Using Dysphonic Voice to Characterize Speaker's Biometry

Pedro Gómez, Eugenia San Segundo, Luis M. Mazaira,
Agustín Álvarez & and Victoria Rodellar

Center for Biomedical Technology, Universidad Politécnica de Madrid &
Universidad Internacional Menéndez Pelayo (UIMP), Madrid, Spain

Abstract. *Phonation distortion leaves relevant marks in a speaker's biometric profile. Dysphonic voice production may be used for biometrical speaker characterization. In the present paper phonation features derived from the glottal source (GS) parameterization, after vocal tract inversion, is proposed for dysphonic voice characterization in Speaker Verification tasks. The glottal source derived parameters are matched in a forensic evaluation framework defining a distance-based metric specification. The phonation segments used in the study are derived from fillers, long vowels, and other phonation segments produced in spontaneous telephone conversations. Phonated segments from a telephonic database of 100 male Spanish native speakers are combined in a 10-fold cross-validation task to produce the set of quality measurements outlined in the paper. Shimmer, mucosal wave correlate, vocal fold cover biomechanical parameter unbalance and a subset of the GS cepstral profile produce accuracy rates as high as 99.57 for a wide threshold interval (62.08-75.04%). An Equal Error Rate of 0.64 % can be granted. The proposed metric framework is shown to behave more fairly than classical likelihood ratios in supporting the hypothesis of the defense vs that of the prosecution, thus offering a more reliable evaluation scoring. Possible applications are Speaker Verification and Dysphonic Voice Grading.*

Keywords: *Phonation, Speaker Recognition, Voice Production, Speech Processing.*

Resumo. *A distorção de fonação deixa marcas relevantes no perfil biométrico de um falante. A produção de voz disfônica pode ser usada como caracterização biométrica. Neste artigo, propõe-se a utilização de aspectos de fonação derivados da parametrização da fonte glótica (FG), após a inversão do trato vocal, para caracterização de voz disfônica em tarefas de verificação de locutor. Os parâmetros derivados da fonte glótical são combinados em um sistema de avaliação forense para definir uma especificação métrica baseada em distância. Os segmentos de fonação utilizados no estudo são derivados de elementos de preenchimento, vogais longas e outros segmentos de fonação produzidos em conversas telefônicas espontâneas. Segmentos de fonação de um banco de dados telefônicos de 100 falantes nativos espanhóis do sexo masculino são combinados em uma tarefa de validação*

cruzada por 10 vezes para produzir o conjunto de medições de qualidade descrito neste artigo. Shimmer, correlato de onda mucosa, desequilíbrio de parâmetro biomecânico de cobertura da prega vocal e um subconjunto dos perfis de cepstrais de FG produzem taxas de precisão de até 99,57 para um largo intervalo (62,08-75,04%). Uma Taxa de Erros Iguais de 0,64% pode ser concedida. Demonstra-se que a estrutura métrica proposta comporta-se de forma mais justa do que a clássica razão de verossimilhança para apoiar a hipótese da defesa vs a do promotor, oferecendo assim um escore de avaliação mais confiável. As aplicações possíveis são Verificação de Locutor e Graduação de Voz Disfônica.

Palavras-chave: *Fonação, Reconhecimento de Falante, Produção de Voz, Processamento de Fala.*

Introduction

Voice Pathology has been profoundly studied and characterized in the past decade (Dejonckere, 2010; Hakkesteegt *et al.*, 2010; Roy *et al.*, 2013). Most of the advances produced in the detection and grading of pathology can be applied in other fields such as forensic speaker recognition. In this article phonation features derived from the parameterization of the glottal source after the vocal tract inversion is proposed for dysphonic voice characterization in speaker verification tasks (Gomez-Vilda *et al.*, 2012), where the glottal source can be seen as a correlate of pressure build up in the glottis.

Phonation is the activity of voice production as a consequence of vocal fold vibration. It is present in speech, in voiced sounds, although speech is composed of both voiced and voiceless sounds, and the latter sounds are not based on phonation. Phonation must be seen as a biometrical mark of the person, similar to other behavior-based activities, such as gait, or writing. It presents several advantages with respect to speech as a study signal, in the sense that the vocal tract transfer function in speech is interfering with phonation biometry by introducing articulation features, which increment intra-speaker variability.

Phonation may be classified into the following overlapping groups:

- Normophonic, which is defined by the presence of a stable fundamental frequency in sustained vowels, stable intensity and long phonation capability, absence of roughness, absence of breathiness, and effortless voice production. Besides, it is characterized by clear and precise open and closed phases of the vocal folds, large Maximum Flow Declination Rate, and good extension of harmonic spectrum, extending over 5 KHz. The instrumental exploration of the larynx must not reveal organic or anatomical defects or lesions.
- Dysphonic, non organic, which is defined by the presence of perceptual acoustical features related to unstable or asymmetric phonation, such as presence of roughness, air in voice or strain, showing an irregular or too short vocal fold closed phase. The extension of the harmonic spectrum may not reach 4 KHz. Nevertheless the instrumental exploration of the larynx does not reveal organic defects or lesions, although anatomical defects may be present, as a certain degree of asymmetry.
- Pathologic, organic, which is defined by perceptual phonation defects affecting stability of fundamental frequency and intensity, shorter phonation capability, and roughness, air in voice, weak voice, and affected short harmonic spectrum,

usually not extending over 2 kHz. Instrumental exploration of the larynx will reveal specific defects or lesions, as nodules, polyps, cysts, edemae, granulomae, sulci, carcinomae, etc.

- Pathologic, neurological, which is defined by perceptual phonation defects as in the organic case, but in this group the instrumental inspection of the larynx will not reveal specific organic defects or lesions, although vocal folds will not show a regular vibration pattern, and many times vocal fold vibration asymmetry is present, affecting one of the vocal folds (unilateral paresis), or both vocal folds. Other forms of irregularity may affect the stability of phonation (spasmodic dysphonia). Frequently the etiology of the irregularity remains unclear.

The burning question is to what extent dysphonic voice may be present in a given speaker. In other words, to what extent normophonic voice is the norm in a sample of a general population. This extent is difficult to assess, and depends on how strict the specification for the term *normophonic* is established. Besides, the phonation capability of a speaker will vary strongly during a lifetime, progressively degrading with age to become a *presbyphonic* voice during the third age in most of the population, characterized by an increment in roughness, breathiness and asthenia, depicting a creaky phonation condition. It must be taken into account that many people suffer from a higher degree of phonation deterioration due to specific habits such as smoking, drinking or drug abuse, or to the consequence of larynx inflammatory processes (flu, cough, and other respiratory diseases), or simply from voice abuse (contact center professionals, actors, speakers, dealers, etc.). Thus, it may be said that phonation conditions are better during youth, and start to degrade with age. Therefore, it is really hard to establish the population percentage corresponding to each group.

It is very important to determine the characteristics of normophonic voice production, since even in that case, small irregularities may be expected in the main features mentioned, as stability in frequency and intensity, regular and symmetric fold vibration, perfect and complete open and closed phases, and timbre spectrum, making phonation a specific personal print. Even under perfect phonation conditions population differences exist, opening the possibility to use phonation features as biometrical marks.

The main phonation features resulting from biometrical differences are due to very specific physiological causes, and can be grouped into these two classes (Gómez *et al.*, 2013):

- Vocal fold vibration asymmetry
- Deficient glottal closure during the closed phase (contact phase)

The physiological reasons conditioning phonation features are summarized in Figure 1.

The template in Figure 1.a shows the vocal folds as two vertical bands united in the anterior side of the cricoid process (upper part of each sketch), separated in the posterior side (lower part of each sketch), leaving a space for the free flow of air to and from lungs. In Figure 1.b the vocal folds are shown together closing the glottis (contact phase), due to the action of the transversal and oblique laryngeal and crico-arytenoid muscles. The flow of air is stopped. In Figure 1.c the vocal folds are still united in the posterior part of the glottis under the action of the laryngeal muscles, but the pressure built up in the lungs has taken them apart (abduction), leaving a glottal space through with air can
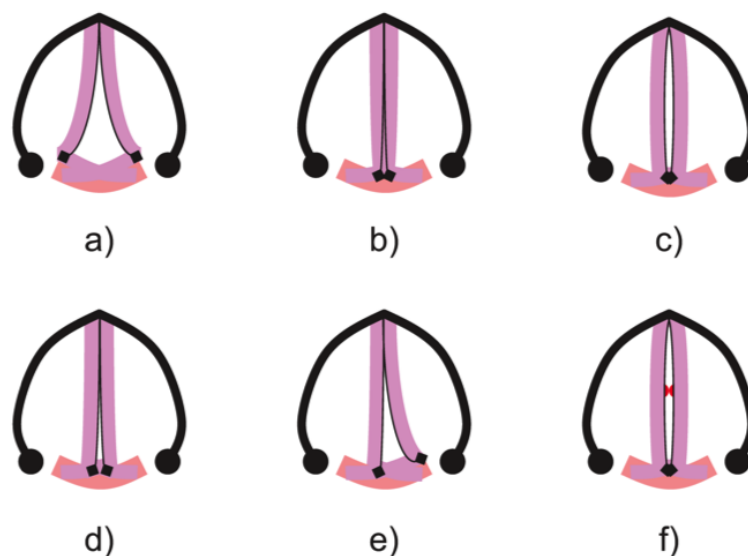
Figura 1. Vocal fold simplified situations: a) Open glottis in breathing; b) closed phase (contact phase) as part of the phonation cycle; c) open phase as part of the phonation cycle; d) deficient closure in the posterior third of the glottis, showing a permanent gap; e) asymmetric contact defect; f) deficient closure in the medial third, due to a bilateral lesion (nodules). Contact defects during the contact phase may be produced by other lesions (unilateral or bilateral). In all the plots the anterior part of the glottis is depicted upwards. (Figures produced by authors.)

