# BioMet®Phon: A System to Monitor Phonation Quality in the Clinics

Pedro Gómez, Victoria Rodellar, Víctor Nieto, Rafael Martínez, Agustín Álvarez
Neuromorphic Speech Processing Lab
Centro de Tecnología Biomédica, U. Politécnica de Madrid
Campus de Montegancedo, 28223 Pozuelo de Alarcón
Madrid, Spain
e-mail: pedro.gomez@ctb.upm.es

Bartolomé Scola[1], Carlos Ramírez[2], Daniel Poletti[1], Mario Fernández[2]
ENT Services
[1]Hospital Universitario Gregorio Marañón
C/ Máiquez 7, 28009 Madrid, Spain
[2]Hospital del Henares
Avda. Marie Curie s/n, 28822 Coslada, Madrid, Spain

***Abstract*— BioMet®Phon is a software application developed for the characterization of voice in voice quality evaluation. Initially it was conceived as plain research code to estimate the glottal source from voice and obtain the biomechanical parameters of the vocal folds from the spectral density of the estimate. This code grew to what is now the Glottex®Engine package (G®E). Further demands from users in laryngology and speech therapy fields instantiated the development of a specific Graphic User Interface (GUI's) to encapsulate user interaction with the G®E. This gave place to BioMet®Phon, an application which extracts the glottal source from voice and offers a complete parameterization of this signal, including distortion, cepstral, spectral, biomechanical, time domain, contact and tremor parameters. The semantic capabilities of biomechanical parameters are discussed. Study cases from its application to the field of laryngology and speech therapy are given and discussed. Validation results in voice pathology detection are also presented. Applications to laryngology, speech therapy, and monitoring neurological deterioration in the elder are proposed.**

***Keywords: speech therapy; voice quality analysis; dysphonia.***

## I. INTRODUCTION

In this paper, we give an overview on an end-user-driven application to study the glottal source and its associated mucosal wave [1] for voice quality assessment, pathology detection and classification. The glottal source may be seen as the pressure build-up in the glottis just above the vocal folds in the laryngeal cavity. It is the result of the phonation cycle, seen as a sequence of openings and closings of the vocal folds under the influence of lung pressure and vocal fold visco-elasticity and air dynamics [2]. The glottal source is expected to follow closely the pattern proposed by G. Liljencrants and G. Fant [3] known as the L-F pattern given in Figure 1. The L-F profile (top blue line) is the result of simulating the flow of air from the lungs to the vocal tract through the glottis as the vocal folds open and close (the equivalent light seen through the glottis is called the *gap* in dash red). Classically the cycle is considered to start at the opening instant (tO), nevertheless, as this instant sometimes is rather inaccurate, the closing instant (t=0) is preferred. The sudden stop of the airflow by vocal folds at contact produces a fast drop of the dynamic pressure from 0 to a minimum (at t=0 and t=tC).

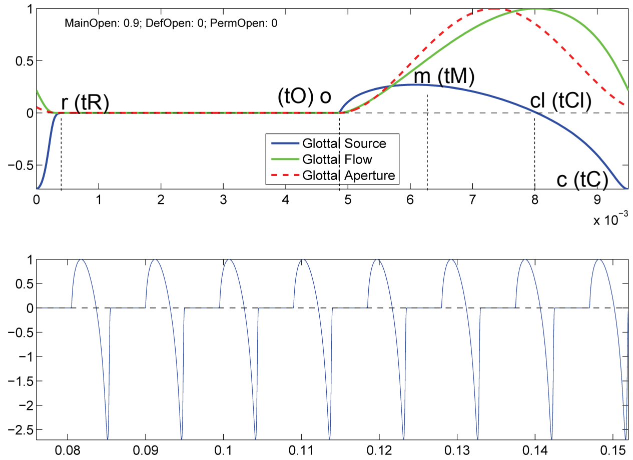


Figure 1 L-F pattern. Top: glottal opening (gap) in dash-red; glottal flow in green; glottal source in blue. Bottom: sequence of L-F patterns for 8 consecutive phonation cycles.

After a time interval (tR: recovery time), the dynamic pressure returns to its quiescent value (0). During the remnant part of the closed phase ending at tO, the vocal folds are supposedly in contact and no airflow is allowed through the glottis. The dynamic pressure remains in its resting value (0). At tO the vocal folds start opening, and a pressure build-up towards a maximum is produced (tM) where the airflow (green line) is in its steepest ascent. As the vocal folds come closer (adduction), the pressure drops crossing the resting value at tCl, and falling to a minimum when both vocal folds produce a complete flow stop (tC). This pattern is repeated each glottal during the phonation process. From what has been said, it may seem clear that the specific profiles of the recovery, closed, open and closing phases will reveal important details of the system biomechanics. A good reconstruction of the glottal source is of most relevance to ensure proper estimates of the system biomechanics. For such, a careful removal of the vocal tract by system inversion is necessary [1]. Biomechanical parameters offer a more relevant semantics of vocal fold physiological structure than classical acoustic parameters derived from voice. Section II presents dysphonic voice biomechanical modeling, in Section III study cases are discussed, Section IV is devoted statistical validation of the methodology, and in Section V conclusions and future work are summarized.

