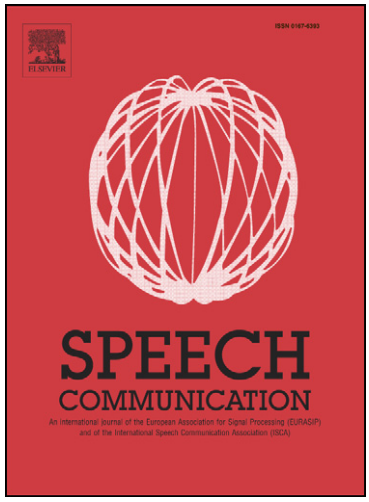
Glottal Source Biometrical Signature for Voice Pathology Detection

Pedro Gómez-Vilda, Roberto Fernández-Baillo, Victoria Rodellar-Biarge, Víctor Nieto Lluis, Agustín Álvarez-Marquina, Luis Miguel Mazaira-Fernández, Rafael Martínez-Olalla, Juan Ignacio Godino-Llorente



Glottal Source Biometrical Signature for Voice Pathology Detection

Pedro Gómez-Vilda, Roberto Fernández-Baillo, Victoria Rodellar-Biarge, Víctor Nieto Lluis, Agustín Álvarez-Marquina, Luis Miguel Mazaira-Fernández, Rafael Martínez-Olalla, Juan Ignacio Godino-Llorente

PII: S0167-6393(08)00137-4
DOI: 10.1016/j.specom.2008.09.005
Reference: SPECOM 1753

To appear in: *Speech Communication*

Received Date: 1 December 2007
Revised Date: 20 May 2008
Accepted Date: 14 September 2008

Please cite this article as: Gómez-Vilda, P., Fernández-Baillo, R., Rodellar-Biarge, V., Lluis, c.N., Álvarez-Marquina, A., Mazaira-Fernández, L.M., Martínez-Olalla, R., Godino-Llorente, J.I., Glottal Source Biometrical Signature for Voice Pathology Detection, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.09.005

Title:

Glottal Source Biometrical Signature for Voice Pathology Detection

Authors:

Pedro Gómez-Vilda[1], Roberto Fernández-Baillo[1], Victoria Rodellar-Biarge[1], Víctor Nieto Lluis[1], Agustín Álvarez-Marquina[1], Luis Miguel Mazaira-Fernández[1], Rafael Martínez-Olalla[1], Juan Ignacio Godino-Llorente[2]

Affiliations:

[1]Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain

[2]Escuela Universitaria de Ingeniería Técnica de Telecomunicaciones, Universidad Politécnica de Madrid, Ctra. de Valencia, Km. 7,000, 28031, Madrid, Spain

Corresponding Author:

Pedro Gómez-Vilda, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain, tel: +34.913367384, fax: +34.913366601, e-mail: pedro@pino.datsi.fi.upm.es

Abstract:

The Glottal Source is an important component of voice as it can be considered as the excitation signal to the voice apparatus. The use of the Glottal Source for pathology detection or the biometric characterization of the speaker are important objectives in the acoustic study of the voice nowadays. Through the present work a biometric signature based on the speaker's power spectral density of the Glottal Source is presented. It may be

shown that this spectral density is related to the vocal fold cover biomechanics, and from literature it is well known that certain speaker's features as gender, age or pathologic condition leave changes in it. The paper describes the methodology to estimate the biometric signature from the power spectral density of the mucosal wave correlate, which after normalization can be used in pathology detection experiments. Linear Discriminant Analysis is used to confront the detection capability of the parameters defined on this glottal signature among themselves and compared to classical perturbation parameters. A database of 100 normal and 100 pathologic subjects equally balanced in gender and age is used to derive the best parameter cocktails for pathology detection and quantification purposes to validate this methodology in voice evaluation tests. In a study case presented to illustrate the detection capability of the methodology exposed a control subset of 24+24 subjects is used to determine a subject's voice condition in a pre- and post-surgical evaluation. Possible applications of the study can be found in pathology detection and grading and in rehabilitation assessment after treatment.

Keywords:

Voice Biometry, Speaker's Identification, Speaker Biometrical Characterization, Voice Pathology Detection, Glottal Source

1. Introduction

The issue of pathology detection from audio recordings of a subject's phonation is one of the most promising technologies in the care of voice. This technique may be helpful in pathology pre-screening and as a complementary inspection routine to detect pathology

before it has fully developed when the first alterations of vocal folds do not yet show physiological visible clues but slight perturbations from normal voicing are already present. Besides, a sensitive pathology detection methodology could help in maintaining good preventive practices for voice health care. Traditionally voice pathology detection has been mainly based on the estimation and monitoring of certain perturbation parameters which are well known correlates to pathology. These can be included in one of the following categories:

- Parameters obtained from the original voice (taken as a whole signal with no other pre-processing than a pre-emphasis enhancement, preserving the resonances imposed by the vocal tract). Typical perturbation parameters are jitter, shimmer and harmonics-to-noise ratios in their different interpretations. This has been the traditional approach and is well documented in the works of different researchers [1][2][3][4][5]. The main objection to this methodology is that the original voice signal is contaminated by phonetic-acoustic information related to specific articulation features, and therefore perturbation parameters derived from this signal are influenced by articulation, making it difficult to grant new advances towards pathology classification.
- Spectral domain parameters derived from the voice signal. Specific correlates among the original voice spectral profile and certain pathologies have been proposed in the literature [6][7]. Although pathology is detectable using specific correlates between harmonics and formants, the acoustic-phonetic contamination poses important limitations to this methodology for the same reasons exposed in the previous point, blurring the observations to a certain extent, making it difficult to improve detection

sensitivity.

- Glottal Source[1] time-domain parameters. Recognizing the difficulties encountered when dealing with spectral estimates the from original voice signal several researchers [8][9][10] have proposed the use of the glottal source parameterization in the time domain for pathology characterization studies, after removing the vocal tract from voice. This line seems to offer rather promising results, especially in improving the capability of detecting pathology at an earlier stage.

- Glottal Source frequency-domain parameters. An alternative methodology is based on the use of glottal source spectral correlates for pathology characterization, these having already been found of interest in voice characterization, as in the detection of gender [11][12].

The present work is oriented to formalize the use of the glottal source power spectral density, adding estimates of the vocal fold biomechanics by model inversion [13][14] to define a glottal source signature for pathology studies. The substantial improvements of this work with respect to the last ones is to be found in the Linear Discriminant Analysis (LDA) used to detect the most relevant parameters in the study, in the detection technique used which is based on Gaussian Mixture Models, in the database selected, consisting in 200 speakers equally balanced by gender and non-pathological/pathological condition and in the presentation of a study case to visualize the potential use of the methodology proposed in a real situation.

[1] In what follows the Glottal Flow Derivative will be addressed as Glottal Source following the source-filter model of G. Fant

A comprehensive review of the characterization of voice pathology from acoustic analysis may be found in [1]. Holmberg et al. using estimates of the L coefficient (energy difference between the first two harmonics H_1-H_2 found in the power spectral density of voice), the ratio between the first and third formant harmonics A_1-A_3, and the cross ratios H_1-A_1 and H_1-A_3, reported interesting findings correlating these parameters with certain specific pathologies [6][7]. Studies in the time domain using the Return Quotient (RQ), the Open Quotient (OQ), the Closed Quotient (CQ) the Closing Quotient (ClQ) or the Normalized Amplitude Quotient (NAQ) offered similar results [8][9]. These last works show the way to establish a more structured study regarding pathology characterization. On one side they point out to the use of time or frequency domain parameters as the basis of the study. On the other side, they deal either with voice or with the Glottal Source as the basic signals for the study. In the present approach the Glottal Source has been selected as the object of the research. Although time and frequency domain studies are related in some way, the work has been carried out to the frequency domain. Taking relations H_1-H_2, A_1-A_3, H_1-A_1 and H_1-A_3, as good correlates to pathology availed by other researchers' results, a generalized signature is proposed on singularities detected on the Glottal Source spectral envelope (peaks and troughs). This generalization is based on the biomechanical dynamics of the vocal folds found on the Glottal Source spectral envelope [15], whose singularities may be shown to be strongly determined by the relations among parameters in well-known *k-mass* models [16][17] once the influence of the vocal tract has been removed. Having this perspective into account these are some relevant questions to be answered by the present study:

a) Is the envelope of the Glottal Source spectral profile a good generalization of partial

relations as H_1-H_2, A_1-A_3, H_1-A_1 and H_1-A_3 used in pathology studies?

b) How this spectral profile can be parameterized to provide an accurate and compact description of the speaker's glottal features?

c) Are parameter distributions sensitive to gender bias effects?

d) If so, which are the most relevant ones by gender groups concerning pathology detection?

The paper is organized as follows: Section 2 is devoted to establish the framework for splitting voice into glottal and vocal components. The estimation of the glottal source and other dynamic correlates used in the study is given in section 3. Questions a) and b) are answered in detail in section 4 where a complete description of the parameterization procedure of the glottal source spectral envelope is given. Questions c) and d) are treated in section 5, where Principal Component Analysis is proposed to offer more compact data sets which can be used in different detection and classification problems, accordingly with voice-pathology related features, and Gaussian Mixture Models are proposed for the separation of non-pathologic samples from pathologic ones. In section 6 the results derived from PCA dimensionality reduction, LDA parameter discrimination and GMM detection are briefly commented. A specific study case selected among others is presented in section 7 to illustrate the capabilities of the methodology exposed as far as pathology detection and grading is concerned. Section 8 is devoted to extract conclusions from the results presented and give hints on future research on the topic.

2. Structured parameterization of voice

Considering the classical source-filter model for the generation of voice composed by an excitation (Glottal Source) and a modulating filter (Vocal Tract) as proposed by Fant [18], it may be expected that the excitation will depend on the less-varying biometric characteristics of the speaker (lungs, larynx and naso-pharynx cavities, etc.) being weakly influenced by the message (text), but strongly conditioned by the production process (physiological and emotional conditions, prosody, tonal height, production gesture, pathology, etc.). Under the methodological point of view it seems that in treating the voice signal following a deconstructive way independent features could be observed. This means that vocal and glottal parameters have to be treated separately with methods more specifically oriented to their respective nature (accordingly to their statistical inter-speaker and intra-speaker characteristic distributions). Voice parameterization procedures should have these specific facts into account, as the parameters estimated on one or the other context will be later subject to evaluation, analysis and classification under clearly different paradigms. The parameterization of voice may be carried out using estimates of one of the following main categories:

- The Voice Power Spectral Density (VPSD), estimated either using traditional FFT or by LPC. The short-time power spectrum is usually coded as Mel-Frequency Cepstral Coefficients (MFCC) by well-known methods [19]. Up to recently this has been the only description used in applications as Pathology Detection [20] or Speaker Identification [21]. Classically *pitch*, *jitter*, *shimmer*, *harmonics-noise ratios*, or *energy* may be evaluated on a pitch-synchronous or asynchronous basis to produce a rich parameter description. Other parameters as the ratios H_1-H_2, A_1-A_3, H_1-A_1 and H_1-A_3, or the OQ, CQ, ClQ and NAQ have also be included in these studies.

- The Vocal Tract Transfer Function Modulus (VTTFM). The VPSD reflects the influence of the Glottal Source spectral envelope as a *1/f* spectral tilt, which distorts Vocal Tract Transfer Function estimates. A decoupling between Vocal Tract and Glottal Source could render better results both in decoding the message (Speech Recognition) as well as in the characterization of the source (Pathology Detection). The accurate estimation of the VTTFM is of interest to grant the careful removal of its influence from voice and to produce reliable estimates of the Glottal Source. It is also useful in certain speaker characterization studies, as age or gender [22].

- The Glottal Source Power Spectral Density (GSPSD). The Glottal Source can be parameterized in the time or in the frequency domain. Time domain methods are based in the well-known Liljiencrants-Fant model [23]. The time domain parameterization is oriented to the estimation of OC, SC, ClQ, RQ and NAQ (Open, Speed, Closing, Return and Normalized Amplitude Quotients). The frequency domain is oriented to the estimation of the ratio H_1-H_2 (which is known to be related to the CQ - Close Quotient), and to the Maximum Flow Declination Rate (MFDR) and the Spectral Slope [12]. Other possible parameterization methods for the Glottal Source in the frequency domain may be based on MFCC or LPCC parameters from the power spectral density of the glottal signals (Glottal Flow derivative, Glottal Source derivative, etc.) similarly to the methodology used to parameterize the VTTFM. Another line of study is related with the establishment of correlates among the Glottal Source frequency envelope and the biomechanical parameters of a k-mass vocal fold model by model inversion as suggested by the authors in previous work [13].

3. Estimation of the Glottal Source

The methodology proposed in the present work may be seen as a frequency domain parameterization of the glottal source power spectral density, with the following distinctive characteristics:

- It is carried out either on the Glottal Source or on the Mucosal Wave Correlate (MWC), which is a signal derived from the Glottal Source removing the Acoustic Average Wave (AAW) from it [24]. The AAW, as it will be later explained, can be seen as the Body Dynamic Component, because it may be associated to the one-mass/one-spring equivalent model of the vocal fold body. The residual left when removing the AAW from the Glottal Source signal is designed as the Mucosal Wave Correlate (also the Cover Dynamic Component or CDC), as it can be associated to higher-order oscillation modes of the vocal folds related mainly with the dynamic behaviour of the fold cover. Both signals can be considered correlates to the body and cover dynamics, and will be referred as such.

- It estimates the singularities of the power spectral density of the Glottal Source or the MWC as a series of peaks and notches in amplitude and frequency relative to the fundamental frequency of voice F_0. Therefore it can be considered as a generalization of the parameters used in [8] and [9], because the ratio H_1-H_2 and the Spectral Slope are part of the parameterization proposed.

The character of this parameterization is typically biometrical, as its inter-speaker statistical variability is mainly conditioned by the personal characteristics of each speaker (gender, age, tension, glottal gesture, pathology, etc). An added value is that its frequency-domain

character may be more robust to estimation errors than other time-domain techniques, as those based on the L-F model in the time-domain. The argument in favour of this assertion resides in that the estimates of OC, SC, ClQ, RQ and NAQ are based on fitting real glottal source patterns against the well-known L-F model. This model is an idealized version of the resulting glottal source behaviour in a standard vocal fold system. Indeed male voice adheres better to the L-F paradigm than female voice, and both deviate from it even when mild pathology (functional, non-organic) is present. In the presence of strong pathology curve fitting of a real Glottal Source estimate to the L-F pattern is not straight forward. Even in the presence of normal real glottal signals standard deviations in time-domain parameters as strong as a 20% over the mean estimate have been reported [25]. If airy or breathy voice is present an added perturbation factor may induce even more intra-speaker variability, rendering less robust estimates. The glottal source spectral features of interest for pathology detection are to be found in the lowest part of the spectrum of the glottal source (usually below 2000 Hz, see the singularity points labelled as rhombi and stars in Figure 2). This part of the spectrum is less exposed to airy or breathy noise corruption as this concentrates mainly in high frequencies. Therefore frequency-domain estimates of the glottal source spectral profile show standard deviations under 10% with respect to their means (for the same vowel and same speaker in sustained phonations of non-pathologic voicing up to 300 msec. long).

In what follows a brief description of the methodology used for both the Glottal Source extraction and inversion will be given for the interested reader. The methodology used for the estimation of the Glottal Source is based on the elimination of the vocal tract influence by inverse filtering using an iterative implementation of well-known methods [26][27], and

the biomechanical estimations are based on the separation of the Glottal Source into the two referred components (AAW and MWC). The background details to produce the results used in the present study may be found in [14]. An example of the glottal signal estimation results from inverse filtering may be seen in Figure 1. These are obtained from quasi-stationary utterances of the vowel /a/ by typical male and female speakers (those closest to the centroids of the respective normal male and female clusters). The presence of specific point-like negative spikes in the glottal source associated with the instant where vocal folds initiate the closed phase (closure spikes) can be clearly appreciated.

The reconstruction of the glottal source is very much inspired in the adaptive version of the iterative inverse filtering developed by Alku et al. [26] where the LPC filters have been implemented by adaptive lattice filters as shown in Figure 3. A brief description of the extraction technique used is as follows:

1. The radiation effects in input voice is first removed by a radiation cancelling filter $H(z)$. The resulting voice $s_l(n)$ will in this way be equalized to inside-lip conditions, and will be referred to as the Radiation-Compensated Voice (RCV). This signal is filtered by a Glottal Pulse Inverse Model $H_g(z)$, which for the first iteration will consist in a K_1-order prediction-error adaptive lattice filter (usually of order 1, 2 or 3 as will be discussed later). This filter will roughly remove the strong glottal formant spectral envelope by placing one, two or three zeroes on the real axis of the unity circle (see Figure 4 and the sequel for a wider explanation). Special care has to be put for it not affecting or cancelling the first true formant of voice. The residual of the filtering $s_v(n)$ is the so-called De-Glottalized Voice (DGV), where the spectral

tilt due to the glottal profile has been removed.

2. This signal is the base for the estimation of the Vocal Tract Transfer Function (VTTF), which is carried out by another prediction-error adaptive master lattice filter $F_v(z)$ of order K_2 (in this case the filter order should be scaled accordingly with the sampling frequency being used; as a reference, for sampling frequencies of 16,000 Hz suitable order filters may be within the range of 12-18).

3. A subordinate paired lattice $H_v(z)$ fed with the filter parameters of $F_v(z)$ will now be used to remove the VTTF from the RCV, thus producing a first estimate of the glottal pulse $s_g(n)$ (or Glottal Source as is designated by G. Fant in his classical source-filter model [18]).

4. The glottal pulse is modelled by another prediction-error adaptive master lattice filter $F_g(z)$ which places a small number of zeroes on the real axis to cancel the rough spectral envelope of the glottal pulse.

5. Another iteration is started using a subordinate lattice filter $H_g(z)$ loaded by $F_g(z)$, and the cycle is repeated. Some 2-3 iterations are usually enough to obtain a good estimate of the glottal source.

The prediction-error filters used in cancelling the glottal tilt must be of low order to implement the removal of the rough glottal source spectral envelope (low resolution) as over-sizing will produce interference with the estimation of the VTTF low formants (crosstalk). This effect may be appreciated in the results given in Figure 4, where two estimates of the glottal source for a specific speaker are produced for two values of $K_1=3$ and 4. The estimates for $K_1=3$ show that the filter zeroes align on the real axis and help in

cancelling the glottal formant and the spectral tilt resulting from it without affecting the low formants in the VTTF. The results for $K_1=4$, on the contrary, show that the resulting cancelling zeroes appear as complex pairs and interfere with the lowest formants capturing some of their power spectral density, which produces a certain crosstalk on the reconstructed glottal source (see Figure 4 bottom right). The VTTF, on its turn has to be estimated with enough resolution to accurately determine its formants, but higher orders are not desirable as oversized filter orders will start picking up harmonics. Other possible strategies to model the spectral envelope of the glottal pulse are bi-spectrum joint estimation and ARMA methods [28][29][30], which can be used in step 4 to obtain a description of the zeroes and poles of the glottal pulse spectral profile (peaks and troughs) because these are good descriptors of the glottal source biometry as will be addressed in the sequel, but special care has to be used in discriminating pole-zero behaviour in the glottal source vs that in the vocal tract transfer function, this study being an open line. The estimation of the glottal source produced by any of these methods is to be used in step 5 to accurately remove glottal influence from RCV and produce good estimates of the VTTF. Another important issue to be considered is the selection of the best time interval for the estimation of the VTTF, as in modal phonation the vocal folds close the vocal tract during part of the phonation cycle (close phase) and leave it open during the remnant part of the cycle (open phase). In the present study the restrictions posed by the open/close conditions of the vocal tract have been overcome by using lattices in the inverse estimation of the vocal tract, as under certain hypotheses these have the property of not only encoding the tube section profile in the estimated reflection coefficients but of reconstructing the forward and backward propagating waves as in a transmission line [19]. Once the forward and

backward components are known, the flow and pressure waves in a given point may be estimated by imposing a termination condition at that point (in our case the tube section where the vocal folds are assumed to be located) and a more precise reconstruction of the glottal source can be achieved. The adaptive implementation of the lattices grants a more accurate temporal estimation under the changing conditions of the system, including changes in the vibration conditions of the vocal fold or phonation gesture (time variance), detected in a phonation-cycle basis. The algorithmic variant used for adaptation in the present study is the one based on the Least Mean Square Error (LMS) [19]. Similar strategies to the one described have been applied showing accurate results [27].

Once the Glottal Source has been reconstructed using the above mentioned procedures its dynamics may be compared with estimations derived from vocal fold modelling [25] by means of a convenient fitting of its power spectral density, as it may be shown that the power spectral density of the Glottal Source is strongly conditioned by the biomechanical parameters of the vocal fold models [15]. This finding may be used in the characterization of the pathologic behaviour of a specific speaker's voice or in the biometric characterization of the speaker. For such, the Glottal Source is decomposed in two parts, one mainly influenced by the low-order vibration of the vocal folds, integrated by the Average Acoustic Wave (AAW) and the second one based on the higher-order vibration modes of the vocal folds, integrated by the Mucosal Wave Correlate (MWC). The AAW is a term coined by Titze [24] (pg. 16, exp. 21-22) to refer to the low-frequency contents of the signal under analysis. In the present case as by [14] (exp. 1-2) the Average Acoustic Wave is defined as a sinusoid

$$s_{gk}(n) = y_{0k}\, sin(\omega_k n\tau); \quad n \in N_k \tag{1}$$

of optimal amplitude y_{0k} evaluated adaptively by minimizing the energy of the difference between the AAW and the glottal source $y_{gk}(n)$ over a generic *k-th* time window given as the set of samples N_k defined on the *k-th* phonation cycle

$$L = \sum_{n \in N_k} \varepsilon_k^2(n) = \sum_{n \in N_k} \left(y_{gk}(n) - s_{gk}(n) \right)^2 \tag{2}$$

as by forcing

$$\frac{\partial L}{\partial y_{0k}} = 0 \quad \Rightarrow \quad y_{0k} = \frac{\sum_{n \in N_k} y_{gk}(n) sin(\omega_k n\tau)}{\sum_{n \in N_k} sin^2(\omega_k n\tau)} \tag{3}$$

where $\omega_k = \pi / T_k$ is the angular frequency associated to double the period of the cycle under study T_k, n is the time index and τ is the sampling period. In this way the AAW would represent a second order system response (one mass + one spring) associated to the vocal fold body. Therefore the AAW is dominated by the dynamics of the vocal fold body, and MWC is mainly contributed by the dynamics of the vocal fold cover. This study is developed in detail in [13]-[15]. Incidentally it may be said that in tense voicing the decay of the glottal source during the closing phase will mimic the shape defined by (1) very closely as it is shown in Figure 12 (bottom-left), demonstrating that the definition of the AAW in (1)-(3) possesses a rich semantics by itself which can be exploited for the interpretation of the nature of the pathologic behaviour expressed in the glottal source, as will be discussed in the sequel. Additional details to produce the results shown in the present study may be found in [28][32].