flow from lungs to pharynx (open phase). The situations described in a), b) and c) are considered normal in the behavior of a healthy larynx. In the lower row some defects are described related with the contact phase. For instance, in Figure 1.d both vocal folds are not completely closed at the posterior side, therefore an air escape is to be expected. In Figure 1.e the incomplete closure is due to an asymmetry affecting mainly one of the vocal folds (unilateral paralysis). In Figure 1.f the contact is compromised by a bilateral lesion in the contact surface of the vocal folds, as in the case of nodules, for instance. The closure is not perfect and an escape of air is to be expected. Pictures of these contact defects from actual endoscopic recordings taken during the contact phase are presented in Figure 2.

The situations described in d), e) and f) produce observable correlates in the air flow and pressure build up in larynx, and propagate to the signal recorded by a microphone as phonated speech. Therefore, the contact defects will leave a biometrical mark in the phonation of a speaker if any of these defects is present to a greater or lesser extent. The behavior of the biometrical mark may be inferred from Fant's source-filter model illustrated in Figure 3 (Fant, 1997).

Voiced speech (phonation) is produced by a glottal excitation model, resulting from vocal fold vibration. The pressure build up in the vocal folds (glottal source) propagates through the vocal tract (or more properly, the oro-naso-pharyngeal tract) to reach the mouth or nostrils (depending on nasalization) to be radiated as a signal $S_r(n)$ reaching a microphone or other recording device. Voiceless speech is produced by frictional air turbulence (turbulent source) resulting from fast airflow in specific parts of the vocal tract (vocal folds, pharynx, tongue, teeth, lips…). Either glottal source, or turbulent

Figura 2. Pictures illustrating contact defects: Left picture: deficient closure in the posterior third of the glottis as a result of bilateral nodules. Middle picture: Unilateral contact defect due to a right vocal fold Reinke's edema. Right picture: bilateral contact defect in an hourglass pattern showing anterior and posterior gaps. Anterior section of larynx upwards (Photos provided by the ENT Services of Hospital Universitario Gregorio Marañón of Madrid.)
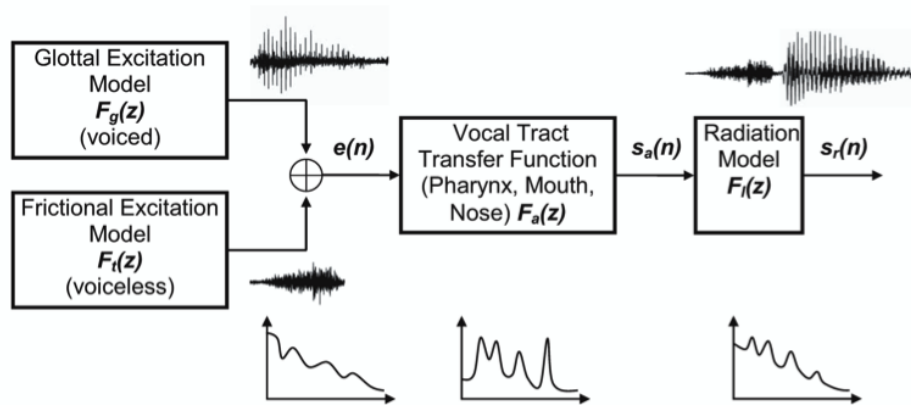
Figura 3. Fant's source-filter model to explain speech production. (Figure produced by authors)

flow, or both, will be the cause of the speech signal radiated. The resulting spectrum of the radiated signal (Figure 3, low row, right) will be the consequence of the application of the vocal tract transfer function (Figure 3, low row, middle) on the source spectrum (Figure 3, low row, left). Fant's model inspires the methodology to reconstruct the glottal source from phonated speech. The methodology consists in removing the influence of the radiation model and the vocal tract transfer function by inverse filtering by different methods. The one used in the present study is described in Gómez-Vilda *et al.* (2009), and is summarized in Figure 4.

The speech signal s(n) is first processed (1) to eliminate the influence of radiation and other undesirable effects due to channel characteristics. The radiation-compensated signal $s_l(n)$ is filtered by a lattice-ladder mirror filter (2) which is designed to remove partially the influence of a hypothesized glottal source, generating a signal $s_v i(n)$ which is mainly characterized by the vocal tract. This signal is modeled (4) to obtain the inverse signature of the vocal tract, which will be applied to the radiation-compensated signal $s_l(n)$ to remove the influence of the vocal tract (5). The resulting signal $s_r i(n)$ will be
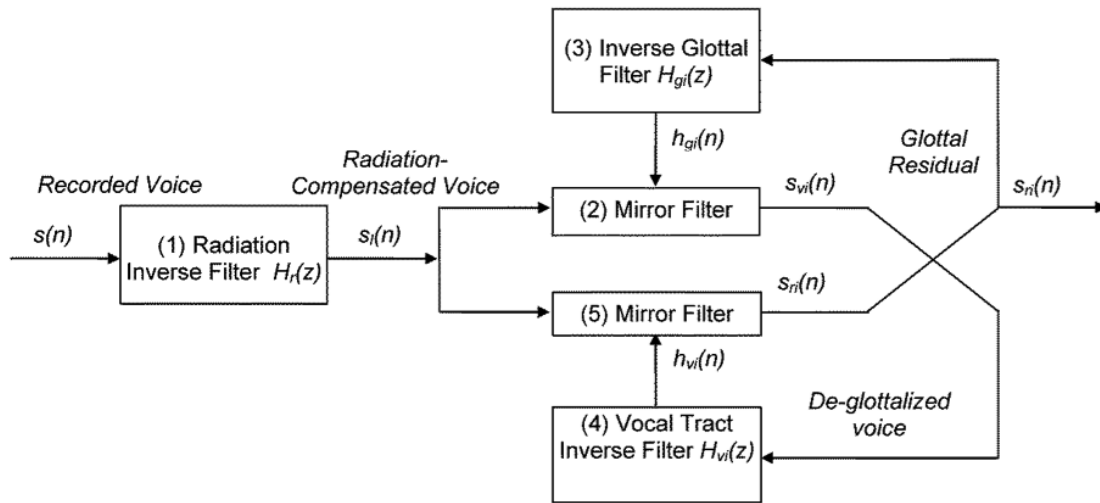
Figura 4. Methodology for the reconstruction of the glottal source from segments of phonated speech by recursive inverse adaptive filtering. (Figure produced by authors.)

dominated by the glottal features, and may be modeled (3) to produce a better inverse estimate of the glottal features, and injected in (2) to produce also a better estimate of $s_v i(n)$. The recursion is iterated a low number of times, and the glottal residual $s_r i(n)$ will be used to produce the glottal source by numerical integration. An example of the glottal source reconstruction is shown in Figure 5.

It may be seen that the reconstructed glottal residual $s_r i(n)$ in Figure 5.b is the result of removing vocal tract resonances found in the original speech signal s(n). In particular the presence of the first resonance (formant) may be seen as a ringing (successive oscillations) taking place during each of the 17 pseudo-periodical glottal cycles extending over slightly more than 180 ms in Figure 5.a. The residual $s_r i(n)$ is numerically integrated to produce the glottal source in Figure 5.c, which shows the main features of the pressure build up in the glottis. The main feature as far as the harmonic spectral contents of speech are concerned, is the maximum flow declination rate (MFDR), which is the negative drop of pressure signaled by red asterisks due to the closing phase. The glottal source is restored to its quiescent value (0) following a recovery pattern to reach a plateau, marking the duration of the contact phase. During the open phase, a pressure increment can be appreciated to reach a maximum, after which a sharp drop to reach the MFDR may be appreciated (closing phase). Finally in Figure 5.d a series of patterns showing the successive glottal flow cycles may be seen.

