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 electromechanical 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
Intra-speaker period-synchronous variations of cover variants of Mm first cepstral coefficients of the mucosal wave Slenderness of the two first “V troughs” Singularities of mucosal wave correlate power spectral Harmonic-Noise and pitch

Intra-speaker period-synchronous variations of body Mm Biomechanical parameters of vocal fold cover (masses, 

Table 1. Description of the parameter set

<table>
<thead>
<tr>
<th>Param.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>x₁</td>
<td>pitch</td>
</tr>
<tr>
<td>x₂</td>
<td>jitter</td>
</tr>
<tr>
<td>x₃-₅</td>
<td>3 variants of shimmer</td>
</tr>
<tr>
<td>x₆</td>
<td>Glottal closure parameters</td>
</tr>
<tr>
<td>x₇-₁₀</td>
<td>Harmonic-Noise and H-H: Ratios</td>
</tr>
<tr>
<td>x₁₁-₁₄</td>
<td>4 first cepstral coefficients of the mucosal wave correlate power spectral density</td>
</tr>
<tr>
<td>x₁₅-₂₃</td>
<td>Singularities of mucosal wave correlate power spectral density (amplitude)</td>
</tr>
<tr>
<td>x₂₄-₃₂</td>
<td>Singularities of mucosal wave correlate power spectral density (frequency)</td>
</tr>
<tr>
<td>x₃₃-₃₄</td>
<td>Slenderness of the two first “V troughs”</td>
</tr>
<tr>
<td>x₃₅-₃₇</td>
<td>Biomechanical parameters of vocal fold body (masses, losses, tensions)</td>
</tr>
<tr>
<td>x₃₈-₄₅</td>
<td>Intra-speaker period-synchronous variations of body biomechanics</td>
</tr>
<tr>
<td>x₄₆-₄₉</td>
<td>Biomechanical parameters of vocal fold cover (masses, losses, tensions)</td>
</tr>
<tr>
<td>x₅₀-₅₃</td>
<td>Intra-speaker period-synchronous variations of cover biomechanics</td>
</tr>
</tbody>
</table>

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:

\[
X_{Mm} = \left[ x_{1m}, \ldots, x_{im}, \ldots, x_{Mm} \right]^T
\]

\[
X_{Mf} = \left[ x_{1f}, \ldots, x_{if}, \ldots, x_{Mf} \right]^T
\]

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

\[
X_p = \left[ X_{Mm}, f, X_{Cm}, f \right]^T
\]

\[
C_p = X_p X_p^T
\]

The matrix (\(E_p\)) of eigenvalues of \(C_p\) is used to project the original observations matrices on the new principal component matrices:

\[
Y_m = X_m E_p
\]

\[
Y_f = X_f E_p
\]

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

\[
p(y_i / \Gamma_{Mm,f}) = \frac{1}{(2\pi)^{D/2} |C_{Mm,f}|^{1/2}} e^{-1/2 (y - \psi_{Mm,f})^T C_{Mm,f}^{-1} (y - \psi_{Mm,f})}
\]

\[
y_i, \psi_i, \text{ and } C_i \text{ being respectively the data vector under test of subject } i, \text{ the centroids of the parameter Gaussians GMM’s and the Covariance Matrices of each observation set, } p \text{ 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:}

\[
Pr(y_i / \Gamma_{Nm,f}) = \sum_w w_k p_k \left( y_i / \Gamma_{km,f} \right)
\]

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:

\[
A_{p}(y_{mi}) = \log \left( p(y_{mi} / \Gamma_{nm}) \right) - \log \left( p(y_{mi} / \Gamma_{mn}) \right)
\]
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 $\theta$ ($A_p(y_{\text{norm}}/T_{\text{norm}})>\theta$ or $A_p(y_{\text{path}}/T_{\text{path}})\leq\theta$) 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.](image)

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]:

$$p(I_u/I_a) = \frac{p(I_u/I_a)p(I_a)}{p(I_a)}$$ (8)
where $I$ is in general de information available from a specific speaker ($I_a$ from the questioned or unasserted speaker, $I_u$ from the asserted or suspect). The above probabilistic model will be formalized by the classical likelihood ratio (LLR) evaluating the Prosecutor’s Hypothesis ($H_p$) against the Defender’s Hypothesis ($H_d$):

$$\Lambda_{u/a} = \log \left( \frac{f(E / H_p, I)}{f(E / H_d, I)} \right)$$

$$= \log \left( \frac{p(I_a / I_u)}{p(I_a)} \right) = \log \left( \frac{p(I_u / I_a)}{p(I_u)} \right) = (9)$$

$$= \log \left( p(x^u / \Gamma_A) \right) - \log \left( p(x^a / \Gamma_B) \right)$$