4. Biometric signature based on the Glottal Source

Power Spectral Signature

Through the present approach a methodology to derive biometrical parameters of the Glottal Source in the frequency domain is proposed. The biometrical parameters may be estimated on the power spectral density of either the Glottal Source or the Mucosal Wave Correlate. The signature obtained from the Mucosal Wave Correlate is more specifically related to the biomechanics of the vocal fold cover, while that from the Glottal Source includes the biomechanics of both the body and the cover of the vocal fold. The estimates based on this last approach are more suitable for biometric applications, the estimates from the Mucosal Wave Correlate are more suitable for studies in vocal fold pathology. In both cases the parameter estimation methodology to be applied similar. The power spectral densities shown in Figure 2 correspond to the Glottal Source from prototype male and female voices. It may be seen that in both cases a common behaviour is observed in the envelopes of the power spectral densities: a fast raise from low frequencies to a maximum, followed by a decay towards higher frequencies at a rate around *12 dB/oct* in male voice, (which may be a little less in female voice). In between, a series of valleys or local minima may be appreciated surrounded by peaks. These "V" grooves (notches or troughs) are strongly related to the biomechanics of the vocal folds. Notches are explained by the anti-resonances in the tissues of the vocal fold structures behaving as systems of lumped masses and springs [15]. In general, the slenderer the notch, the smaller the value of mass-linking springs in *k-mass* vocal fold models (see [17] for an explanation of the biomechanical foundations of this effect).

In Figure 5 the envelope of the glottal source power spectral density of the male prototype has been extracted for clarity. The behaviour detected in the mentioned case may be summarized as a fast raise from low frequencies to a first maximum of amplitude T_{M1} found at a frequency f_{M1} which is followed by a descent to a minimum T_{m1} in f_{m1} and to a new maximum T_{M2} at a frequency f_{M2}. This first notch is a very important one, as if the Glottal Source has been used in the estimations it gives a picture of the coupling between the fold body and cover structures. In case that the signal under analysis was the MWC a description of the coupling between the two lips (subglottal and supraglottal) of the vocal fold would be obtained. This type of notch may appear several more times along the decay of the power spectral density. These troughs are present in all the speakers: in normal ones their shape shows a certain sharpness, in over-tense pathologic notch sharpness it diminishes, in certain pathologies as Reinke's Edema it may be even augmented, therefore sharpness deviation from the normal pattern may be associated to pathology. Taking all these facts into consideration a glottal signature of voice may be established detecting each notch by estimating the amplitude and position of its singularity points and its slenderness factor. This signature may be used in voice pathology studies, in speaker's identification and characterization tasks as well as in forensic studies. For the present paper the first two notches will be included in the biomechanical signature. The estimations of the singularities on the power spectral density of the MWC for the first notch are normalized to the first maximum found $\{T_{M1}, f_{M1}\}$ as

$$\tau_{m1} = T_{m1} - T_{M1}; \quad \varphi_{m1} = \frac{f_{m1}}{f_{M1}} \tag{4}$$

$$\tau_{M2} = T_{M2} - T_{M1}; \quad \varphi_{M2} = \frac{f_{M2}}{f_{M1}} \tag{5}$$

The definitions for the first notch may be extended to any other one in the spectral profile of Figure 5 provided that each minimum at f_{mq} follows a maximum at $f_{Mq}<f_{mq}$ as given by

$$\left.\begin{aligned} \tau_{Mq} &= T_{Mq} - T_{M1} \\ \tau_{mq} &= T_{mq} - T_{M1} \end{aligned}\right\}; \quad 1 \le q \le Q \tag{6}$$

$$\left.\begin{aligned} \varphi_{Mq} &= \frac{f_{Mq}}{f_{M1}} \\ \varphi_{mq} &= \frac{f_{mq}}{f_{M1}} \end{aligned}\right\}; \quad 1 \le q \le Q \tag{7}$$

where q is the notch index and Q is the number of notches included in the study, therefore implicitly $\tau_{M1}=0$ and $\varphi_{M1}=1$. This normalization in amplitude and frequency is a guarantee to small intra-speaker variability. Correspondingly, the slenderness factor of the notch may be defined as

$$\sigma_{mq} = \frac{f_{Mq}\left(2T_{mq} - T_{Mq+1} - T_{Mq}\right)}{2\left(f_{Mq+1} - f_{Mq}\right)}; \quad 1 \le q \le Q \tag{8}$$

The slenderness is strongly related with the value of the springs linking the corresponding masses on the *k-mass* equivalent biomechanical model originating the peaks and notches, and is a measure of the general tension in voicing.

Biomechanical Signature

It has been shown in previous work [13] that reliable estimates of the relative values of fold body masses and tensions could be obtained from the power spectral density of the *average acoustic waveform*. The estimation technique used was the adaptive fitting of the AAW

power spectral density against the transfer function of the *1-mass* model as explained before. The work hypothesis is based on the assumption that the AAW is determined by the fold body dynamic component, therefore the power spectral density of the AAW is directly related with the square modulus of the input admittance derived from the *1-mass* model as

$$T_b(\omega) = |Y_b|^2 = \left|\frac{V_x(\omega)}{F_x(\omega)}\right|^2 = \left[\left(\omega M_b - \omega^{-1} K_b\right)^2 + R_b^2\right]^{-1} \quad (9)$$

where ω is the angular frequency in rad/sec given as $\omega=2\pi f$ and M_b, K_b and R_b are respectively the parameters associated with the lumped mass, elasticity and losses of the *1-mass* model when only the body of the vocal fold is taken into account following the dimensional reduction of the Story-Titze model [25].

The robust estimation of the model parameters is based in the selection of two points on the power spectral density of the AAW, these being $\{T_{b1}, \omega_1\}$ and $\{T_{b2}, \omega_2\}$. The lumped body mass may be estimated then as

$$M_b = \frac{\omega_2}{\omega_2^2 - \omega_1^2}\sqrt{\frac{T_{b1} - T_{b2}}{T_{b1} T_{b2}}} \quad (10)$$

The selection of the most adequate points for $\{T_{b1}, \omega_1\}$ and $\{T_{b2}, \omega_2\}$ is highly related with the accuracy and robustness of the estimation procedure. A good candidate for $\{T_{b1}, \omega_1\}$ is the position of the main (resonant) peak in the amplitude of the power spectral density of the dynamic correlate. A good candidate for $\{T_{b2}, \omega_2\}$ is the position of the third harmonic from the peak position, as the time series shows odd symmetry. These two points have shown to be robust enough in all the cases studied, some data on intra-speaker variability having been supplied in Table 3.

Once the mass has been estimated, the elastic parameter (body stiffness) K_b may be obtained from the precise determination of the position of the maximum associated to the resonant peak, this being $\{T_r, \omega_r\}$

$$K_b = M_b \omega_r^2 \tag{11}$$

The parameter of body losses can be estimated (but for a scale factor G_b) as

$$R_b = \frac{G_b}{\sqrt{T_r}} \tag{12}$$

where T_r stands for the value of the square modulus of the input admittance in eq. (9) at the frequency of resonance ω_r associated to the first maximum in the Glottal Source power spectral density.

Similar derivations may be defined for the biomechanical parameters of the vocal fold cover using in its case the spectral density of the MWC, as the influence of the body dynamics has been removed implicitly on separating the AAW from the Glottal Source, reducing the problem to a single mass model. In this way the application of the same methodology to the cover biomechanics may follow essentially the same steps in a similar way. Estimates of the biomechanical parameters for the body and cover structures are given in Figure 6 and Figure 7 from the reference male and female speakers.

Other strategies for the estimation of biomechanical parameters by spectral matching are also possible, as functional approximations [33] or adaptive curve fitting [14], and are currently under study.

Biomechanical parameter unbalance

It has been considered for the purpose of spectral estimation and fitting that both vocal folds were symmetric. This assumption does not stand in most of the cases either if dysphonic or non-dysphonic voice is involved. Asymmetry will result in the unbalance of the biomechanical parameters estimated for neighbour phonation cycles. It seems reasonable to think that this unbalance will be larger in cases where vocal fold pathology is present than in normal cases. Unbalance in vocal fold vibration will leave an effect on biomechanical parameter estimations from a given subject when comparing results between neighbour cycles. It is generally accepted that the presence of unbalance is a correlate to vocal fold pathology (as unbalance is related in a certain way with *jitter* and *shimmer*). Unbalance between neighbour phonation cycles may be appreciated in Figure 6 and Figure 7 where the cycle-synchronous estimates show variations which may be around 10% for that specific male subject and under 2% for the female one. As the estimations of mass, stiffness and losses are produced on a phonation cycle-frame basis, the (intra-speaker) unbalance of these parameters (μ_b: Body Mass Unbalance; σ_b: Body Losses Unbalance; γ_b: Body Stiffness Unbalance) may be defined as

$$\begin{aligned} \mu_{bk} &= \left(\hat{M}_{bk} - \hat{M}_{bk-1}\right)/\left(\hat{M}_{bk} + \hat{M}_{bk-1}\right) \\ \rho_{bk} &= \left(\hat{R}_{bk} - \hat{R}_{bk-1}\right)/\left(\hat{R}_{bk} + \hat{R}_{bk-1}\right) \\ \gamma_{bk} &= \left(\hat{K}_{bk} - \hat{K}_{bk-1}\right)/\left(\hat{K}_{bk} + \hat{K}_{bk-1}\right) \end{aligned} \tag{13}$$

where $1 \leq k \leq K$ is the index of the phonation cycle window and $\hat{M}_{bk}, \hat{R}_{bk}, and\ \hat{K}_{bk}$ are the *k-th cycle* estimates of mass, losses and stiffness on a given voice sample (for a single specific subject, i. e., intra-speaker).

Definition of the complete glottal signature for pathology detection

The estimation of the spectral profile singularity parameters may be carried out in different ways for a given frame. One possibility would be using specific frames of N_f samples each sliding a given time interval (every N_s samples) estimating the power spectral density by FFT on the sliding windows as already mentioned. Another possible parameterization strategy would be clipping voicing frames pitch-synchronously in segments aligned with the pitch cycle, from one closing point (closure spike) to the next one as shown for example in Figure 4 (top and bottom right). In this way a different estimation would be produced for each phonation-cycle frame. In the present study voice segment durations of *0.2* sec. long are used, which will include different numbers of phonation-cycle frames for male and female voice (typically 20 for a male voice with F_0=100 Hz and 40 for a female voice with F_0=200 Hz). The number of pitch cycles being used is designated generically as N_k, which will vary from speech segment to speech segment depending on pitch as said.

In a practical case the biometrical signature is estimated from the FFT power spectral density of both dynamic correlates: the AAW and the MWC defined in (1)-(3) to obtain the envelope singularities in the following steps:

- The corresponding dynamic correlate (AAW or MWC) is windowed in specific N_k-sample frames and the power spectral density of each window is estimated by FFT in dB for prototype male and female voice.
- The envelopes of the power spectral densities of these short-time power spectra are estimated.
- The maxima (*) and minima (◊) found on the respective envelopes are detected and their amplitudes and frequencies collected as two lists of ordered pairs: $\{T_{Mq}, f_{Ma}\}$ and

$\{T_{mq}, f_{mq}\}$, with q the singularity ordering index.

- The first (and usually the largest of all maxima: T_{M1}, f_{M1}) is used as a normalization reference both in amplitude and in frequency as given by (4)-(7).
- The reference points in the dynamic correlate power spectral density $\{T_{b1}, \omega_1\}$ and $\{T_{b2}, \omega_2\}$ are estimated.
- The mass, stiffness and losses for the body and cover are estimated following expressions (10)-(12).
- The biomechanical unbalances are estimated according to expressions (13).

The complete biometric signature for pathology detection is composed with the different parameter estimates as follows:

- Pitch, which is assigned to p_1.
- Classical perturbation estimates are assigned to the signature parameters as p_2 (jitter, estimated as the ratio of the difference between neighbour periods with respect to its average value for the voice segment), p_3 (amplitude shimmer estimated as the ratio of the difference between neighbour maximum amplitudes with respect to their average value for the voice segment), p_4 (slenderness shimmer estimated as the ratio of the difference between the acuteness of neighbour closure spikes with respect to their average value for the voice segment), p_5 (area shimmer estimated as the ratio of the difference between neighbour glottal source areas with respect to their average value for the voice segment), p_6 (ratio of the difference between the closure spike amplitude of neighbour cycles with respect to their average value for the voice segment), p_7 (ratio of

the difference between the slenderness of neighbour closure spikes with respect to their average value for the voice segment), p_8 (ratio between the energy of the MWC with respect to the AAW), p_9 (ratio between the frequency position of the MWC second harmonic and the fundamental frequency), p_{10} (ratio between the amplitude of the second harmonic of the MWC relative to the amplitude of the first harmonic).

- Glottal Source spectral parameters, including the maxima and minima of the two first V-troughs and their frequency positions, and the value and position its upper limit, which are assigned to variables p_{18}, p_{19}, p_{21}, p_{22}, p_{23}, p_{27}, p_{28}, p_{30}, p_{31} and p_{32}, plus the notch slenderness parameters assigned to p_{33} and p_{34} as

$$\left.\begin{array}{lll} p_{17}=T_{M1}; & p_{18}=\tau_{m1}; & p_{19}=\tau_{M2}; \\ p_{21}=\tau_{m2}; & p_{22}=\tau_{M3}; & p_{27}=\varphi_{m1}; \\ p_{28}=\varphi_{M2}; & p_{30}=\varphi_{m2}; & p_{31}=\varphi_{M3}; \\ p_{32}=\tau_{Nf}; & p_{33}=\sigma_{m1}; & p_{34}=\sigma_{m2}; \end{array}\right\} \tag{14}$$

- Biomechanical parameters from the Vocal Fold Body and Cover dynamic correlates (AAW and MWC), consisting in estimations of the body dynamic mass, losses and tensions, assigned to p_{35}, p_{36} and p_{37}, the cover equivalent parameters assigned to p_{41}, p_{42} and p_{43}, and their respective unbalances evaluated cycle by cycle, assigned to p_{38}, p_{39} and p_{40} (body), and p_{44}, p_{45} and p_{46} (cover) as

$$\left.\begin{array}{lll} p_{35}=M_b; & p_{36}=R_b; & p_{37}=K_b; \\ p_{38}=\mu_b; & p_{39}=\rho_b; & p_{40}=\gamma_b; \\ p_{41}=M_c; & p_{42}=R_c; & p_{43}=K_c; \\ p_{44}=\mu_c; & p_{45}=\rho_c; & p_{46}=\gamma_c; \end{array}\right\} \tag{15}$$

As each parameter was estimated on a phonation-cycle basis, for a prototype male voice

(with pitch around *100 Hz*) an average of *N=20* values was obtained, which for a prototype female voice (with a typical pitch of *200 Hz*) should be around *N=40*. In this way *J=46* means and standard deviations of each observation parameter p_{in} over $1 \le k \le K$ phonation cycles following (6) can be estimated as

$$x_{ij} = \frac{1}{K} \sum_{k=1}^{K} p_{ijk} \tag{16}$$

$$\sigma_{ij} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(p_{ijn} - x_{ij} \right)^2} \tag{17}$$

where $1 \le j \le J$ and $1 \le i \le I$ are respectively the parameter and speaker indices assuming reasonable stationary conditions along the frame duration (considering that a stable vowel is being produced). The biometric signature used in the study is summarized in Table 1. The calibration of the estimation algorithmics has been carried out on two speakers, one male (#536) and one female (#452) selected by their stable phonation characteristics, which are given in Table 2. Both speakers were inspected by video-endoscopy to disregard any organic anomaly, and GRBAS evaluated [36]. Their phonation stability was evaluated on phonation-cycle estimates of pitch, this being around ±0.101 Hz in an average of 202.74 Hz for the female speaker and ±0.097 Hz in 106.49 Hz for the male speaker over a 60-200 msec frame battery of estimation experiments in increments of 20 msec (for frame sizes of 60, 80, 100, 120, 140, 160, 180, 200, 220, 240, 260, 280 and 300 msec, which included between 12-40 and 6-20 phonation cycles, respectively). Average estimations of the jitter were under 0.5% with standard deviations of 0.36% and 0.31%. Shimmer was lower for the male (average: 0.97%, std. dev.: 0.85%) than for the female speaker (average: 2.62%, std. dev.: 1.94%). The Noise to Harmonics Ratio was smaller for the female voice (average:

5.03%, std. dev.: 0.36%) than for the male voice (average: 6.50%, std. dev.: 0.26%). The calibration results for the most relevant spectral parameters (x_{19}-x_{22}) and the body (x_{35-37}) and cover (x_{41-43}) biomechanical parameters accordingly to the experimental battery explained before are given in Table 3. It may be seen that the estimates of the biomechanical parameters behave much better than those of the glottal spectral profile in general, with the exception of x_{43}, which exhibits the strongest variability among the biomechanical ones. In general the standard deviation of biomechanical parameters is under 1%, which is the variability supposedly introduced by the estimation method, as the deviations strictly imputable to true intra-speaker variability as assessed from the pitch estimates is well under that figure (around 0.1%) for the same set of experiments. This allows to conclude that biomechanical estimates following (9)-(12) are much more robust than time-domain or spectral profile estimates, and may be doubly useful as they convey important semantics regarding biomechanical alterations of the vocal fold system.

5. Materials and Methods

The methodology proposed for pathology detection from voice recordings was intended to increase specificity and reliability in speaker recruiting (collecting general data bases separating patients by groups of gender and age) as well as in signal processing (using separate glottal correlates as the glottal source, the AAW and the MWC instead of the original voice signal). The strategy for separating glottal correlates into the AAW and MWC has been fully justified by the need to obtain body and cover biomechanical estimates. The need for gender-specific recordings is justified as well in the differences observed between the glottal correlates detected for both genders. A similar assertion

should stand as far as age is concerned, although the present study is concentrated in adult voice only. It is well established through medical literature that gender-sensitive issues have to be taken into account when creating databases for medical applicability [34], as the biological differences between male and female subjects result in differentiations of physiological structures and functions of body organs. The case of voice is not an exception, therefore, methodologically speaking, subjects under test should be evaluated against control groups carefully selected by gender. This is also true when age is considered: children voice is completely different from adult voice and databases should be captured and modelled having this fact into account [22].

Based on these considerations a corpus of 200 equally distributed normal and pathologic subjects of both genders was randomly selected from a wider database recorded during the lifetime of project MAPACI [35] oriented to the study of speech pathology. The corpus contained 50 normal (FN) and 50 pathologic (FP) female cases and 50 normal (MN) and 50 pathologic (MP) male cases. Speaker ages ranged from *19* to *56*, with an average of *30.02* years and a standard deviation of *9.94* years. The normal phonation condition of speakers was determined by electroglottography, video-endoscopy and GRBAS evaluation [36]. Pathologic sets comprised approximately the same amounts of mild (functional) and mid-severe cases, ranging from defective closure to polyps, nodules and Reinke´s Edemae. The recordings consisted in three utterances of the vowel */a/* of about *3 sec* per record. A *0.2 sec* frame from the record centre was used in the estimations.

Two related studies were conducted using this database. In the first study a control set of 100 cases equally balanced by gender and pathology condition (25 FN + 25 FP + 25 MN + 25 MP) referred as S_c was selected. It was further decomposed into a male (25 MN + 25

MP) and a female (25 FN + 25 FP) control set, respectively designated as S_{cm} and S_{cf}, to take into account gender influence on pathology condition assessment. These sets were used to determine the best parameters for pathology detection using Linear Discriminant Analysis based on Fisher's Discriminant Ratio

$$fdr_j = \frac{\left(x_{mj} - x_{fj}\right)^2}{\sigma_{mj}^2 + \sigma_{fj}^2}; \quad 1 \le j \le J \tag{18}$$

where (x_{mj}, σ_{mj}) and (x_{fj}, σ_{fj}) are the means and standard deviations of S_{cm} and S_{cf} distributions for parameter j. The results of the comparison studies are given in Figure 8 and Figure 9 and will be discussed in the sequel.

A second study was carried out using the remnant 100 cases to test the discrimination capability of the detection algorithms. These cases, organized as well as (25 MN + 25 MP) and (25 FM + 25 FP) will be referred to as S_{tm} and S_{tf}. The detection algorithms are based on Principal Component Analysis [37][38] for dimensional reduction and Gaussian Mixture Models (GMM's) [28] for detection in itself. The methodology used in this case is based on the following steps:

1. Estimates of observation parameter j for the respective speakers $1 \le i \le I$ in the sets S_{cm} and S_{cf} are stacked as a column vector from (16)

$$\boldsymbol{x}_{cmj} = \left[x_{1j}, x_{2j} \ldots x_{ij}, \ldots x_{Ij}\right]^T; \quad \forall i \in S_{cm} \tag{19}$$

$$\boldsymbol{x}_{cfj} = \left[x_{1j}, x_{2j} \ldots x_{ij}, \ldots x_{Ij}\right]^T; \quad \forall i \in S_{cf} \tag{20}$$

2. The estimations for the whole set of parameters $1 \le j \le J$ are piled as two observations matrices

$$\boldsymbol{X}_{cm} = \left[\boldsymbol{x}_{cm1}, \ldots \boldsymbol{x}_{cmj}, \ldots \boldsymbol{x}_{cmJ}\right] \quad (21)$$

$$\boldsymbol{X}_{cf} = \left[\boldsymbol{x}_{cf1}, \ldots \boldsymbol{x}_{cfj}, \ldots \boldsymbol{x}_{cfJ}\right] \quad (22)$$

3. Principal Component Analysis is applied to this dataset as described in [38]. The set of eigenvalues and eigenvectors $\{\lambda_i, \boldsymbol{e}_i\}$ of the covariance matrices $\boldsymbol{C}_{cm}$ and $\boldsymbol{C}_{cf}$ of $\boldsymbol{X}_{cm}$ and $\boldsymbol{X}_{cf}$ are estimated. The sets of observation parameters are re-evaluated in terms of principal components as

$$\boldsymbol{y}_{cmj} = \boldsymbol{X}_{cm}\boldsymbol{e}_{mj}; \quad 1 \le j \le J \quad (23)$$

$$\boldsymbol{y}_{cfj} = \boldsymbol{X}_{cf}\boldsymbol{e}_{fj}; \quad 1 \le j \le J \quad (24)$$

where the column vectors $\boldsymbol{y}_{mfj}$ and $\boldsymbol{y}_{cfj}$ contain the new parameters (principal components) for each speaker in the list $1 \le i \le I$ their variance diminishing with component order, according to their respective eigenvalues $\{\lambda_i\}$, provided that $\lambda_i \ge \lambda_{i+1}$. This means that after a certain point, let's suppose it be $j=r \ll J$, the residual variance contained in the remaining components can be considered negligible, which allows truncating the component set to the first r column vectors, thus reducing the size of the data set substantially. In this practical case $r=12$ grants that at least 99% of the variability of S_c is represented within the reduced component matrix $\boldsymbol{Y}_{\boldsymbol{c}}$ formed piling up the component vectors as in (23).