## II. MODELLING DYSPHONIA WITH BIOMET®PHON

The computer routines providing the inversion of the vocal tract and the reconstruction of the glottal source are encapsulated in a software package referred to as the Glottex®Engine (G®E), which is built as a C++ package generated from MATLAB® [4]. It produces a set of 65 parameters including distortion, cepstral, spectral, biomechanical, temporal, contact and tremor obtained from the glottal source following the methodology in Figure 3. This requires the inversion of the vocal tract and the vocal fold biomechanics, as explained in [1]. A good example of the glottal source reconstruction from a normophonic male subject, non-smoker, pathology-free condition assessed by objective endoscopy, is shown in Figure 2.

The reconstructed glottal source and flow are quite realistic and resemble the simulated pattern in Figure 1. The time references (tR1, tR2, tO1 and tO2) and contact defects (ContGap, AdducGap and PermGap) are also given. In the specialized literature there are other methodologies and products which also estimate parameter sets from voice and use them in the assessment of pathology [5]. The most important feature that the parameters estimated by BioMet®Phon convey when compared with these other approaches is their ability to fill is the *semantic gap* between acoustics and structure. This concept addresses the ability of the biomechanical structural parameters provided by BioMet®Phon to describe the etiologic characteristics of the dysphonia in contrast to other methodologies. For instance, it is well known that many times deviations in the behavior of *jitter* and *shimmer* or *HNR* (harmonics-noise ratio) may point to the presence of pathology in voice [6], but one cannot go any further on investigating which kind of pathology may be behind this behavior. Special care has been devoted in G®E technology to define the biomechanical parameters adding description capabilities of physiological structures.

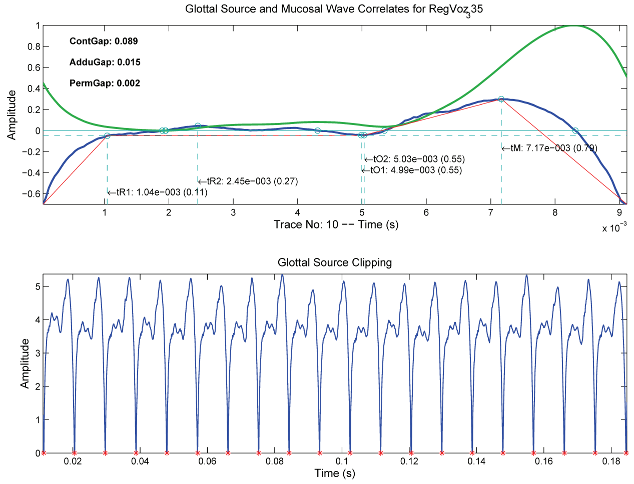


Figure 2 Typical glottal source. Top: a glottal cycle spanning from a closing instant to the next closing instant. Bottom: Sequence of glottal cycles in an interval of 183 ms.

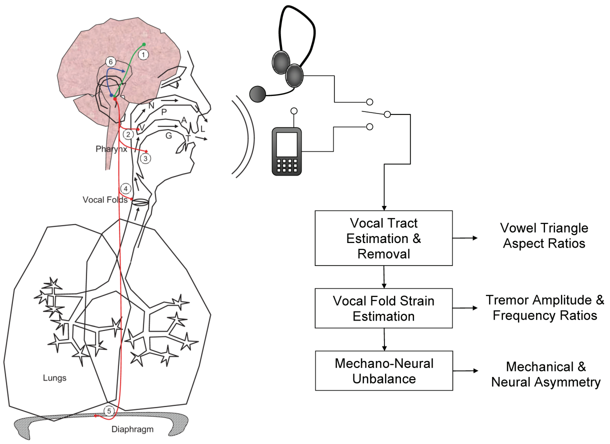

1
6
N
P
A
V
L
G
T
2
3
4
5
Pharynx
Vocal Folds
Lungs
Diaphragm
Vocal Tract Estimation & Removal
Vowel Triangle Aspect Ratios
Vocal Fold Strain Estimation
Tremor Amplitude & Frequency Ratios
Mechano-Neural Unbalance
Mechanical & Neural Asymmetry

Figure 3 Model inversion to estimate vocal tract, biomechanical and neurological parameters from voice.

