A Methodology for Monitoring Emotional Stress in Phonation

V. Rodellar, D. Palacios, P. Gómez, E. Bartolomé
Center for Biomedical Technology, Universidad Politécnica de Madrid, Campus de Montegancedo, 28223 Pozuelo de Alarcón, Madrid, Spain
victoria@pino.datsi.fi.upm.es, daniel.palacios.alonso@gmail.com, pedro@fi.upm.es, elenabartolome.bm@gmail.com

***Abstract*— Stress in phonation is mainly shown in the signature of the fundamental frequency. The proposed methodology is based on the estimation of the vocal fold biomechanics in terms of the distribution of the dynamic mass and the mechanical tension of the vocal fold structure. These parameters are derived from the reconstruction of the glottal source by inverse filtering. The vocal fold mechanical tension correlates (stress and strain), are used as the bases for tremor estimation. The correlates of tension and tremor are used to characterize the spontaneous speech of a database of 40 speakers of both genders (20 male and 20 female). Spontaneous speech consists in short interviews of 20 s of duration where the speakers have to express opinions on hot issues with which they are in agreement (*pro*) or in disagreement (*con*) following Arciuli's methodology. The emotional stress is estimated from the biomechanical correlates expressed above (tension and tremor). The null hypothesis formulated as the insensitivity of the speaker to *pro* and *con* situations has to be disregarded in view of the results for both genders. Interesting open questions are to be raised regarding the possibility of speakers consciously hiding their true opinion based on political correctness. The discussion will offer different hypotheses to further exploit the objective of detecting self-congruence in spoken messages.**

Keywords - Speech; phonation; self-congruence; stress detection.

I. INTRODUCTION

Speech is the main and most direct way of interaction among human beings. It does not only contain linguistic information (message) but also other meta-linguistic data on the biometry of the speaker, and their emotional and health status as well. In the last decades, an intensive research has been conducted in the topics of automatic speech and speaker recognition and more recently in the identification and classification of emotional states expressed in speech as a major focus of attention ([1]-[3]). The identification of emotions in speech could provide more effective and natural computer interfaces [4]. This topic has many potential applications in different fields, as security, marketing, medical diagnosis, video and computer games, intelligent toys, aids for handicapped, elder people surveillance, etc. The key to this research is to understand how people express, characterize and identify emotions, and how to estimate them for the development of the above mentioned applications. Classically emotions have been characterized by prosody and energy, among other clues related with voice quality [2] but research is still far to identify a complete set of biomarkers clearly separating the number of discrete emotions one can distinguish.

The aim of the present paper is more restricted, concentrating in the detection of emotional stress present in voice regarding the self agreement of the speaker with the message produce as accurately as possible in the line proposed by Arciuli ([5]-[7]). The objective of the research, rather than estimating markers of basic emotions (anger, boredom, disgust, fear... etc.), is concentrated in detecting if running speech is produced under an emotion-induced stress or if it is neutral, and if stress is induced by self disagreement with the message. The main hypothesis, following [6] is that in emitting a spoken message the speaker has to face two situations: either they are in self agreement with the message (pro) or in disagreement (con). Going one step further these two situations could be associated with "telling the truth" or "lying", although our approach does not intend to go that far. Obviously, the speaker could suffer emotional stress both in *pro* and in *con* situations, induced by basic emotions related with the positive or negative valence of the emotion associated, but the intention is not to disclose which emotion is present rather that what is the amount of emotional activation expressed in the message.

The paper is organized as follows: The data collection and methods used are described section II. In section III the results obtained for the database are presented. Section V is devoted to discuss the results, extract conclusions and comment future lines.

II. MATERIALS AND METHODS

The problem, as is being posed, can be based on the detection of fillers from a database of running speech taken from short interviews on an experimental framework described by Arciuli as "low-stakes laboratory-elicited lies" [5], in which the same speakers were asked to give a truthful account of their views regarding one topic and to provide an untruthful account their views for a different topic. The topics included hot issues regarding controversial social questions. The speakers set comprised 20 male and 20 female speakers of Spanish in the first run, with ages between 22 and 60 years, with an average of 24.3 for males and 23.5 for females. The main hypothesis is based in the following chain of reasoning: a speaker is differentially stressed when producing a *pro* statement than when is producing a *con* statement. This stress is manifested in their phonation quality following [2], in which open and closed vocal fold quotients are detected and used as markers. But in the present research changes in the biomechanical parameters of their vocal folds were used instead. As the database was recorded in Spanish the main hypothesis was tested on vocalizations of type /ɛ/ and /e/ present in common words as /de/ or /que/ either as a vowel lengthening or as a filler. Preliminary results from the processing of the database

were published in [8]. Further research on this topic is to be found in [9] and [10].

The biomechanics of the vocal fold system is seen in Fig. 1. Template a) gives the transversal section of the vocal folds, extending in the plane normal to the template. The vocal folds are two tiny muscular structures (body, or *musculus vocalis*) stretching between the thyroid and arytenoids cartilage processes, linked by visco-elastic ligaments to the epithelium or cover. Both vocal folds come together under the action of a set of laryngeal muscles on the occipital side of the larynx, known as the transversal and oblique laryngeals. Once the vocal folds come together they close the space between them (glottis), stopping the air of flow from the subglottal (lung side) to the supraglottal (pharyngeal side) chambers. Under this closure the subglottal pressure raises till it can force both vocal folds to come apart and a puff of air is expelled to the supraglottal chamber (glottal flow during the open phase). This aperture results in a decrement of subglottal pressure and both vocal folds come together again, while supraglottal pressure experiences a sharp decay (maximum flow declination ratio during the closing phase). This cycle is repeated around 110 times/s in male voice as an average, and around 210 times/s in female voice, and is the basis of fundamental frequency of voicing in both genders.

