Glottal Parameter Estimation by Wavelet Transform for Voice Biometry

Pedro Gómez Vilda, Ph.D.
Member, IEEE

Cristina Muñoz Mulas, MSc.
Ph.D. Student

Luis M. Mazaira Fernández, M.Sc.
Ph.D. Student

Victoria Rodellar Biarge, Ph.D.
Member, IEEE

Rafael Martínez Olalla, Ph.D.

Agustín Álvarez Marquina, Ph.D.

UPM: Universidad Politécnica de Madrid, Facultad de Informática, Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain. Tel: +34.91.3367384, Fax: +34.91.3366601, e-mail: pedro@fi.upm.es

Abstract - Voice biometry is classically based on the parameterization and patterning of speech features mainly. The present approach is based on the characterization of phonation features instead (glottal features). The intention is to reduce intra-speaker variability due to the 'text'. Through the study of larynx biomechanics it may be seen that the glottal correlates constitute a family of 2-nd order gaussian wavelets. The methodology relies in the extraction of glottal correlates (the glottal source) which are parameterized using wavelet techniques. Classification and pattern matching was carried out using Gaussian Mixture Models. Data of speakers from a balanced database and NIST SRE HASR2 were used in verification experiments. Preliminary results are given and discussed.

Index Terms — Glottal excitation, Voice Biometry, Inverse Filtering, Larynx Biomechanics

I. INTRODUCTION

Since its early infancy voice biometry is being dominated by speech features, mainly supported by the information conveyed by the vocal tract and related articulation organs [1]. The variability introduced by this approximation is quite large due to the effects of the message contents (usually referred as "the text"). Therefore text independence is much a desired objective for most applications. The intra-speaker variability resents from this approach, as the dispersion of the different articulation gestures is strong even for the same speaker. In the search of other complementary biometric features the characteristics of voicing may be exploited. If only the phonated fragments of speech are used, the glottal excitation may be estimated using accurate inverse filtering [2]. The main characteristics of the glottal excitation can be parameterized in terms of its spectral and cepstral components. Another set of parameters may be derived from the open and close phases of the phonation cycle. These can be further exploited by the use of wavelet transforms. This technique has been already used in voice pathology studies, as for example in [3] where the authors used a method based in the application of wavelets to full speech (instead to only voice) by combining wavelet sub-band energy and entropy parameters classified with Support Vector Machines. In the present paper a method based on wavelet transform of the glottal excitation is presented to model the scale-temporal evolution of the phonation cycle.

II. GLOTTAL-SOURCE WAVELET DESCRIPTION

Voice can be seen as the part of speech which is contributed by the vibration of the vocal folds. Accordingly with the well-known source-filter model of G. Fant for voice production [2] the glottal source is the basic excitation signal which when filtered by the articulation organs (naso-pharyngeal and oral cavities) produces voice. Going to the physiology of phonation, the glottal source can be seen as the dynamic pressure near the vocal folds in the oropharyngeal side (supraglottal), as given in Fig. 1.

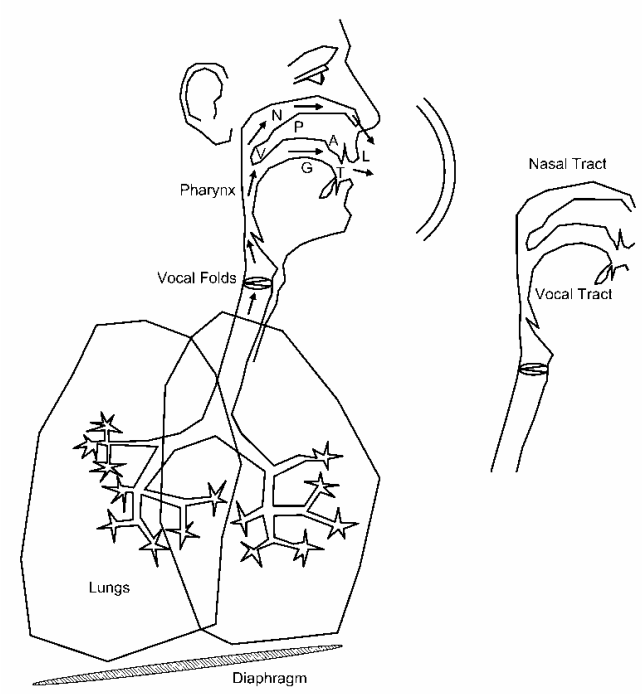

Fig. 1 Left: Schematic view of the phonation organs. Right: oversimplified view of the glottal, pharyngeal, nasal and oral tracts.

The mechanical equivalent to the pressure-building process is given in Fig. 2 (top) and works as follows:

- The diaphragm fills the lungs with air during its contractive operation. The lungs act as mere air repositories.
- During diaphragm relaxation pressure is put to the lungs by the diaphragm working as a piston. The air tends to

escape through the vocal folds if these are open (abduction, as when breathing). In case of close contact of the vocal folds induced by cartilage activation (adduction, as when phonating) the air finds an obstacle.