Once the glottal source has been reconstructed it is being parameterized according to different techniques in the time as well as in the frequency domain. The parameters are evaluated for each of the phonation cycles in the speech segment being analyzed (typically between 50 and 200 ms long). For male voice, between 5-20 glottal cycles are to be found in such an interval. Cycle-synchronous estimations of each parameter are stored in an array, average values and standard deviations are also evaluated. In what follows a brief description of these techniques and the resulting parameters is given:
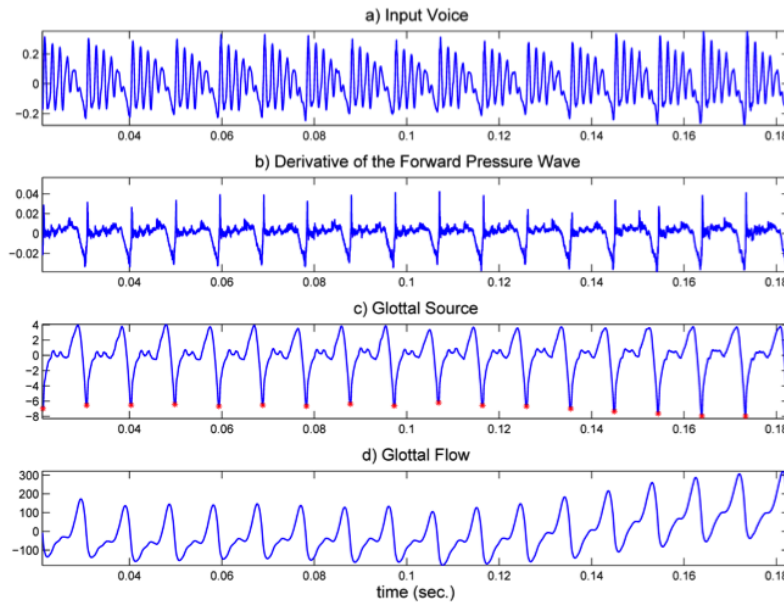
Figura 5. Example of glottal source and flow reconstruction from phonated speech: a) original speech signal (s); b) glottal residual $s_r i$(n), or derivative of the forward pressure wave; c) reconstructed glottal source (correlate of the pressure build up in the glottis); d) reconstructed glottal flow. (Figure produced by authors.)

• Perturbation parameters. These are a group of time-domain parameters related with voice quality, as the fundamental frequency f0, the jitter (relative fluctuations of the glottal source period), the shimmer (relative fluctuations of the glottal source amplitude for each glottal cycle), the absolute minimum sharpness (value of the MFDR), the noise to harmonic energy contents (HNR), or the ratio between the higher glottal source components to the first-order glottal source component (MAE). These parameters are given in Table 1.

Perturbation parameters
1. Absolute Pitch
2. Abs. Norm. Jitter
3. Abs. Norm. Ar. Shimmer
4. Abs. Norm. Min. Sharp (MFDR)
5. Noise-Harm. Ratio (NHR)
6. Muc./AvAc. Energy (MAE)

Tabela 1. Perturbation parameters.

• Cepstral parameters. This group consists in a collection of 14 parameters directly estimated from the cepstral description of the glottal source. The estimation process consists in generating the Fourier power spectrum of the glottal source. The cosine transform is applied to the logarithm of this spectrum and the first 14 resulting parameters are selected. Some of these parameters are extremely sensitive to certain factors such as gender or age (Muñoz, 2014). The parameters are listed in Table 2.

Cepstral Parameters

7. MWC Cepstral 1
8. MWC Cepstral 2
9. MWC Cepstral 3
10. MWC Cepstral 4
11. MWC Cepstral 5
12. MWC Cepstral 6
13. MWC Cepstral 7
14. MWC Cepstral 8
15. MWC Cepstral 9
16. MWC Cepstral 10
17. MWC Cepstral 11
18. MWC Cepstral 12
19. MWC Cepstral 13
20. MWC Cepstral 14

Tabela 2. Cepstral parameters.

- Spectral parameters. The spectral profile of the glottal source is conditioned by the biomechanical behavior of the vocal folds, especially the visco-elastic link between the fold body (*musculus vocalis*) and the epithelial cover and conjunctive tissues in Reinke's space. The envelope of the harmonic spectrum of the glottal source shows peaks and valleys which are influenced by this biomechanical behaviour. Anomalous relations among these peaks and valleys may serve as biometrical markers. The first group of parameters given in Table 3 are amplitude estimates of the peaks and valleys (21-27). The second group give their relative positions in frequency (28-32). Parameters 33 and 34 give the depth of the two first valleys relative to their frequency span (slenderness).

Spectral Parameters

21. MW PSD 1st Max. ABS.
22. MW PSD 1st Min. rel.
23. MW PSD 2nd Max. rel.
24. MW PSD 2nd Min. rel.
25. MW PSD 3rd Max. rel.
26. MW PSD End Val. rel.
27. MW PSD 1st Max. Pos. ABS.
28. MW PSD 1st Min. Pos. rel.
29. MW PSD 2nd Max. Pos. rel.
30. MW PSD 2nd Min. Pos. rel.
31. MW PSD 3rd Max. Pos. rel.
32. MW PSD End Val. Pos. rel.
33. MW PSD 1st Min NSF
34. MW PSD 2nd Min NSF

Tabela 3. Spectral parameters.

- Biomechanical parameters. The spectral behavior of the glottal source is directly related to the distribution of mass and visco-elasticity of the vocal fold body and cover. A methodology to estimate the distribution of mass and stiffness of each structure is possible using spectral matching techniques (Gómez-Vilda *et al.*, 2007). The most significant estimates are the mass and stiffness of the vocal fold body and cover, the ratio of energy losses due to viscid and turbulent flow behavior, and their respective unbalances. These are estimated using relative comparisons of mass, stiffness and losses from neighbor glottal cycles. The list of estimated parameters is given in Table 4.

Biomechanical Parameters
35. Body Mass
36. Body Losses
37. Body Stiffness
38. Body Mass Unbalance
39. Body Losses Unbalance
40. Body Stiffness Unbalance
41. Cover Mass
42. Cover Losses
43. Cover Stiffness
44. Cover Mass Unbalance
45. Cover Losses Unbalance
46. Cover Stiffness Unbalance

Tabela 4. Biomechanical parameters.

- Temporal parameters. The glottal cycle is divided into a closed phase and an open phase. The time instants associated with the start of the closed and open phase, as well as the time required to reach the quiescent pressure (recovery) and the maximum amplitude of the glottal source relative to the MFDR are estimated as important parameters in the time domain. Due to irregularities in the glottal source time profile, the recovery and open instants are estimated twice to produce more robust results. The open and closed instants, as well as the start of the closing phase are also estimated on the flow signal. The list of temporal parameters is given in Table 5.

- Glottal gap parameters. This set of parameters is designed to evaluate the contact defects, directly on the flow, calculating the ratio of air escape during the contact phase relative to the air escape during the open phase (59), or on the glottal source, in which case the defects are differentiated as contact, adduction or permanent ones, depending to which phase of the glottal cycle they affect. The list of the parameters is given in Table 6.

- Tremor parameters. The stiffness of the vocal fold body (*musculus vocalis*) is directly influenced by the neuromotor action of the laryngeal muscles, therefore, many neurological pathologies may be characterized from the estimates of this stiffness (parameter 37). Hypo-tonic or hyper-tonic deviations of this parameter are important correlates in Parkinson's Disease, for instance, as well as tremor.

Temporal Parameters

47. Rel. Recov. 1 Time
48. Rel. Recov. 2 Time
49. Rel. Open 1 Time
50. Rel. Open 2 Time
51. Rel. Max. Ampl. Time
52. Rel. Recov. 1 Ampl.
53. Rel. Recov. 2 Ampl.
54. Rel. Open 1 Ampl.
55. Rel. Open 2 Ampl.
56. Rel. Stop Flow Time
57. Rel. Start Flow Time
58. Rel. Closing Time

Tabela 5. Temporal parameters.

Glottal GAP Parameters

59. Val. Flow GAP
60. Val. Contact GAP
61. Val. Adduction GAP
62. Val. Permanent GAP

Tabela 6. Glottal GAP parameters.

A set of six parameters is devoted to track this disease. The first three give a description of the tremor in terms of its autoregressive modeling (63-65). The last ones give the tremor frequency in cycles/s (66), the reliability of this estimate (67), or the tremor amplitude in root mean square relative to the vocal fold body average amplitude (68). The list of tremor parameters is given in Table 7.

Tremor Parameters

63. 1st. Order Cyc. Coeff.
64. 2nd. Order Cyc. Coeff.
65. 3rd. Order Cyc. Coeff.
66. Tremor Frequency
67. Estimation Reliability
68. Tremor rMS Amplitude

Tabela 7. Tremor parameters.

The interested reader can find a more detailed description of each parameter meaning and distribution in Gómez *et al.* (2013).

Materials and methods

The purpose of the present research was to describe a methodology to parameterize the glottal source in terms of dysphonic voice and to study how to apply these parameters in speaker verification tasks. For this purpose a database of GSM-quality recordings from telephone conversations by 100 male speakers was used. Speech was recorded

at 8 KHz 16 bits and mu-law. Each conversation lasted between 5 and 30 min., fillers and long vowels were extracted from them. These long vowels were samples of vowels [a], [æ], [ɛ] and [e]. For classification purposes, the first two groups were labelled as /a/, whilst the last two groups were labelled as /e/. This last group covers most of the fillers which may be found in Spanish, consisting in lengthening of words as "de" or "que", or spontaneous insertions of /e/. An average of 6-8 of these fillers may be found in recordings of hesitating statements along a duration of 1-2 minutes. Fillers and long vowels were segmented as 100 ms fragments, and 68 parameters were obtained from each glottal cycle in the fragment. The resulting feature database is a matrix referred to as $\mathbf{Z}_t$.