This makes the biomechanical parameters by far the most interesting parameter set to assess the dysphonic conditions of a patient in relation with a specific etiology. The biomechanical parameters are defined from a 2-mass model of the vocal folds [7] as the one depicted in Figure 4. The template (a) illustrates the physiological structure of the vocal folds as a body composed by the *musculis vocalis*, and a cover or *lamina propria* and the conjunctive tissues in Reinke's space and the visco-elastic ligament giving support to the folds.

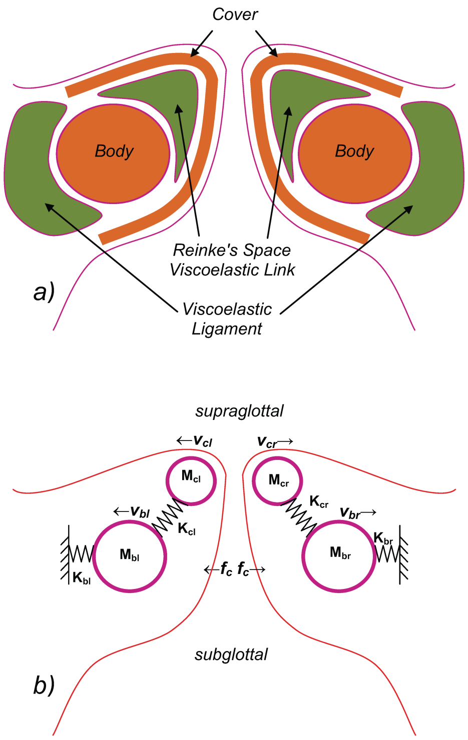


Figure 4 Vocal fold 2-mass biomechanical model assumed in G®E. a) Stuctural description of vocal folds. b) Model equivalent in masses and viscoelasticities.

The biomechanical model in (b) shows that the massive structures of the cover and Reinke's space have been included in the cover masses $M_{cl}$ and $M_{cr}$ for the left (l) and right (r) vocal folds. Masses $M_{bl}$ and $M_{br}$ account for the body and visco-elastic ligaments. It must be kept in mind that these masses are not tissular distributed masses, but dynamic point-like ones (a dynamic mass is simply a relation between force and acceleration, and is only a fraction of the tissular mass). Visco-elastic parameters $K_{cl}$ and $K_{cr}$ explain the relations between tissue compression and acting forces on the cover and Reinke's space. Parameters $K_{bl}$ and $K_{br}$ bear also the same meaning regarding the body and visco-elastic ligament. Visco-elastic parameters account for the behavior of the conjunctive tissues under compression as well as for the contribution of the *lamina propria* and *musculus vocalis* along its length due to the stretching forced by the crico-arytenoid muscles on the vocal folds during adduction and contact phases. Besides, the visco-elastic parameter encloses also the losses, which account for energy dissipation in heat, radiation and turbulence. Having this description in mind, the subset of biomechanical parameters is composed of the following correlates:

- Parameter 35: Dynamic mass associated to the body, as an average of $M_{bl}$ and $M_{br}$.
- Parameter 37: Stiffness parameter associated to the body averaged on the left and right folds ($K_{bl}$ and $K_{br}$).
- Parameter 38: Unbalance of dynamic body mass per each two neighbor cycles.
- Parameter 40: Unbalance of body stiffness per each two neighbor cycles.
- Parameter 41: Dynamic mass associated to the cover averaged on the left and right folds ($M_{cl}$ and $M_{cr}$).
- Parameter 43: Stiffness parameter associated to the cover averaged on the left and right folds ($K_{cl}$ and $K_{cr}$).
- Parameter 44: Unbalance of dynamic cover masses per each two neighbor cycles.
- Parameter 46: Unbalance of cover stiffness per each two neighbor cycles.

The estimation of the above parameters is carried out by inverting the 2-mass model in Figure 4 in the spectral domain as described in [1]. Examples of estimates from each parameter on a balanced database of 50 male and 50 female normophonic speakers collected and evaluated by endoscopy at Hospital Universitario Gregorio Marañón are given in Figure 5 and Figure 6. It may be seen that parameter 35 (body mass) is differentially distributed for males and for females, being larger for males, as expected. The distribution for parameter 37 (body stiffness) is distributed differentially but reciprocally (larger for females than for males), as well as parameter 43 (cover stiffness). On the other hand, cover masses (parameter 41) do not show gender differences. Regarding unbalance parameters (38, 40, 44 and 46) all the distributions concentrate towards low values with a few exceptions (outliers). This means that large unbalance may be an indication of dysphonic or pathological behavior. The irregular behavior of these parameters bears a clear semantics on possible dysphonia of pathological etiology.