- Airflow may be forced through the vocal folds if enough pressure is applied resulting in partial abduction (opening).
- The partial opening results in an escape of air reducing the lung pressure in a small amount. This allows the vocal folds to join again due to cartilage-muscle activity resulting in a new duct closure (adduction). The phonation cycle starts again.

In the electrical equivalent in Fig. 2 (bottom) the mechanical operation of the system is modeled as follows:

- The action of the diaphragm and lungs is seen as a current generator injecting airflow (u_l) in the trachea.
- The trachea elastic walls absorb the airflow during the closed phase (usually some 2-6 ms), behaving as a compliance C_l.
- The vocal folds interact with airflow as a conductance G_g, ranging between *0* (closure) and G_{gmax} (maximum opening). Its profile in time is given as a hunchback curve in time.
- The oro- and naso-pharyngeal (articulation) cavities are modeled as an inductance L_t expressing the inertial behavior of the air column present in the cavities.
- The pressure build-up in the trachea is given as p_l.
- The pressure resulting immediately after the vocal folds (supraglottal, or lips side) is given as p_g. This is to be identified with the glottal source within some first order approximation.
- The pressure at the lips is assumed to be the steady atmospheric value p_0.
- The glottal airflow through the articulation cavities is given as u_g.

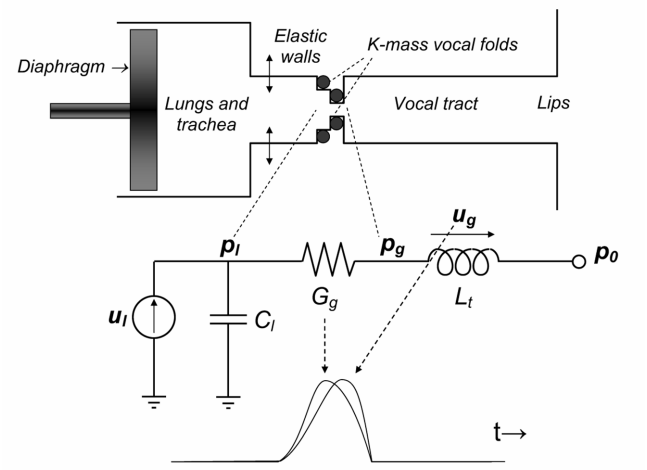

Fig. 2 Top: Mechanical equivalent of the phonation system. Bottom: Electrical model used in the study.

The main relations sustaining model dynamics are the following:

$$u_l(t) = C_l \frac{\partial p_l(t)}{\partial t} + u_g(t) \quad (1)$$

$$p_g(t) - p_0 = L_t \frac{\partial u_g(t)}{\partial t} \quad (2)$$

$$u_g(t) = \left[p_l(t) - p_g(t)\right] G_g(t) \quad (3)$$

It may be seen that the main complexity of the model comes from its strong nonlinear nature due to the dependency of conductance G_g with time. To understand in full de operation of such a system some of simulations have been conducted using MATLAB®. The results are given in Fig. 3 and below.

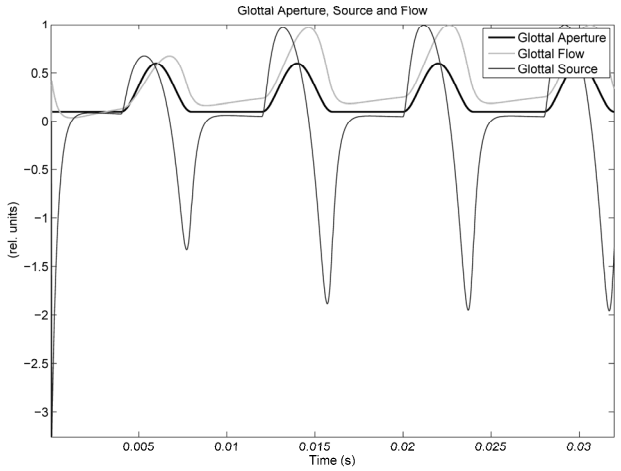

Fig. 3 Results from simulations for a simple abduction-adduction process. Thick dark line: glottal aperture (conductance). Thin dark line: glottal source. Grey line: glottal flow. The four initial cycles from simulation are shown.

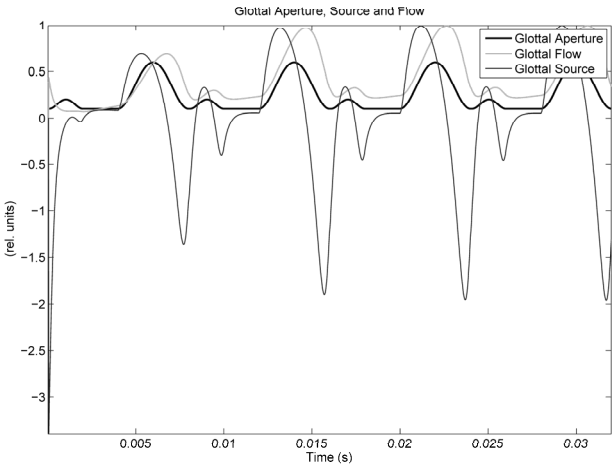

Fig. 4 Results from simulations for a defective abduction-adduction process showing a spurious opening during closed phase. Thick dark line: glottal aperture (conductance). Thin dark line: glottal source. Grey line: glottal flow. The four initial cycles from simulation are shown.

