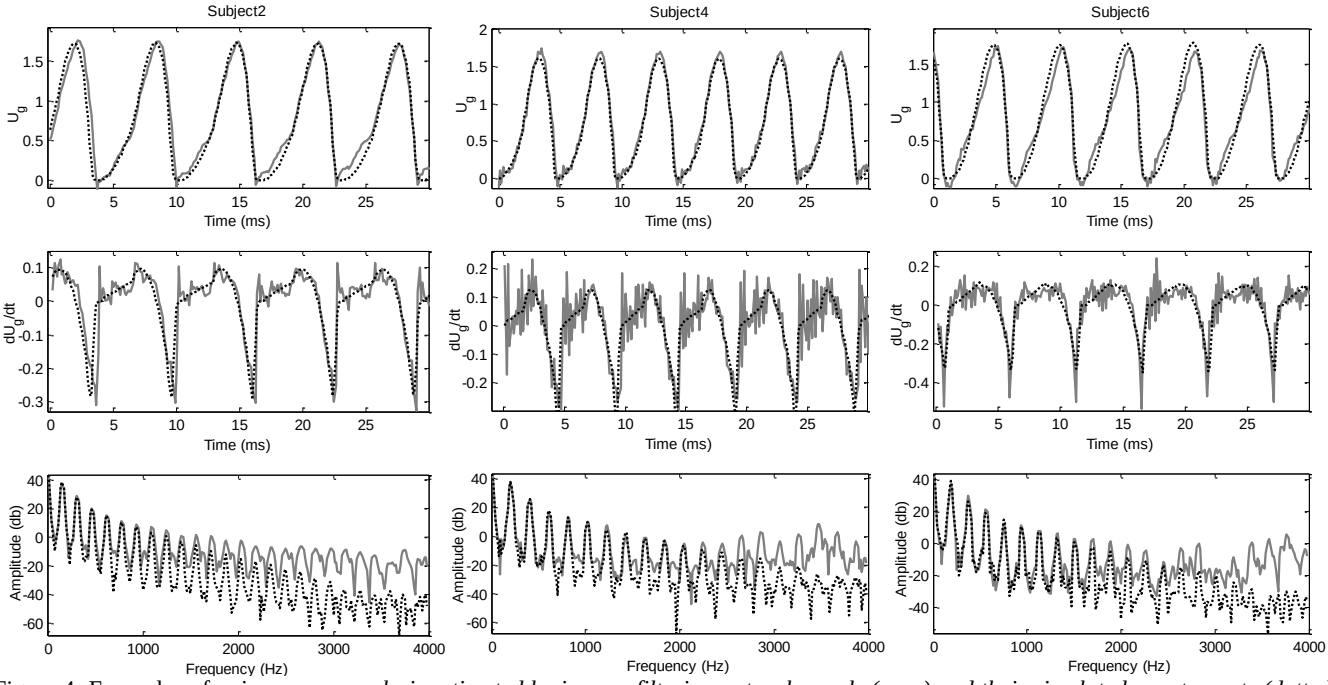

Figure 4: Examples of *voice source analysis estimated by inverse filtering natural vowels (gray) and their simulated counterparts (dotted black). Upper panels: glottal flow. Middle panels: time derivative of the glottal flow. Lower panels: power spectrum of the glottal flow.*

For Subject 4 (Figure 4, middle column), the flow given by the model shows again a good fit with the waveform estimated by inverse filtering. The closing phase (see the time derivative of the flow) is well reproduced as well as its maximum negative peak. Again, noise is more manifested in the closed phase and the beginning of the opening. In the case of Subject 6 (Figure 4, right column), a larger difference is observed throughout the glottal opening phase. However, the closing phase shows a good fit and the derivatives have similar waveforms.

On the whole, the proposed biomedical modeling approach was able to yield a good match for the glottal flows computed by inverse filtering natural vowels from several speakers. The GA reached a solution using one hundred generations and the simplex method refined the solution decreasing the error. The differences between the simulated and estimated flows may be explained by the presence of stochastic noise components in the estimated signals. In addition, it is possible that the glottal flows extracted from natural utterances involved jitter and shimmer, which resulted in increased mismatch, because the deterministic modeling cannot reproduce these common perturbations. Moreover, glottal inverse filtering might have caused artifacts, such as insufficient cancellation of formants, which might have given rise to a noise component that cannot be modeled by the biomechanical model. Finally, limitations of the optimization procedure, such as local convergence, may have been an additional cause for the mismatch in the modeling.