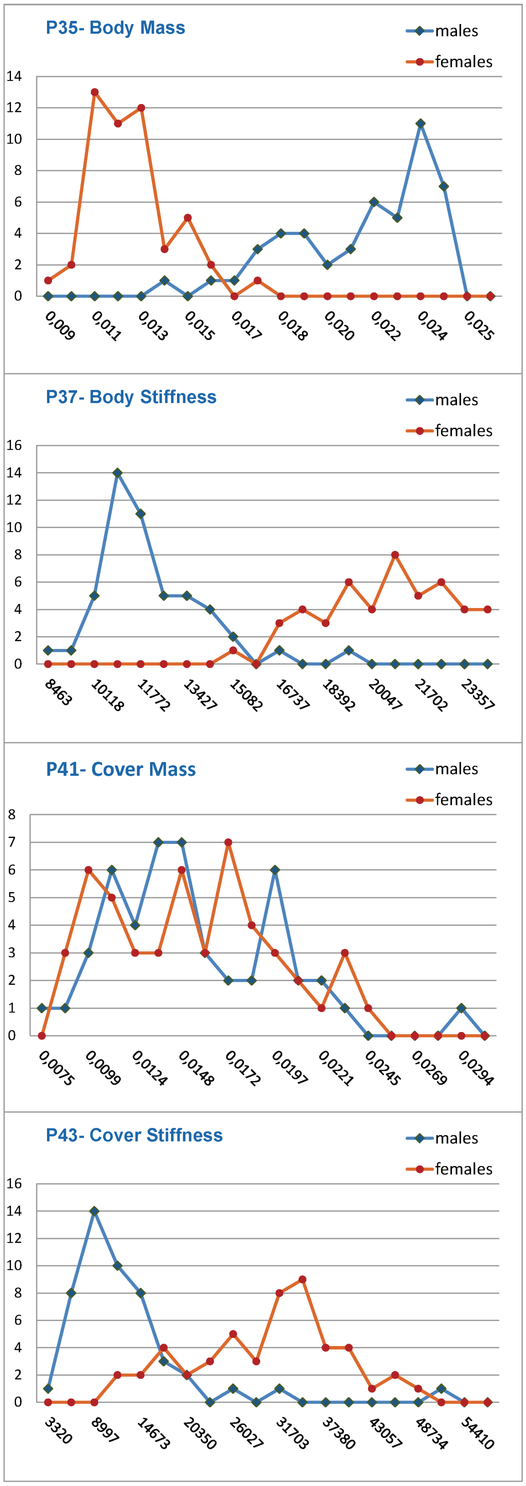


Figure 5 Histograms of the biomechanical parameters (dynamic masses and stifnesses) for normophonic male and female datasets. In abscisae masses are given in g, stifnesses given in g.s$^{-2}$. Ordinates give number of subjects.

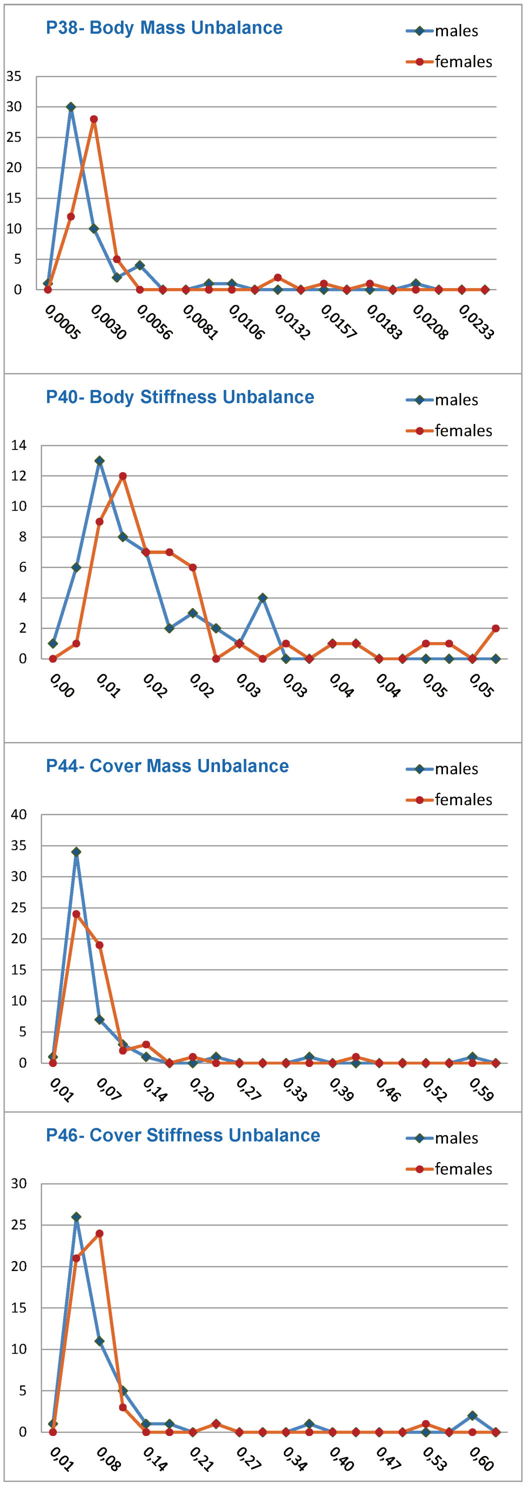
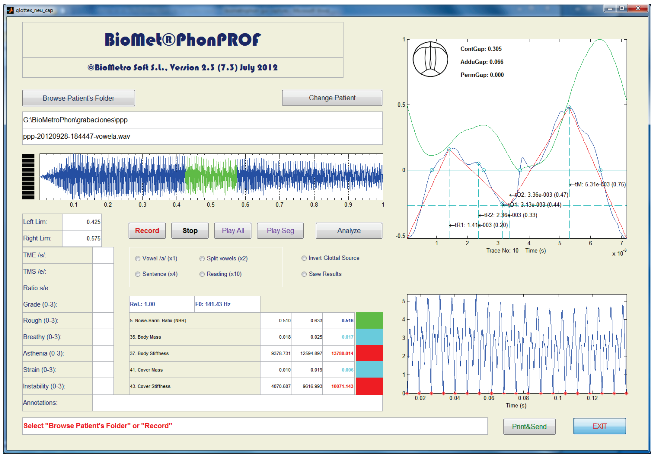