The main conclusions derived from the inspection of the above figures are the following:

- Each main opening (see Fig. 3) produces a characteristic signature in the glottal source, which appears as a classical Liljencrants-Fant (L-F) pattern [5] showing a positive hunchback anticipating the opening (approx. at t=0.012 s), a sharp decay reaching a minimum peak synchronized with the closing instant (approx. at t=0.016 s), and a steady plateau stabilizing prior to the new opening instant (approx. at t=0.02 s).
- Spurious openings (see Fig. 4) may appear as delayed and reduced versions of the main L-F pattern.
- Spurious opening signatures may be seen as wavelet versions of the main opening. Thus wavelet representations may play a most relevant role in the detection of spurious openings and closings in the glottal source.

As it happens that spurious fluctuations of this kind are specific and personal of each individual independent of articulation and modality of phonation, they may serve as biometrical markers after being characterized using wavelets. This is the main hypothesis supporting the present research.

For such a general glottal opening may be defined in the discrete time domain (t=nτ, τ being the sampling interval) as:

$$G_g(n) = G_0 + G_1\alpha\frac{n-n_o}{n_c-n_o}sin\left\{2\pi\frac{n-n_o}{n_c-n_o}\right\} \tag{4}$$

where G_0 is the permanent opening (defective or imperfect closure of the vocal folds found in many speakers as a result of modal phonation mainly when this is breathy or whispery), G_1 being the amplitude of the dynamic opening and n_o signaling the opening instant. This pattern may be seen as a modified slant bell-shape ancestor of the opening wavelet as presented in the sequel.

The methodology proposed consists in estimating the wavelet transform of the glottal source $p_g(t)$ which may be described in the continuous time domain as:

$$W_{\psi p_g}(s,d) = \frac{1}{\sqrt{|s|}}\int_{-\infty}^{\infty}\psi\left(\frac{t-d}{s}\right)p_g(t)dt \tag{5}$$

where s and d are the continuous scale (dilation) and position coefficients, respectively.

There are many possible wavelets to approximate the main pattern of the glottal source as given in Fig. 3 (see [6]). The ones considered in the present study are Haar and low-order Daubechies. The description for the Haar wavelet in the discrete time domain is (see [7]):

$$\psi_{00}^{g}(n) = \begin{cases} 1; & 0 \le n < \frac{n_c}{2} \\ -1; & \frac{n_c}{2} \le n < n_c \end{cases}; \qquad \psi_{jk}^{g}(n) = 2^{-j/2}\psi_{00}^{g}(2^{-j}n-k); \tag{6}$$

where n_c is the sample index associated to the closing instant (cycle duration), and j and k are the indices of binary (dyadic) dilation and position. The glottal source could be defined in terms of the Haar wavelets as:

$$p_g(n) = \sum_{j,k=-\infty}^{\infty} c_{jk}\psi_{jk}^{g}(n) \tag{7}$$

the weights of the linear combination being estimated as:

$$c_{jk} = W_{\psi p_g}(2^{-j}, k2^{-j}) = W_{\psi p_g}(j,k) \tag{8}$$

From wavelet decomposition the closed and open phase gap correlates (γ_c, γ_o) and efficiencies (ε_c, ε_o) may be estimated (the gap correlate being a parameter which is null for perfect closure, and non-null if there is a spurious opening during the closed phase). These definitions are given in terms of the average energy of the wavelets in certain dilation and delay indices as:

$$\gamma_c = \frac{L_c}{Lr} \tag{9}$$

$$\gamma_o = \frac{L_o}{Lr} \tag{10}$$

$$\varepsilon_c = 1-\gamma_c \tag{11}$$

$$\varepsilon_o = 1-\gamma_o \tag{12}$$

where:

$$L_c = \sum_{j=0}^{J}\sum_{k=0}^{k_o}\left|W_{\psi p_g}(j,k)\right|^2 \tag{13}$$

$$L_o = \sum_{j=j_1}^{J}\sum_{k=k_o}^{n_c}\left|W_{\psi p_g}(j,k)\right|^2 \tag{14}$$

$$L_r = \sum_{j=0}^{j_1-1}\sum_{k=0}^{n_c}\left|W_{\psi p_g}(j,k)\right|^2 \tag{15}$$

J being the largest scale order used (typically 5), j_1 being a lower estimation threshold in scale to determine the limits between the main and spurious openings, and k_0 being the delay associated to the opening instant. The approximation and detail counterparts of the glottal source example in Fig. 4 are given in Fig. 5 as a reference of the wavelet decomposition obtained using the proposed approach, in terms of scale and position.

