Dispersion of Vocal-Fold Biomechanical-Parameter Estimates

Pedro Gómez, Roberto Fernández-Baíllo, Víctor Nieto, Victoria Rodellar, Luis M. Mazaira, Cristina Muñoz, Agustín Álvarez, Rafael Martínez

Grupo de Informática Aplicada al Procesado de Señal e Imagen
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain
pedro@pino.datsi.fi.upm.es

Abstract

Modeling the vocal fold biomechan ical system is relevant for several fields in s peech and voice s tudies, as in voice production, natural speech sy nthesis, em otion estim ation or voice pathology detection. The key stone to good phonation models is the availability of reliable estimates. An indirect method for the estimation of the biomechanical parameters of the voice production model is presented. The availability of previously-normalized voice databases for pathology studies allows the evaluation of the methodology , and the establishment of the distributi on profiles of the parameters under study . The paper illustrates the validation process and the descriptive statistics of the biomechanical parameters.
Index Te rms: voice production, speech synthesis, speech biomechanics, gender-sensitive parameter-distributions

1. Introduction

During the last seven y ears inte rest in the Liljencrants-Fant voice-production source-filter m odel [1] has been retaken in the sense of establishing be tter m odels and estim ation methods of the glottal excitation. For long this pattern had been referred as a 1/f power spectrum signal considering it as a useless signal with nothing else to offer. Nevertheless since the pioneering work of P. Alku [2] contributing to better estimate the glottal source by recursive inverse filtering, a growing interest in this specific signal has contributed to a hatching of different studies, as in the m easurement of the open-close cycle, or the power s pectral profile [3], the correct reconstruction of its causal and anti-causal components [4] or the use of its spectral profile singularities in pathology detection [5]. What is intended in the present work is to derive estimations for the m ost relevant biomechanical parameters, as dynamic masses and elas tic ratios , thes e being es pecially important in model building as we ll as in pathology studies or in voice education and rehabilitation, as well as in natural speech synthesis. The paper is divided into five m ain sections besides the present one. The next section is devoted to present the biom echanical m odel us ed in the study . Section 3 is intended to explain the indirect param eter estim ation methodology. Through section 4 the statistical description of the estimates is given for four different groups of 50 speakers each: m ale and fem ale norm ophonic and dy sphonic, respectively. Section 5 presents and analy zes the res ults obtained for the study sets. Conclusions are presented in section 6.

2. Source-Filter Production Model

The Voice Production Model is depicted in Figure 1 For phonated speech (voiced) the lungs inject a flow of air through the vocal folds to the phary nx, nose and lary nx. The vocal folds (presented in cross-section) vibrate under the action of the forces exerted by the differential pressure between the subglottal and supraglottal sides.