Figure 6 Histograms of the biomechanical parameter unbalance for nomophonic male and female datasets (given in rel. values). Abscisae give unbalance relative to unity (for instance, 0.01 is 1%). Ordinates give number of subjects.

Accordingly to a full study of specific pathologic cases treated at Hospital Universitario Gregorio Marañón the most frequent behaviors may be classified within one of the following groups:

- If body mass and stiffness are significantly increased above normality in a given case, this may be taken as an indication to a possible pathology affecting the vocal fold body. For instance, non-reciprocal body over-stiffness may point out to vocal fold paresis in one or both vocal folds. On the contrary, an increment in the body stiffness accompanied by a reciprocal reduction of mass or vice-versa may point out to a modal variation in pitch as in prosodic intonation or in singing.
- An increment in the cover stiffness not accompanied by a reciprocal reduction of mass most probably will point out to lesions affecting the *lamina propria* or Reinke's space, especially if an increment in mass is also observed. For instance, non-reciprocal over-stiffness in the cover may be a clear indication that a lesion (nodules, polyps, cysts, or unilateral paresis) may be under development or already installed affecting the cover and possibly Reinke's space.
- An unbalance, expressed mainly in the body stiffness may point out also to unilateral or asymmetric vocal fold paresis.
- An unbalance in the fold cover may be associated with unilateral lesions affecting the cover or Reinke´s space (polyps, cysts).

The statistics of the biomechanical parameters for the normophonic sets described have been incorporated into G®E. The package is embedded into a Graphic User Interface (see Figure 7) as the application BioMet®Phon, designed for use in Voice Quality Analysis by Laryngologists or Speech Therapists. The GUI is rather simple: a new voice recording, analysis and its automatic report in Adobe®pdf, and Excel® may be generated in less than 10 s by three button clicks. The GUI allows the handling of a small patient's database. Once a patient is selected either a new recording may be obtained and analyzed or an old one may be processed. A sketch of the glottal source is presented in the upper right window of Figure 7.



G:\BioMetroPhon\grabaciones\ppp

ppp-20120928-184447-vowela.wav

Left Lim: 0.425

Right Lim: 0.575

○ Vowel /a/ (x1) ○ Split vowels (x2) ○ Invert Glottal Source

○ Sentence (x4) ○ Reading (x10) ○ Save Results

| | | | | | | |
|---|---|---|---|---|---|---|
| TME /s/: | | | | | | |
| TMS /e/: | | | | | | |
| Ratio s/e: | | | | | | |
| Grade (0-3): | | Rel.: 1.00 | F0: 141.43 Hz | | | |
| Rough (0-3): | | 5. Noise-Harm. Ratio (NHR) | 0.510 | 0.633 | 0.516 | 🟩 |
| Breathy (0-3): | | 35. Body Mass | 0.018 | 0.025 | 0.017 | 🟦 |
| Asthenia (0-3): | | 37. Body Stiffness | 9378.731 | 12594.897 | 13780.014 | 🟥 |
| Strain (0-3): | | 41. Cover Mass | 0.010 | 0.019 | 0.006 | 🟦 |
| Instability (0-3): | | 43. Cover Stiffness | 4070.607 | 9616.993 | 10071.143 | 🟥 |
| Annotations: | | | | | | |

Select "Browse Patient's Folder" or "Record"

Figure 7 GUI of BioMet®Phon.