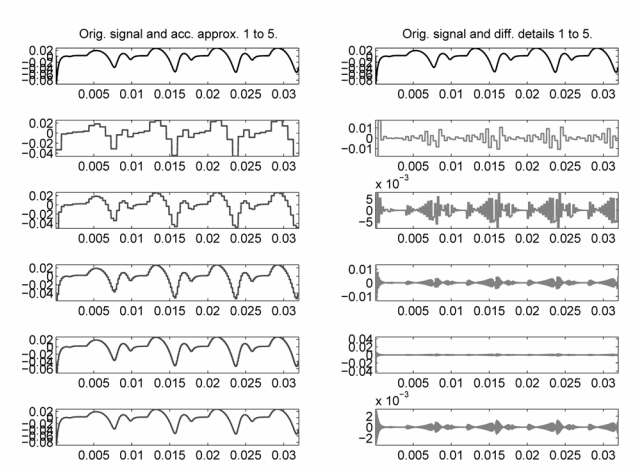

Fig. 5 Scale and position results from Haar-Daubechies decomposition of the glottal source in Fig. 4 given in logarithmic amplitude by descending scale (larger *j's* from top to bottom). Left column: Original signal and approximations by scale. Right column: Original signal and details by scale.

What is most significant from the description given in the above figure is that the lower-order scale approximations (larger index) give an account of the main opening, while the larger order details give an account of the presence of spurious openings. The accumulated energy estimates of smaller and larger indices are selectively used in (13)-(15) to estimate the gaps and efficiencies.

III. PATTERN MATCHING METHODS

Voice Biometry may be characterized using different strategies, classically mel-cepstrum parameterization and GMM (Gaussian Mixture Models) or SVM (Support Vector Machine) classification [8]. Nevertheless the use of mel-cepstral coefficients on the whole voice signal, although efficient, lacks semantics, i.e., it is really difficult to infer which factors convey to successful characterization, this being a major aim in the field far from being completed. A different approach is that one based on parameter sets directly related with behavioral singularities of the glottal source or vocal fold biomechanical parameters as dynamic masses, tensions or time-domain efficiency as derived in previous sections. This approach has been used in the recent past yielding interesting results ([9]-[11]). The combination of specific parameter cocktails may yield quite accurate results. The methodology relies in the selection of a set of control speakers which may be considered the reference or background model [1]. This set is the key to the correct score normalization. Speakers need to be recruited and modeled separately for both genders, as morphologic differentiations between male and female are meaningful [8]. From the inversion of the Liljencrants-Fant source-filter model the glottal source (excitation) is reconstructed [2]. Advanced parameterization techniques are used for the estimation of observation vectors, where each speaker *i* is represented by a parameter vector:

$$\mathbf{x}_i = \left[x_{i1}, x_{i2}, \ldots x_{iJ}\right] \tag{16}$$

composed of *J* values x_{ij} produced from a 200 msec. segment of voice corresponding to a sustained utterance of /a/ accordingly with the description given in [12]. Once the reference male (*m*) and female (*f*) sets are completed the model observation matrices are produced:

$$\begin{aligned} \mathbf{X}_{Mm} &= \left[\mathbf{x}_{1m}, \ldots \mathbf{x}_{im}, \ldots \mathbf{x}_{Im}\right]^T \\ \mathbf{X}_{Mf} &= \left[\mathbf{x}_{1f}, \ldots \mathbf{x}_{if}, \ldots \mathbf{x}_{If}\right]^T \end{aligned} \tag{17}$$

Similarly the control observation matrices $\mathbf{X}_{Cm}$ and $\mathbf{X}_{Cf}$ are produced using observations from the dysphonic male and female sets. The PCA projection is based on the joint model-control covariance matrix (see [13], [14]):

$$\begin{aligned} \mathbf{X}_P &= \left[\mathbf{X}_{Mm,f}^T, \mathbf{X}_{Cm,f}^T\right]^T \\ \mathbf{C}_P &= \mathbf{X}_P \mathbf{X}_P^T \end{aligned} \tag{18}$$

The matrix (E_P) of eigenvalues of C_P is used to project the original observations matrices on the new principal component matrices:

$$\begin{aligned} \mathbf{Y}_m &= \mathbf{X}_m \mathbf{E}_P \\ \mathbf{Y}_f &= \mathbf{X}_d \mathbf{E}_p \end{aligned} \tag{19}$$

A GMM for each gender is produced (Γ_m for the male set and Γ_f for the female one). The mean vectors $\boldsymbol{\psi}_{Mm}$ and $\boldsymbol{\psi}_{Mf}$ as well as the corresponding covariance matrices $\mathbf{C}_{Mm}$ and $\mathbf{C}_{Mf}$ are estimated. The GMM is built using Gaussian multivariate functions as:

$$p(\mathbf{y}_{ti} / \Gamma_{Mm,f}) = \frac{1}{(2\pi)^{Q_m/2} \left|\mathbf{C}_{Mm,f}\right|^{1/2}} \times \\ \times e^{-1/2\left(\mathbf{y}_{ti} - \boldsymbol{\psi}_{Mm,f}\right)^T \mathbf{C}_{Mm,f}^{-1} \left(\mathbf{y}_{ti} - \boldsymbol{\psi}_{Mm,f}\right)} \tag{20}$$