Three experiments are described in this paper, the first oriented to provide full compatibility of parameter distributions of phonations from /a/ against phonations from /e/. This experiment is described in this section. The second experiment is designed to select a database of normative speakers from telephone quality recordings based in /e/ by contrasting the available telephone recordings with a normative database from high quality recordings. The selected normative speakers will be used as a control group in future work. The third experiment is designed to match telephone-quality /e/ recordings from the normative speakers against themselves to test the forensic matching capability of the methodology and to produce sensitivity and specificity estimates for the matching protocol.

The first experiment consisted in confronting the distributions of each parameter in $\mathbf{Z}_t=[\mathbf{Z}_t a\ \mathbf{Z}_t e]$ from the /a/-group $\mathbf{Z}_t a$ and the /e/-group $\mathbf{Z}_t e$ to check their degree of equivalence. The null hypothesis consisted in assuming the equivalence of distributions. The histograms for the fundamental frequency f0, jitter, shimmer, body mass and stiffness, and cover mass and stiffness are given in Figures 6 to 9.
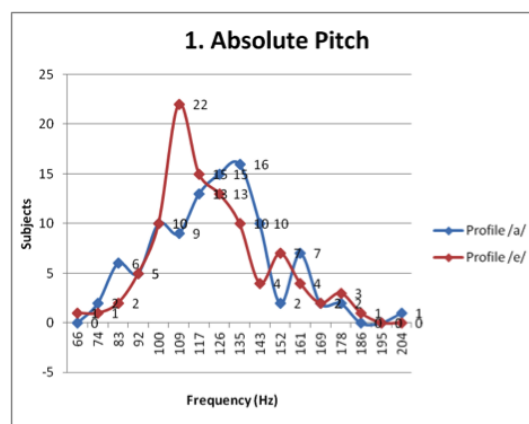
Figura 6. Comparison of the histograms of f0 from the /a/-group vs the /e/-group: The null hypothesis cannot be rejected given the distribution overlap (Figure produced by authors.)

The second experiment consisted in dividing the speakers in the database $\mathbf{Z}_t a$ in two subsets of 50 speakers each ($\mathbf{Z}_t an$ and $\mathbf{Z}_t ad$) according to the degree of dysphonia present in their phonations confronting the whole speaker set with a normative set of 50 normophonic speakers selected and inspected at the ear, neck and throat service of Hospital Gregorio Marañón in Madrid. Normophonic speakers were inspected by video-
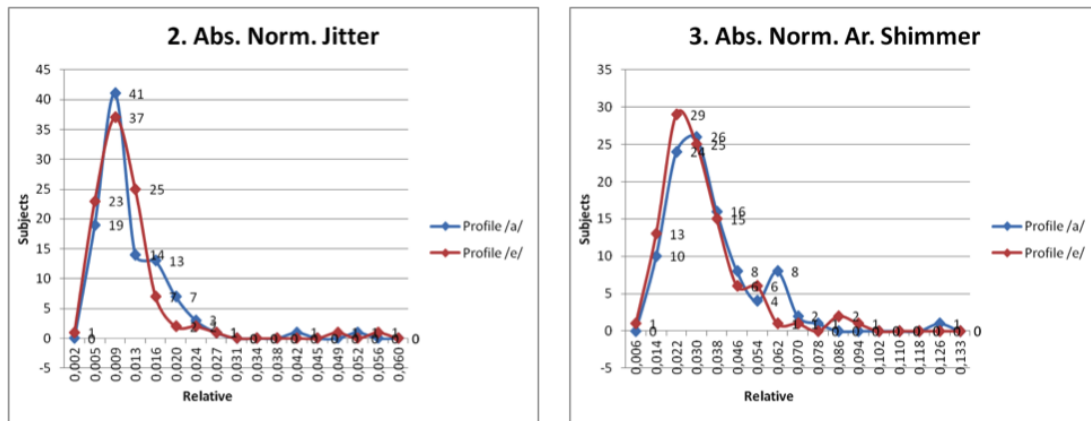
Figura 7. Comparison of the histograms of jitter and shimmer from the /a/-group vs the /e/-group: The null hypothesis cannot be rejected given the distributions overlap. (Figure produced by authors.)
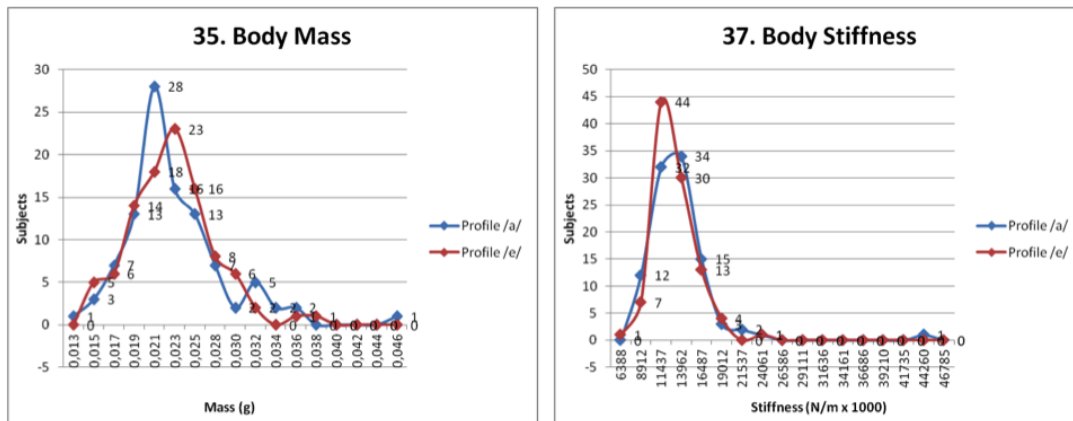

Figura 8. Comparison of the histograms of body mass and stiffness from the /a/-group vs the /e/-group: The null hypothesis cannot be rejected given the distributions overlap. (Figure produced by authors.)
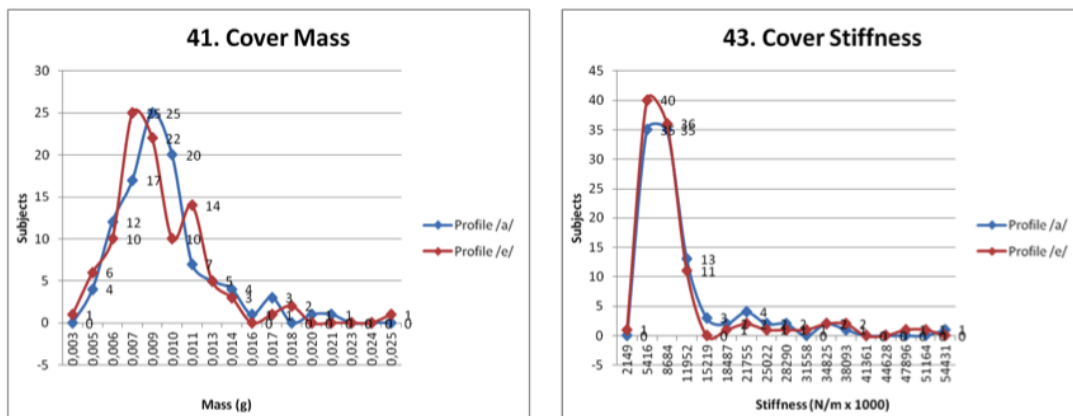

Figura 9. Comparison of the histograms of cover mass and stiffness from the /a/-group vs the /e/-group: The null hypothesis cannot be rejected given the distributions overlap. (Figure produced by authors.)

endoscopy to discard any organic problem in their larynx, and their non-dysphonic condition was assessed by the GRBAS test (Hirano, 1981). Fragments of phonations of vowel /a/ lasting 200 ms from the normative set of speakers taken at 44100 Hz and 16 bits were parameterized and used as a normative model ($\mathbf{Z}_m an$) in the task of grading the /a/-group from GSM-quality recordings.

The third experiment was to match the features from each speaker in the subset of 50 normophonic males of the /e/-grou and telephone quality ($\mathbf{Z}_t en$) against his own feature set as target, and against the other 49 as imposters using the matching methodology to be described in what follows. The fillers from each speaker used in the matching as questioned tokens and the target set used as suspects' set were generated from two different recording sessions.