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-252-3

A set of five selected parameters are presented in comparative windows (mid bottom) showing normality limits and ticketing the results as green (within normality), blue (under normality) or red (above normality). This code allows a fast semantic interpretation by the laryngologist or speech therapist.

## III. PRE-POST-TREATMENT STUDY CASES

In what follows, a typical study case will show how BioMet®Phon may be used in assessing voice quality improvement after treatment. A specific case of pre-treatment compared with three post-treatment inspections are presented and discussed. It corresponds to a female patient 65 years-old who suffered from post-Thyroidectomic Vocal Fold Recurrent Paralysis (pTVFRP). The treatment consisted in infiltration of fat from the patient in the vocal folds. The patient's voice was examined for almost a year (2011) once before the intervention (pre: March) and three times after the intervention (post1: May; post2: September; and post3: November). The behavior of twelve parameters from the set of 65 is plotted in Figure 8. The 8 most relevant parameters for dysphonic voice evaluation are listed in TABLE I. Classically *2-Jitter*, *3-Shimmer* and *5-HNR* are parameters used very often in voice quality evaluation, as they are known to be well correlated with dysphonic voice [6]. Nevertheless these parameters lack structural semantics, as they do not allow producing hypotheses on possible etiological circumstances. On the contrary biomechanical parameters as the subset left (*38-Body Mass Unbalance*, *40-Body Stiffness Unbalance*, *41-Cover Mass*, *43-Cover Stiffness*, *44-Cover Mass Unbalance* and *45-Cover Stiffness Unbalance*) allow casting hypotheses on possible etiological implications based on their specific definitions.

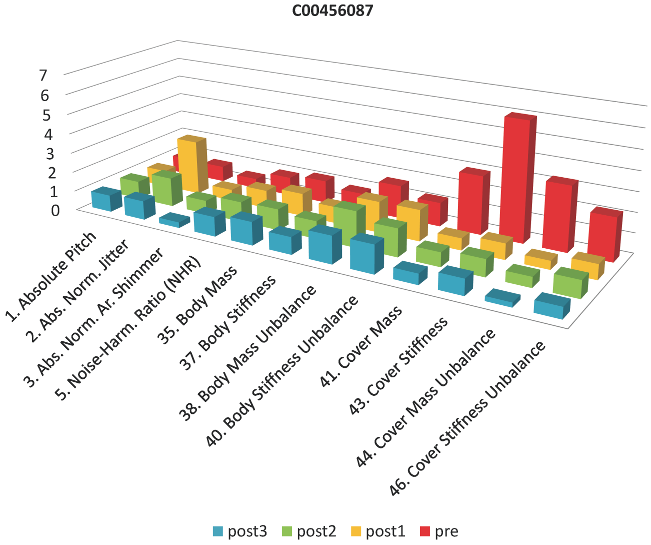


Figure 8 Results of pre- and post-treatment for a specific case of pTVFRP normalized on the reference female set medians.

TABLE I. RESULTS OF PRE- AND POST-TREATMENT FOR A SPECIFIC CASE (PTVFRP) ON A SET OF SELECTED PARAMETERS.

| Parameter | Pre | Post1 | Post2 | Post3 |
|---|---|---|---|---|
| *2-Jitter (%)* | 2.8 | 5.4 | 0.6 | 0.6 |
| *3-Shimmer (%)* | 10.5 | 3.3 | 1.5 | 1.0 |
| *38-Body M. Unb. (%)* | 4 | 21 | <1 | <1 |
| *40-Body S. Unb. (%)* | 10 | 30 | 1 | 1 |
| *41-Cover M. (mg)* | 26 | 8 | 8 | 6 |
| *43-Cover S. (g.s$^{-2}$)* | 91,746 | 24,228 | 14,175 | 11,808 |
| *44-Cover M. Unb. (%)* | 47 | 14 | 2 | 1 |
| *46-Cover S. Unb. (%)* | 43 | 26 | 3 | 3 |

It may be seen that jitter (2) correlates more with fold body unbalance (38, 40), whereas shimmer (3) is more related to cover parameters (41, 43, 44, 46). As jitter and body unbalance suffer an increment after intervention (in *post1* relative to *pre*) contrary to shimmer and cover parameters, it seems that the intervention affected the fold body in a different way than to the cover. It is like if initially after treatment the fold body suffered a regression to pathological behavior, which disappeared later (possible after fat assimilation by surrounding fold tissues). This observation demonstrates the superior introspective power of the biomechanical parameters compared to classical acoustical ones regarding physiological semantics.

