GLOTTAL BIOMETRIC FEATURES: ARE PATHOLOGICAL VOICE STUDIES APPLIABLE TO VOICE BIOMETRY?

P. Gómez-Vilda, A. Álvarez-Marquina, L. M. Mazaira-Fernández, R. Fernández-Baíllo, V. Rodellar-Biarge, V. Nieto-Lluis.
[1] Grupo de Informática Aplicada al Procesado de Señal e Imagen
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain
e-mail: pedro@pino.datsi.fi.upm.es

Abstract: **The purpose of the present paper is to introduce a methodology successfully used already in voice pathology detection for its possible adaptation to biometric speaker characterization as well. For such, the behavior of the same GMM classifiers used in the detection of pathology will be exploited. The work will show specific cases derived from running speech typically used in NIST contests against a Universal Background Model built from the population of normophonic subjects in specific vs general evaluation paradigms. Results are contrasted against a set of impostors derived from the same population of normophonic subjects. The relevance of the parameters used in the study will also be discussed.**

Keywords: **Speaker Characterization, Glottal Source, GRBAS, Voice Pathology Grading, Gaussian Mixture Models**

1. INTRODUCTION

The purpose of the present paper is to explore to which extent results in voice pathology detection and grading studies can be extended to give an accurate description of the speaker's voice biometry. In past studies our group has proposed new sets of parameters derived from the glottal component of voice which have been shown to be highly resolving in the detection and grading of pathology [4][9]. These may be grouped into three different classes:

- Glottal Source Spectral Profile Features (GSSPF), which are produced pinpointing the singularities of the Glottal Source Power Spectral Density (GSPSD), specifically the first two "V-grooves" resulting from anti-resonances in the vocal fold biomechanical behavior [3].

- Vocal Fold Biomechanical Parameter Descriptors, which result from the inversion of the electro-mechanical equivalents of the vocal folds when the power spectral profiles are fit to the transfer functions associated to the biomechanical parameters of the vocal folds [3][4].

- Glottal Phonation Cycle Features, which result from the parameterization of the time-domain behavior of the reconstructed Glottal Source. Open, Close and Return Quotients are among the most widely used ones [14], although others based on the vocal gap are introduced as well [4].

The importance of some of these parameters in voice pathology detection is more than evident. Some of the parameters showing better correlation to voice pathology are highly sensitive and mark the presence of pathology with high accuracy. The intention of the present work is to explore if these parameters or others alike may be applied as well to determine personality features or biometric markers of a speaker with a similar degree of accuracy.

The paper is divided in the following sections: an overview of the methodology used in voice pathology detection is given in section 2; section 3 is devoted to the formulation of this study for voice biometry; section 4 is intended to describe the materials and methods for voice biometrical differentiation in intra- and inter-speaker experiments; results are discussed in section 5, and finally conclusions drawn from the present study are presented.

2. VOICE PATHOLOGY DETECTION

Voice Pathology may be detected using different strategies, classically mel-cepstrum parameterization and GMM (Gaussian Mixture Models) classification [8]. Nevertheless the use of mel-cepstral coefficients on the whole voice signal, although efficient, lacks semantics, i.e., it is really difficult to infer which factors convey to successful detection, and from this point it seems really difficult to infer which are the clues to successful classification of pathologies, this being a major aim in the field far from being completed. A different approach is that one of biometric and biomechanical parameter extraction based on the glottal excitation, which produces parameter sets directly related with spectral singularities or vocal fold parameters as dynamic masses or tensions. This approach has been used in the recent past yielding interesting results [9][10]. The combination of specific parameter cocktails may yield quite accurate results not only in voice pathology detection, but in estimating the degree of pathology as well, mimicking the objective estimation of GRBAS [6]. The methodology relies in the accurate determination of a set of individuals which may be considered "healthy" or "pathology-free" from examinations including electroglottogram and endoscopy of the vocal folds. This set of "normophonic" speakers is the key to the correct evaluation of pathology. Normophonic speakers need to be recruited for both genders, as morphologic differentiations between male and female are meaningful [15], and a normophonic male subject may appear as dysphonic if contrasted against a female database. From the inversion of the Liljencrants-Fant source-filter model the glottal source (excitation) is reconstructed [1]. Advanced parameterization techniques are used for the estimation of observation vectors, where each speaker *i* is represented by a parameter vector:

$$\mathbf{x}_i = [x_{i1}, x_{i2}, \dots x_{iJ}] \qquad (1)$$

composed of *J* values x_{ij} produced from a 200 msec. segment of voice corresponding to a sustained utterance of /a/ accordingly with the description given in [5].