$\mathbf{y}_{ti}$, $\boldsymbol{\psi}_n$, and $\mathbf{C}_n$ being respectively the data vector under test of subject *i*, the centroids of the parameter Gaussians GMM's and the Covariance Matrices of each observation set, *p* being the conditional probability of an observation vector being a member of the specific set represented by the specific Gaussian.Qm,f on their turn are the dimensions of the observation vectors in (17). If the model GMM is composed by a certain number of Gaussians the joint probability will be expressed as:

$$p_T(\mathbf{y}_{ti} / \Gamma_{Nm,f}) = \sum_k w_k \, p_k(\mathbf{y}_{ti} / \Gamma_{km,f}) \tag{21}$$

where w_k are the weights of the linear combination generating the overall likelihood. In the present case mono-Gaussian Models show to be accurate enough. Finally vector membership to a model may be scored as the Log-Likelihood Ratio (LLR) of the odds:

$$\Lambda_p(\mathbf{y}_{tmi}) = \log\{p(\mathbf{y}_{tmi}/\Gamma_{nm})\} - \log\{p(\mathbf{y}_{tmi}/\Gamma_{\bar{n}m})\} \quad (22)$$

This score is based on distance metrics, and it may be used for assigning the subject a given membership using classical ROC-DET (Receiver Operator Characteristics or Detection Error Trade-Off) plots depending if the LLR is over or under a given threshold θ: $\Lambda_p(y_{tmi}/\Gamma_{nm})>\theta$ or $\Lambda_p(y_{tfi}/\Gamma_{nf})<\theta$.

IV. APPLICATION TO VOICE BIOMETRY

The main problem in applying the above conclusions to voice biometrical studies is the intra-speaker variability. In other words: to which extent the parameters obtained for a given speaker under a given phonation modality are similar to the speaker's other phonation modalities and distinct at the same time to the parameters obtained from other speakers' phonations. To answer this "burning question" one has to take into account the sources of intra- and inter-speaker variability. For intra-speaker studies these may be the main sources of variability:

- Modality of the phonation, being normal (modal), over-pressed or under-pressed. The modal phonation is associated with the relaxed (emotion-less) speaker, whilst the over-pressed corresponds to emotional excitation (anger, exultation, wrath...), and the under-pressed has to see with anguish, fatigue, depression, etc.
- Vocalization. The decomposition of the voice into the glottal source and vocal tract (filter-source model) is highly dependent on the last one. Therefore the results will be different for open or close vowels, and for voicing consonants. This characteristic has to see with articulation or acoustic-phonetic issues.
- Prosody. The stress and emphasis of the phonation in running speech is of most importance. Raising or lowering the pitch reduces or adds duration to the glottal phonation cycle and to parameterization. The raising or lowering of pitch in speech can produce quite different results in the parameter description of the glottal source in interrogative, declarative or imperative sentences.

In what follows examples will be given from voice samples corresponding to different articulation and prosody cases regarding speaker recognition studies. The relevance of the speaker's emotional state will be left for further elaboration. The study is conducted in terms of the Prosecutor's vs Defender's approach as a classical Log-Likelihood Ratio (LLR) estimation by the specificity-typicality paradigm [15]:

$$p(I_u/I_a) = \frac{p(I_a/I_u)p(I_u)}{p(I_a)} \quad (23)$$

where I is in general de information available from a specific speaker (I_u from the questioned or unasserted speaker, I_a from the asserted or suspect). The above probabilistic model is formalized as the LLR evaluating the Prosecutor's Hypothesis (H_p) against the Defender's Hypothesis (H_d):

$$\Lambda_{u/a} = \log\left\{\frac{f(E/H_p,I)}{f(E/H_d,I)}\right\} = \log\left\{\frac{p(I_u/I_a)}{p(I_u)}\right\} = \log\left\{\frac{p(I_a/I_u)}{p(I_a)}\right\} = \log\left\{p(\mathbf{x}_i^u/\Gamma_A)\right\} - \log\left\{p(\mathbf{x}_i^u/\Gamma_B)\right\} \quad (24)$$

E, H_p and H_d being respectively the Evidence, the Prosecutor and the Defender Hypotheses. The general speaker information, composed by the set of observations (parameter medians of the set of parameters in TABLE I from the asserted or suspect (a) and the unasserted or questioned (u) observations are defined as:

$$\mathbf{x}_i^a = \left[x_{1i}^a, x_{2i}^a, \ldots x_{mi}^a\right]^T$$
$$\mathbf{x}_i^u = \left[x_{1i}^u, x_{2i}^u, \ldots x_{mi}^u\right]^T \quad (25)$$

The Universal Background Gaussian Model (UBGM) Γ_B will be composed by the covariance matrix $\mathbf{C}_B$, and mean vector ψ_B for the reference population data set. The Asserted Gaussian Model Γ_A is to be built in a similar way from all the data available from the suspect, resulting in $\mathbf{C}_A$, and ψ_A. The evaluation of the membership of a given questioned frame $\mathbf{y}_i^u$ with respect to the UBMG or the AGM will be estimated in terms of the conditioned probability:

$$p(\mathbf{y}_i^u/\Gamma_{A,B}) = \frac{1}{(2\pi)^{Q_{A,B}/2}\left|\mathbf{C}_{A,B}\right|^{1/2}} \times e^{-1/2\left(\mathbf{y}_i^u-\boldsymbol{\psi}_{A,B}\right)\mathbf{C}_M^{-1}\left(\mathbf{y}_i^u-\boldsymbol{\psi}_{A,B}\right)^T} \quad (26)$$

Once the relative membership probabilities are produced, the LLR of the Prosecutor's vs the Defender's Hypothesis given in (24) will be estimated.