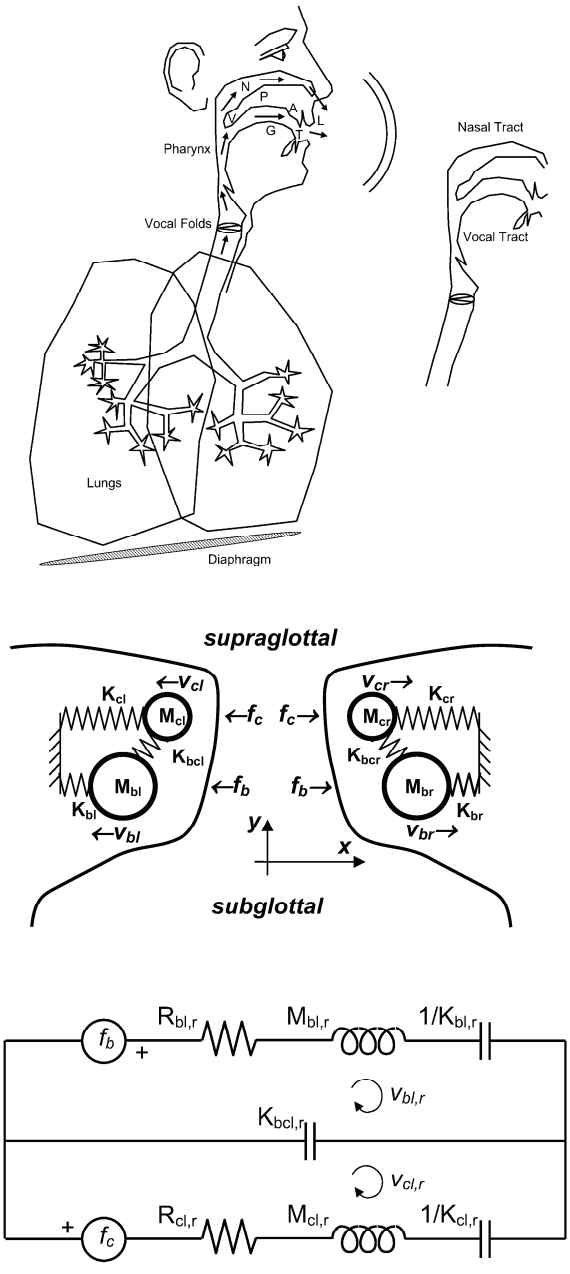

Figure 1 Top: The Voice Production System. Middle: Biomechanical 2-m ass m odel of the vocal folds . Bottom: Electromechanical equivalent.

The system of body (b) and cover (c) masses ($M_{b,cl,r}$) of the left (l) and right (r) vocal folds move against the elastic elements represented by the springs ($K_{b,cl,r}$). The system behaves as the electromechanical equivalent in the bottom section of the figure. This linear model can not cope with nonlinear effects present in the vibration, but is accurate enough to represent the overall behavior of the glottal source (pressure wave exciting the vocal tract as a results of vibration) in the frequency domain (power spectral density). Having in mind that this is an over-simplified model, it must be stressed that its main interest is to be found in the relative feasibility of its inversion by numeric methods for most of the voiced segments of interest in speech studies. The articulation acoustic model is presented in Figure 2 showing the models representing the glottal (voicing) and turbulent (unvoicing) excitations, the articulatory organs (vocal tract) which could be summarized in a transfer function $F_v(z)$ and the radiation effects. is the transfer function of the vocal tract and $g_s(n)$ is the glottal excitation (glottal source) during the voicing segments of speech (unvoiced segments are not considered here as the study is exclusively concentrated in voicing). Classically the glottal source is conceived as a pressure wave resulting from the opening and closing of the vocal folds following the natural vibration of a pair of masses linked elastic tissues to the walls of the larynx. To reconstruct the glottal excitation the influence of the vocal tract has to be removed from voice following the system in Figure 2

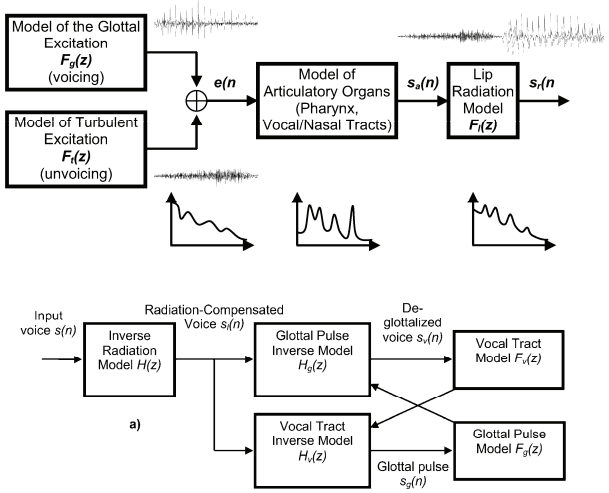

Figure 2 Top: Source-Filter model of Gunnar Fant (see text) widely used in acoustic phonetic studies. The glottal excitation (top trace) enters the chain of tubes modulated by the articulatory organs (vocal tract) and is radiated as phonated speech. If the excitation is not harmonic (bottom trace) the result is unvoiced speech. Bottom: Iterative extraction of the glottal pulse. The voice input is compensated in radiation. A first estimate of the glottal pulse is inverted and used to remove the glottal influence in the input. The resulting de-glottalized voice is modelled to extract the vocal tract model $F_v(z)$, which is used to remove the vocal tract influence on the input voice, giving birth to the glottal source estimation. This signal is modelled to be removed from the input voice by $H_g(z)$, in a criss-crossing iteration which is repeated 2-3 times.