Table 1. Description of the parameter set

Param.	Description
x_1	*pitch*
x_2	*jitter*
x_{3-5}	3 variants of *shimmer*
x_{6-7}	Glottal closure parameters
x_{8-10}	Harmonic-Noise and H_2-H_1 Ratios
x_{11-14}	4 first cepstral coefficients of the mucosal wave correlate power spectral density
x_{15-23}	Singularities of mucosal wave correlate power spectral density (amplitude)
x_{24-32}	Singularities of mucosal wave correlate power spectral density (frecuency)
x_{33-34}	Slenderness of the two first "V troughs"
x_{35-37}	Biomechanical parameters of vocal fold body (masses, losses, tensions)
x_{38-40}	Intra-speaker period-synchronous variations of body biomechanics
x_{41-43}	Biomechanical parameters of vocal fold cover (masses, losses, tensions)
x_{44-46}	Intra-speaker period-synchronous variations of cover biomechanics

The observations derived from a given speaker are not used as such, but transformed according to Principal Component Analysis procedures [5] (PCA projection). The reasons are two-fold: on one side the reduction of correlation among the observations improves the data inversion process and results in more stable GMM's; on the other side the dimensions of the vectors can be reduced, thus implying less computational expenses. Once the normophonic male (*m*) and female (*f*) sets are completed the model observation matrices are produced:

$$\begin{aligned} \mathbf{X}_{Mm} &= [\mathbf{x}_{1m}, \dots \mathbf{x}_{im}, \dots \mathbf{x}_{Im}]^T \\ \mathbf{X}_{Mf} &= [\mathbf{x}_{1f}, \dots \mathbf{x}_{if}, \dots \mathbf{x}_{If}]^T \end{aligned} \qquad (2)$$

Similarly the control observation matrices $\mathbf{X}_{Cm}$ and $\mathbf{X}_{Cf}$ are produced using observations from the dysphonic male and female sets. The PCA projection is based on the joint model-control covariance matrix [12]:

$$\begin{aligned} \mathbf{X}_P &= \left[\mathbf{X}_{Mm,f}^T, \mathbf{X}_{Cm,f}^T\right]^T \\ \mathbf{C}_P &= \mathbf{X}_P \mathbf{X}_P^T \end{aligned} \qquad (3)$$

The matrix ($\mathbf{E}_P$) of eigenvalues of $\mathbf{C}_P$ is used to project the original observations matrices on the new principal component matrices:

$$\begin{aligned} \mathbf{Y}_m &= \mathbf{X}_m \mathbf{E}_P \\ \mathbf{Y}_f &= \mathbf{X}_d \mathbf{E}_p \end{aligned} \qquad (4)$$

Once the enrolment of enough normophonic individuals of both genders is available, a GMM for each gender set is produced (Γ_m for the male set and Γ_f for the female one). For such the mean vectors $\boldsymbol{\psi}_{Mm}$ and $\boldsymbol{\psi}_{Mf}$ as well as the corresponding covariance matrices $\mathbf{C}_{Mm}$ and $\mathbf{C}_{Mf}$ are estimated. The GMM is defined by a set of Gaussian multivariate functions of the kind:

$$p(\mathbf{y}_{ti} / \Gamma_{Mm,f}) = \frac{1}{(2\pi)^{Q_m/2} \left|\mathbf{C}_{Mm,f}\right|^{1/2}} e^{-1/2 (\mathbf{y}_{ti} - \boldsymbol{\psi}_{Mm,f})^T \mathbf{C}_{Mm,f}^{-1} (\mathbf{y}_{ti} - \boldsymbol{\psi}_{Mm,f})} \qquad (5)$$

$\mathbf{y}_{ti}$, $\boldsymbol{\psi}_n$, and $\mathbf{C}_n$ being respectively the data vector under test of subject *i*, the centroids of the parameter Gaussians GMM's and the Covariance Matrices of each observation set, *p* being the conditional probability of an observation vector being a member of the specific set represented by the specific Gaussian. As a generalization, if the normophonic GMM is composed by a certain number of Gaussians the joint probability will be expressed as:

$$p_T(\mathbf{y}_{ti} / \Gamma_{Nm,f}) = \sum_k w_k p_k(\mathbf{y}_{ti} / \Gamma_{km,f}) \qquad (6)$$

where w_k are the weights of the linear combination generating the overall probability. In the present case mono-Gaussian Models show to be accurate enough. Finally the issue of voice pathology detection may be stated in terms of a score usually given as a Log-Likelihood Ratio (LLR) of the odds:

$$\begin{aligned} &\Lambda_p(\mathbf{y}_{tmi}) = \\ &log\{p(\mathbf{y}_{tmi} / \Gamma_{nm})\} - log\{p(\mathbf{y}_{tmi} / \Gamma_{\bar{n}m})\} \end{aligned} \qquad (7)$$

This score is based on distance metrics as shown in Figure 1, and it may be used for detecting the pathological condition of the subject using classical ROC-DET (Receiver Operator Characteristics or Detection Error Trade-Off) plots. Depending if the LLR is over or under a given threshold θ ($\Lambda_p(\mathbf{y}_{tmi}/\Gamma_{nm})>\theta$ or $\Lambda_p(\mathbf{y}_{tfi}/\Gamma_{nf})<\theta$) the voice of the subject under test is considered normal or dysphonic.