TABLE I
GLOTTAL SOURCE PARAMETER DESCRIPTION (see [8])

Param.	Description
x_1	*pitch*
x_2	*jitter*
$x_{3\text{-}5}$	3 variants of *shimmer*
x_6	Noise/Glottal parameter
$x_{7\text{-}20}$	Glottal Source Spectral Density cepstral parameters
$x_{21\text{-}26}$	Singularities of mucosal wave correlate power spectral density (amplitude)
$x_{27\text{-}32}$	Singularities of mucosal wave correlate power spectral density (frecuency)
$x_{33\text{-}34}$	Slenderness of the two first “V troughs”
$x_{35\text{-}37}$	Biomechanical parameters of vocal fold body (masses, losses, tensions)
$x_{38\text{-}40}$	Intra-speaker period-synchronous variations of body biomechanics
$x_{41\text{-}43}$	Biomechanical parameters of vocal fold cover (masses, losses, tensions)
$x_{44\text{-}46}$	Intra-speaker period-synchronous variations of cover biomechanics
$x_{47\text{-}55}$	Glottal Source time-domain relative intervals and amplitudes
$x_{56\text{-}58}$	Glottal closed and open efficiencies and gap

Results for a practical study case will illustrate this technique in the next section.

V. RESULTS AND DISCUSSION

For the purposes of the present study a set of 30 male speakers from [16] will be used in the experiments. Part of this subset, specifically 20 speakers will serve as the Universal Background Model Set, and 10 speakers more will be used as imposters for T-norm contrast. The questioned and suspect frames have been obtained from a 300-sec. record of running speech (test 4, channel a) from the last NIST SRE10 HARS1 competition [17] selecting 12 frames where the utterance /ah/ or /uh/ have been produced, either in long vowels or in fillings, as listed in TABLE II.

TABLE II
FRAMES USED IN THE STUDY

Frame start	Frame end	Frame #
9.2	9.4	4009
28.7	28.9	4028
43.95	44.20	4043
201.55	201.75	4201
213.85	214.00	4213
232.30	232.55	4232
243.55	243.85	4243
248.80	249.35	4248
267.00	267.35	4267
276.00	276.20	4276
289.95	290.25	4289
291.30	291.60	4291

The whole set of 20+10+12 frames taken at 8kHz are parameterized and PCA projected. The corresponding Model, Control and Test sets are described in TABLE III.

TABLE III
MODEL, CONTROL AND TEST SETS USED IN THE EXPERIMENTS

Set	Frames
Model	15 271 274 314 333 334 335 347 353 361 362 363 366 368 372 383 397 399 400 406
Control	4009 4028 4043 4201 4213 4232 4243 4248 4289 4291
Test	408 416 417 419 422 427 429 432 443 464 4267 4276

It may be seen that the PCA projection will be carried out on the Model and Control Sets, the first constructed exclusively from 20 frames of different speakers. The Control Set is integrated by 10 frames from the same speaker. The Test set includes 10 frames from different normophonic speakers and 2 more frames from the questioned speaker. The aim is twofold: on one side to determine if the samples taken at different time instants from the same speaker present some similarity among themselves, on the other side to determine if they can be differentiated from a Universal Background Model resumed in certain parameters selected using Fisher's Discriminant Ratios (FDR), as given by:

$$fdr_j = \frac{\left(\mu_{Mj} - \mu_{Cj}\right)^2}{\sigma_{Mj}^2 + \sigma_{Cj}^2}; \quad 1 \le j \le J \tag{27}$$

The set of parameters under study has been selected among the most resolving ones. These are the Noise/Glottal (x_8), the second cepstral (x_{10}) and the 2nd minimum in the GSPSD (Glottal Source Power Spectral Density: x_{21}). The Model (o), Control (◊) and Test (*) Sets given by matrices $\mathbf{X}_M$, $\mathbf{X}_C$ and $\mathbf{X}_T$ in (17) are shown in Fig. 6.