4. The reduced component matrices $\boldsymbol{Y}_{\boldsymbol{cm}}$ and $\boldsymbol{Y}_{\boldsymbol{cf}}$ are modeled by a GMM system following [19]. Each model will be referred as

$$\Gamma_i = \{w_i, \boldsymbol{\psi}_i, \boldsymbol{C}_i\} \quad 1 \le i \le G \quad (25)$$

where w_i, $\boldsymbol{\psi}_i$ and $\boldsymbol{C_i}$ are respectively the weight with which that specific model contributes to the global probability model, the centroid of the model and the covariance matrix of the set used for modelling. G is the number of Gaussians used in the modelling. The index i is used to label each specific Gaussian distribution used in a particular modelling. In our case, a specific distribution will model the set of male non-pathologic sample group, another distribution will model the female non-pathologic group, and for pathology classification other similar distributions could model pathologic sample groups (according to each specific pathology), either as single distributions or as a combination of them. In the present case, where the objective is detection the use of two Gaussian distributions for male and female non-pathologic sample grouping is enough to grant good detection scores.

5. Steps 1 and 2 are also repeated with the speaker set S_t. Two reduced component matrices $\boldsymbol{Y_{tm}}$ and $\boldsymbol{Y_{tf}}$ are produced.

6. The log-likelihood of each speaker's template is then evaluated as

$$\Lambda(\boldsymbol{y}_{tmi}) = log\left[p(\boldsymbol{y}_{tmi} / \Gamma_{nm})\right] - log\left[p(\boldsymbol{y}_{tmi} / \Gamma_{\bar{n}m})\right] \quad 1 \le i \le I \tag{26}$$

$$\Lambda(\boldsymbol{y}_{tfi}) = log\left[p(\boldsymbol{y}_{tfi} / \Gamma_{nf})\right] - log\left[p(\boldsymbol{y}_{tfi} / \Gamma_{\bar{n}f})\right] \quad 1 \le i \le I \tag{27}$$

where n and $\bar{n}$ are referring to normal (non-pathologic) and non-normal (pathologic) labels, therefore nm, nf, $\bar{n}m$ *and* $\bar{n}f$ refer respectively to the normal male, normal female, pathologic male and pathologic female distributions. In the present case specific models are produced for normal voice for each gender, these being Γ_{nm} and $\Gamma_{nf,}$ therefore the generation probability of pathologic voice is defined by the complementary probability of normal voice. The generation probability for a given

template from the test sets S_{tf} and S_{tm} will then be evaluated as

$$p(\mathbf{y}_{tmj}/\Gamma_{nm}) = \frac{1}{(2\pi)^{G/2}\left|\mathbf{C}_{nm}\right|^{1/2}} e^{-1/2\left(\mathbf{y}_{tmj}-\boldsymbol{\psi}_{nm}\right)^T \mathbf{C}_{nm}^{-1}\left(\mathbf{y}_{tmj}-\boldsymbol{\psi}_{nm}\right)}; \quad 1 \le i \le I \tag{28}$$

$$p(\mathbf{y}_{tfj}/\Gamma_{nf}) = \frac{1}{(2\pi)^{G/2}\left|\mathbf{C}_{nf}\right|^{1/2}} e^{-1/2\left(\mathbf{y}_{tfj}-\boldsymbol{\psi}_{nf}\right)^T \mathbf{C}_{nf}^{-1}\left(\mathbf{y}_{tfj}-\boldsymbol{\psi}_{nf}\right)}; \quad 1 \le i \le I \tag{29}$$

7. The log-likelihood ratio is compared with a threshold θ. Depending if $\Lambda(\mathbf{y}_{tmi}/\Gamma_{nm})>\theta$ or $\Lambda(\mathbf{y}_{tfi}/\Gamma_{nf})<\theta$ the voice of the subject whose template under test is considered normal or pathological. The same decision is taken with the female test set. The decision threshold is then set to adjust the trade-off between labelling pathological voices as normal (False Rejection Rate or FRR) or accepting normal voices as pathological (False Acceptance Rate or FAR). The threshold used is the same for both genders, and results are given in Figure 11 in terms of the number of False Detections as a function of the threshold θ. In the top template the False Negatives (FRR: ○) and False Positives (FAR: Δ) are plotted vs θ. It may be seen that as the rate of False Negatives descends the rate of False Positives ascends up to a point where both rates come even. This is the Equal Error Rate (ERR) point, which for our case is around 3% for a Normalized Threshold value of 71. This means that one out of 33 pathologic cases would be labelled normal at the same time that the same ratio of normal samples is labelled as pathologic. Usually there are other two ways for giving detection results: in terms of the Receiver's Operating Characteristic curve (ROC), which is produced plotting the True Positives vs the False Positives (see the curve labelled with rhombi in the middle template of Figure 11: ◊) or as the Detection-Error Trade-off curve plotting the False Negatives vs the False Positives (see the cuve

labelled with circles in the same template: ○). The smaller the area enclosed by the DET curve against the x-y axes, the better the detection process. Finally, in the bottom template of Figure 11 the ROC and DET curves are given for the detection experiment if male and female samples were treated as a single set, i. e. when gender is not taken into account. It may be seen that the EER is around 11%, which means that almost 1/9 pathologic cases will be labelled as normals if 1/9 normal cases are labelled as pathologic. It can be concluded that the gender-sensitive methodology renders much better results in this case.

6. Results and discussion

From the results presented in Figure 8 and Figure 9 it may be seen that the most resolving parameters as far as pathology is concerned are in order of relevance x_{22}, x_{21}, x_{45} and x_{19} for the female set, whereas these come to be x_{22}, x_{19}, x_{42}, and x_{45} for the male set as given in Table 4. A first inspection of the results tabulated show that albeit some parameters are present in both male and female groups discriminating pathologic from normal cases, their relevance are different, and some of the most important parameters are not shared by both sets. In general the discrimination capability is larger for female than for male cases. This may be due to the larger statistical dispersion shown by female glottal parameters already detected in earlier gender studies [32]. The most resolving parameter for both genders, x_{22} is the height of the third maximum relative to the first one, and therefore is related to mass alterations on the vocal fold cover, whereas x_{19} is a similar parameter related with mass alterations in the fold body, and x_{45} is related with the unbalance in energy losses between consecutive phonation cycles. Parameter x_{22} is generally associated to harmonics ranging

from 5-10, which correspond to vibration modes appearing on the glottal source revealing an anomalous behaviour of the fold cover, and are well in agreement with perturbations in the A_1-A_3 ratio. A similar explanation could be found for x_{19}, responsible in this case for harmonics in the glottal source ranging from 3-6 although this parameter seems to be less significant. The meaning associated to x_{42} and x_{45} has to see with the presence and unbalance of energy losses on the fold cover. This will indicate that vocal folds with anomalous losses of energy and unbalance are being affected by some pathological process. Another complementary view pointing to this conclusion can be extracted from 3D plots of the data sets displayed in terms of the three most relevant parameters after LDA analysis, as in Figure 10 for sets S_{tm} (top) and S_{tf} (bottom) showing the 12-to-3 down-dimensional projection of the statistical sample distributions of the average template matrices X_{cm} and X_{cf}. A careful analysis of the distributions for male subjects shows that normal case groupings are associated with low-valued 2nd and 3rd maxima (x_{19} and x_{22}) and low cover losses (x_{42}), while for female subjects the 3rd maximum and the 2nd minimum (x_{22} and x_{21}) as well as cover losses unbalance (x_{45}) are the most relevant parameters to separate pathology when their values are above certain limits. In general it may be said that normal cases tend to cluster near small values of the discriminating parameters, the pathologic ones spreading over larger areas far from the clusters of normals. Interesting conclusions can be derived also from the analysis of Figure 11 (top) where false detections are plotted vs the discrimination threshold θ. The existence of a given value (near 71) for which the number of false acceptances (o) and false rejections (Δ) come to a minimum indicate the possibility of a very acute detection rates around 97% and above (at a cost of some increment in the FAR). This is better clarified in Figure 11 (middle) where the equivalent ROC and DET

curves for separate-gender pathology detection are given. As a contrast ROC and DET curves for joint-gender pathology detection are also given using the same cohorts of speakers for both training and testing. It must be emphasized in Figure 11 (top) that the distribution of normal speakers is more tightly packed than the distribution of pathologic ones, as indicated by the gracious sloping down of the False Rejection Rate (Pathologics detected as Normals) which degrades smoothly to the right. This is due to the balance between mild and severe pathological cases included the study, which reveals the sensitivity of the methodology to detect even mild pathology. This is especially important for the early detection of pathology before it could develop into a serious problem, and reinforces the pre-screening potential of the methodology proposed. The contrast between separate and joint gender detection curves in Figure 11 (middle) and (bottom) shows that the detection capability of gender-splitting methods can be larger than methodologies based on joint distributions. This result is coherent with the strong differences in larynx physiology and biomechanics found between both genders. This research was conducted on adult persons, leaving the study of children and the elderly for a future study.

7. A study case

To further illustrate the potential use of the methodology presented in the detection of normality and pathology, a specific study case has been taken from the bench test. It is based on data from a 34-year old female, non-smoker, theatre actress, reporting chronic dysphonia, vocal fatigue, changes in loudness and soaring during speaking or singing as a result of a polyp on the right vocal fold as shown in Figure 12 (top). The acoustic analysis of the Glottal Source showed normal ratios for jitter (under 2%), shimmer (under 2%) and

HNR (under 6%). The Glottal Source, AAW and MWC extracted from voice recordings of the patient before and after surgical removal of the polyp (3 months later) are given in the bottom (left and right) templates of the same figure respectively. It may be seen from comparing both figures that the time-domain glottal coefficients CQ, ClQ and SQ (speed coefficient, see [8]) seem to be closer to normal condition after treatment than before as given in Table 5, and that the high-frequency ringing present in the Glottal Source have almost disappeared. Especially important is the study of the MWC, as before surgery the behaviour of the Glottal Source during the Closing Phase showed a tendency to neatly follow the AAW (the vocal fold was so tense that it behaved as a single body-cover structure), therefore the MWC almost disappeared between 3.2 and 4.9 msec (see Figure 12 bottom-left). On its turn, after surgery the MWC during the Closing Phase was clearly restored (see Figure 12 bottom right).

In a further step ahead to check the capability of the classification method proposed, the respective Glottal Signatures before (labelled as *#0E8*) and after treatment (labelled as *#2DC*) were introduced in the database as if produced by two different speakers for their comparison against normal and pathological cases as shown in Figure 13. For the sake of clarity the control set was reduced to 24 FN + 24 FP. The consequence of the comparison is rather interesting (see Figure 13–bottom). The glottal signature labelled as *#0E8* extracted from pre-surgery data (encircled in dash line) was labelled by the clustering algorithm as member of the subset of mild pathological cases (○). After surgery the situation changed essentially, as the glottal signature #2DC associated to post-surgery data was clearly allocated inside the group of normal phonation subjects, labelled as (▼). The arrow shows the change in the patient's condition from the pathologic to the normal groups. This was

confirmed by the by the strong changes observed on the respective spectral signatures of the glottal source as derived from pre- and post-surgery voice records, given in Figure 14.a and b. It may be appreciated there that the spectral contents of the glottal source changed drastically from before to after surgery conditions. In Figure 14.a the harmonic structure of the glottal source between 1500 and 3200 Hz is almost inexistent, whereas this band has been completely restored in Figure 14.b. This experiment shows that the proposed methodology may detect and describe pathology, this capability to be extended to objective pathology level grading in further studies.

8. Conclusions

First of all a consideration on the technique used to extract the glottal source by model inversion is due at this point. A discussion in full on the existence and uniqueness of the solutions found for glottal source reconstruction would take the issue far beyond the scope of this work, which is intended to offer some semantics to parameters used in pathology detection by connecting them with vocal fold dynamics that classical perturbation parameters as jitter, shimmer and other observables do also convey, although less explicitly. Briefly it may be felt that this issue is linked to an inverse problem with two parts: on one hand the precise estimation of the vocal tract and its decoupling from the glottal source, rendering an estimation of the vocal tract transfer function and an estimate for the glottal source free from formant cross-talk; on the other hand the estimation of vocal fold biomechanical parameters from the recovered glottal source. Neither the first inversion problem solution, nor the second can be granted to be accurate and unique by formal modelling. As the emphasis of the present paper is placed on pathology detection the issue

of the accuracy and uniqueness of the solutions contributed is left open. This does not mean that the problem is ignored, but on the contrary that the consistency of the solutions is assessed by evaluating the accuracy and stability of the estimates obtained under statistical coherence criteria. The first criterion to check the validity of the results is that the estimates of the glottal source shows the benchmarks of vocal fold dynamics on its spectral behaviour, in that specific troughs or zeroes produced by anti-resonances of multiple mass-spring dynamical equivalents to the body and cover subsystems are present on it. Apart from glottal biomechanical studies [16][17] this assert is very much availed by direct measurements of vocal fold dynamics shown in [39]. The second criterion means that if the calibration procedures of the method render stable and reliable results showing a degree of intra-speaker variability well below inter-speaker variability for stable and reliable voice samples, these estimates may be found reasonably acceptable and put to indirect validation by other means. That is precisely the reason to introduce intra- and inter-speaker consistency tests as Fisher's Discriminant Ratio. Another indirect test regarding the uniqueness of the solution is based in checking if there is proportionality and linearity in the estimates produced from the same speakers under the same and different conditions. This fact is related as well with the biometric identity of the speaker, and can be formulated in statistical terms of estimate means and confidence intervals.

To summarize, a good model inversion method to estimate the glottal source must jointly determine the VTTF, and the separation lines must meet these criteria:

- The glottal source reproduced from normal voice should show the main patterns of the LF model: a sharp closing spike, a recovery phase, a closed phase, and an open phase.

- The estimation of the zeroes of the vocal tract must not crosstalk with the zeroes of the glottal source (as these are biomechanical and are to be expected in all cases, except in pathologically over-tense voicing).
- Statistical stability and low dispersion (intra-speaker variability) should be observed both in the spectral profiles of the glottal source and the vocal tract as well as in the time-domain parameters estimated on the glottal source when stable segments of sustained vowels under modal phonation produced by healthy subjects are used (i. e. when using test frames ranging from 50-250 msec. of stable voice small dispersion with standard deviation under 1% could be observed).

Secondly a reflection on the use of spectral and biomechanical parameters proposed by the present study is needed. It may seem that what is proposed here is that biometrical should override classical perturbation parameters as jitter, shimmer or HNR, but indeed this is not the intention of the work presented. What is proposed is to combine both kinds of parameters to improve detection scores, and what is even more important, to provide pathology classification. Essentially classical perturbation parameters are observables which may serve as indices to abnormal vocal fold behaviour, but not always. Jitter in its first definition may be seen as a change in the timing conditions of vibration between neighbour cycles, affecting mainly each of two of them (most commonly), and in case of strong pathology distorting completely the pseudo-periodicity of phonation. Jitter is mainly associated with pathologies compromising both vocal folds asymmetrically, for example in case of cysts, polyps, or unilateral vocal fold paralysis. But many other pathologies, as Reinke's Edema or nodules may not manifest by strong jitter as far as the balance between both vocal folds can be forced by the speaker or when a strong control of phonation is

exerted by the patient [40][41] as in the study case presented. Regarding shimmer the same situation holds, one can expect strong deviations from normality when asymmetry is present, but not if the pathology affects both folds and the speaker knows how to control voice production. HNR is associated with defective closure, the patient does not succeed in producing a complete closure and some air can escape during the moments where a complete stop should be expected. This can be manifested as an increment in the turbulence during the pseudo-closing phase due to gas escape, or as a modulation of the glottal source at a frequency in the 4-10 harmonic expressing the successive approximations to unsuccessful closure, and by a reduction in the amplitude and slenderness of the closing spike. But turbulence and low-order modulations may also be present in breathy voice near to whispering, and spike amplitude and slenderness may be affected by non-modal phonation. Therefore classical perturbation parameters can not be always associated to pathology. This means that certain pathologies may show high jitter, shimmer and HNR and others not, whereas non pathologic voice can exhibit HNR and low-order modulations. Needless to say this means that there is not a bi-univocal relation among these indices and the presence of pathology. Or to put it otherwise there will be pathologies capable of disguising under the scope of these indices. Glottal spectral profile and biomechanical parameters convey a completely different semantics, which can complement classical acoustic indices. Evidence tells that if the Glottal Source is carefully extracted the peaks and troughs in its spectral profile can be associated to the dynamic behaviour of the vocal folds acting as masses linked by springs (the k-mass models). Model-based curve fitting may be used in inverting the system and offering a method for the estimation of the biomechanical parameters associated (masses and springs). Although the problem of

uniqueness is still present, it is clear that the alterations and unbalance of masses and tension can be related to distortions in the depth and position of troughs, and vice-versa, as that can be taken for granted. Therefore non-asymmetric pathologies as nodules, edemae, etc will leave their influence in the glottal signature derived from biomechanical estimates. A reduction in the springs linking cover masses will result in deepening the troughs, and vice-versa, an increment in fold tension will reduce trough depth. Incidentally many non-asymmetric pathologies produce a reduction in the tension (for instance edemae) and others produce an increment, as nodules, polyps, cysts, etc. Other pathologies as sulci will result in a lack of linking among masses, producing even deeper troughs. Another interesting biomechanical parameter is the factor of losses which is associated with an inefficient use of energy. This is the case with edemae or sulci. The presence of losses can be associated with a widening of peaks and troughs, therefore these pathologies can not escape to the glottal spectral signature going unnoticed. The combination of spectral and biomechanical parameters may offer a deeper insight into pathology than perturbation parameters alone by adding a semantics which classical perturbation parameters do not offer. The the presence of some unbalance parameters as x_{39}, x_{44} or x_{45} among the most relevant ones after LDA is not coincidental (see Figure 8 and Figure 9), but expressing a kind of actual meaning worth of being further interpreted. Incidentally it may be seen that jitter is also considered a relevant parameter in the detection experiments presented, contrary to shimmer, which does not appear to be that conclusive.

Coming back now to the questions posed in section 1 the study presented can offer the following answers to them:

- The short-time period-synchronous power spectral density of the Glottal Source can be

parameterized to give a general description of specific harmonic-harmonic relations as H_1-H_2, A_1-A_3, H_1-A_1 and H_1-A_3 in a more consistent and formal way (Power Spectral Signature).

- The parameterization of the envelope of the glottal source or mucosal wave correlate power spectral densities retains the basic ratios between harmonics found in the glottal signals when the vocal tract influence has been removed. Moreover one can associate peaks in the power spectral density with specific vibration modes of the vocal fold cover and body, therefore generalizing the concept of harmonic-formant relations. The ratios expressed in the glottal signature are of the type H_m-H_1, where H_m is any harmonic corresponding to a peak or a trough, and H_1 is the harmonic associated to the first peak, both in dB.

- The general relations between the maximum H_1 to the valleys G_m, reveal the coupling between different cover masses, and are a good indicator to pathology when differences of the kind H_1-G_m are small due to excessive cover stiffness. For such reason first and second notch slenderness parameters x_{33} and x_{34} have been added, although their relevance is lower than the relative notch values in themselves. It must be observed that x_{19}, x_{21}, x_{22}, x_{42} and x_{45} are among the parameters most sensitive to pathology.

- Genders show different parameter dispersions, these being larger in female voice, which bay be a beneficial factor for specific classification studies. Having confirmed that the sets of most sensitive parameters to pathology are different for the sets of male and female voice it may be concluded that parameter distributions are sensitive to Gender Bias, and that this fact has to be taken into account when dealing with

pathology detection. Besides, it has been confirmed that provided that Gender Bias has been taken into account, pathology detection can be thesholded using a common criterion, which is equally sensitive for both genders.

- PCA is useful in helping to implement pathology detection based on Gaussian generative models and in visualizing general results when used to reduce the dimensionality of the data sets.

As a general conclusion it may be said that a structured treatment of voice is a real need for pathology detection, as specific and clearly differentiated information is present in the glottal components of voice, independently from features observed in vocal tract features. Therefore splitting voice into vocal and glottal components is a reasonable technique to be used when articulation and pathology are two different objectives, as is the case in pathology detection. It may be also concluded that the glottal signature is sensitive to certain biometric features of the speaker as gender, other issues as age pending on a further study. Voice pathology has to take this conclusion into account, implementing detection and classification methodologies accordingly to the patient's gender. This is especially important as far as the False Acceptance Rates in pathology detection applications are critical to determine the suitability of voice screening in e-health environments. The methodology presented may be generalized to the study of other speaker features as age, voice profile, emotional features and others alike. It also could be of high interest in other areas, as for example in the production and care of the singing voice. Another important study pending is related to the inclusion of the time-domain parameters corresponding to OQ, CQ, ClQ and NAQ within the present methodology.

9. Acknowledgments

This work is being funded by grants TIC2003-08756 and TEC2006-12887-C02-01/02 from Plan Nacional de I+D+i, Ministry of Education and Science, 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. The authors want to express their most thanks to the anonymous reviewers helping to produce a better conceptualized and understandable manuscript.

10. References

[1] Boyanov, B., Hadjitodorov, S. Acoustic analysis of pathological voices. IEEE Engineering in medicine and biology July (1997) 74-82.

[2] Ritchings, T., McGillion, M., Moore, C. Pathological voice quality assessment using artificial neural networks. Medical Egineering and Physics 24 (8) (2002) 561-564.

[3] Hadjitodorov, S., Boyanov, B. and Teston, B. Laryngeal pathology detection by means of class-specific neural maps. IEEE Trans. Inform. Technol. Biomed. 4 (2000) 68–73.

[4] Parsa, V. and Jamieson, D. G. Identification of pathological voices using glottal noise measures. J. Speech, Language and Hearing Res. 43-2 (2000) 469–485.

[5] Godino, J. I. and Gómez, P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. IEEE Trans Biomed. Eng. 51 (2004) 380-384.

[6] Holmberg, E. B. et al. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. J. Acoust. Soc. Am. 84 (2) (1988) 511-529.