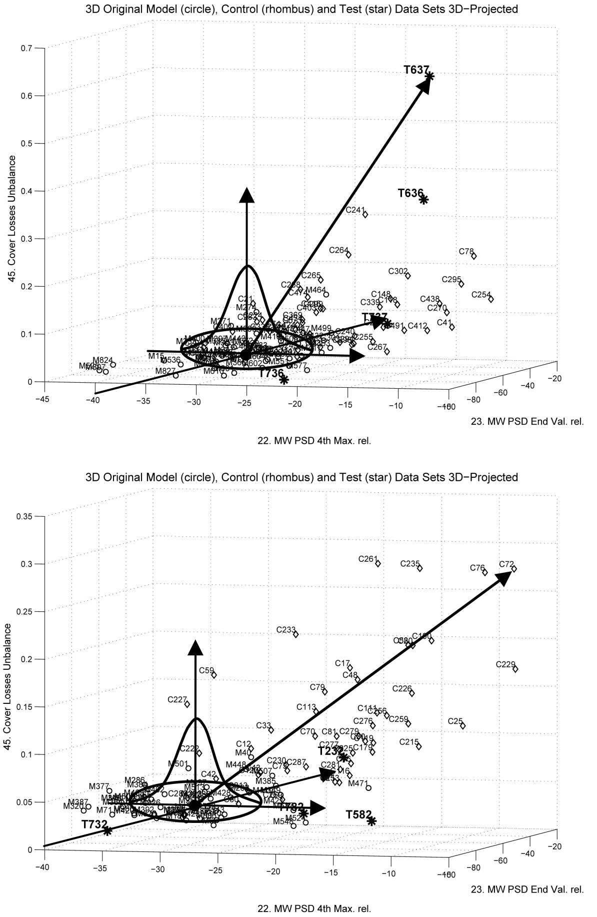

Figure 1. Top: Male cluster set with joint normophonic and pathologic distributions. The distance to the normal distribution may serve as a measure of the pathology grade. Bottom: Idem for the female cluster set.

The figures give an idealized idea on how each respective GMM quantifies the membership probability of each subject relative to its respective model set (Universal Background Model) plotted on the three parameters with largest FDR's. The normalized distance of each subject to the respective model centroid is used as a voice quality evaluation factor (grade) for each individual (g_i) [9]. This distance is marked by arrows for each set farthest cases.

3. Application to Voice Biometry

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

- The modality of the phonation, this being normal (modal), over-pressed or under-pressed. The modal phonation is considered to be associated with the relaxed (emotion-less) speaker, whilst the over-pressed corresponds to emotional excitation (anger, exultation, wrath...), and the under-pressed has to see with anguish, fatigue, depression, etc. Thus modality is very much related to the speaker's emotional state.

- Vocalization. Usually the decomposition of the voice under the source-filter model into the glottal source and vocal tract transfer function is highly dependent on this last pattern. Therefore the results will be different for open than for close vowels, and for voicing consonants. This characteristic has to see with articulation or acoustic-phonetic issues.

- Prosody. The stress and emphasis of the phonation in running speech is of most importance. Raising or lowering the pitch reduces or adds duration to the glottal phonation cycle, and consequently to the resulting parameterization. This situation is similar to the study of voice in singing, as prosody may be considered as the "music" of running speech. The raising or lowering of pitch in speech can produce quite different results in the parameter description of the glottal source in interrogative, declarative or imperative sentences.

With all this information in mind examples will be given from voice samples corresponding to different articulation and prosody cases, and consequences will be drawn regarding their use in speaker recognition studies. The relevance of the speaker's emotional state will be left for further elaboration.

The study will be conducted in terms of the well-known Prosecutor's vs Defender's approach as a classical Log-Likelihood Ratio (LLR) estimation by the specificity-typicality two-stage paradigm [11]:

$$p(I_u / I_a) = \frac{p(I_a / I_u)\, p(I_u)}{p(I_a)} \qquad (8)$$