## IV. VALIDATION RESULTS AND DISCUSSION

The ultimate objective of an application to evaluate voice quality is to produce accurate results in detecting dysphonic voice from normal. Therefore, a validation of the application was carried out using a database of 200 subjects collected at Hospital Universitario Gregorio Marañón divided into two subsets of 100 subjects equally balanced by gender, and these on their turn comprising half normophonic and half dysphonic subjects. Therefore, the set used in the study consisted in 50+50+50+50 subjects balanced by gender and voicing condition. The age span covered from 20 to 60 years, the medians in 35 for male and 34 for females. Sustained phonation emissions of vowel /a/ were recorded in three different sessions. Samples 200 ms long of each emission were used in the extraction of a set of 65 parameters for each phonation cycle. Estimations of medians (Q2), first (Q1) and third quartiles (Q3) were used as distribution descriptors for each emission. Medians from each emission were used in the study, to evaluate the probability of a given patient observation $\mathbf{x_q}$ being associated to the respective gender normophonic set:

$$\Pr(\mathbf{x_q} \mid \mathbf{\Gamma_m}) = \frac{1}{(2\pi)^{P/2}\left|\mathbf{C_m}\right|^{1/2}} \iiint_{(-\infty,\mathbf{x_q})} e^{-1/2(\zeta-\chi_m)^T \mathbf{C}_m^{-1}(\zeta-\chi_m)} \mathbf{d}\zeta$$
$$\Pr(\mathbf{x_q} \mid \mathbf{\Gamma_f}) = \frac{1}{(2\pi)^{P/2}\left|\mathbf{C_f}\right|^{1/2}} \iiint_{(-\infty,\mathbf{x_q})} e^{-1/2(\zeta-\chi_f)^T \mathbf{C}_f^{-1}(\zeta-\chi_f)} \mathbf{d}\zeta \qquad (1)$$

where $\mathbf{x_q}$ is the *P*-dimensional feature vector for subject *q*, and $\mathbf{\Gamma_m}=\{\mathbf{C_m}, \boldsymbol{\chi}_\mathbf{m}\}$ and $\mathbf{\Gamma_f}=\{\mathbf{C_f}, \boldsymbol{\chi}_\mathbf{f}\}$ are the respective Gaussian models for male (m) female (f) datasets. The means $\boldsymbol{\chi}_\mathbf{m}$ and $\boldsymbol{\chi}_\mathbf{f}$ and the Covariance Matrices $\mathbf{C_m}$ and $\mathbf{C_f}$ are estimated on each gender set. The likelihood of each subject given a label *v* as normophonic (*n*) or dysphonic (*d*) relative to his/her gender set will be compared to a certain threshold *θ*:

$$\lambda_{\mathrm{m}}(\mathbf{x_q}) = \log\frac{\Pr(\mathbf{x_q} \mid \boldsymbol{\Gamma}_{\mathbf{m}})}{1 - \Pr(\mathbf{x_q} \mid \boldsymbol{\Gamma}_{\mathbf{m}})};\quad \nu_{\mathrm{m}}(\mathbf{x_q}) = \begin{cases} n & if \quad \lambda_{\mathrm{m}} \ge \theta \\ d & if \quad \lambda_{\mathrm{m}} < \theta \end{cases}$$
$$\lambda_{\mathrm{f}}(\mathbf{x_q}) = \log\frac{\Pr(\mathbf{x_q} \mid \boldsymbol{\Gamma}_{\mathbf{f}})}{1 - \Pr(\mathbf{x_q} \mid \boldsymbol{\Gamma}_{\mathbf{f}})};\quad \nu_{\mathrm{f}}(\mathbf{x_q}) = \begin{cases} n & if \quad \lambda_{\mathrm{f}} \ge \theta \\ d & if \quad \lambda_{\mathrm{f}} < \theta \end{cases} \tag{2}$$

The database was processed using a ten-time cross-validation procedure replacing 5 subjects each time out of 50 within a ten-time scale, thus producing 1000 scores per gender set. The results are plotted in Figure 9 and Figure 10 for each respective gender set as Tippet plots, ROC (Receiver Operator Characteristic), Reliability Functions and DET (Detection-Error Trade-off) curves [8].