[7] Kuo, J. et al. Discriminating Speakers with Vocal Nodules Using Aerodynamic and Acoustic Features. Proc. of the ICASSP'99 1 (1999) 77-80.

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

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

[10] De Oliveira Rosa, M., Pereira, J. C. and Grellet, M. Adaptive estimation of residue signal for voice pathology diagnosis. IEEE Transactions On Biomedical Engineering 47 (1) (2000), 96–104.

[11] Nickel, R. M. Automatic Speech Character Identification. IEEE Circuits and Systems Magazine 6 (4) (2006) 8-29.

[12] Price, P. J. Male and female voice source characteristics: Inverse filtering results. Speech Comm. 8 (1989) 261-277.

[13] Gómez, P. et al. Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters. J. Voice 21 (4) (2007) 450-476.

[14] Gómez, P. et al. Voice pathology detection by vocal fold biomechanical parameter estimation. Lecture Notes in Computer Science 3817 (2006) 242-256.

[15] Gómez, P. et al. Biomechanical Parameter Fingerprint in the Mucosal Wave Power

Spectral Density. Proc. of the ICSLP'04 (2004) 842-845.

[16] Story, B. H. and Titze, I. R. Voice simulation with a bodycover model of the vocal folds. J. Acoust. Soc. Am. 97 (1995) 1249–1260.

[17] Berry, D. A. Mechanisms of modal and non-modal phonation. J. Phonetics 29 (2001) 431-450.

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

[19] Deller, J. R., Proakis, J. G. and Hansen, J. H. L. Discrete-Time Processing of Speech Signals. Macmillan, NY (1993).

[20] Godino, J. I., Gómez, P. and Blanco, M. Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters. IEEE Trans. on Biomed. Eng. 53 (10) (2006) 1943-1953.

[21] Bimbot, F. et al. A tutorial on text-independent speaker verification. Eurasip Journal on Applied Signal Processing (4) (2004) 430–451.

[22] Whiteside, S. P. Sex-specific fundamental and formant frequency patterns in a cross-sectional study. J. Acoust. Soc. Am. 110 (1) (2001) 464–478.

[23] 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 (2004) 95-108.

[24] Titze, I. R. Summary Statement. Workshop on Acoustic Voice Analysis, National Center for Voice and Speech (1994).

[25] Arroabarren, I. and Carlosena, A. Glottal Source Parameterization: A comparative

study. Proc. VOQUAL'03 (2003) 29-34.

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

[27] Akande, O. O. and Murphy, P. J. Estimation of the vocal tract transfer function with application to glottal wave analysis. Speech Communication 46 (1) (2005) 1-13.

[28] Jackson, L. B. Non-Causal ARMA Modeling of Voiced Speech. IEEE Trans. on Acoustics, Speech and Signal Proc. 10 (37) (1989) 1606-1608.

[29] Shalvi, O. and Weinstein, E. New Criteria for Blind Deconvolution of Non-Minimum Phase Systems (Channels). IEEE Trans. Infor. Theory 36 (1990) 312-321.

[30] Doval, B. and d'Alessandro, C., "The Voice Source as a Causal/Anticausal Linear Filter," Proc. of VOQUAL'03 (2003) 16-20.

[31] Godino, J. I., Aguilera, S. and Gómez, P. Automatic Detection of Voice Impairments due to Vocal Misuse by means of Gaussian Mixture Models. Proc. of the IEEE Engineering, Medicine and Biology Conference (2001) 4253-4258.

[32] Gómez, P., et al. Detecting Pathology in the Glottal Spectral Signature of Female Voice. Proc. of MABEVA'07, Florence, Italy, (2007) 183-186.

[33] Rodellar, V. et al. A numerical method based on Pade's Approximation to simulate and design a low-cost auditory filter for speech processing. Simulation: Practice and Theory 1 (1) (1993) 17-19.

[34] Ruiz, M. T. Gender Bias: A polarized view of the Human Gender. J. Epidemiol. Comm. Health 51 (1997) 106-109.

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

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

[37] Johnson, R. A. and Wichern, D. W. Applied Multivariate Statistical Analysis. Prentice-Hall, Upper Saddle River, NJ (2002).

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

[39] Švec JC, Horáček J, Šram F, Veselý J. Resonance properties of the vocal folds: *In vivo* laryngoscopic investigation of the externally excited laryngeal vibrations. J. Acoust. Soc. Am. 108 (4) (2000) 1397-1407.

[40] Fernández-Baíllo, R., et al. Pre-Post Surgery Evaluation based on the Profile of the Glottal Source. Proc. of the MAVEBA07, Florence, Italy (2007) 65-68.

[41] Fernández-Baíllo, R. and Gómez, P. Study of the Mucosal Wave Correlate in Non-Pathological Voice. Poster at the VII Pan European Voice Conference. Groningen, the Netherlands. August 29th-September 1st (2007) 94.

Figure Captions

Figure 1. Examples of reconstructed glottal signals from vowel /a/ for typical male and female normal speakers. In each four templates from top to bottom: input voice, glottal residual, source and flow (four top templates: male prototype; four bottom templates: female prototype). Horizontal axes are given in sec. for a sampling frequency of 11,050 Hz.

Figure 2. Power spectral density of the glottal source from vowel /a/ for the male and female speakers in Figure 1The spectral singularities are labelled as: *-maxima; ◊-minima. Relative amplitudes are given in dB. Horizontal axes are given in Hz for a 512-samples window and sampling frequency of 11,050 Hz.

Figure 3 a). Iterative estimation of the vocal tract transfer function $F_v(z)$ and the glottal pulse residual $s_g(n)$. Blocks $F_g(z)$, $H_g(z)$, $F_v(z)$ and $H_v(z)$ are implemented by successive chains of adaptive paired-lattice filters. b) Paired lattice. A chain of K of these structures will result in a *K-th* master prediction-error filter (upper part of the structure), and in a *K-th* order subordinate FIR filter (lower part of the structure), f, g, p and q being the forward and backward propagation signals in the filter (see [19] for a general explanation). The reflection coefficient c_k is evaluated adaptively in the master lattice and used by itself and by the subordinate one to implement cross-counter blocks as $F_g(z) \rightarrow H_g(z)$ or $F_v(z) \rightarrow H_v(z)$.

Figure 4. Top (left): polar plots for the Glottal Source (*) and Vocal Tract (x) Inverse Models $H_g(z)$ and $H_v(z)$ derived for $K_1=3$ and $K_2=24$. Top (right): associated Glottal Source, AAW and MWC. The *3* zeroes of $H_g(z)$ are aligned on the real axis (as well as a zero from $H_v(z)$). Bottom (left and right): Similar results for $K_1=4$ and $K_2=24$. The *4* zeroes of $H_g(z)$ are arranged as complex pairs between zeroes *1-2* and *3-4* of $H_v(z)$. It may be seen

that in this last case a certain crosstalk from the first four formants has been introduced into the glottal source estimate.

Figure 5. Power spectral density envelope of the glottal source for speaker #185 showing the first notch profile $\{T_{M1}, f_{M1}\}$, $\{T_{m1}, f_{m1}\}$ and $\{T_{M2}, f_{M2}\}$, and the meaning of 10 of the singularity parameters used in the study {p_{17}, p_{18}, p_{19}, p_{21}, p_{22}, p_{27}, p_{28}, p_{30}, p_{31} and p_{32}}. Relative amplitude is given in dB. Horizontal axes are given in Hz.

Figure 6. Phonation Cycle-Synchronous Estimates of the Biomechanical Parameters of the Vocal Fold Body for subject #185 (male prototype). The estimates for each phonation cycle are given on the left hand column. On the right their respective statistical distributions for the frame are plotted. It may be seen that the dynamic mass involved in the phonation oscillates between *6.4* and *7.9* mg, and the average tension is below *2,800* dyn.cm^{-2} A slight unbalance between vocal folds marked by oscillations between neighbour phonation cycles may also be appreciated.

Figure 7. Phonation Cycle-Synchronous Estimates of the Biomechanical Parameters of the Vocal Fold Body for subject #158 (female prototype) on the same basis as the ones given in Figure 6. The dynamic mass involved in the phonation oscillates in this case between *5.1* and *5.6* mg while the average tension is around *9,400* dyn.cm-1. Fold unbalance can also be appreciated.

Figure 8. Top: Statistical distribution of normal against pathologic female samples. Normal cases present lower-valued and less disperse z-scored parameter values. The distributions in pathologic cases tend to be more skewed than in normal cases. It may be seen that the statistical overlap is low in parameters x_{19}, x_{21}, x_{22} and x_{45}. Bottom: The values of Fisher's

Discriminant Ratio for the same parameters confirm the observations in the top template.

Figure 9. Top: Statistical distribution of normal against pathologic male samples. Normal cases present lower-valued and less disperse z-scored parameter values. The distributions in pathologic cases tend to be more skewed than in normal cases. It may be seen that the statistical overlap is low in parameters x_{19}, x_{22}, x_{41}, x_{42} and x_{45}. Bottom: The values of Fisher's Discriminant Ratio for the same parameters confirm the observations in the top template.

Figure 10. 3D plot of the statistical dispersion of normal (o) vs pathologic (∇) samples in terms of the 3 most relevant parameters after LDA (different for each gender). Top: Male set S_{tm} plotted vs x_{19}, x_{22} and x_{42}. Bottom: Female set S_{tf} plotted vs x_{21}, x_{22} and x_{45}. Normal subjects are related to low values of the parameters detected by LDA (left hand side of both plots), pathologic ones spreading over larger values of the discriminating parameters.

Figure 11. Top: Pathology Detection Performance of the method proposed showing the percentage of False Detections (False Rejection Ratio: o - Pathologics detected as Normals and False Acceptance Ratio: Δ - Normals detected as Patologics). The threshold has been nonlinearly expanded to better show the crossing point (Equal Error Rate) and to illustrate the smooth degradation affecting to False Rejection: cases showing strong pathology have been grouped to the left hand side whilst mild pathological (usually detected as functional or pre-physiological) are clearly spread over between threshold values ranging from *10* to *70*. Middle: Receiver Operating Characteristic (ROC) and Detection-Error Trade-Off (DET) curves for separate gender distribution detection showing an Equal Error Rate around a 3%. Bottom: ROC and DET curves for the joint gender distribution detection

showing an EER of around 11%. The same sets of samples were used in both cluster formation and detection experiments.

Figure 12. Study case. Top: The left and right templates show images of pre- and post-surgery vocal folds in removing a gelatine-type polyp (pointed by arrows). Bottom: The left and right templates show the glottal source and mucosal wave analysis corresponding to the same pre- and post- surgery case. The main differences between both profiles are that the pre-surgery is more noisy and close to the prediction of a one-mass model (the glottal source sticks to the AAW during the closing phase, therefore the MWC is very small during this phase indicating that the body and cover masses stick tightly together by more tense cover springs). On the contrary the post-surgery shows a much smoother behaviour and the MWC during the closing phase is restored. The comparison of the MWC during the recovery and close phase shows an agreement in the average pattern, altered by the noisy behaviour of the pre-surgery case.

Figure 13. Top: Clustering results of the case study (pre: #0E8 and post: #2DC) against a group control of 24 FN + 24 FP in terms of three perturbation parameters (x_2: jitter, x_{21}: second trough minimum, x_{42}: cover losses). Normal phonation is clustered in the left lower hand side corner (minimum jitter, depth and losses). Bottom: Close-up view showing the re-settling effect of surgery (labelled by dot circles and arrow) restoring the pathological case to the cluster of normal cases.

Figure 14. Glottal Source Power Spectral Signature for a pathological case. Top: pre-surgery. The harmonic structure between *1500* and *3000* Hz is completely altered. Bottom: post-surgery. The harmonic structure between *1500* and *3000* Hz is clearly restored.

Horizontal axes given in Hz.

Tables

Table 1. Phonation-cycle average parameters used as biometric signature for the study.

Param.	Description
x_1	*Pitch*
x_2	*Jitter*
x_{3-5}	3 different estimations of *shimmer*
x_{6-7}	parameters related with glottal closure
x_{8-10}	3 parameters related with HNR
x_{11-14}	MWC power spectral density in 4 bins
x_{15-23}	Amplitude of the MWC PSD singularities as described in (14)
x_{24-32}	Position of the MWC PSD singularities as described in (14)
x_{33-34}	Slenderness of the first and second "V" notches as described in (14)
x_{35-37}	Estimations of the vocal fold body biomechanical parameters as described in (15)
x_{38-40}	Estimations of the vocal fold cover biomechanical parameters as described in (15)
x_{41-43}	Vocal fold body biomechanical parameter unbalance as described in (15)
x_{44-46}	Vocal fold cover biomechanical parameter unbalance as described in (15)

Table 2. Main features of reference speakers. Average values for pitch, jitter, shimmer and HNR are derived from the analysis of a 200 msec. modal phonation frame of a sustained /a/. Standard deviations are given between parenthesis.

Speaker	Age	Gender	Pitch (Hz)	Jitter (%)	Shim. (%)	HNR (%)	GRBAS
#452	47	Female	202.74 (0.101)	0.42 (0.36)	2.62 (1.94)	5.03 (0.36)	00000
#536	35	Male	106.49 (0.097)	0.45 (0.31)	0.97 (0.85)	6.50 (0.26)	00000

Table 3. Average estimations and standard deviations for spectral (x_{19}, x_{21} and x_{22}) and biomechanical parameters (body: $x_{35\text{-}37}$; cover: $x_{41\text{-}43}$). Percent of std. dev./average ratios are given between parenthesis. The estimates are drawn from processing a set of variable-size frames (60, 80, 100, 120, 140, 160, 180, 200, 220, 240, 260, 280 and 300 msec. long) centered around 1.15 sec. after the voice onset from female (#452) and male (#536) non-pathologic subjects.

Speaker	**#452**			**#536**		
Parameter	**Average**	**Std. Dev.**	**(%)**	**Average**	**Std. Dev.**	**(%)**
x_{19} *(dB)*	-30.23	1.53	5.05	-31.43	1.74	5.52
x_{21} *(dB)*	-40.98	1.19	2.90	-41.88	1.82	4.35
x_{22} *(dB)*	-32.54	0.79	2.42	-34.99	1.21	3.45
x_{35} *(g)*	0.0120	0.0001	0.73	0.0231	0.0002	0.82
x_{36} $(g.sec^{-1})$	4.64	0.02	0.40	4.63	0.03	0.70
x_{37} $(g.sec^{-2})$	19,544	154	0.79	10,331	88	0.85
x_{41} *(g)*	0.0099	0.0001	0.85	0.0153	0.0001	0.83
x_{42} $(g.sec^{-1})$	16.250	0.1243	0.76	13.956	0.0878	0.63
x_{43} $(g.sec^{-2})$	23,971	490	2.05	10,868	448	4.12

Table 4. Most relevant parameters from FDR (male: left; female: right)

Parameter index and name	**Relevance**	**Parameter index and name**	**Relevance**
22. GS PSD 3rd. Max. Rel.	1.3439	22. GS PSD 3rd. Max. Rel.	2.3629
19. GS PSD 2nd. Max. Rel.	1.1928	21. GS PSD 2nd. Min. Rel.	1.8862
42. Cover Losses	0.7187	45. Cover Losses Unbalance	1.4950
45. Cover Losses Unbalance	0.6891	19. GS PSD 3nd. Max. Rel.	1.4325

Table 5. Time-domain glottal coefficients before and after treatment

Case/Coefficient	CQ	ClQ	SQ
#0E8 (before)	0.57	0.30	0.43
#2DC (after)	0.46	0.18	2.00

Figure-1-top

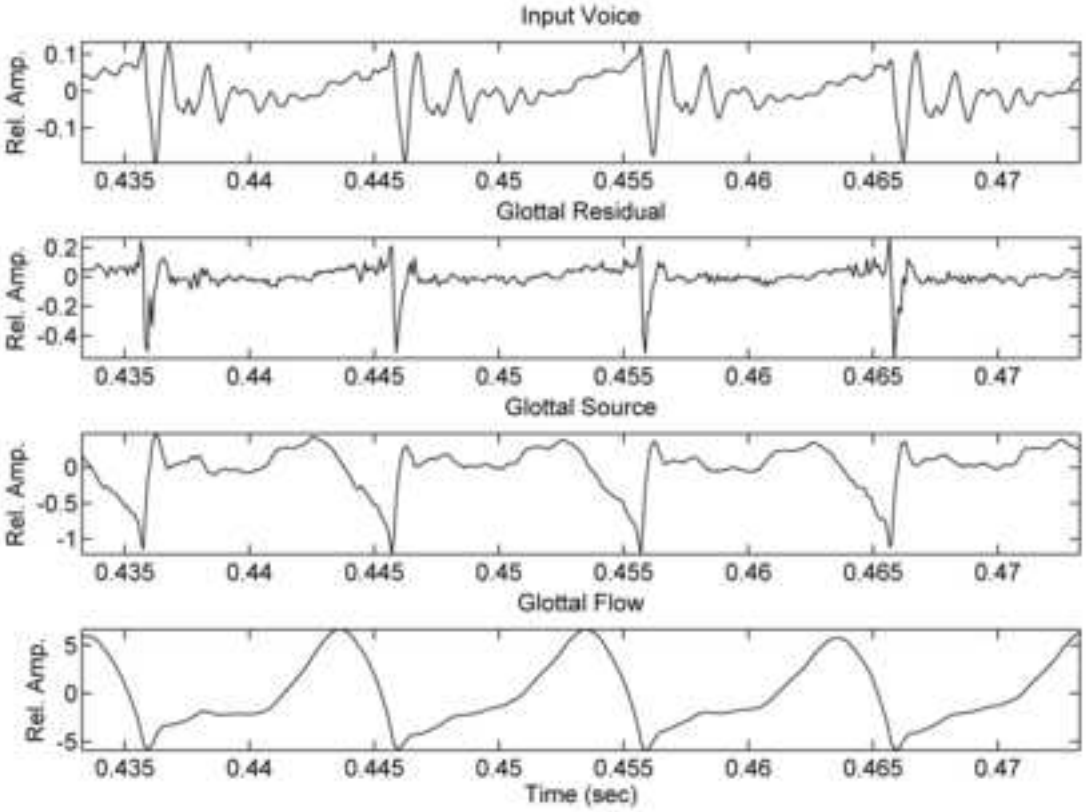


Figure-1-bottom

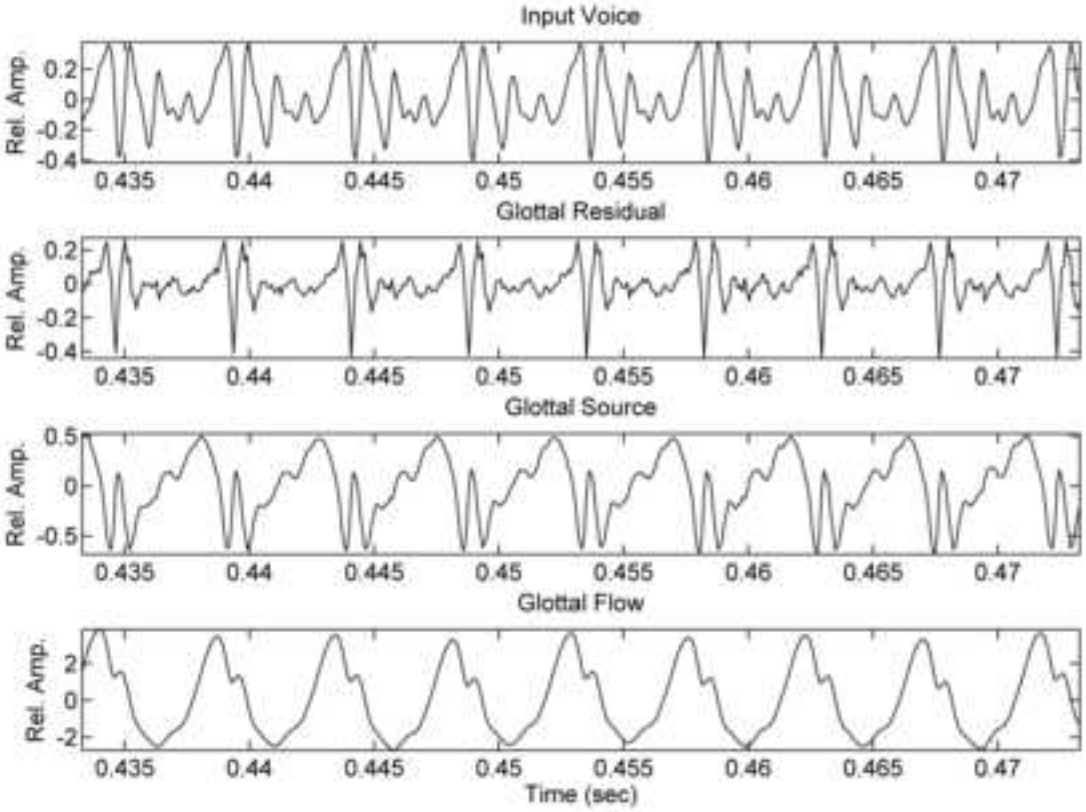


Figure-2-top

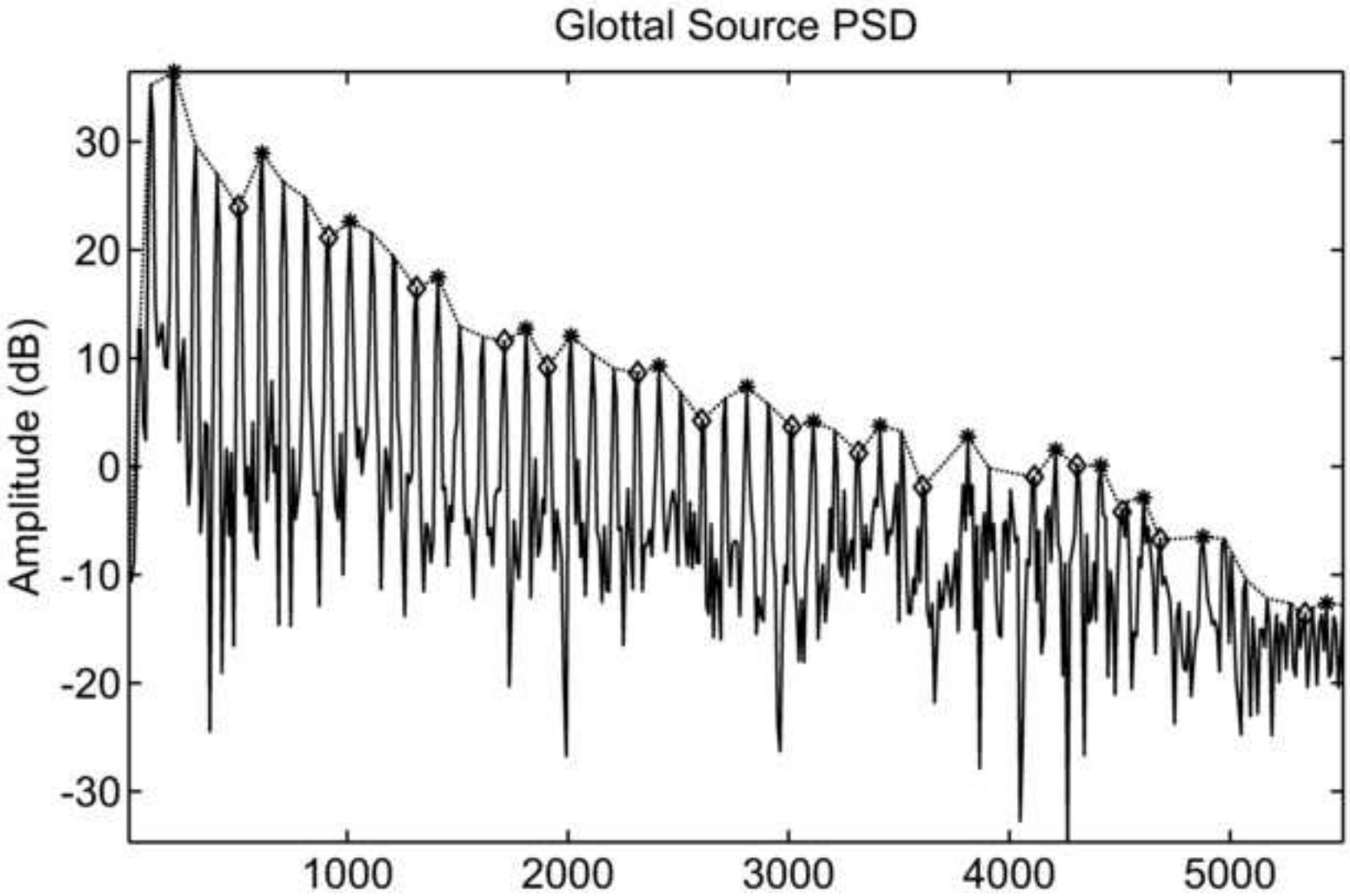

Glottal Source PSD
Amplitude (dB)
30
20
10
0
-10
-20
-30
1000
2000
3000
4000
5000