$E$, $H_p$, and $H_d$ being respectively the Evidence, the Prosecutor and the Defender Hypotheses. The general speaker information, composed by the set of observations (parameter medians of the set of parameters in Table 1) from the asserted or suspect (a) and the unasserted or questioned (u) observations are defined as:

$$x^u = [x^u_1, x^u_2, ..., x^u_m]^T$$

$$x^a = [x^a_1, x^a_2, ..., x^a_m]^T$$

The Universal Background Gaussian Model (UBGM) $\Gamma_B$ will be composed by the covariance matrix $C_{B}$, and mean vector $\psi_B$ for the reference population data set. The Asserted Gaussian Model $\Gamma_A$ is to be built in a similar way from all the data available from the suspect, resulting in $C_{A}$, and $\psi_A$. The evaluation of the membership of a given questioned frame $y^u$ with respect to the UBMG or the AGM will be estimated in terms of the conditioned probability:

$$p(y^u / \Gamma_{A,B}) = \frac{1}{(2\pi)^{D/2} |C_{A,B}|^{1/2}} e^{-1/2 (y^u - \psi_{A,B})^T C_{A,B}^{-1} (y^u - \psi_{A,B})}$$

Once the relative membership probabilities are produced, the LLR of the Prosecutor’s vs the Defender’s Hypothesis given in (9) will be estimated. Results for a practical study case will illustrate this technique in the next section.

4. RESULTS AND DISCUSSION

For the purposes of the present study a set of 30 normophonic male speakers from [7] will be used in the experiments. Part of this subset, specifically 20 speakers will serve as the Universal Background Model Set, and 10 speakers more will be used as imposters for T-norm contrast. The questioned and suspect frames will be obtained from a 300-sec. recording of running speech (test 4, channel a) from the last NIST SRE10 HARS1 competition [13] selecting 12 frames where the utterance /ah/ or /uh/ have been produced, either in long vowels or in fillings, these frames being given in Table 2.

<table>
<thead>
<tr>
<th>Frame start</th>
<th>Frame end</th>
<th>Frame #</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.2</td>
<td>9.4</td>
<td>4009</td>
</tr>
<tr>
<td>28.7</td>
<td>28.9</td>
<td>4028</td>
</tr>
<tr>
<td>43.95</td>
<td>44.20</td>
<td>4043</td>
</tr>
<tr>
<td>201.55</td>
<td>201.75</td>
<td>4201</td>
</tr>
<tr>
<td>213.85</td>
<td>214.00</td>
<td>4213</td>
</tr>
<tr>
<td>232.30</td>
<td>323.55</td>
<td>4232</td>
</tr>
<tr>
<td>243.55</td>
<td>243.85</td>
<td>4243</td>
</tr>
<tr>
<td>248.80</td>
<td>249.35</td>
<td>4248</td>
</tr>
<tr>
<td>267.00</td>
<td>267.35</td>
<td>4267</td>
</tr>
<tr>
<td>276.00</td>
<td>276.20</td>
<td>4276</td>
</tr>
<tr>
<td>289.95</td>
<td>290.25</td>
<td>4289</td>
</tr>
<tr>
<td>291.30</td>
<td>291.60</td>
<td>4291</td>
</tr>
</tbody>
</table>

The whole set of 20+10+12 frames taken at 8kHz are parameterized and PCA projected. The corresponding Model, Control and Test sets are described in Table 3.

| Frames under study |

<table>
<thead>
<tr>
<th>Frame start</th>
<th>Frame end</th>
<th>Frame #</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.2</td>
<td>9.4</td>
<td>4009</td>
</tr>
<tr>
<td>28.7</td>
<td>28.9</td>
<td>4028</td>
</tr>
<tr>
<td>43.95</td>
<td>44.20</td>
<td>4043</td>
</tr>
<tr>
<td>201.55</td>
<td>201.75</td>
<td>4201</td>
</tr>
<tr>
<td>213.85</td>
<td>214.00</td>
<td>4213</td>
</tr>
<tr>
<td>232.30</td>
<td>323.55</td>
<td>4232</td>
</tr>
<tr>
<td>243.55</td>
<td>243.85</td>
<td>4243</td>
</tr>
<tr>
<td>248.80</td>
<td>249.35</td>
<td>4248</td>
</tr>
<tr>
<td>267.00</td>
<td>267.35</td>
<td>4267</td>
</tr>
<tr>
<td>276.00</td>
<td>276.20</td>
<td>4276</td>
</tr>
<tr>
<td>289.95</td>
<td>290.25</td>
<td>4289</td>
</tr>
<tr>
<td>291.30</td>
<td>291.60</td>
<td>4291</td>
</tr>
</tbody>
</table>

The whole set of 20+10+12 frames taken at 8kHz are parameterized and projected. The corresponding Model, Control and Test sets are described in Table 3.