The inversion system is divided in a section to compensate lip-radiation, a second block to model and remove the inverse vocal tract transfer function $H_v(z)$ and a third block to reconstruct the glottal source $g_s(n)$ from the residual signal left at the output of $H_v(z)$: $r_s(n)$. Usually this structure is refurbished as a recursion, as once an estimate of the glottal source is available its inverse sequence may be used to remove the influence of the glottal source in the radiation-compensated speech, thus producing a speech signal which is easier to invert to estimate a good vocal-tract inverse. All these structures: lip-radiation, vocal tract and glottal source cancellers may be implemented as lpc lattices [6]. The main problem now is how to better determine their respective filter orders. Classically lip-radiation cancellers can be first-order lattices. The issue of vocal tract modeling is a little bit more complex, as many possibilities are at hand, the two most popular ones being to select the filter order in the order of the sampling frequency f_s divided by 1,000. This may be enough for the purposes of voice coding, but for an accurate glottal source reconstruction for pathology detection, the adequate order has been established in around twice this order [5].

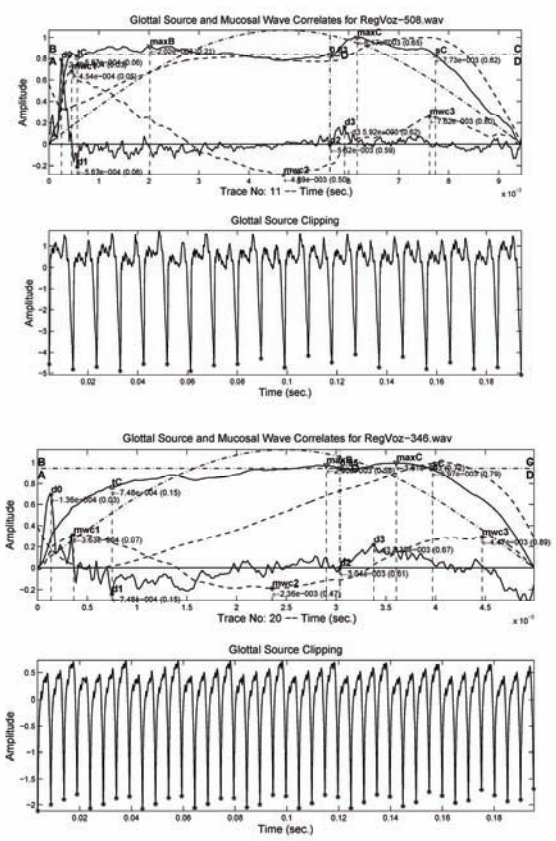

Figure 3 Typical glottal source excitations (thick filled trace) and clipped frames. The spikes mark the instant of closure. Male prototype (top). Female prototype (bottom).