Figure-2-bottom

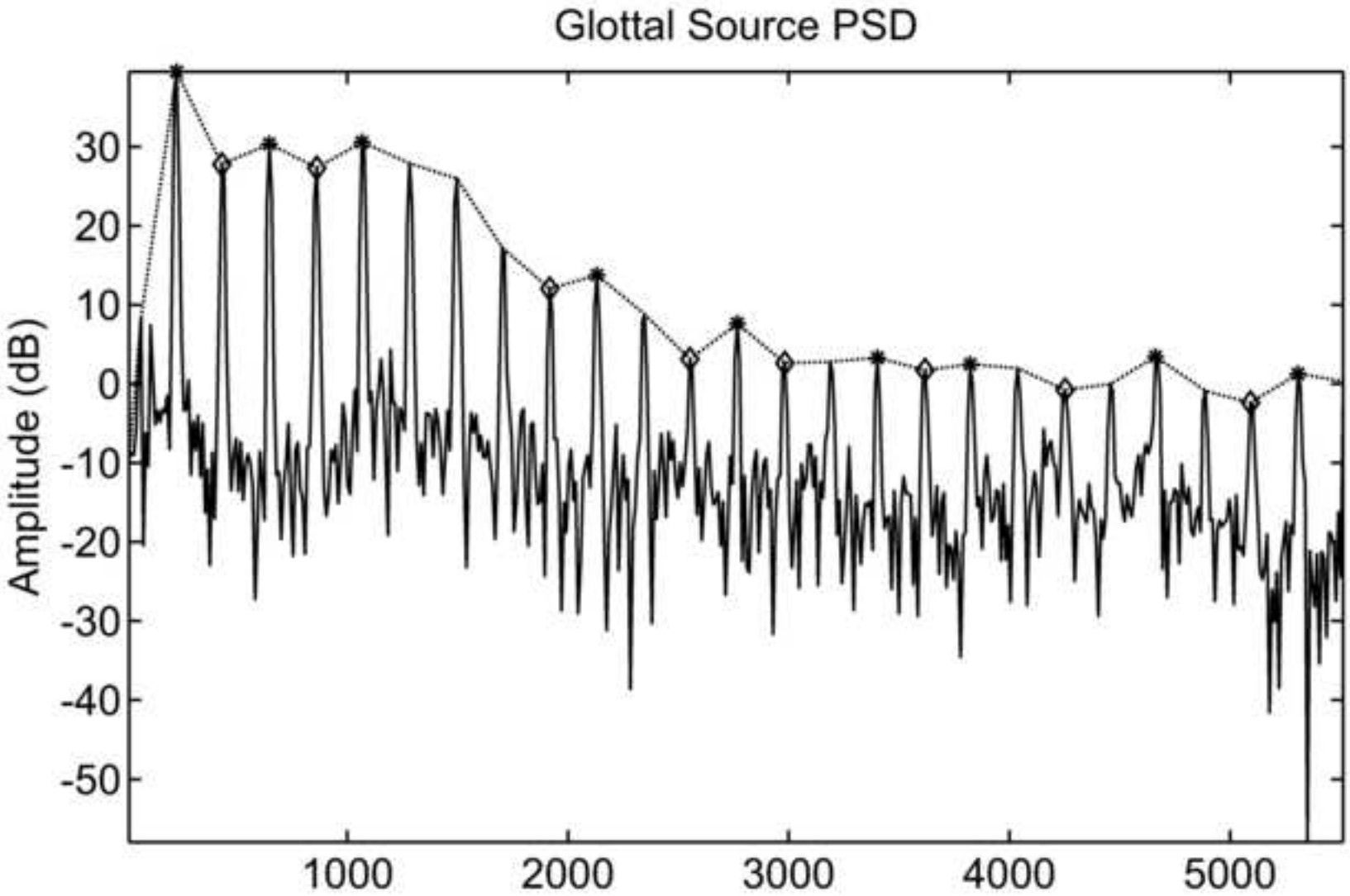


Figure-3a

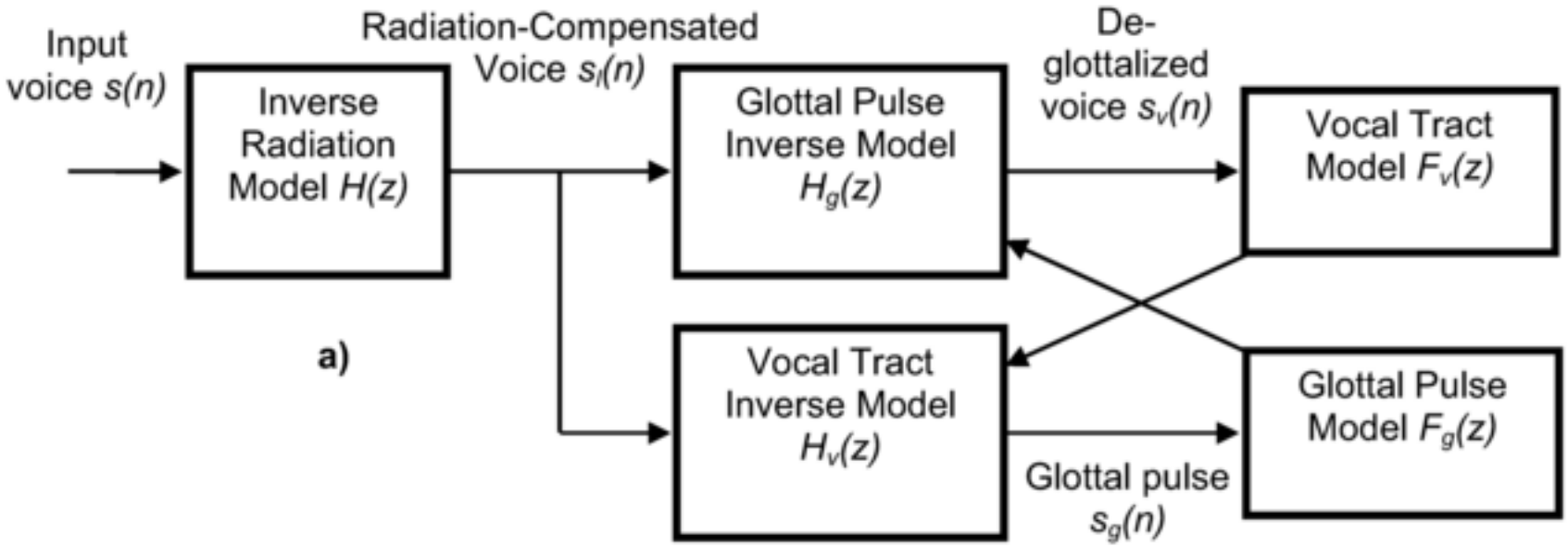

Input
voice s(n)
Inverse
Radiation
Model H(z)
Radiation-Compensated
Voice $s_l(n)$
Glottal Pulse
Inverse Model
$H_g(z)$
De-
glottalized
voice $s_v(n)$
Vocal Tract
Model $F_v(z)$
a)
Vocal Tract
Inverse Model
$H_v(z)$
Glottal Pulse
Model $F_g(z)$
Glottal pulse
$s_g(n)$