5. RESULTS AND DISCUSSION

For the purposes of the present study a set of 30 normophonic male speakers from [7] will be used in the experiments. Part of this subset, specifically 20 speakers will serve as the Universal Background Model Set, and 10 speakers more will be used as imposters for T-norm contrast. The questioned and suspect frames will be obtained from a 300-sec. recording of running speech (test 4, channel a) from the last NIST SRE10 HARS1 competition [13] selecting 12 frames where the utterance /ah/ or /uh/ have been produced, either in long vowels or in fillings, these frames being given in Table 2.
terms of possible dysphonia reflected in certain parameters. As a side objective, the parameters reflecting the dysphonic condition are to be determined. This last objective is achieved using Fisher's Discriminant Ratios (FDR), as given by:

\[
fdr_j = \frac{(\mu_{Mj} - \mu_{Gj})^2}{\sigma_{Mj}^2 + \sigma_{Gj}^2}; \quad I \leq j \leq J
\]  

(12)

the resulting estimations are given in Figure 2. The set of parameters being investigated have been selected among the most resolving ones, although not all of them are sensitive enough. The three most resolving are the HNR \((x_d)\), the \(f_{M1}-f_{M2} (x_d)\) and the 2nd minimum in the GSPSD (Glottal Source Power Spectral Density: \(x_{j2}\)). It is interesting to see that shimmers \((x_{j,a})\) are more resolving than jitter \((x_d)\), and that the spectral singularities of the GSPSD \((x_{j2};x_{j3})\) show also important discriminating capabilities in this case. Thus a good balance between classical distortion parameters and biometrical ones is expected to enhance discrimination results. The Model \((o)\), Control \((\bar{o})\) and Test \((\Lambda)\) Sets given by matrices \(X_M\), \(X_C\) and \(X_T\) in Table 3 are shown in Figure 3.

As seen in the table, for each frame in the Test Set (first column to the left) the pathologic LLR in (7) is given. It may be seen in the second column that according to the data available the two suspect/questioned frames can not be considered pathological, this fact being reinforced by the objective grade \(G\) (third column), which results null, this corresponding with no dysphonia. The fourth column gives the squared Mahalanobis distance from each sample corresponding with no dysphonia. The fourth column gives the same figure for each sample relative to the Test Centroid estimated from frames extracted from the question speech segment. The fifth column gives the relative membership probability of each Test frame to both the Control and Model Sets. The membership probability of the upper ten frames relative to the UBGM is clearly larger than their respective membership probability relative to the AGM. This results in negative LLR's favoring the \(H_0\). On the contrary, the last two frames show membership probabilities larger for the AGM than for the UBGM. The respective LLR's are positive and similar, favoring the \(H_p\) in their case.

5. Conclusions
Interesting consequences may be derived from the present study. First of all it seems that parameters classically derived for the study of voice pathology as shimmer, HNR or $H_2-H_1$ can be used for the biometrical characterization of the speaker as well. This conclusion is very important, as these parameters have clear semantics as far as the characterization of a speaker is concerned. A second conclusion is that the Glottal Source Power Spectral Distribution (GSPSD), and especially its singularities (peaks and troughs) are relevant for the biometrical characterization of the speaker. It is known that the Glottal Source may be altered by articulation as well as by modality, vocalization or prosody, as explained in section 3. The frames selected from the running speech segment were not especially conditioned by any factor except by vowel coloring (in fact most of them correspond to the kind of fillers /uh/’s and /ah/’s, which are spontaneously produced by Native Speakers of English). The modality is different in most of them, as well as the prosody (some present questioning or surprise marks). Nevertheless, the system identified clearly all of them as being different from the Model Set, selected from sustained vowels, and similar among themselves. This fact may indicate that the parameters selected are robust to modal information and sensitive to biometrical differences. Of course the work is still far from being completed. Massive tests on model and test running speech segments as the ones proposed in the last NIST SRE HARS2 contest need to be processed. For such automatic vowel selection and framing is to be put into work and the discussed methodology applied on blind tests to measure its capability in speaker characterization tasks.

ACKNOWLEDGMENTS

This work has been funded by grants TIC2003-08756, TEC2006-12887-C02-01/02 and TEC2009-14123-C02-03 from Plan Nacional de I+D+i, Ministry of Science and Technology, by grant CCG06-UPM/TIC-0028 from CAM/UPM, and by project HESPERIA (http://www.proyecto-hesperia.org) from the Programme CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain.

REFERENCES