For the second experiment the membership of each speaker to the normophonic or dysphonic group was assessed using the log likelihood ratio between the conditioned probability of membership of a specific speaker s_i with feature set $\mathbf{z}_t ai$ relative to the normative Gaussian mixture model (GMM) defined as $\Gamma_{man} = \mathbf{w}_{man}, \mu_{man}, C_{man}$ built on the normative feature dataset $\mathbf{Z}_{man}$, where $\mathbf{w}_{man}$, μ_{man} and $\mathbf{C}_{man}$ are the mixture weights, the average vector and the covariance matrix of the dataset. The definition of the normophonic membership log likelihood may be estimated as:

$$\lambda_{tai} = \log\left\{\Pr(\mathbf{z_{tai}} \mid \mathbf{\Gamma_{man}})\right\} - \log\left\{1 - \Pr(\mathbf{z_{tai}} \mid \mathbf{\Gamma_{man}})\right\} \tag{1}$$

where the conditioned membership probability will be given as:

$$\Pr(\mathbf{z_{tai}} \mid \mathbf{\Gamma_{man}}) = \sum_k w_k^{man} \frac{1}{(2\pi)^{m_k/2} \left|\mathbf{C}_k^{man}\right|^{m_k}} e^{-1/2\left(\mu_{tai} - \mu_k^{man}\right)^T \left[\mathbf{C}_k^{man}\right]^{-1}\left(\mu_{tai} - \mu_k^{man}\right)} \tag{2}$$

where k is the order of the GMM, and mk is the size of each Gaussian cluster.

In turn, the speaker matching methodology used in the third experiment was designed to estimate to which extent acoustic evidence from speaker s_i ($\mathbf{z}_{tei}$, considered the questioned evidence) against acoustic evidence from speaker s_j (Γ_{tej}, built on the suspect's evidence $\mathbf{z}_{tej}$) can modify the degree of conviction (gain of belief) in favour or against the suspect in relation with the case. This gain of belief is formulated as a log likelihood between the conditioned probability of $\mathbf{z}_{tei}$ being produced by the GMM model Γ_{tej} relative to the conditioned probability of $\mathbf{z}_{tei}$ being produced by any foil speaker from a line-up set characterized by the GMM model Γ_{ten}. This log likelihood ratio is a rephrasing of the balanced reasons method established by C. S. Peirce (1878), formulated as the conditioned probability of the prosecutor's hypothesis vs the defender's hypothesis (see Taroni *et al.* 2006; Gomez-Vilda *et al.* 2012:

$$V_{pt} = L_{pd} \times V_{pr} = \frac{\Pr(E \mid H_p, I)}{\Pr(E \mid H_d, I)} \times V_{pr} \tag{3}$$

where E is the evidence (questioned), H_p is the prosecutor's hypothesis (questioned evidence being produced by the suspect), and H_d is the defender's hypothesis (questioned

evidence being produced by any other speaker). In this way the a priori probability V_{pr} in favour of H_p will be amplified or attenuated by the gain of belief L_{pd} (likelihood ratio) to produce the a posteriori probability V_{pt}. The log likelihood ratio may be estimated as:

$$\lambda_{pd} = \lambda_{teij} = \log\left\{Pr(\mathbf{z_{tei}} \mid \mathbf{\Gamma_{tej}})\right\} - \log\left\{Pr(\mathbf{z_{tei}} \mid \mathbf{\Gamma_{ten}})\right\} \tag{4}$$

and the conditioned probabilities evaluating the prosecutor's and defender's hypotheses are given as:

$$Pr(\mathbf{z_{tei}} \mid \mathbf{\Gamma_{tej}}) = \sum_k w_k^{tej} \frac{1}{(2\pi)^{m_k/2} \left|\mathbf{C_k^{tej}}\right|^{m_k}} e^{-1/2\left(\mu_{tai}-\mu_k^{tej}\right)^{\mathrm{T}}\left[\mathbf{C_k^{tej}}\right]^{-1}\left(\mu_{tai}-\mu_k^{tej}\right)} \tag{5}$$

$$Pr(\mathbf{z_{tei}} \mid \mathbf{\Gamma_{ten}}) = \sum_k w_k^{ten} \frac{1}{(2\pi)^{m_k/2} \left|\mathbf{C_k^{ten}}\right|^{m_k}} e^{-1/2\left(\mu_{tai}-\mu_k^{ten}\right)^{\mathrm{T}}\left[\mathbf{C_k^{ten}}\right]^{-1}\left(\mu_{tai}-\mu_k^{ten}\right)} \tag{6}$$

It must be noted that in the third experiment the questioned and the suspect evidence were derived from individual speakers in the /e/-group normative feature set $\mathbf{Z}_{ten}$, whereas the line-up feature set was generated using the whole feature set $\mathbf{Z}_{ten}$. The results of the second and third experiments will be commented on in the section entitled "Validation and Sample Matching Results".

Another relevant aspect has to do with the selection of the parameters considered most relevant for dysphonia assessment or speaker matching. This procedure will be a premise to be incorporated into any of these procedures prior to the conditional probability estimation. The feature selection carried out was based on the evaluation of Fisher's discriminant ratios (Kim *et al.*, 2005), defined as:

$$C_{Fi} = \frac{\mu_{ki} - \mu_{kj}}{\sqrt{\dfrac{\varsigma_{ki}^2}{n_i} + \dfrac{\varsigma_{kj}^2}{n_j}}} \tag{7}$$

where μ_{ki} and μ_{kj} are the sample averages of subsets i and j for parameter k, ζ_{ki} and ζ_{kj} are the sample standard errors of subsets i and j, also for parameter k, and n_i and n_j are the respective subset sample sizes. To select the most relevant features a comparison of subset distributions is carried out, and only the most relevant features are included in the posterior analysis. An example is given in Figure 10.

Finally the issue of speaker match metrics is to be addressed. When estimating log likelihood ratios following (4), (5) and (6), if feature datasets can be grouped in a low number of clusters, log likelihood ratios can be expressed in terms of normalized distances among the questioned (test), suspect (control) and line-up (model) centroids, as shown in Figure 11.
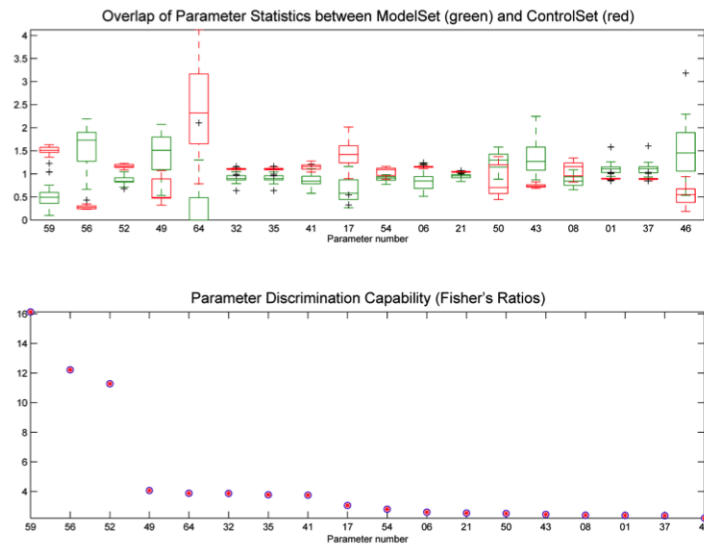
Figura 10. Parameter selection based on Fisher discriminant analysis: Upper template: boxplots of the most relevant parameters comparing the normophonic feature subset (green) against the dysphonic one (red). It may be seen that most of the distributions show low overlap, and small extent (being the conditions to produce a large Fisher's ratio as given by (7)). Lower template: Values of Fisher's ratios. (Figure produced by authors.)
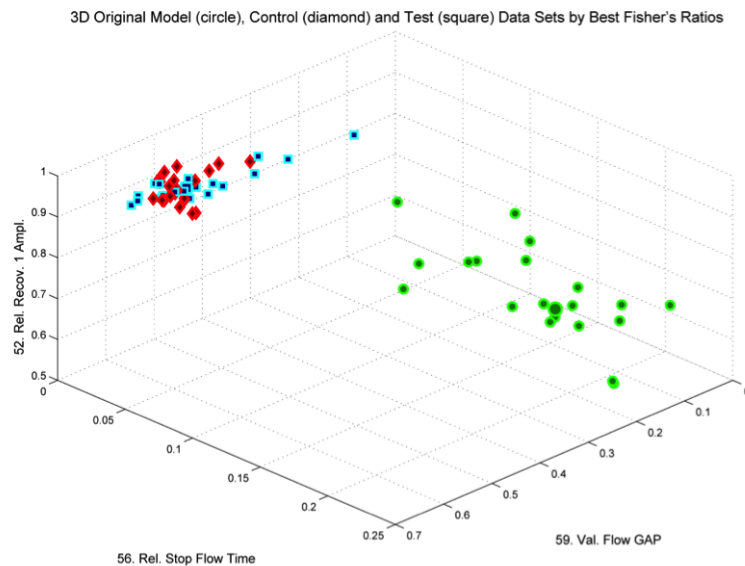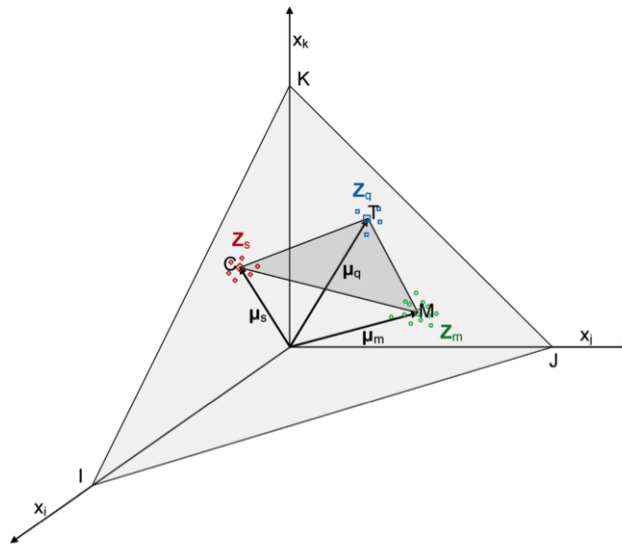

Figura 11. 3D description of evidence matching from a practical case in terms of the three most relevant features derived from Fisher's analysis in Figure 10: The questioned evidence is grouped as the test subset (blue squares). The suspect's evidence is grouped as the control subset (red diamonds). The line-up data is grouped as the model subset (green circles). Each subset centroid is signaled by a larger circle, diamond or square. A simple visual inspection allows inferring that the clusters of questioned and suspect evidence are much closer between themselves than to the line-up cluster. (Figure produced by authors.)