Figure-3b

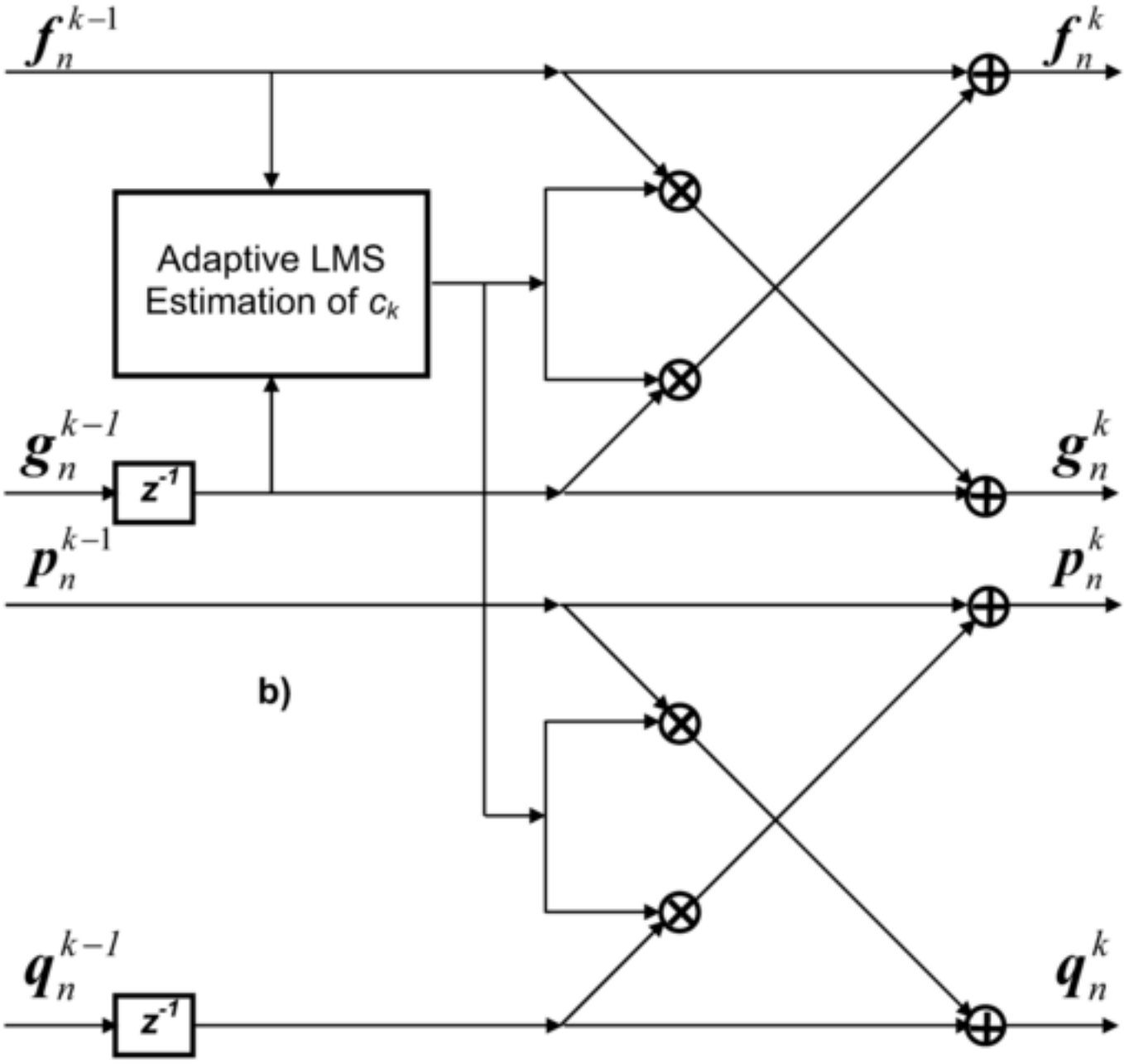


Figure-4-top-left

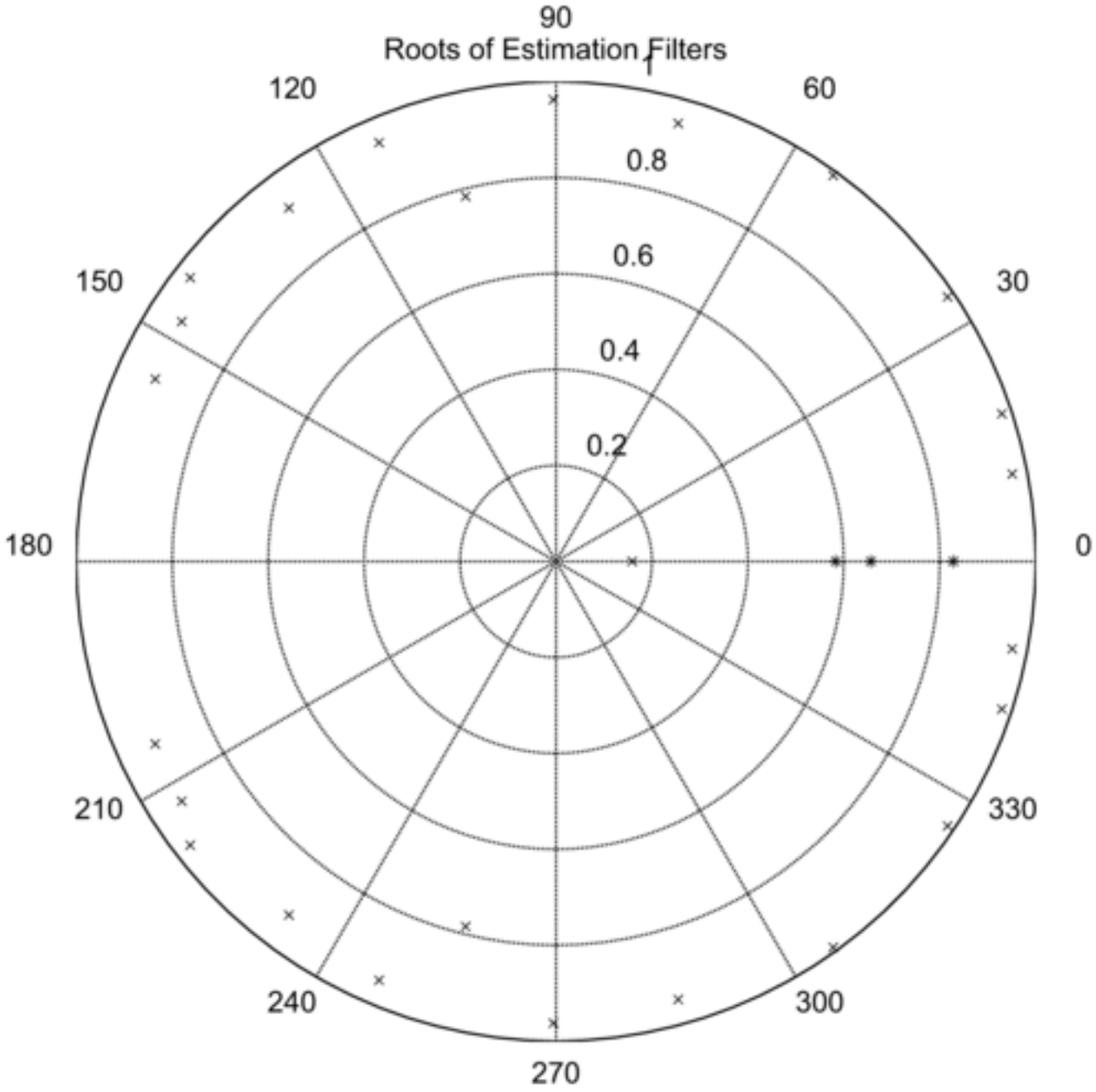


Figure-4-top-right

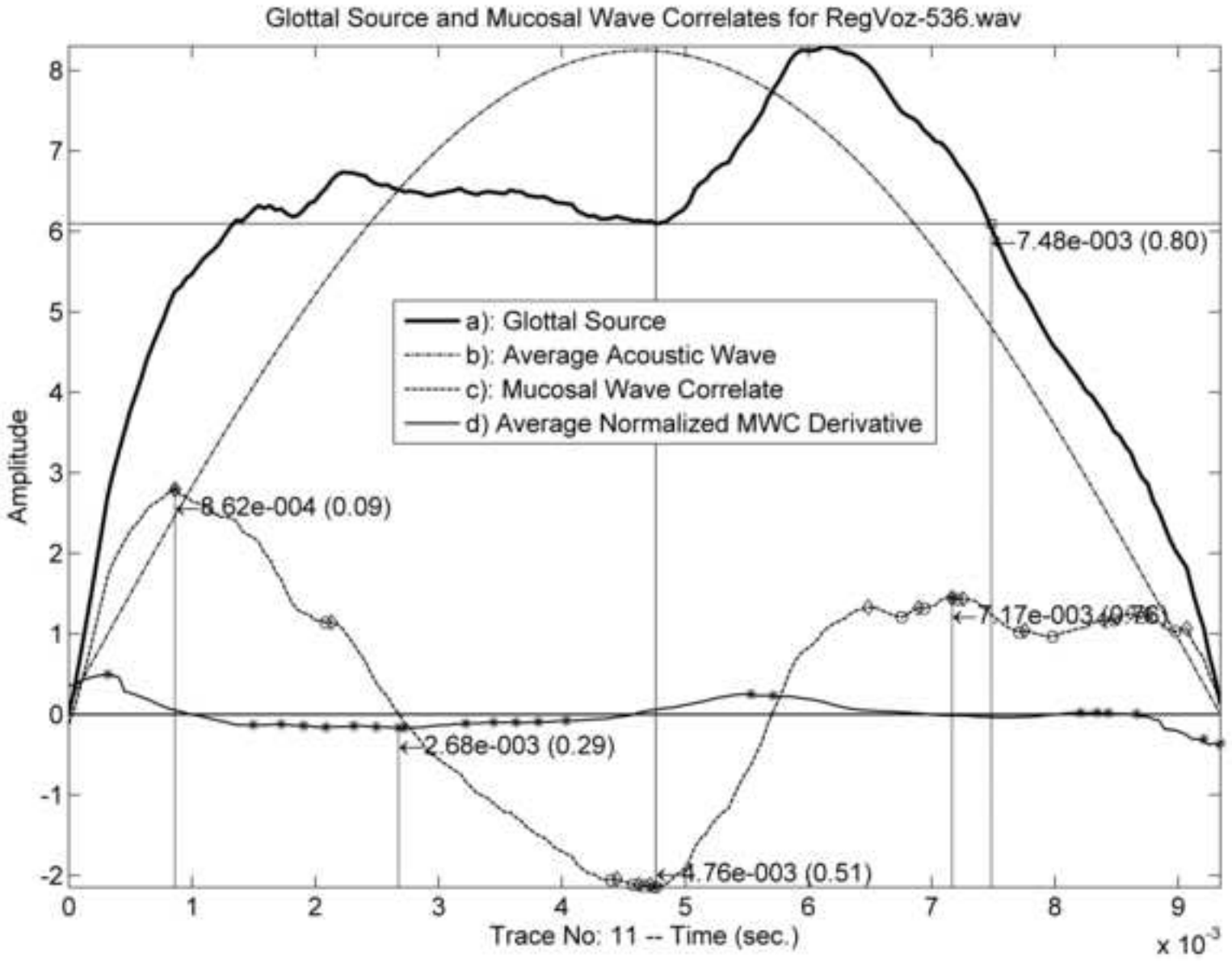


Figure-4-bottom-left

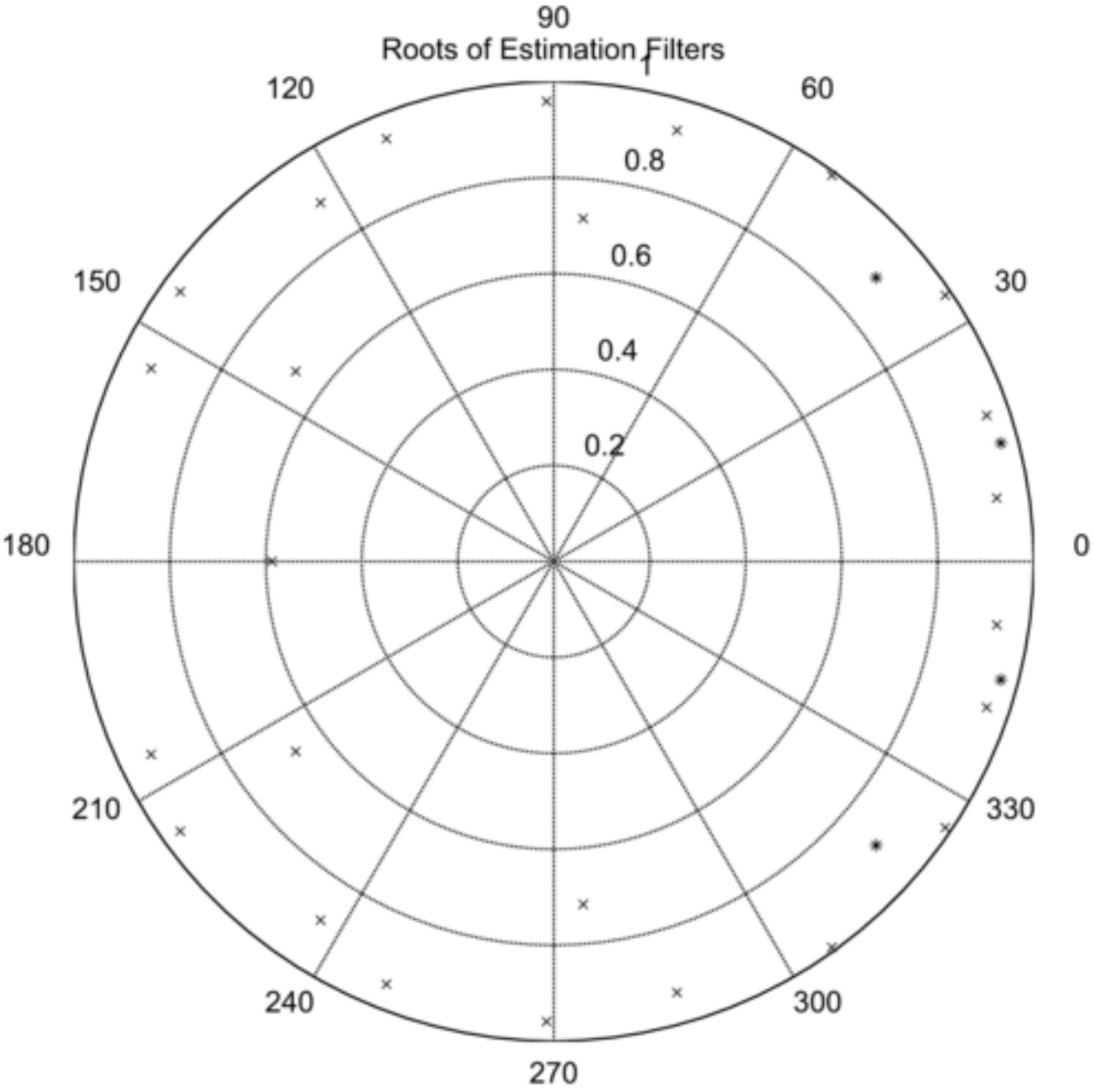


Figure-4-bottom-right

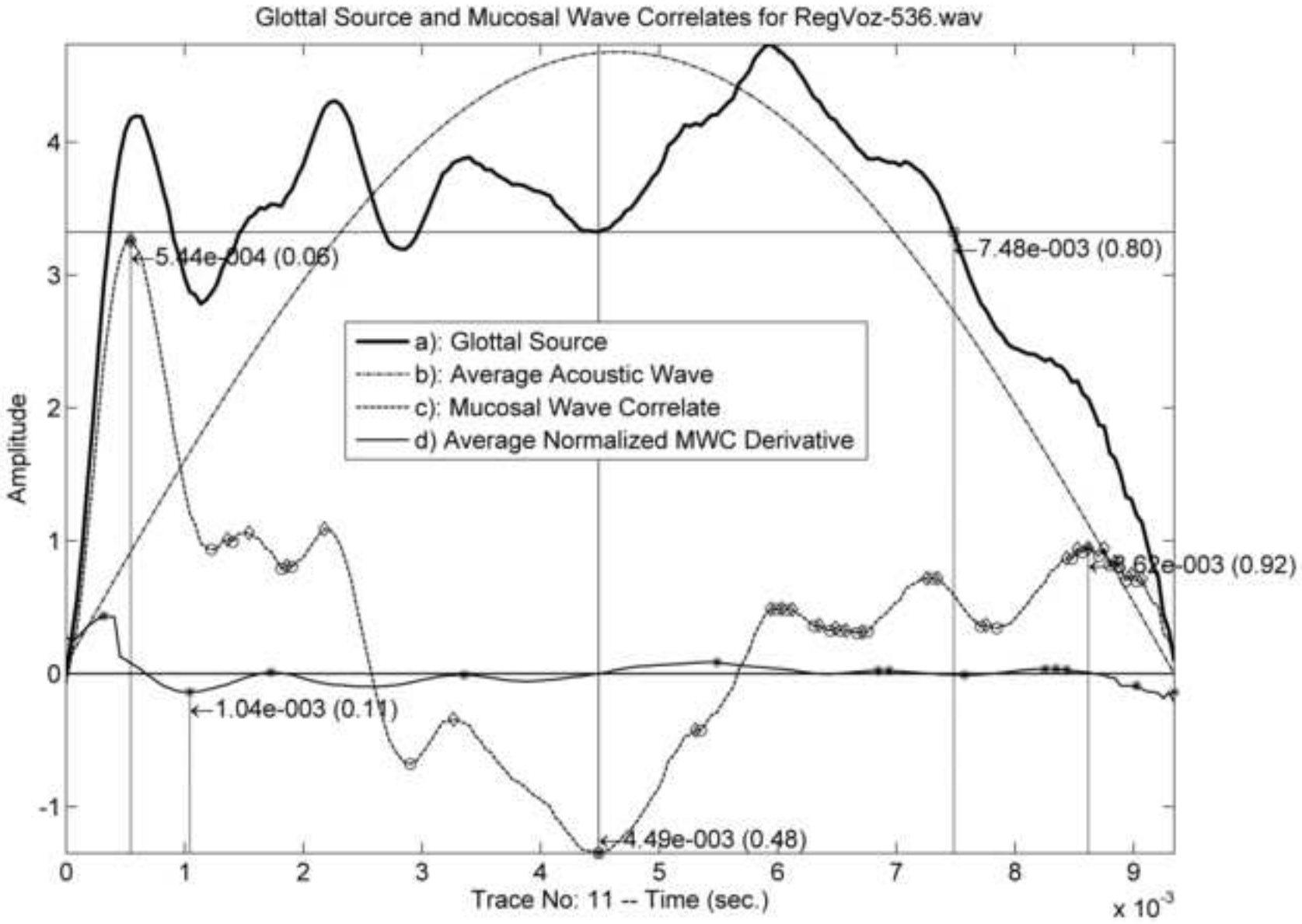


Figure-5

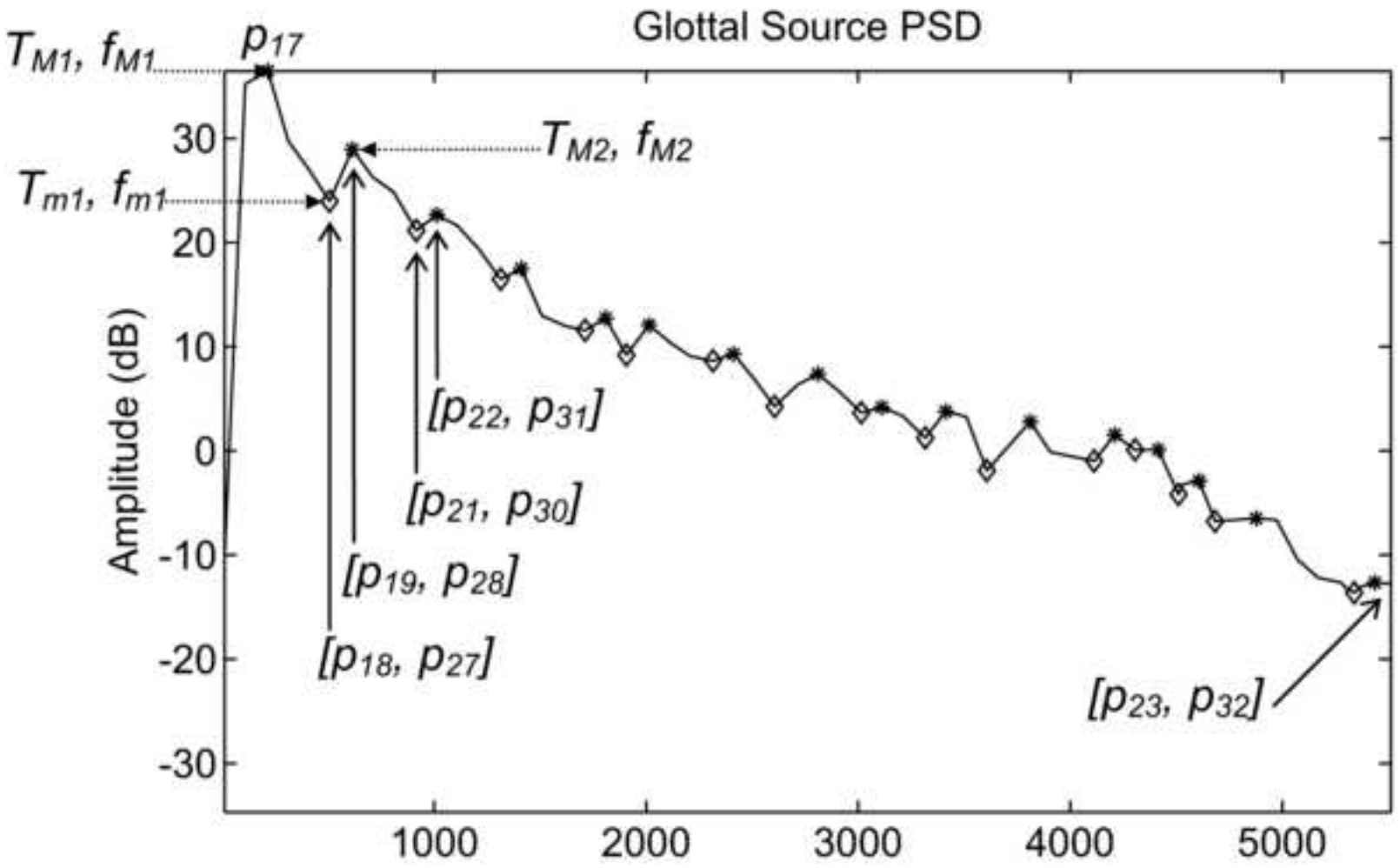

Glottal Source PSD
T_{M1}, f_{M1}
p_{17}
T_{M2}, f_{M2}
T_{m1}, f_{m1}
$[p_{22}, p_{31}]$
$[p_{21}, p_{30}]$
$[p_{19}, p_{28}]$
$[p_{18}, p_{27}]$
$[p_{23}, p_{32}]$
Amplitude (dB)
30
20
10
0
-10
-20
-30
1000
2000
3000
4000
5000

Figure-6

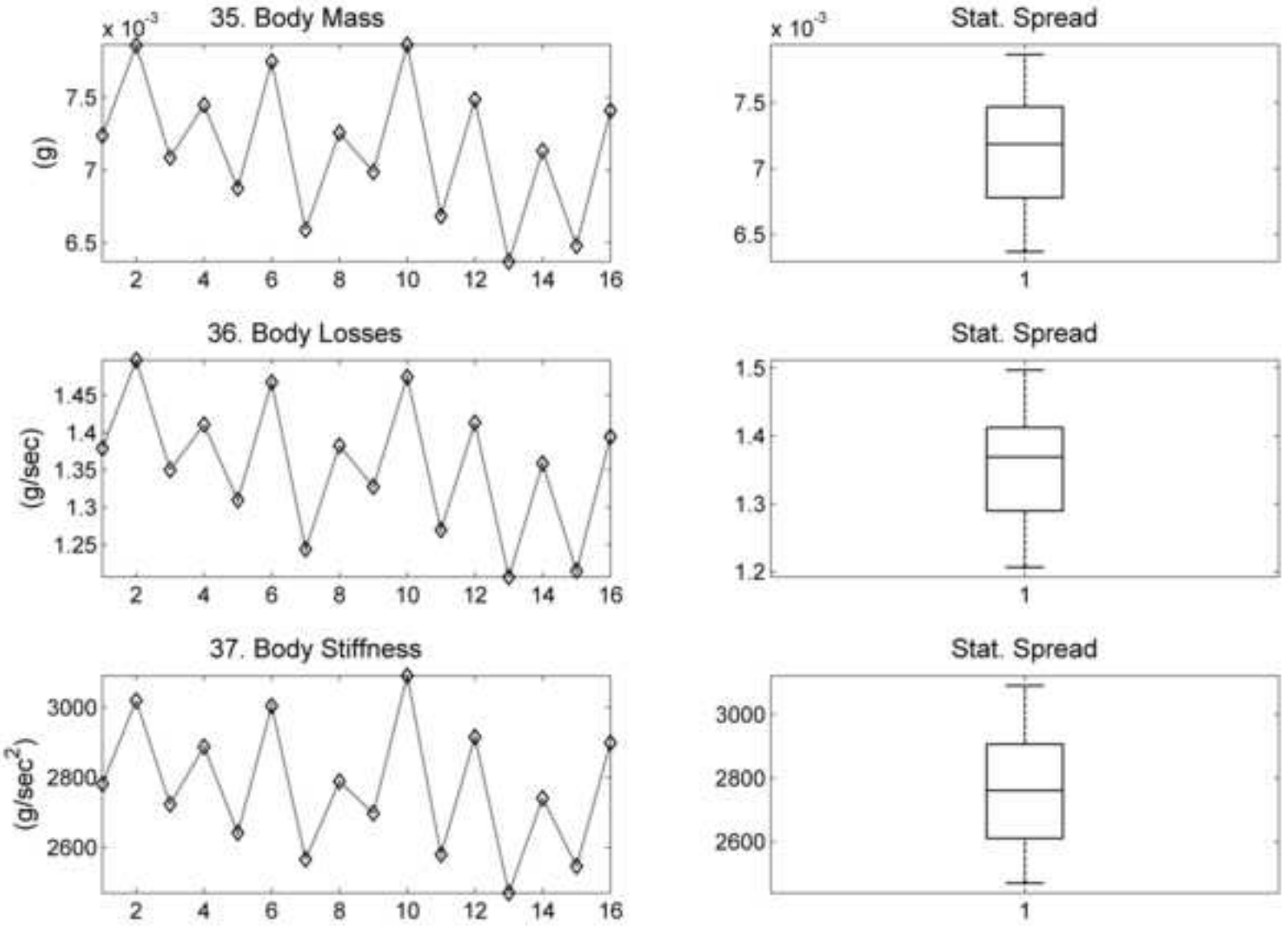

35. Body Mass
x 10^-3
(g)
7.5
7
6.5
2
4
6
8
10
12
14
16
Stat. Spread
x 10^-3
7.5
7
6.5
1
36. Body Losses
(g/sec)
1.45
1.4
1.35
1.3
1.25
Stat. Spread
1.5
1.4
1.3
1.2
37. Body Stiffness
(g/sec^2)
3000
2800
2600
Stat. Spread
3000
2800
2600

Figure-7

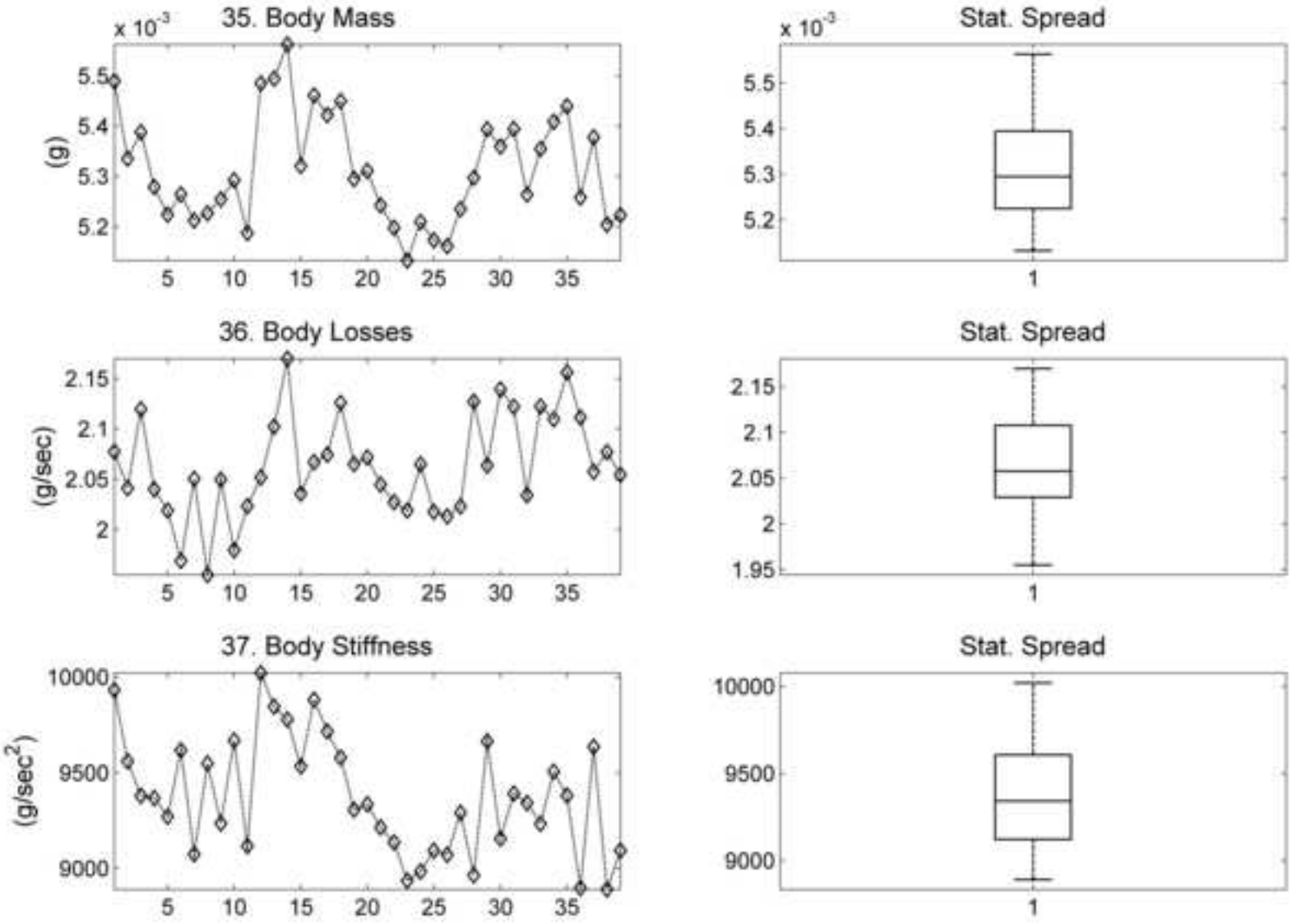

35. Body Mass
x 10^-3
(g)
5.5
5.4
5.3
5.2
5 10 15 20 25 30 35
Stat. Spread
x 10^-3
5.5
5.4
5.3
5.2
1
36. Body Losses
(g/sec)
2.15
2.1
2.05
2
5 10 15 20 25 30 35
Stat. Spread
2.15
2.1
2.05
2
1.95
1
37. Body Stiffness
(g/sec^2)
10000
9500
9000
5 10 15 20 25 30 35
Stat. Spread
10000
9500
9000
1

Figure-8

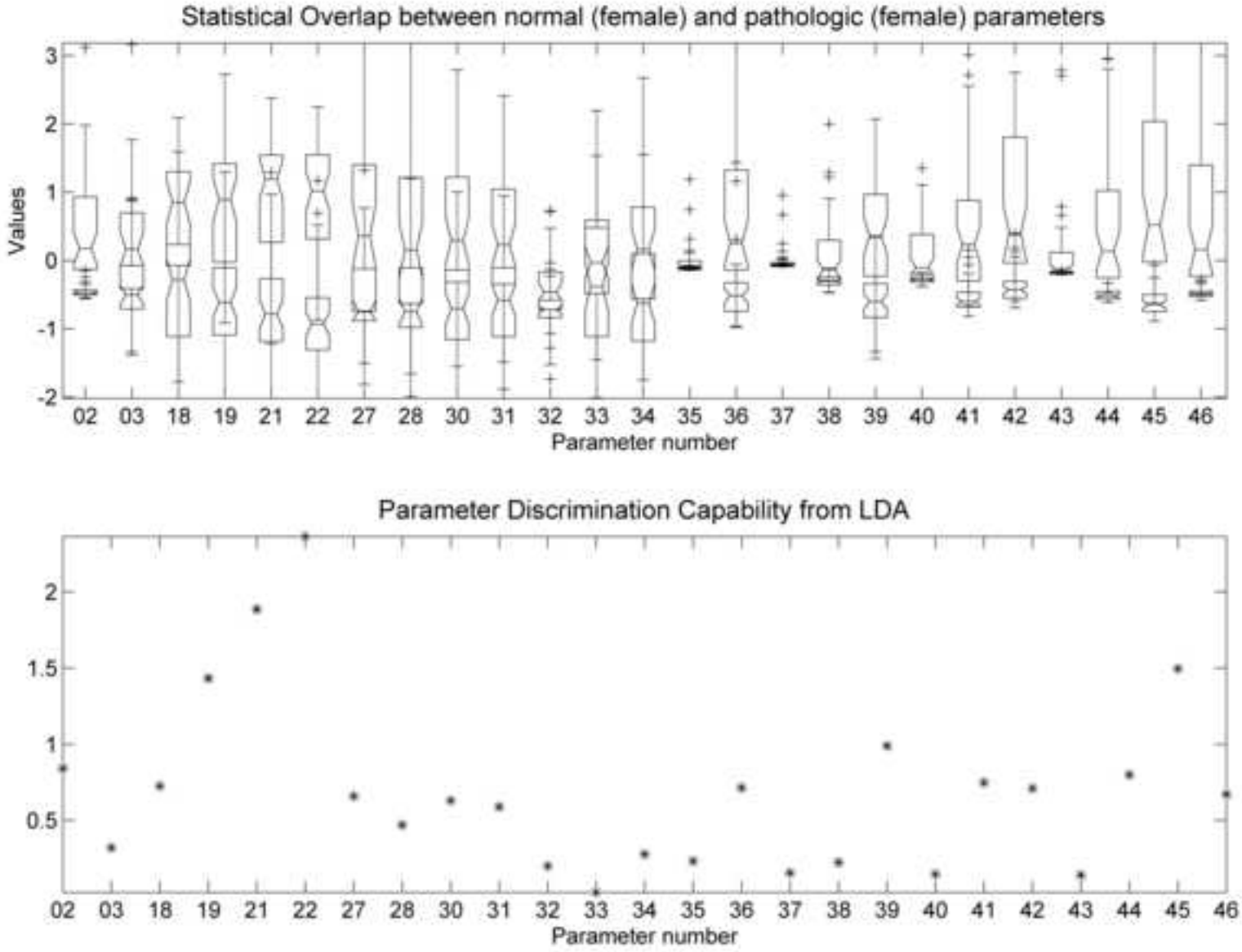

Statistical Overlap between normal (female) and pathologic (female) parameters
Values
3
2
1
0
-1
-2
02 03 18 19 21 22 27 28 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
Parameter number
Parameter Discrimination Capability from LDA
2
1.5
1
0.5
02 03 18 19 21 22 27 28 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
Parameter number