Finally, Figure 5 compares the estimated glottal flows (gray curve) and the modeled ones (black curve) in terms of two objective parameters, NAQ and H1-H2. The mean NAQ, averaged over all speakers, was 0.122 and 0.138 for the extracted and modeled glottal flows, respectively. The mean H1-H2 was 10.64 and 9.96 for the extracted and modeled glottal flows, respectively. These numerical data on H1-H2 indicate that the proposed modeling technique is able to capture the main frequency characteristics of the glottal flow pulses accurately. The larger difference in NAQ values is explained by a slight over-smoothing in the flow derivative waveforms at the instants of glottal closure in the model outputs.

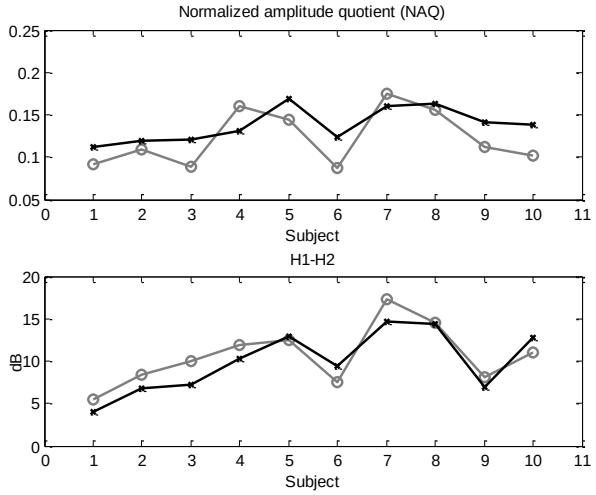

Figure 5: *Glottal flow parameters NAQ (upper panel) and H1-H2 (lower panel) for the ten speakers analyzed. Parameterization was conducted from the waveforms estimated from natural speech (gray) and from simulated flows (black).*

4. Conclusion

In this study, a two-mass model of the vocal folds was utilized together with an optimization process in order to model the excitation of voiced speech, the glottal flow. Parameters of the model were optimized with a genetic algorithm in order to reproduce flow waveforms that were estimated with glottal inverse filtering from natural vowels produced by ten speakers.

The results show that the proposed method can simulate glottal flow signals estimated from natural utterances with a good performance. In employing a single personal computer and a microphone, the procedure benefits from a simple experimental setup. In addition, the computational cost of the procedure is low. The limitations of the technique are related mainly to the optimization procedure, which does not ensure a global optimum. Moreover, the matching between the modeled and estimated flow is affected by artifacts of inverse filtering, such as formant ripple or the use of recording equipment with non-linear phase response.

5. References

[1] Mergell, P., Herzel, H. and Titze, I. R., "Irregular vocal-fold vibration - High-speed observation and modeling", J. Acoust. Soc. Am., 108(6):2996-3002, 2000.

[2] Baer, T., Lofqvist, A. and McGarr, N. S., "Laryngeal vibrations: A comparison between high-speed filming and glottographic techniques", J. Acoust. Soc. Am., 73(4):1304-1308, 1983.

[3] Wong, D., Markel, J. and Gray, A., "Least squares glottal inverse filtering from the acoustic speech waveform", IEEE Trans. Acoust. Speech Signal Process., 27(4):350-355, 1979.

[4] Alku, P., "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering", Speech Communication, 11(2-3):109-118, 1992.

[5] Alku, P., Tiitinen, H. and Näätänen, R., "A method for generating natural-sounding speech stimuli for cognitive brain research", Clinical Neurophysiology, 110(8):1329-1333, 1999.

[6] Titze, I. and Sundberg, J., "Vocal intensity in speakers and singers", J. Acoust. Soc. Am., 91(5):2936-2946, 1992.

[7] Alku, P., Bäckström, T. and Vilkman, E., "Normalized amplitude quotient for parameterization of the glottal flow", J. Acoust. Soc. Am., 112(2):701-710, 2002.

[8] Ishizaka, K. and Flanagan, J. L., "Synthesis of voiced sounds from a two-mass model of the vocal cords", Bell Syst. Tech. J., 51(1):1233–1268, 1972.

[9] Rosa, M. and Pereira, J. C., "Aerodynamic study of three-dimensional larynx models using finite element methods", J. Sound Vibrat., 311(1-2):39-55, 2008.

[10] Berry, D. A., Herzel, H. and Titze, I. R., "Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions", J. Acoust. Soc. Am., 95(6):3595-3604, 1994.

[11] Goldberg, D. E. Genetic Algorithms in Search, Optimization, and Machine Learning.Addison-Wesley; 1989.