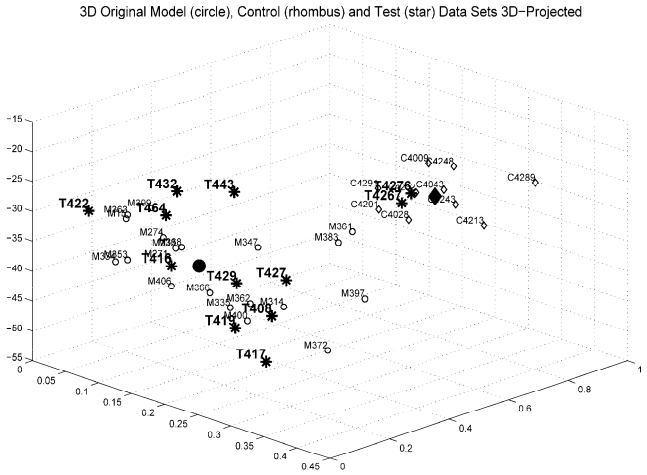

Fig. 6 3D Projection of the Data Set used in the experiments: Model (white circles), Control (white diamonds) and Test (star) samples. Centroids of the Model and Test sets are given by a filled circle and diamond.

The projection is given in terms of the three most resolving parameters as by the values of FDR from [17]. The Model Set frames (labeled as Mxxx-o) are located around a well defined cluster (except for M361 and M383). It may be seen also that the Test Set frames (labeled as Txxx-*) corresponding to impostors are grouped themselves in the neighborhood of the Model Set. Clearly the Control Set frames (labeled Cxxxx-◊) are grouped apart mixed with the two Test frames taken from the questioned speaker (T4267 and T4276). This points out to the questioned frames as being produced also by the suspect (evidence would favor H_p in detriment of H_d). This is more clearly expressed by the data given in TABLE IV (Rec#: Number of the frame record; sDo: Square of Norm. Distance to the Model Set Centroid; sD◊: Id. to the Control Set; $p(y/\Gamma_B)$: Probability of membership to the Model Set; $p(y/\Gamma_A)$: Id. to the Control Set; $\Lambda_{u/a}$: Likelihood Ratio referred to the Prosecutor's Hypothesis vs the Defender's Hypothesis).

TABLE IV
RESULTS FROM THE DETECTION PROCESS

Rec#	sDo	sD◊	$p(y/\Gamma_B)$	$p(y/\Gamma_A)$	$\Lambda_{u/a}$
408	2,26	40,69	4,33E-07	1,93E-13	-14,62
416	3,44	33,15	2,40E-07	8,36E-12	-10,26
417	21,68	416,24	2,63E-11	5,45E-95	-192,69
419	8,33	211,35	2,08E-08	1,69E-50	-96,92
422	18,98	63,07	1,01E-10	2,66E-18	-17,45

427	2,72	50,10	3,43E-07	1,75E-15	-19,09
429	6,46	167,51	5,30E-08	5,58E-41	-75,94
432	7,58	102,44	3,03E-08	7,55E-27	-42,84
443	12,22	37,38	2,98E-09	1,01E-12	-7,99
464	5,61	51,86	8,13E-08	7,26E-16	-18,53
4267	**23,52**	**15,03**	1,04E-11	7,20E-08	**8,84**
4276	**35,00**	**26,71**	3,36E-14	2,09E-10	**8,74**

As seen in the table, for each frame in the Test Set (first column to the left) the LLR in TABLE III is given. The second column gives the squared Mahalanobis distance from each sample to the Model Centroid. The two frames showing a larger value (in bold) are the ones extracted from the questioned speech segment. The third column gives the same distance for each sample relative to the Test Centroid estimated from frames extracted from the suspect speaker speech segment (T4267 and T4276, which happen to coincide with the questioned one in the present experiment). In this case the frames showing a smaller distance (in bold) are the ones from the suspect. The two closer frames to the Test Centroid are again the ones extracted from the questioned speech segment (T4267 and T4276). The next two columns give the relative membership probabilities of each Test frame to both the Control and Model Sets. The membership probability of the upper ten frames relative to the UBGM is clearly larger than their respective membership probability relative to the Test Model. This results in negative LLR's favoring the H_d. On the contrary, the last two frames show membership probabilities larger for the Test Model than for the UBGM. The respective LLR's are positive and similar, favoring the H_p in their case.

IV. CONCLUSIONS

Interesting consequences may be derived from the present study. First of all it seems that parameters classically derived for the study of voice pathology as *Noise/Glottal* can be used for the biometrical characterization of the speaker as well. This conclusion is very important, as these parameters have clear semantics as far as the characterization of a speaker is concerned. A second conclusion is that the Glottal Source singularities (peaks and troughs) are relevant for the biometrical characterization of the speaker. It is known that the Glottal Source may be altered by articulation as well as by modality, vocalization or prosody, as explained in section 3. The frames selected from the running speech segment were not especially conditioned by any factor except by vowel coloring (in fact most of them correspond to the kind of fillers /uh/'s and /ah/'s, which are spontaneously produced by Native Speakers of English). The modality is different in most of them, as well as the prosody (some present questioning or surprise marks). Nevertheless, the system identified clearly all of them as being different from the Model Set, selected from sustained vowels, and similar among themselves. This fact may indicate that the parameters selected are robust to modal information and sensitive to biometrical differences. Of course the work is still far from being completed. Massive tests on model and test running speech segments as the ones proposed in the last NIST SRE HARS2 contest need to be processed. For such automatic vowel selection and framing is to be put into work and the discussed methodology applied on blind tests to measure its capability in speaker characterization tasks.