Figure-9

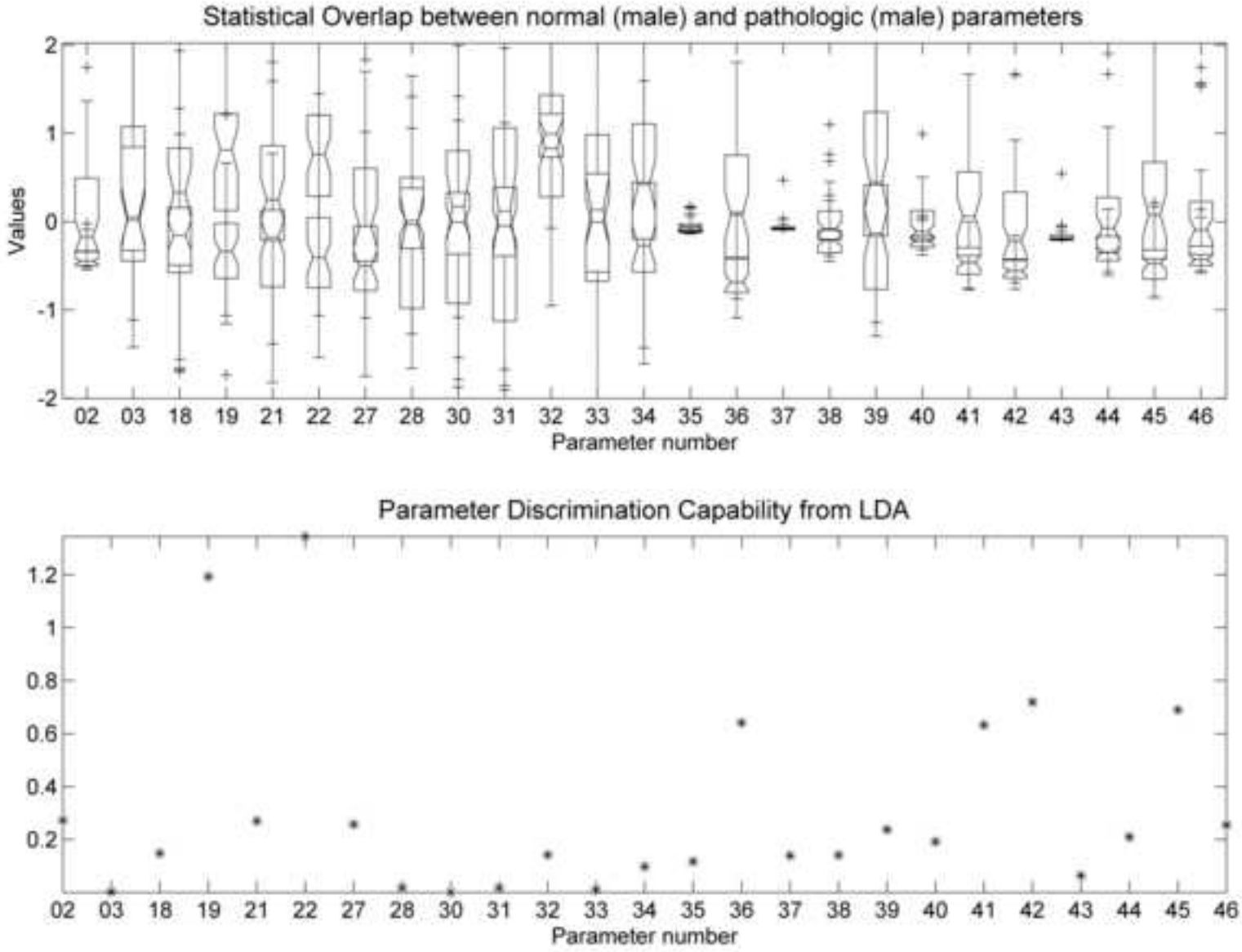

Statistical Overlap between normal (male) and pathologic (male) parameters
Values
2
1
0
-1
-2
02 03 18 19 21 22 27 28 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
Parameter number
Parameter Discrimination Capability from LDA
1.2
1
0.8
0.6
0.4
0.2
02 03 18 19 21 22 27 28 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
Parameter number