The 3D plot may be seen as the expression of the projection from an 18-dimensional vector space defined in terms of the 18 selected features to a 3-dimensional subspace in terms of the 3 most relevant ones. Reducing clusters to centroids allows defining the log likelihood as a normalized distance balance given by:

$$\lambda_{pd} = \frac{D_{TM}^2 - D_{TC}^2}{2} \qquad (8)$$

where $D_T M$ is the normalized distance between the centroids of the questioned evidence set to the model set, and $D_T C$ is the distance between the centroids of the questioned and suspect evidence. The centroids of the three sample sets (test, control and model) define the match triangle CTM as depicted in Figure 12.


Figura 12. Match triangle defined by the test, control and model centroids on the 2D plane projection of a 3D description of evidence matching in similar terms to the one given in Figure 11. (Figure produced by authors.)

It may be seen that the centroids of clusters T (questioned), C (suspect) and M (line-up) define a plane intersecting the three feature axes x_i, x_j and x_k at the points I, J and K. This property allows summarizing the matching results in a balanced chart as the one given in Figure 13.

The Mahalanobis normalized distances between each two centroids C, T and M defining the match triangle, MDTC, MDTM and MDCM as seen in Figure 13 can be used to establish the relationship between questioned, suspect and model evidence. It is clear that the vertical axis in the figure is the place of all possible solutions which share the condition of $D_{TC}=D_{TM}$, for which the log likelihood will be null ($\lambda_{pd}=0$: neutral decision). The right hand plane defined by the vertical axis will define the place of all possible solutions where $D_{TC}>D_{TM}$, therefore the log likelihood will be negative ($\lambda_{pd}<0$: decision favoring the defender's hypothesis). The left hand plane defined by $D_{TC}<D_{TM}$ will correspond to positive log likelihood ratios ($\lambda_{pd}<0$: decision favoring the prosecutor's hypothesis). Nevertheless, the decision cannot be based on just crossing the vertical line
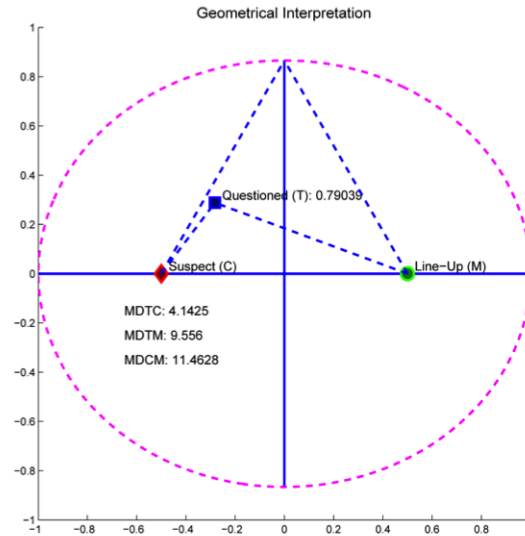
Figura 13. Balanced chart summarizing the match between questioned and suspect evidence relative to the line-up model. (Figure produced by authors.)

to accept the prosecutor's hypothesis, as this threshold would be unfair with respect to the guarantees due to the legal defense of the suspect. A more conservative threshold decision should be used. Accordingly with Daubert rules (see U.S. Supreme Court, 1993) accepting and evaluating the strength of evidence should be left to the Court. However, it is generally accepted by the European Network of Forensic Science Institutions that this kind of scale would be of useful application to grade the strength of the evidence to help the decision of the Court. A reasonable scale can be found in Lucy (2005), and is reproduced in Table 8.

There is another important detail regarding expression (8) and Figure 13, which concerns situations where $D_{TC} >> D_{CM}$ and $D_{TM} >> D_{CM}$. This ill-conditioned case happens when questioned and suspect evidence are far apart from the line-up data, and would indicate a bad selection of the line-up. In this unfair situation accepting results in the left hand side ($D_{TC} < D_{TM}$) would break the guarantee of a fair evaluation, helping to produce a decision in favor of the prosecutor's hypothesis although the line-up has not been well selected. For this reason, the boundary signaled by the pink dash ellipse corresponding to the place of the points meeting the condition:

$$\frac{D_{TM} + D_{TC}}{2} \leq D_{CM} \tag{9}$$

has been defined as a protection boundary. No match should be accepted as valid if the questioned centroid appears beyond the limits of the guarantee boundary thus defined.

Detection and Matching Results

In the present section an account of the results obtained for the second and third experiments, as described in the above section will be given. The summarized characteristics and objectives of each experiment are given below.

Second experiment description:

Range (decimal log)	Range (natural log)	Statement
$\vartheta_{lg} < 0$	$\vartheta_{ln} < 0$	Likelihood **unconditionally** supports the hypothesis that the questioned and the suspect evidence have not been produced by the same speaker (favoring defender's hypothesis)
$0 \; \vartheta_{lg} < 1$	$0 \; \vartheta_{lg} < 2,3026$	Likelihood **weakly** supports the hypothesis that the questioned and the suspect evidence have been produced by the same speaker (favoring prosecutor's hypothesis)
$1 \; \vartheta_{lg} < 2$	$2,3026 \; \vartheta_{lg} < 4,6052$	Likelihood **mildly** supports the hypothesis that the questioned and the suspect evidence have been produced by the same speaker (favoring prosecutor's hypothesis)
$2 \; \vartheta_{lg} < 3$	$4,6052 \; \vartheta_{lg} < 6,9078$	Likelihood **moderately** supports the hypothesis that the questioned and the suspect evidence have been produced by the same speaker (favoring prosecutor's hypothesis)
$3 \; \vartheta_{lg} < 4$	$6,9078 \; \vartheta_{lg} < 9,2103$	Likelihood **strongly** supports the hypothesis that the questioned and the suspect evidence have been produced by the same speaker (favoring prosecutor's hypothesis)
$\vartheta_{lg} \geq 4$	$\vartheta_{lg} \geq 9,2103$	Likelihood **very strongly** supports the hypothesis that the questioned and the suspect evidence have been produced by the same speaker (favoring prosecutor's hypothesis)

**Tabela 8. Strength of evidence according to Lucy (2005).**

- Splitting the 100 male speakers into two equal-sized subsets according to their normophonic condition.
- Using a normative database validated by Hospital Gregorio Marañon in Madrid with samples of /a/ (50 male speakers).
- Log likelihood ratios according to (1) and (2) estimate the conditional probability of a given sample being normophonic or dysphonic (10-fold cross-validation, taking 47 subjects, leaving 3 in each set of normophonics and dysphonics per run).

Second experiment objectives:

- Estimate the discrimination accuracy of the methodology and the most relevant parameters.
- Produce two reference subsets from GSM quality from the /e/-group of use in Spanish.

Second experiment results:

- The normophonic vs dysphonic cumulants, sensitivity, specificity and accuracy, and Detection-Error Trade-off plots are given in Figure 14.
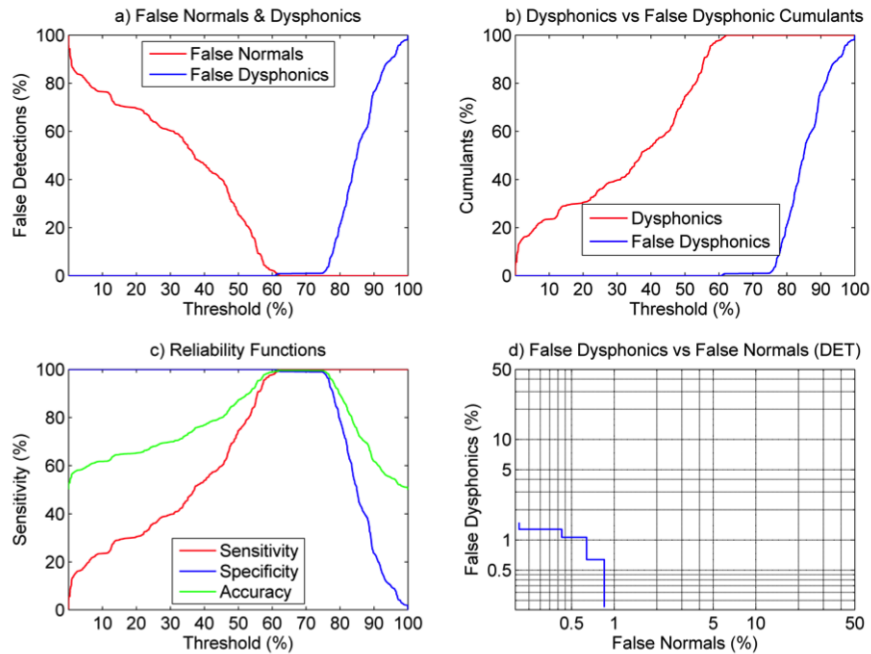
Figura 14. a) False normal vs false dysphonic cumulants. b) Associated Tippet plots. c) Sensitivity, specificity and accuracy for the second experiment. d) Detection-error trade-off curve. (Figure produced by authors.)