The issue of better modeling the glottal source is a complex one. A first approach would use an order-one or two lattice to reproduce the inverse behavior of the glottal source, but it must be considered that such signal, as presented in Figure 3 has two specific parts: a closed segment which is minimum phase, and an opening segment which is not minimum phase [4]. Therefore Linear Predictive methods will lack precision when reconstructing this last segment. To have this effect into account the order of the lattice modeling the inverse of the glottal source should be at least order-two. Larger orders are not advisable as the zeroes of the glottal source cancellers could lock to the largest poles of the vocal tract. This would result in cross-talk between the estimates of the vocal tract and the glottal source. Reasonable results can be obtained with K_r=1, K_s=2, K_v=32 for a sampling frequency of 16,000 Hz, these being the respective orders of the radiation, source and

tract invers e cancellers . The firs t part of the excitation in Figure 3 is a recovery to the average neutral s ituation (top horizontal line). The start of the opening phase is marked by the thin vertical line, approx. in 4.7 m sec for the m ale cas e, and at 2.4 m sec for the fem ale one. Relative opening happens earlier in the phonation cycle of normophonic female subjects. The maximum opening is marked by the maximum amplitude in the pulse after which the closing phase starts, which is complete at the end of the cy cle. Other reference traces as the average acoustic wave, the m ucosal wave, its first derivative, and the glottal flow are represented as well.

3. Biomechanical parameter estimation

The key point to reconstruct th e glottal source is to use the glottal residual after the rem oval of the vocal tract by $H_v(z)$. The resulting signal g $_r(n)$ can be cons idered as the first derivative of the glottal source g $_s(n)$, therefore the glottal source may be obtained by direct integration of the residual. Once the glottal signal is rec onstructed its power spectral density (psd) may be obtained by taking the pitch-synchronous modulus of the glottal source Fourier Transform cycle by cycle. The behavior of this psd can be seen in Figure 4and is that of a general *1/f* decay function with specific peaks and troughs which are more noticeable in the 0-2 kHz interval. It may be shown that these have to see with the res onances and anti-resonances of the electromechanical equivalent of the vocal folds model [7]. Therefore a direct relations hip between the main peak in the glottal source psd and its gentle *1/f* decay may be established with respect to the mass and elasticity parameters of a biomechanical model as the one in Figure 1.

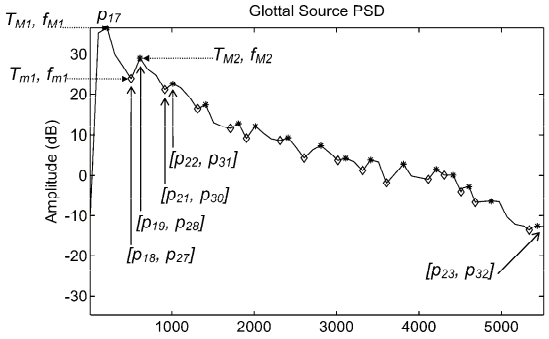

Figure 4 Power spectral density of the glottal source for a phonation cy cle. The singularities of the spectral profile are estimated in term s to their am plitude and position relative to the main peak (T_{M1}, f_{M1})

The process of biom echanical parameter estimation from the glottal source power spectral density is covered by the following steps

$$M_{bl,r} = \frac{\omega_2}{\omega_2^2 - \omega_r^2}\left[\frac{T_r - T_2}{T_r T_2}\right]^{1/2} \quad (1)$$

ω_r being the resonance frequency given by

$$\omega_r^2 = \frac{K_{bl,r}}{M_{bl,r}} \quad (2)$$

where the square modulus of the psd is given by

$$T(\omega) = \frac{1}{\left[R_{bl,r}^2 + \varpi^2 M_{bl,r}^2\right]^2} \quad (3)$$

with the frequency relative to the resonance point (maximum)

$$\varpi = \frac{\omega^2 - \omega_r^2}{\omega} \quad (4)$$

and

$$T_r = T(\omega = \omega_r) = \frac{1}{R_{bl,r}^2} \quad (5)$$
$$T_2 = T(\omega = 2\omega_r)$$

The estimation procedure must detect the value of pitch, which is used to evaluate ω_r. The determination of T_r and T_2 is carried out on the power spectral density of the glottal source spectral profile. This leads to the determ ination of the losses from (5) and to the m ass (1) a nd stiffne ss (2). $M_{bl,r}$ are the equivalent (dynamic average) masses of the vocal fold body , and $K_{bl,r}$ are the (dy namic average) elas tic coefficients of the vocal fold (body), estimated pitch-sy nchronously at the k-th phonation cycle.