Figure-10-top

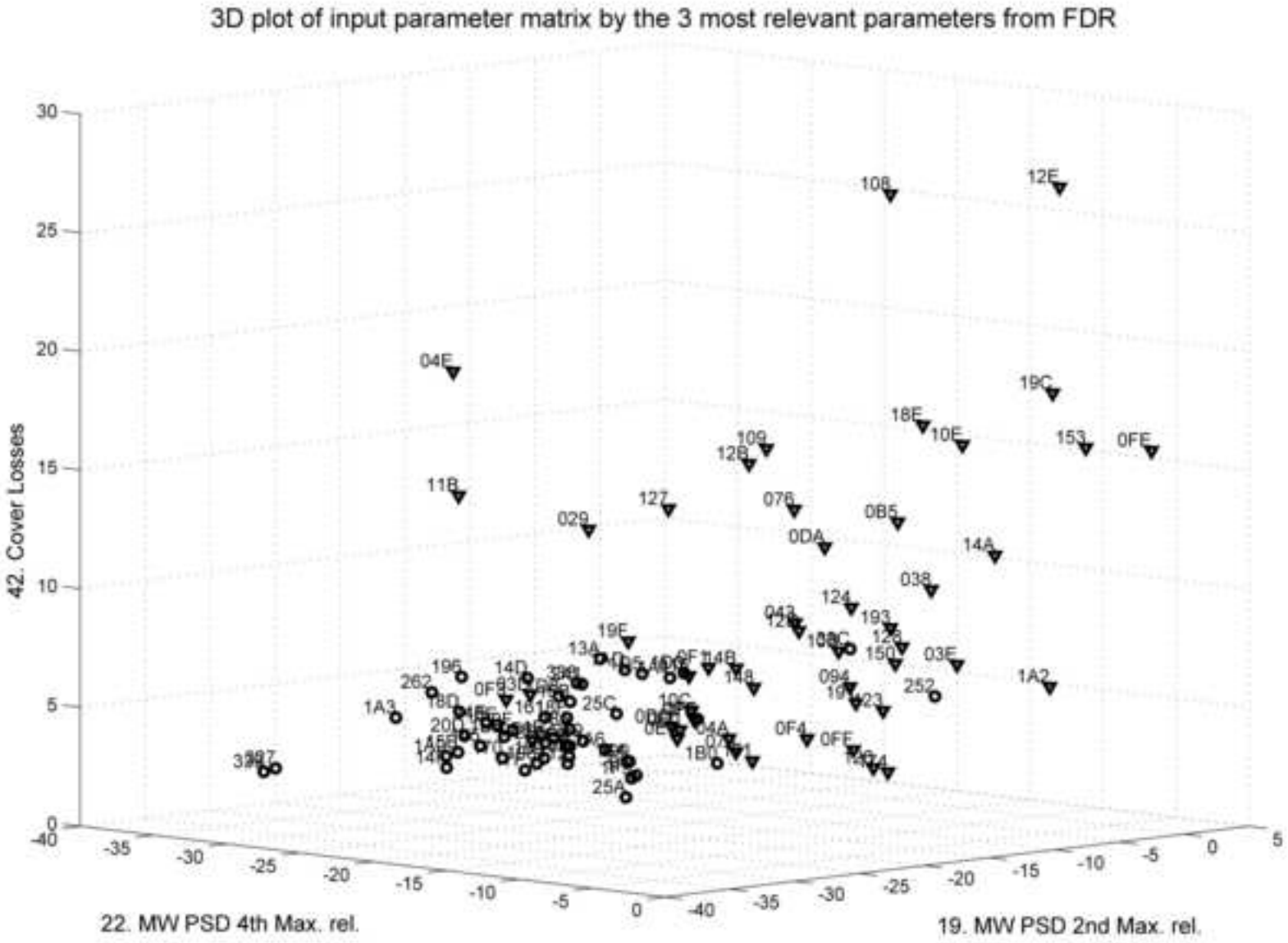


Figure-10-bottom

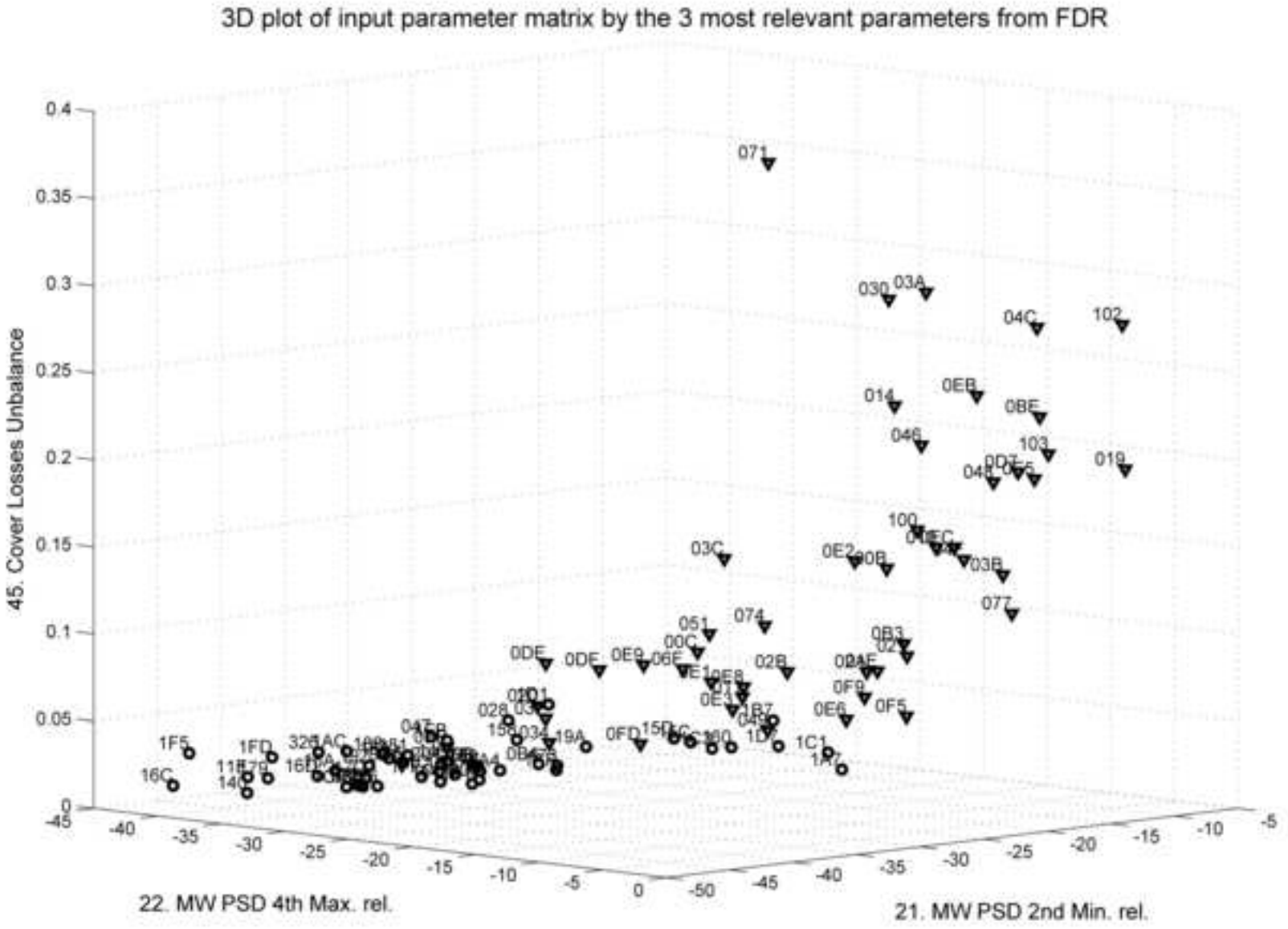


Figure-11-top

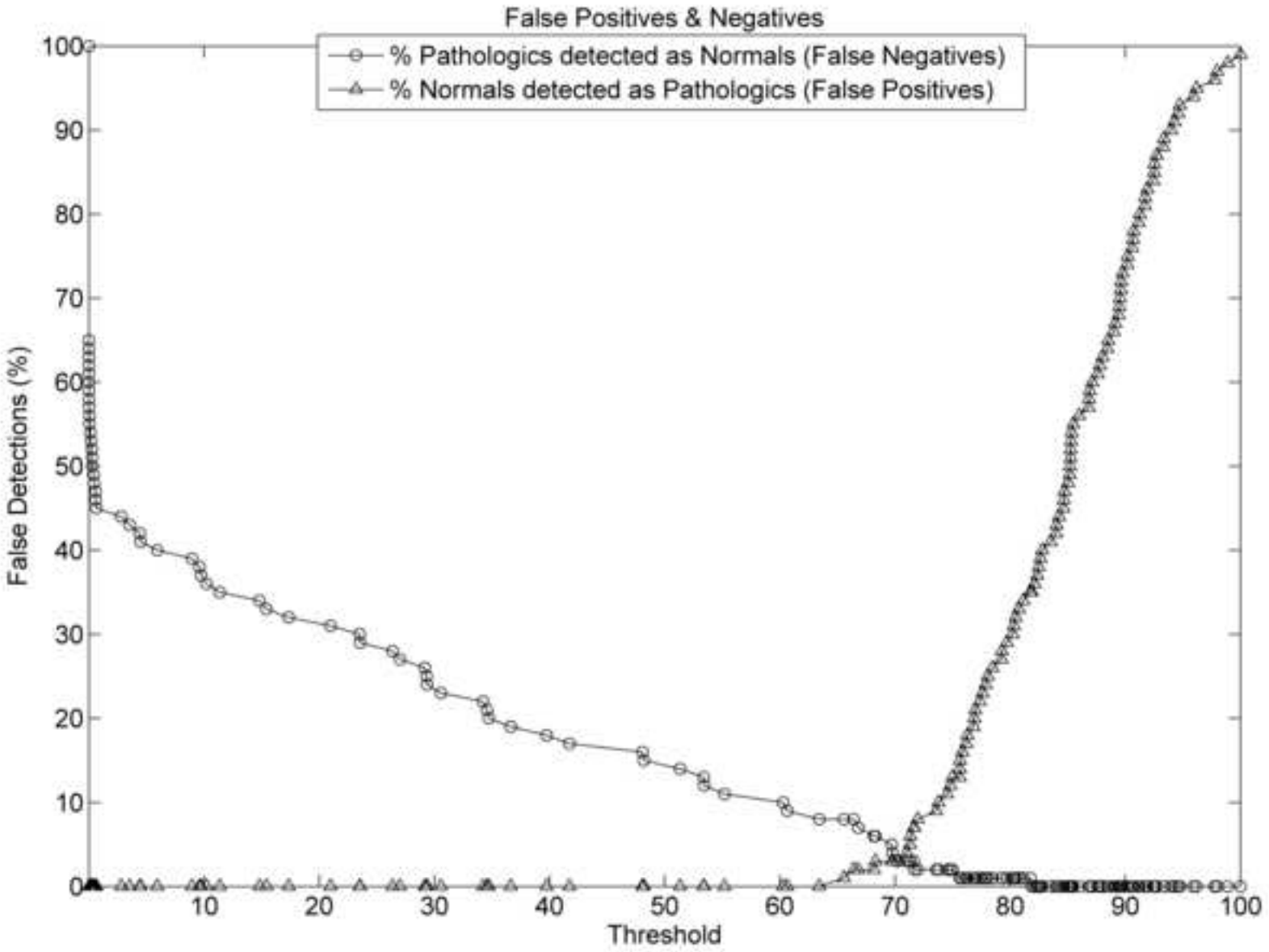


Figure-11-middle

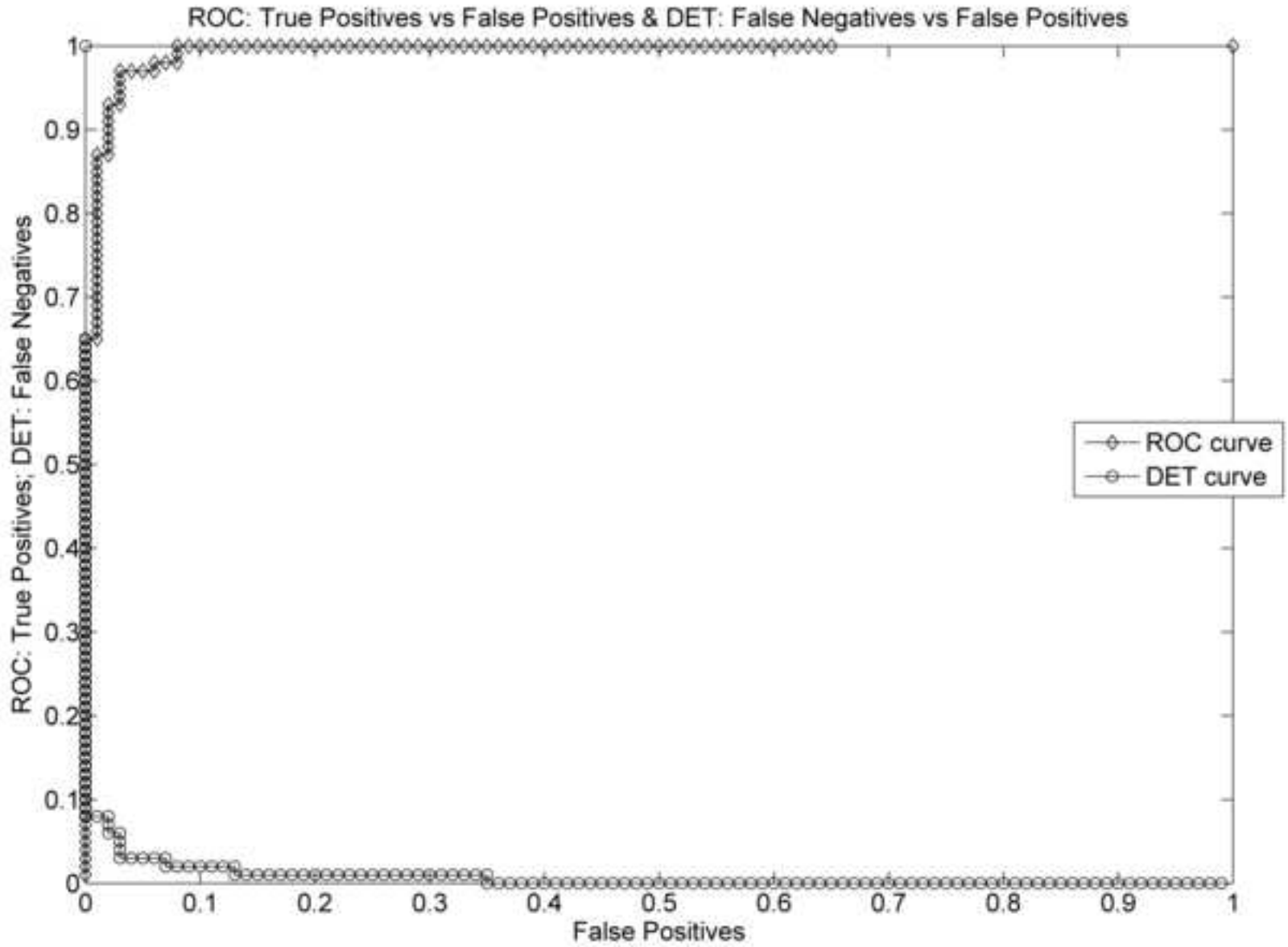


Figure-11-bottom

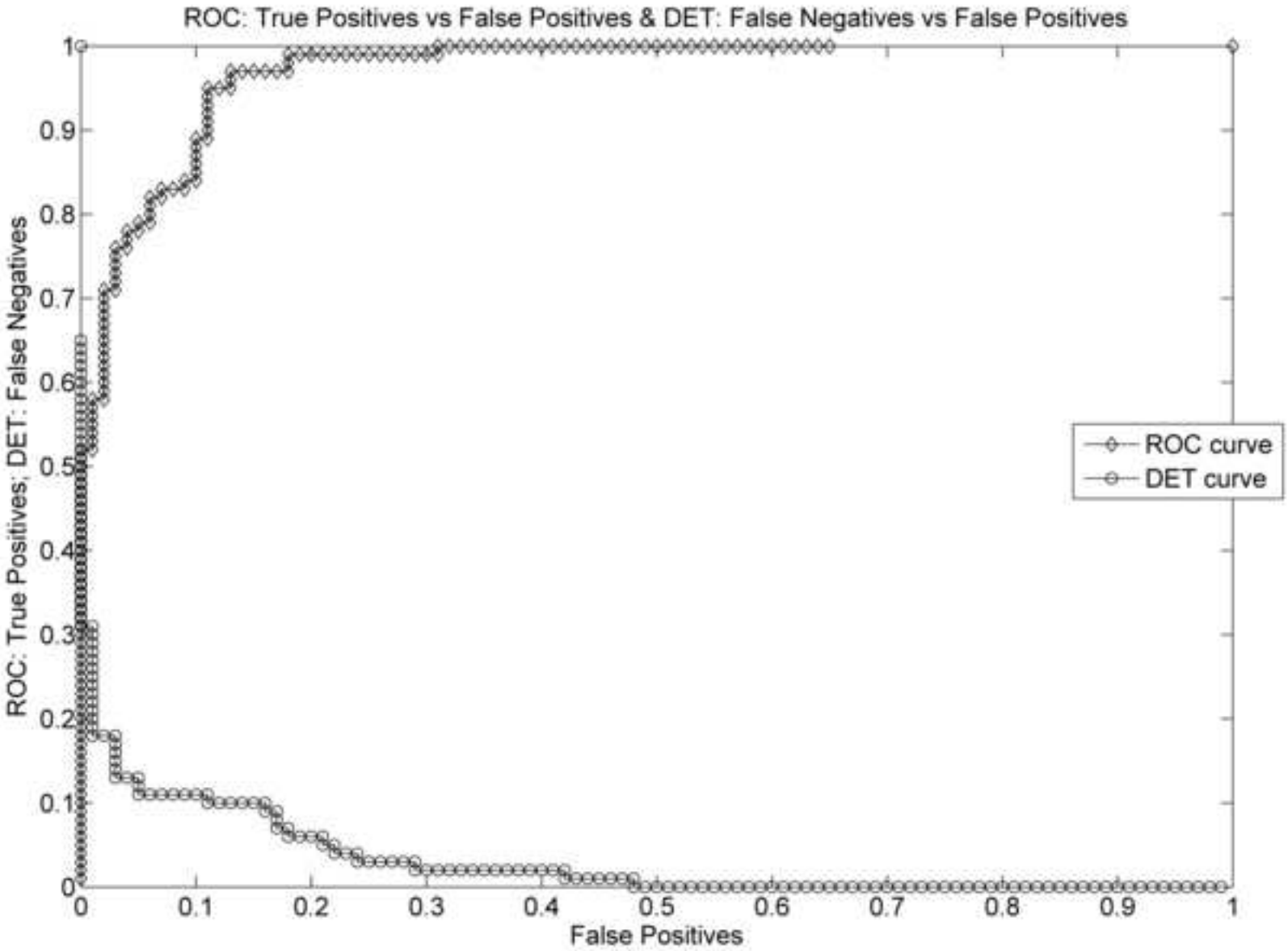


Figure-12-top

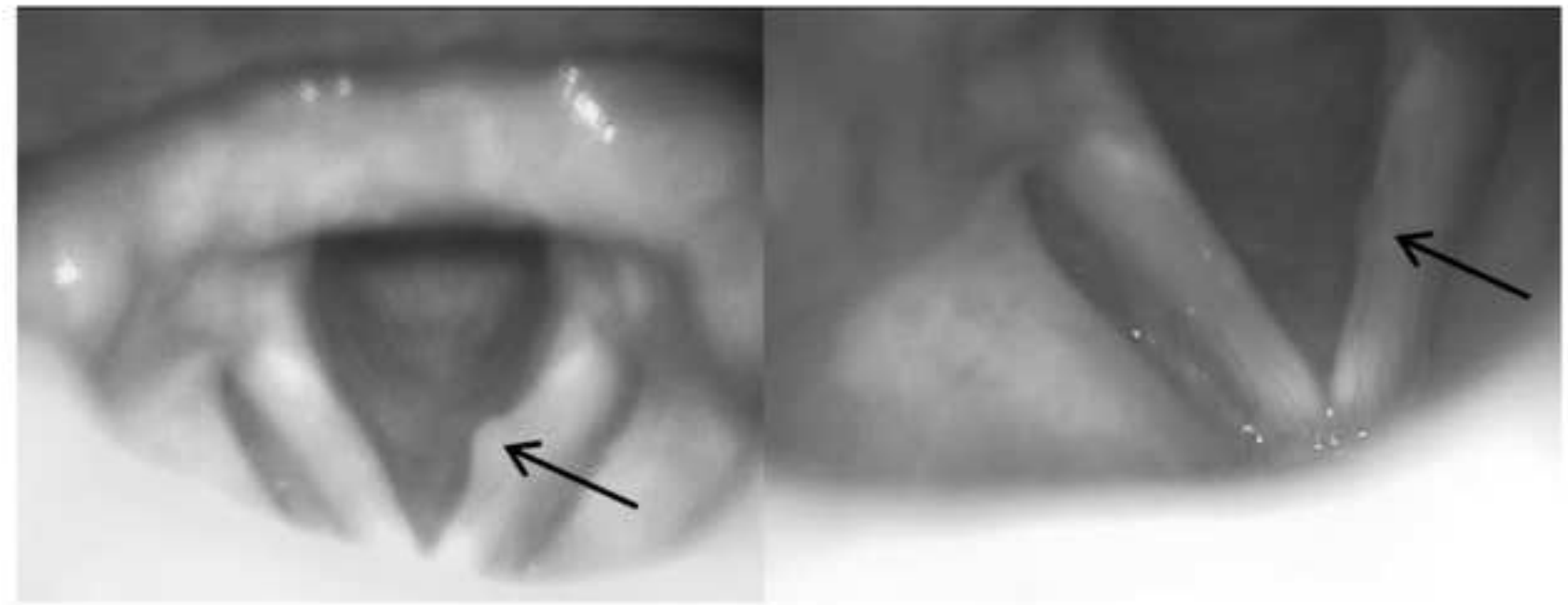


Figure-12-bottom-left

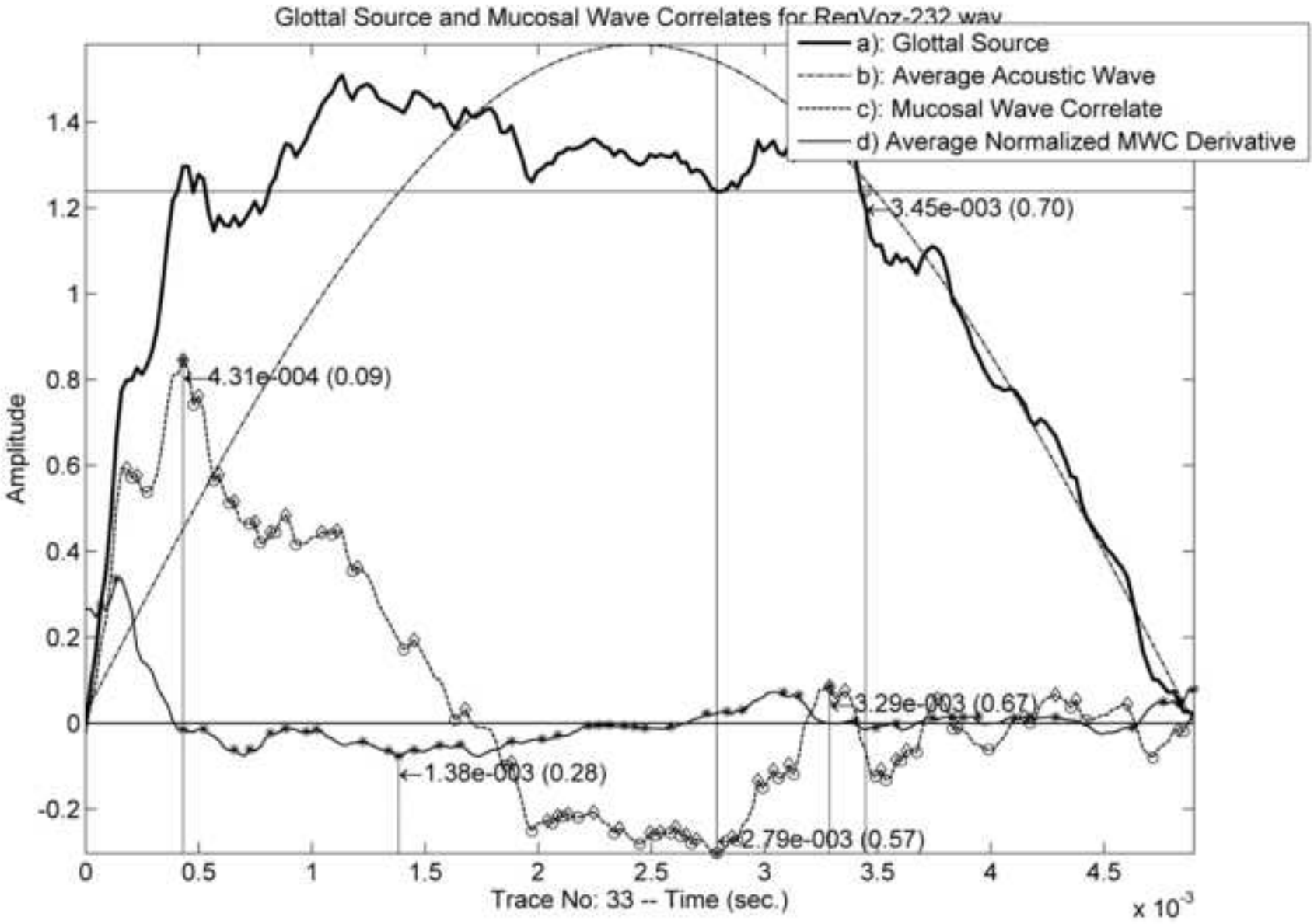


Figure-12-bottom-right

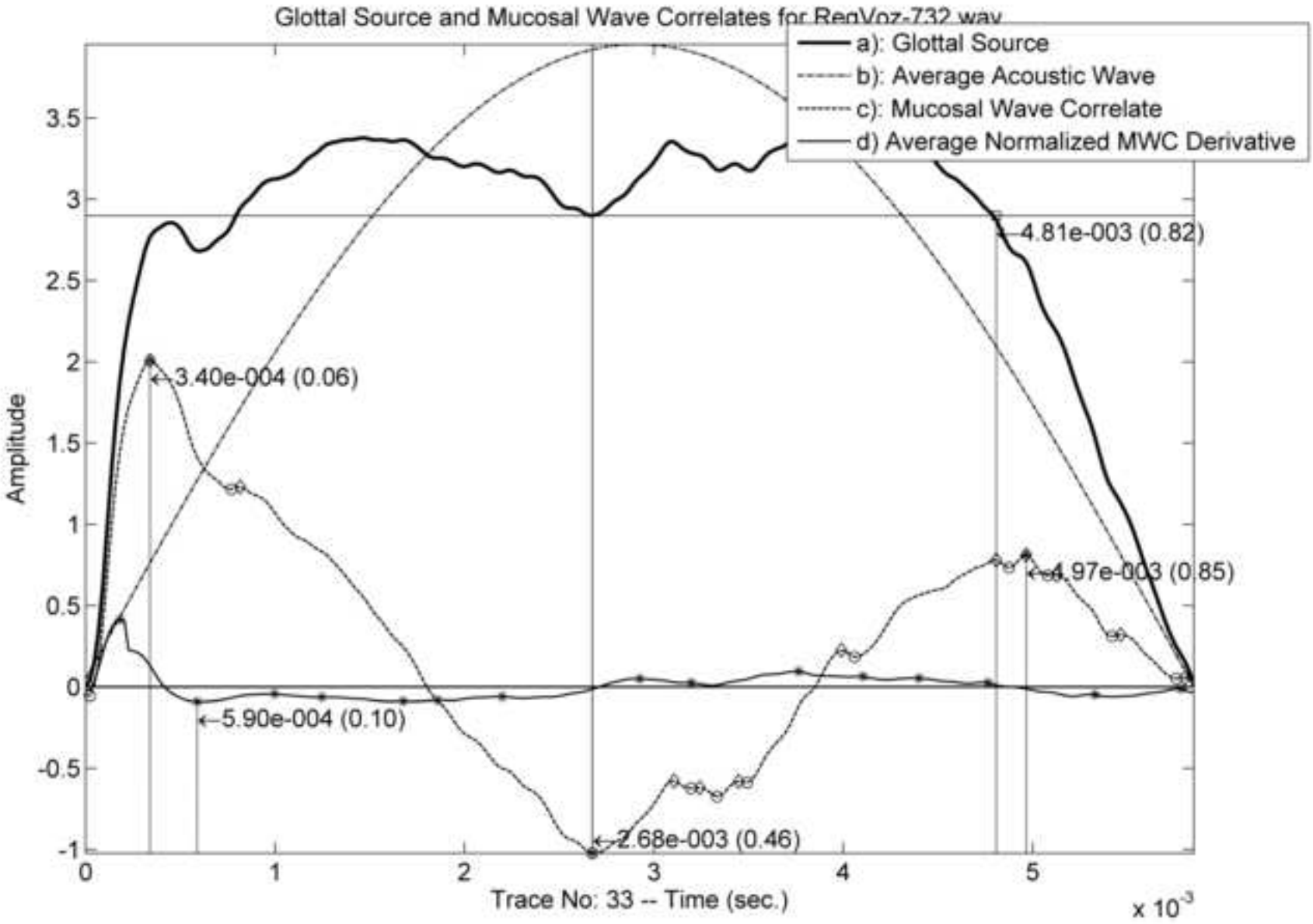


Figure-13-top

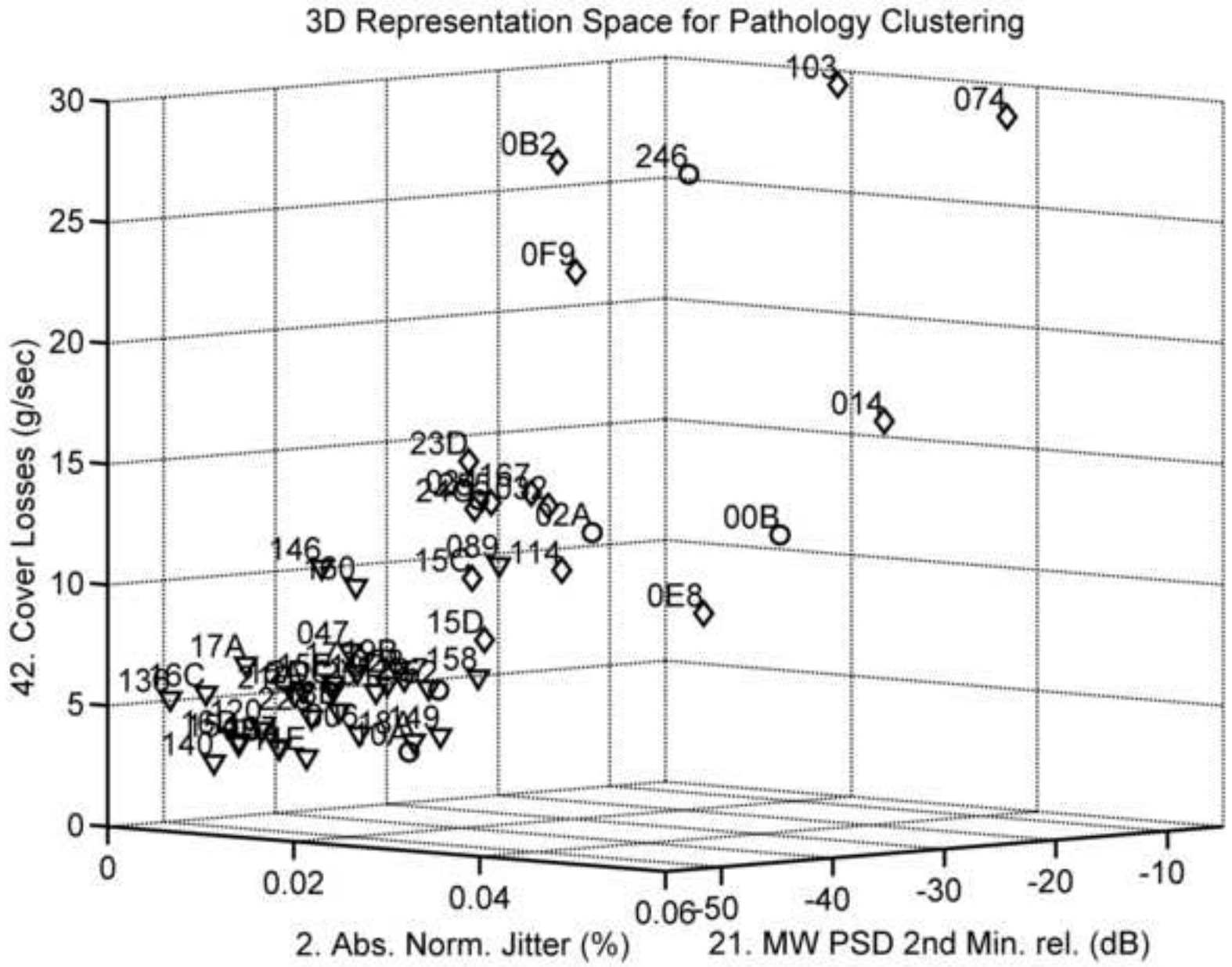


Figure-13-bottom

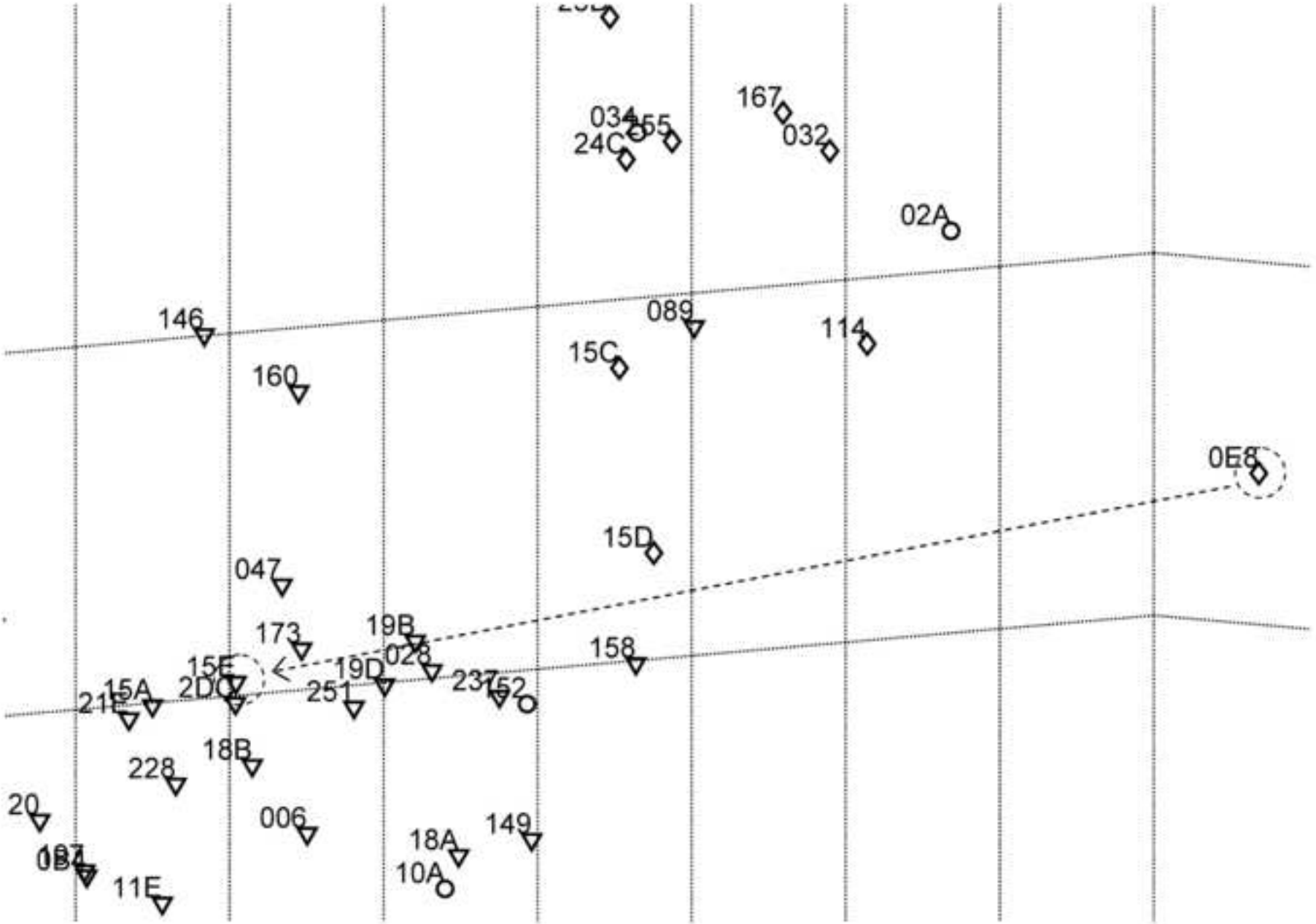


Figure-14-top

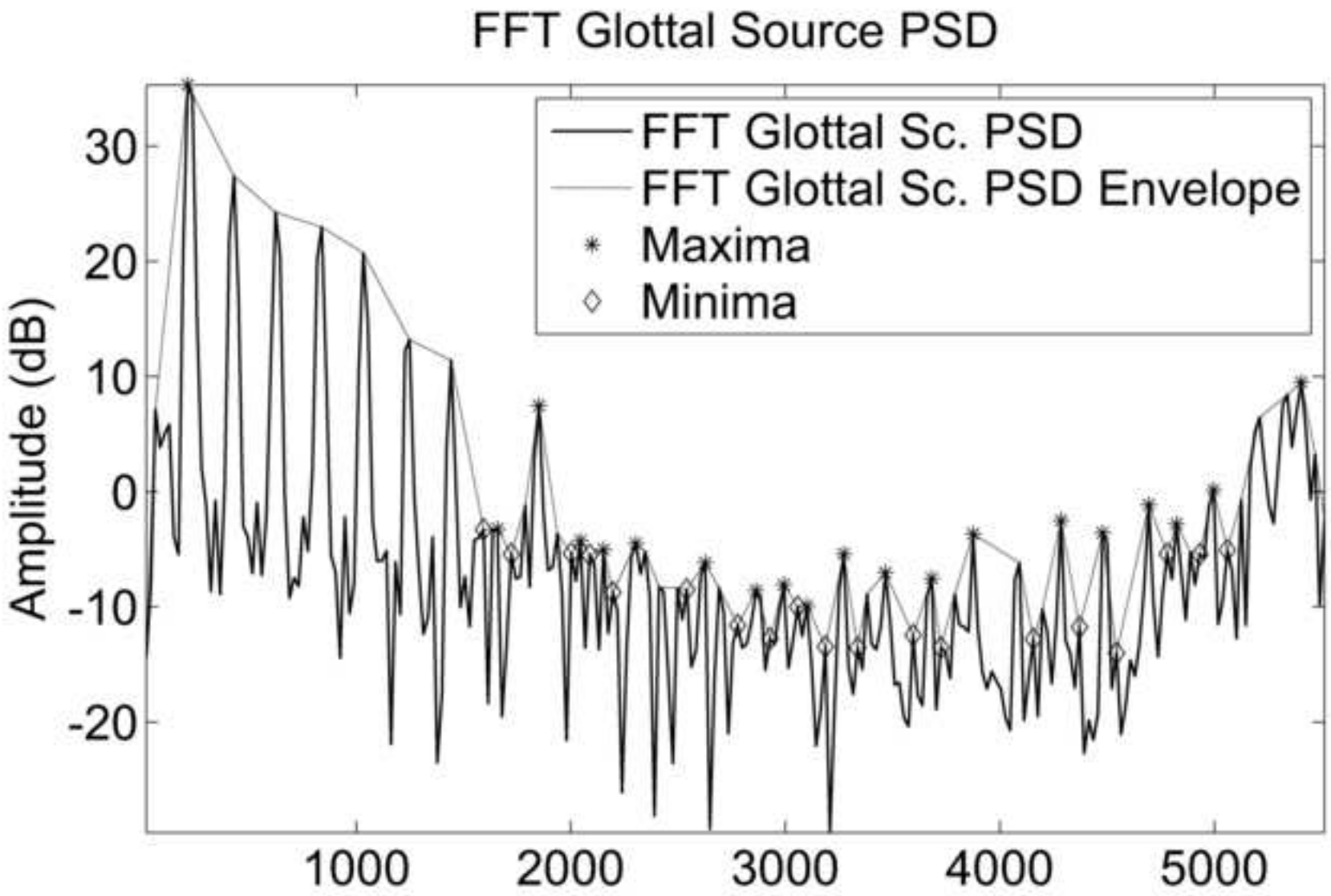


Figure-14-bottom