The detection procedure consists in generating a vector with the log likelihood ratios generated for each sample, and their assumed condition of normophonic or dysphonic. The log likelihood span is normalized as a percentage, and a moving threshold scans it from 0 to 100%. For each value scanned the number of false normophonics (samples annotated as dysphonic but quoted as normophonic because their log likelihood is over the threshold) and false dysphonics (samples annotated as normophonic but quoted as dysphonic because their log likelihood is under the threshold) is annotated and plotted. See that the number of false normophonics diminishes as the threshold moves rightwards in Figure 14.a to reach the point 1, where the number of false normophonics is very low (only 3 cases out of 470 possible ones), whereas the number of false dysphonics is still 0. At point 8 this number starts raising to 4 out of 470, whereas the number of false normophonics has decreased to 0, indicating that the optimal detection conditions are somewhere between 1 and 8, keeping both false detections at a minimum value simultaneously. This fact indicates that there are two different distributions for each population (false normophonics and false dysphonics), whose accumulated distributions are given in Figure 14.b, known as Tippett plots. Based on these distributions, the plots in Figure 14.c give the well-known variables of sensitivity, specificity and accuracy, according to the following relations:

where TP, FP, TN and FN are the number of true dysphonics, false dysphonics, true normophonics and false normophonics, respectively. These three functions are plotted in Figure 14.c, where the optimum detection point is the one where the accuracy is the maximum. If the number of false dysphonics is plotted vs the number of false normophonics

$$Sn = \frac{TP}{TP + FN};$$

$$Sp = \frac{TN}{TN + FP};$$

$$Ac = \frac{TP + TN}{TP + FN + TN + FP}$$

(10)

the result is the template in Figure 14.d, which is known as the detection-error trade-off plot, because the specific situations combining false positives vs false negatives is confronted for a set of critical threshold values. The number of these situations is 8, and they are signaled in the plot. In fact, apart from the two points already analyzed (1 and 8), the rest of the cases is as follows:

2. False dysphonics jump up to 1/470, false normophonics do not change.
3. False dysphonics do not change, false normophonics drop to 2/470.
4. False dysphonics jump to 2/470, false normophonics do not change.
5. False dysphonics do not change, false normophonics drop to 1/470.
6. False dysphonics jump to 3/470, false normophonics do not change.
7. False dysphonics do not change, false normophonics drop to 0.

The optimal case is point 3, where the rates of false dysphonics and normophonics are equal to 0.638% (equal error rate). The detection accuracy function is at its maximum value of 99.57% at this point, for a threshold range between 62.08% and 75.04%, which implies a reasonably wide noise margin.

Third experiment description:

- Matching each normative speaker's sample (questioned) against every other normative sample (suspect: one target sample vs 49 non-target samples). Eliminating repetitions, these settings imply 50 target detections vs 1,225 non-target detections.
- Using as model set (line-ups) the set of 50 normative speakers, to grant condition (9) as much as possible.
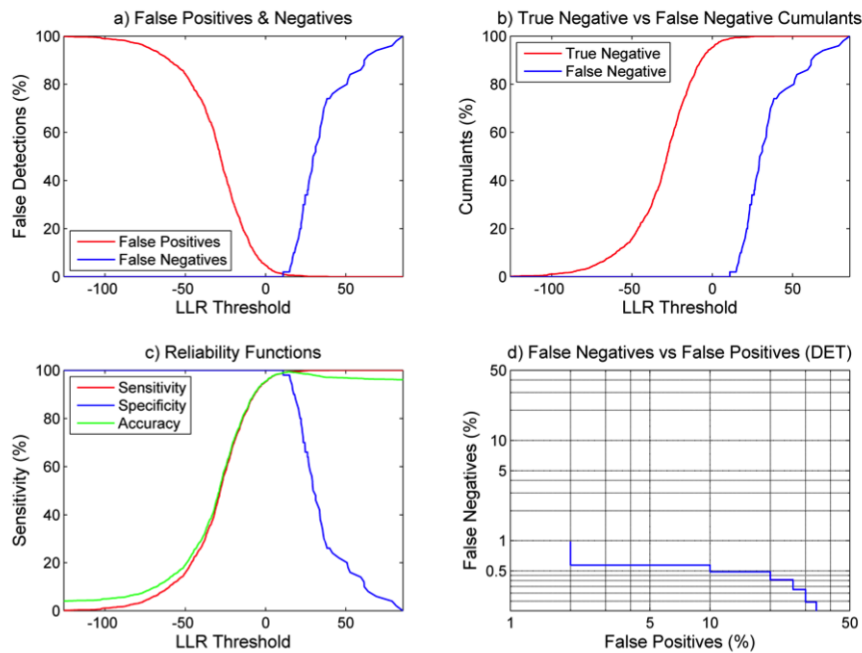
Third experiment objective:

- Estimate the discrimination accuracy of the sample matching methodology in target vs non-target detection tasks.

Third experiment results:

- False target vs false non-target detection cumulants, sensitivity, specificity and accuracy functions, and detection-error trade-off plots given in Figure 15.

As before, the detection procedure consists in generating a vector with the log likelihood ratios (LLR) generated for each sample, and their assumed condition of target or non-target. No normalization of the threshold span has been carried out in this case. The information provided by Figure 15 once the experimental conditions are fixed, can be summarized as follows:

Figura 15. a) False positive vs false negative cumulants. b) Associated Tippet plots. c) Sensitivity, specificity and accuracy for the third experiment. d) Detection-error trade-off curve. (Figure produced by authors.)

a. The rate of false positives (in red) gives the evolution of the non-target cases detected equivocally as targets, as the detection threshold for the log likelihood ratio is moving from left to right. Given the relatively large number of non-target cases (1,225) the evolution of this curve is a smooth decay (inverted sigmoid), expressing that its distribution function will be bell-shaped. On the contrary the low number of target cases (50) given by the blue curve shows slight jumps as the threshold is moving, incorporating new targets as if they were non-targets (false negatives). Both curves cross at the threshold value of 10.88. This is the point of maximal accuracy, on the sixth interval reflected in 0, in the margin where the evidence supports strongly the prosecutor's vs the defender's hypothesis. The detection methodology is maximally accurate just at the beginning of that interval, availing the guarantee of the test. The value of the accuracy function at that point is 99.29%.

b. The graphics given in template a) are given now as Tippett plots. They do not provide any more information to what has been commented up to now, except stressing the fact that the overlap between the two accumulated distributions is very low, granting that at the cross-point the residual tail probabilities (p-values) are under 0.02 and 0.0057, well below the significance level of 0.05.

c. The sensitivity (number of non-targets detected as targets over the total non-targets), specificity (number of targets detected as non targets over the total targets) and accuracy (number of the total targets and non-targets detected as such over the total cases) are plotted as a function of the threshold. The accuracy is very large for the margin of strong support of the prosecutor's hypothesis vs the defender's, with a maximum at 99.29% and not being below 96.08% in any case.

d. The detection-error trade-off curve is shown with a staircase pattern given the specific design of the test. The log likelihood ratios and the corresponding false positive and negative rates are given in Table 9.

Points in the intersection interval between the false positive and false negative curves			
Point	False Positive Rate (%)	False Negative Rate (%)	Log Likelihood Ratio
1	2.0	0.98	10.92
2	2.0	0.90	11.85
3	2.0	0.82	13.06
4	2.0	0.73	13.35
5	2.0	0.65	13.67
6	2.0	0.57	14.84
7	4.0	0.57	15.36
8	6.0	0.57	16.41
9	8.0	0.57	16.79
10	10.0	0.57	17.53
11	10.0	0.49	17.98
12	12.0	0.49	18.61
13	14.0	0.49	19.49
14	16.0	0.49	20.46
15	18.0	0.49	20.91
16	20.0	0.49	21.48
17	20.0	0.41	21.96
18	22.0	0.41	22.05
19	24.0	0.41	22.59
20	26.0	0.41	22.88
21	26.0	0.33	22.94
22	28.0	0.33	23.33
23	30.0	0.33	23.45
24	30.0	0.24	24.33
25	32.0	0.24	24.55
26	34.0	0.24	24.66

Tabela 9. List of points in the intersection interval between the false positive and false negative curves.