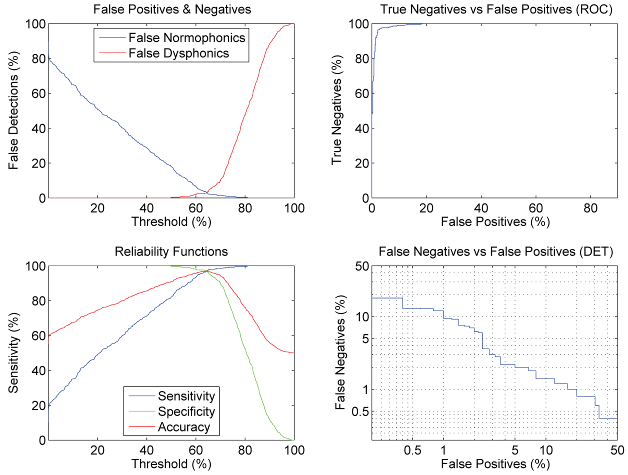


Figure 9 Validation results for male normophonic and dysphonic sets. Top left: Complementary Tippett plots. Top right: ROC curves. Bottom left: Sensitivity, Specificity and Accuracy curves. Bottom right: Equivalent DET plot.

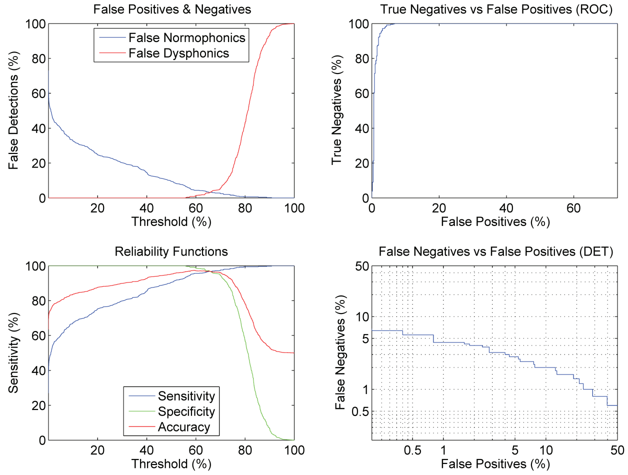


Figure 10 Respective validation results for the female sets (see Figure 9).

The results show that BioMet®Phon provides fairly similar detection capabilities for both genders. Tippett plots (upper left) show the evolution of false positive and negative detections as a function of the threshold $\theta$. DET curves (lower right) scaled in logarithmic axes offer a clear view of the Equal Error Rate point (EER), which is the point of the curve where the rate of False Positives and Negatives equal. This can be taken as a merit factor, which is around 2.7% for the male set and 3.2% for the female set. These curves allow considering different detection scenarios. For instance, to reduce the rate of False Negatives to 1% in the male set a rate of 15% False Positives should be admitted. As the emission of a False Negative in the detection of dysphonic voice is far more critical that the emission of a False Positive, it should be admitted that around 1 out of 6 subjects with normal voice should be labeled in exchange of less than 1 out of 100 dysphonic being labeled as normophonics.

## V. Conclusions and Future Work

The present paper introduced a software package for the extraction of semantic information from the glottal source obtained from phonation has been introduced under the name of Glottal®Engine. Based on this technology, a specific GUI to be applied in voice quality analysis in the clinics has been developed under the name of BioMet®Phon. The validation of this application for laryngological and speech therapeutic purposes has been tested on specific study cases, one of which has been discussed to a certain extent. Data from the validation tests have also been presented and discussed, showing the capabilities of the technology in the detection of dysphonia and in hypothesizing its possible causes. Future work foresees the extension of this methodology to neurological disease monitoring and emotional characterization from voiced speech.

## Acknowledgments

This research has been possible with the aid of grants TEC2009-14123-C04-03 and TEC2012-38630-C04-04 from the Ministry of Economy and Competitiveness of Spain.

## References

[1] Gómez, P., Fernández, R., Rodellar, V., Nieto, V., Álvarez, A., Mazaira, L. M., Martínez, R, and Godino, J. I., “Glottal Source Biometrical Signature for Voice Pathology Detection”, *Speech Comm.*, (51) 2009, pp. 759-781.

[2] Titze, I. R,. *Principles of Voice Production*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[3] Fant, G. and Liljencrants, J., “A four parameter model of the glottal flow”, *STL-QPSR*, Vol. 26, No. 4, 1985, pp. 1-13.

[4] http:\\www.glottex.com, retrieved Dec. 1, 2012.

[5] Sáenz, N., Godino, J. I., Osma, V., and Gómez, P., “Methodological issues in the development of automatic systems for voice pathology detection”, *Biomedical Signal Processing and Control,* (1) 2006, pp. 120-128.

[6] Baken, R. J., and Orlikoff, R. F., *Clinical Measurement of Speech and Voice*, Singular Pub. Group, San Diego, CA, 2000.

[7] Berry, D. A., “Modal and nonmodal phonation”, *J. Phonetics*, (29) 2001, pp. 431-450.

[8] Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M.: The DET curve in assessment of detection task performance. *Proc. Eurospeech 1997*, Rhodes, pp. 1895–1898 (1997).