4. Validation: materials and methods

For the validation of the methodology a data set of 100 normophonic speakers of both genders have been selected from a previously -tested database used in pathology studies [8]. These speakers have been inspected by endoscopy to discard organic alterations of the vocal folds, their phonation cycle has been inspected using stroboscopic illum ination to check its apparent norm ality, have been GRBAS graded [9] and automatically checked by an ins pection tool us ing acoustic analy sis (classical dist ortion m easurements as jitter, shimmer, or harmonics -to-noise ratio: HNR) [5]. As a control a set of 100 dy sphonic speakers of both genders have been used for contrast. The inspection methodology is the following:

- 200 msec. of phonations of the vowel /ah/ have been selected from each speaker
- The glottal s ource is recons tructed for each speaker following Figure 2.
- Phonation cy cles are being detected, which for male speakers yield some 20 arches, this being about twice for the typical female speaker.
- The glottal source psd is estimated for each cycle.
- Jitter, shimmer and HNR are evaluated in a per-cy cle basis.
- The first, second and third p eaks in the glottal source psd profile are detected, as well as the two first troughs.
- The average dy namic m ass, elas ticity and los s are estimated for each cycle.
- The unbalances between neighbor phonation cy cles for each parameter are also evaluated.

For inspection purposes the c entroids of the respective four clusters are evaluated (MN: male normophonic; MD: male dysphonic; FN: female normophoni c; FD : female dy sphonic) and the respective Mahalanobis distances of each sam ple to the gender-respective normophonic centroids are estimated:

$$D_{ti} = \left[(y_{ti} - \mu_M)^T C_M^{-1} (y_{ti} - \mu_M)\right]^{1/2} \quad (6)$$

Where μ_M and C_M are the centroid and the Covariance m atrix of the Model Sets (normophonic subjects). Figure 5 shows the respective histograms of the Mahalanobis distance of each subject respective to its respective gender normophonic cluster.

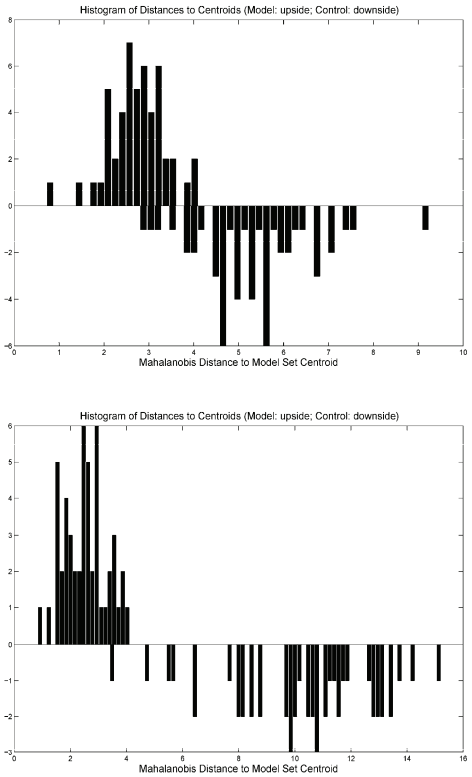

Figure 5 Top: Histograms of Mahalanobis distances of normophonic (upside) and dysphonic (downside) male subjects to normophonic centroid (model set). Bottom: Id. for female subjects.

It may be seen that both genders are clearly differentiated according to the statistical distributions of their parameters.

5. Results and discussion

The statistical description of each parameter distribution is given as well in Table 1 in terms of means and standard deviations.