The equal-error-rate is not easily determined in this case due to the abrupt staircase behavior of the transition interval in the case of false positive rate. Nevertheless several merit figures may be inferred, for instance, it will be possible to sustain a rate of 2% false positives with a rate of 0.57% false negatives (point 6). This means that accepting an error of one negative in 50 taken as positive grants an error of one positive in 175 taken as negative. The merit figures of both the second and third experiments are given in Table 10.

Conclusions

The process of speaker recognition from speech is a complex matter, as far as the co-articulation involved in message coding expands the limits of intra-speaker variability.

Experiment	Samples	No. Tests	Accuracy LLR (%)	EER		p-values
First	50N + 50D	90 Samples vs Model x 10 times cross-val. = 900	99.57	NA	0.638 (3)	0.00638, 0.00638
Second	50N + 50D	50 Samples vs each: 51*50/2 = 1275 (50 target + 1225 non-target)	99.29	10.88	NA	0.02, 0.0057

Tabela 10. Summary of results for the second and third experiments.

This problem can be alleviated if biometrical markers are defined in relation with phonation, as this phenomenon is less variable for a given speaker, depending only on phonatory settings (creaky, modal, pressed, falsetto, etc). Phonation may experience changes from aging as well as from hormonal status, tobacco, drugs or alcohol consumption, vocal abuse, infections, allergies, other health status conditions, and even circadian cycles (phonation late in the evening is not the same as during the first hours after waking up). It must be assumed that no forensic voice analysis system can realistically manage all this variability, as most of the times the questioned evidence is just a segment of poor quality conversation, and not much more. Regarding the modeling of suspect's evidence, it would be possible sometimes to obtain speech samples under different conditions and in different sessions, but this is not possible most of the time. Our group has conducted multisession tests in very specific collaborative situations such as twins' voice studies (San Segundo and Gómez, 2013; San Segundo, 2014), but indeed that is not a realistic forensic scenario. Nevertheless, this factor has been taken into account as far as parameter selection is concerned. Our study is based on 68 phonation parameters, from which some are very variable with phonation modality and condition, while others are almost invariant to the alterations described. The parameters used in the forensic phonation match have been previously selected according to prior knowledge: for instance jitter, shimmer, noise-harmonic ratios, certain cepstral parameters, glottal source spectral profile, closure and contact defects, and low order tremor are not very sensitive to temporal alterations, and can be safely used in these studies. Focussing on phonation biometrical markers does not necessarily reduce the recognition capability of the methodology, as happens with fingerprints. It is well known that fingerprint matching does not use the whole information available in a fingerprint image; on the contrary, only specific biometrical markers, known as minutias are involved in pattern matching. In this way the process of fingerprint matching becomes more efficient, accurate, robust and less computationally expensive (Jain *et al.*, 1997). The application of this deconstructive methodology to speech implies focussing on phonated speech, rather than in the whole set of voiced and unvoiced patterns. Furthermore, from phonated speech only long vowels close to the axis /a/-/e/ were considered in the present study. These are some of the conclusions derived from the experimental setup used in the study:

- The detection of dysphonic voicing from normophonic seems viable using parameterizations of phonation based on the reconstruction of the glottal source.
- The sensitivity, specificity and accuracy in detecting dysphonic phonation are large enough to grant using phonated segments of speech as long vowels and fillers in forensic voice matching over sufficiently wide detection spans.
- The parameterizations of /a/ and /e/ groups of vowels are interchangeable to a paired test extent, to be used in cross-matching tests with no significant statistical differences.
- The accuracy of target vs non-target sample phonation matches grants the applicability of these tests to real forensic cases.
- The margin of optimal log likelihood ratios granting the strength of phonation evidence over 4 in Lucy's scale (Lucy, 2005) allows its applicability under robust conditions.
- The matching of questioned vs suspect's evidence in reference to line-ups may be summarized in meaningful 2D plots of simple and easy interpretation, granting the reliability and security of the procedure regarding court standards.
- Hybridizing scores from speech and phonation standards as MFCC's and glottal source derived parameters may attain competitive low equal error rates over telephone-quality speech (Khoury *et al.*, 2013).

The proposed methodology for voice pathology detection and monitoring, as well as for forensic voice inspection is being used by police services in Spain and other academic and private institutions (Gomez-Vilda *et al.*, 2012).

Acknowledgements

References

Dejonckere, P. H. (2010). Assessment of voice and respiratory function. In M. Remacle and H. E. Eckel, Eds., *Surgery of Larynx and Trachea*, 11–26. Berlin: Springer-Verlag.

Gomez-Vilda, P., Olalla, R. M., Fernandez, L. M. M., Biarge, M. V. R., Mulas, C. M., Marquina, A. A., Hierro, J. A. H. and Salinero, R. N. (2012). Distance metric in forensic voice evidence evaluation using dysphonia-relevant features. In *Proceedings of the VI Meeting of Biometric Recognition of Persons*, 169–178: Ed. Universidad de Las Palmas de Gran Canaria.

Gómez, P., Rodellar, V., Nieto, V., Martínez, R., Alvarez, A., Scola, B., Ramírez, C., Poletti, D. and Fernández, M. (2013). BioMet®Phon: A system to Monitor Phonation Quality in the Clinics. In *Proceedings of the 5th International Conference on e-Health, Telemedicine and Social Medicine*, 253–258, Nice, France.

Gómez-Vilda, P., Fernández-Baillo, R., Rodellar-Biarge, V., Lluis, V. N., Álvarez Marquina, A., Mazaira-Fernández, L. M., Martínez-Olalla, R. and Godino-Llorente, J. I. (2009). Glottal source biometrical signature for voice pathology detection. *Speech Communication*, 51(9), 759–781.

Gómez-Vilda, P., Fernández-Baíllo, R., Nieto-Altuzarra, A., Díaz-Pérez, F., Fernández-Camacho, F. J., Rodellar-Biarge, V., Álvarez Marquina, A. and Martínez-Olalla, R. (2007). Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters. *Journal of Voice*, 21(4), 450–476.

Hakkesteegt, M. M., Brocaar, M. P. and Wieringa, M. H. (2010). The applicability of the dysphonia severity index and the voice handicap index in evaluating effects of voice therapy and phonosurgery. *Journal of Voice*, 24(2), 199–205.

Hirano, M. (1981). *Psycho-acoustic evaluation of voice.* New York: Springer-Verlag.

Jain, A., Hong, L. and Bolle, R. (1997). On-line fingerprint verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4), 302–314.

Khoury, E., Vesnicer, B., Franco-Pedroso, J., Violato, R., Boulkcnafet, Z., Mazaira Fernandez, L. M., Diez, M., Kosmala, J., Khemiri, H., Cipr, T., Saeidi, R., Gunther, M., Zganec-Gros, J., Candil, R. Z., Simoes, F., Bengherabi, M., Alvarez Marquina, A., Penagarikano, M., Abad, A., Boulayemen, M., Schwarz, P., Van Leeuwen, D., Gonzalez-Dominguez, J., Neto, M. U., Boutellaa, E., Gomez Vilda, P., Varona, A., Petrovska-Delacretaz, D., Matejka, P., Gonzalez-Rodriguez, J., Pereira, T., Harizi, F., Rodriguez-Fuentes, L. J., El Shafey, L., Angeloni, M., Bordel, G., Chollet, G. and Marcel, S. (2013). The 2013 speaker recognition evaluation in mobile environments. In *Proceedings of the 6th IAPR International Conference on Biometrics*, Madrid, Spain.

Kim, S. J., Magnani, A. and Boyd, S. (2005). Robust fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, 659–666. MIT Press.

Lucy, D. (2005). *Introduction to Statistics for Forensic Scientists.* Hoboken, NJ: Wiley.

Muñoz, C. (2014). *Speech signals Feature Extraction Model for a Speaker's Gender and Age Identification System.* Phd thesis, Center for Biomedical Technology, Universidad Politécnica de Madrid, Madrid.

Peirce, C. S. (1878). The probability of induction. *Popular Science Monthly*, 12, 705–718.

Roy, N., Barkmeier-Kraemer, J., Eadie, T., Sivasankar, M. P., Mehta, D., Paul, D. and Hilman, R. (2013). Evidence-based clinical voice assessment: A systematic review. *American Journal of Speech-Language Pathology*, 22, 212–226.

San Segundo, E. (2014). *Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics.* Phd thesis, Universidad Internacional Menéndez Pelayo.

San Segundo, E. and Gómez, P. (2013). Voice biometrical match of twin and non-twin siblings. In *Proceedings of MAVEBA 2013*, 253–256, Florence, Italy: Firenze University Press.

Taroni, F., Aitken, C., Garbolino, P. and Biedermann, A. (2006). *Bayesian Networks and Probabilistic Inference in Forensic Science.* Hoboken, NJ: Wiley.

U.S. Supreme Court, (1993). *Daubert v. Merrell Dow Pharmaceuticals, Inc.* 509 US 579, 589".