V. ACKNOWLEDGEMENTS

This work has been funded by grants 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.

VI. REFERENCES

[1] Fazel, A., Chakrabartty, S., "An Overview of Statistical Pattern Recognition Techniques for Speaker Verification", *IEEE Circ. & Sys. Mag.* Vol. 11, No. 2, 2011, pp. 62-81.

[2] Alku, P.: Parameterisation Methods of the Glottal Flow Estimated by Inverse Filtering. *Proc. of VOQUAL'03*, 81-87 (2003)

[3] Behroozmand, R., Almasganj, F., "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis", *Computers in Biology and Medicine*, Vol. 37, 2007, pp. 474-485.

[4] Fant. G., *Theory of Speech Production*, Mouton, The Hague, Netherlands (1960).

[5] Fant, G., Liljencrants, J., "A four-parameter model of glottal flow", *KTH STL-QPSR*, Vol. 26, No. 4, 1985, pp. 1-13.

[6] Addison, P. S., "Wavelet transforms and the ECG: a review", *Physiological Measurement*, Vol. 26, 2005, pp. R155-199.

[7] Mallat, S., *A wavelet tour of signal processing*, Academic Press, London, 1999.

[8] Fraile, R., Sáenz, N., Godino, J. I., Osma, V., Fredouille, C.: Automatic Detection of Laryngeal Pathologies in Records of Sustained Vowels by Means of Mel-Frequency Cepstral Coefficient Parameters and Differentiation of Patients by Sex. *Folia Phoniatrica et Logopaedica*, 61, 146-152. (2009)

[9] Godino, J. I., Osma, V., Sáenz, N., Gómez, P., Blanco, M., Cruz, F.: The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders. *J. of Voice*, 24, (1), 47-56. (2010)

[10] Gómez. P. et al., "Detecting Pathology in the Glottal Spectral Signature of Female Voice", *Proc. of MAVEBA07*, Firenze, Italy, 183-186 (2007)

[11] Gómez, P. et al.: Voice Pathology Grading by Gaussian Mixture Models: Study Cases. *Proc. of MAVEBA09*, Firenze, Italy, 45-48 (2009)

[12] Gómez, P., Fernández, R., Rodellar, V., Nieto, V., Álvarez, A., Mazaira, L. M., Martínez, R, Godino, J. I., "Glottal Source Biometrical Signature for Voice Pathology Detection", *Speech Communication*, Vol. 51, 2009, pp. 759-781.

[13] Gómez. P. et al.: PCA of Perturbation Parameters in Voice Pathology Detection. *Proc. of INTERSPEECH'05*, 645-648 (2005)

[14] Johnson, R. A. and Wichern, D. W. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Upper Saddle River, NJ (2002).
[15] González, J., et al.: Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. *IEEE Trans. on ASLP*, 15 (7), 2104-2115 (2007)
[16] Project MAPACI: http://www.mapaci.com.
[17] http://www.itl.nist.gov/iad/mig//tests/sre/2010/index.html, NIST SRE10 Evaluation Workshop, 24-25 June, Brno, Czech Republic (2010)

VIII. VITA

Pedro Gómez graduated from Universidad Politécnica de Madrid in 1983 with a Ph.D. degree. He worked for the Nuclear Research Council and a private communications company. He obtained the permanent professorship from Universidad Politécnica de Madrid in 1986. Since then he has carried out different research activities in the medical applications of speech processing and speaker recognition. He is a member of the IEEE since 1985.

Cristina Muñoz graduated from Universidad Rey Juan Carlos in 2005 with a MSc degree. She is currently developing her PhD Thesis in the field of speaker gender, age and phonation style under the guidance of Prof. Rafael Martínez.

Luis Miguel Mazaira graduated from Universidad Politécnica de Madrid in 2005 with a MSc degree. He is currently developing his PhD Thesis in the field of speaker recognition under the guidance of Prof. Agustín Álvarez.

Rafael Martínez graduated from Universidad Politécnica de Madrid in 2002 with a Ph.D. degree. He obtained the permanent professorship from Universidad Politécnica de Madrid in 2006. Since then he has carried out different research activities in signal processing and pattern recognition for speech processing and speaker recognition.

Agustín Álvarez graduated from Universidad Politécnica de Madrid in 2001 with a Ph.D. degree. He had a contract in Universidad Rey Juan Carlos as Assistant Professor. He obtained the permanent professorship from Universidad Politécnica de Madrid in 2006. Since then he has carried out different research activities in signal processing and pattern recognition for speech processing and speaker recognition.

Victoria Rodellar graduated from Universidad Politécnica de Madrid in 1983 with a Ph.D. degree. She obtained the permanent professorship from Universidad Politécnica de Madrid in 1987. Since then he has carried out different research activities in hardware/software co-design of platforms for speech processing and speaker recognition. She is a member of the IEEE since 1985.