Table 1. Statistical description of the biomechanical parameter estimates. MN: Male Normophonics; MD: Male Dysphonics; FN: Female Normophonics; FD: Female Dysphonics

Group	Ave. Pitch (Hz)	Std. Pitch (Hz)	Ave. M_b (mg)	Std. M_b (mg)	Ave. K_b (dyn.cm^{-1})	Std. K_b (dyn.cm^{-1})
MN	113.8	11.8	21.7	2.2	11101	1101
MD	113.4	22.8	22.0	4.2	11251	2347
FN	203.4	19.6	12.0	1.2	19682	1897
FD	189.0	40.3	14.0	4.0	19397	8690

The first interesting result is that the average pitch is almost the same for the normophonic and dysphonic male sets, but the dispersion is almost twice for the dysphonic set. Concerning the female sets the pitch is larger for normophonics than for dysphonics, and the dispersion is clearly larger for dysphonic than for normophonic. This result is consistent with classical assumptions. Regarding the fold body dynamic mass, this is slightly larger than 20 mg both for normophonic and dysphonic male subjects, although the dispersion is again larger in this last case. For female subjects the average mass estimates are 12 and 14 mg, which is consistent with what was expected from anatomic expectations. The dispersion in this case is much larger for female than for male dysphonics, which is also consistent with classical assumptions. Regarding elastic parameter estimates, again the difference between male normophonics and dysphonics is not significant as far as averages are concerned, dysphonics being twice more disperse. In the case of females, the elastic parameter is larger, which implies a tighter vocal fold, as expected, the dispersion being also much larger in dysphonics than found in males.

6. Conclusions

The methodology used for the determination of the biomechanical parameters for both genders is consistent with classical assumptions: male speakers present average dynamic body masses almost twice those of the female group, whereas the elastic tension is larger in female than in male in a proportion of around 80%. Dispersion is larger in the female group in general. The conclusion is that the indirect estimations are in good agreement with both gender characteristics. The contrast of indirect measurements with direct evaluation is not possible in living subjects, the possibility of doing this same study with excised larynges being out of the scope of the present research. The results give interesting hints on plausible values to be used in 1-mass biomechanical models, the methodology being extensible to 2-mass models as well. These models are especially interesting for voice synthesis with more realistic glottal excitations in voice synthesis. This is also the case for artificial larynges after radical laryngectomy. Other application field is the biometric description of the speaker.

7. Acknowledgements

This work has been funded by grants TIC2003-08756, TEC2006-12887-C02-01/02 and TEC2009-14123-C04-03 from Plan Nacional de I+D+i, Ministry of Science and Technology, by grant CCG06-UPM/TIC-0028 from CAM/UPM, and by project HESPERIA (http.//www.proyecto-hesperia.org) from the Programme CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain.

8. References

[1] Fant, G., et al., "A four-parameter model of glottal flow", STL-QSPR 4 (1985) 1-13. Reprinted in *Speech Acoustics and Phonetics: Selected Writings*, G. Fant, Kluwer Academic Publishers, Dordrecht, 95-108, 2004.

[2] Alku, P., "An Automatic Method to Estimate the Time-Based Parameters of the Glottal Pulseform", Proc. of the ICASSP'92, II/29-32, 1992.

[3] Orr, R. et al., "An investigation of the parameters derived from the inverse filtering of flow and microphone signals", Proc. of VOQUAL'03, Geneva, 35-40, 2003.

[4] Doval, et al., "The voice source as a causal/anticausal linear filter", Proc. of VOQUAL03, 15-19, 2003.

[5] Gómez. P. et al., "Glottal Source Biometrical Signature for Voice Pathology Detection", Speech Communication, (51): 759-781, 2009.

[6] Deller, J. R., et al., Discrete-Time Processing of Speech Signals, Macmillan, NY, 1993.

[7] Gómez, P., et al., "Biomechanical Parameter Fingerprint in the Mucosal Wave Power Spectral Density", Proc. of the ICSLP'04, 2004, pp. 842-845.

[8] Project MAPACI: http://www.mapaci.com.

[9] M. Hirano et al., Acoustic analysis of pathological voice. Some results of clinical application, Acta Otolaryngologica 105 (5-6): 432-438, 1988.