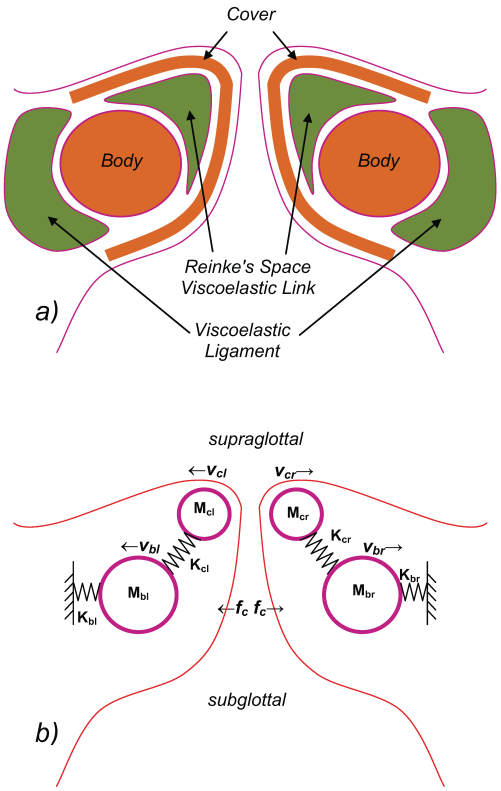

Fig. 1 Vocal fold 2-mass biomechanical model assumed in the study. a) Structural description of vocal folds seen in transversal section. b) Equivalent model in masses and visco-elasticities.

Template b) on its turn gives a biomechanical model of the structures depicted in template a). The average dynamic mass of the body contributing to vibration of the body is represented by masses M_{br} and M_{bl}, for the right and left vocal folds. These masses are attached to the rigid walls of the larynx cartilages by springs with stiffness given as K_{br} and K_{bl}, respectively. The complex structure of the cover and visco-elastic ligaments in Reinke's space are represented by masses M_{cr} and M_{cl}, and the linking stiffness springs K_{cr} and K_{cl}. Transversal forces f_c resulting from the pressure difference between the subglottal and supraglottal chambers, as well as from the flow of air (intra-glottal forces) act on both systems. As a result the masses move with transversal velocities v_{cl}, v_{cr}, v_{bl} and v_{br} along the horizontal axis (there is also a vertical component, assumed null in the present model). It may be shown that in this biomechanical model the relation between acting forces and the transversal velocities, for the case of the body (*musculus vocalis*) can be expressed by the trans-admittance functional given as

$$T_b(\omega_b, \mu_b, \xi_b, \sigma_b) = \left[\left(\omega \mu_b - \omega^{-1} \xi_b \right)^2 + \sigma_b^2 \right]^{-1} \qquad (1)$$

where μ_b, σ_b and ξ_b stand for the estimates of the massive, viscous and elastic parameters of the vocal fold body biomechanical model, corresponding to $R_{bl,r}$, $M_{bl,r}$ and $K_{bl,r}$ respectively. It may be shown [12] that under certain assumptions the modulus of this functional can be associated with the power spectral density of the *glottal source* $s_r(t)$, a sound wave corresponding with the supraglottal pressure just at the point where the glottal flow is injected.

$$\left\| S_r(\omega) \right\| = \left| \int_{-\pi}^{\pi} s_r(t) e^{-j\omega t} dt \right| \qquad (2)$$

That association can be exploited to estimate the biomechanical parameters μ_b, σ_b and ξ_b minimizing the cost function

$$L(\omega, \mu_b, \xi_b, \sigma_b) = \\ = \oint_{2\pi} \left(\left\| S_r(\omega) \right\| - \left\| T_b(\omega, \mu_b, \xi_b, \sigma_b) \right\| \right)^2 d\omega \qquad (3)$$

The numerical estimation of the biomechanical parameters, and particularly the stiffness induced by the neuro-motor activity on the transversal and oblique laryngeal muscles controlling phonation (considered proportional to ξ_b) can be carried out using different approaches. A possible one will be to produce an estimate of the *glottal source* $s_r(t)$ from inverse filtering as explained in [9], where the interested reader will find relevant information. To determine the parameters of the body and cover dynamics the electromechanical equivalent, represented by the system given in Fig. 2, is used.

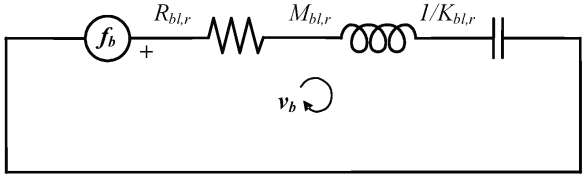

Fig. 2 Electromechanical equivalent model of the body dynamic mass, losses and stiffness when stimulated by intra-glottal forces. A similar model stands for cover dynamics.

It will be assumed that the power spectral density of the *glottal source* will be related (within a scale factor) to the square modulus of the input admittance of each electromechanical equivalent given in (1). The estimation of the body parameters from the *glottal source power spectral density* is not complicate, because this magnitude is smooth and predictable. The process of estimation is as follows

$$M_{bl,r} = \frac{\omega_2}{\omega_2^2 - \omega_r^2}\left[\frac{T_r - T_2}{T_r T_2}\right]^{1/2} \quad (4)$$

ω_r being the resonance frequency given by

$$\omega_r^2 = \frac{K_{bl,r}}{M_{bl,r}} \quad (5)$$

where the square modulus of the admittance in (1) is rewritten as

$$T(\omega) = \frac{1}{\left[R_{bl,r}^2 + \varpi^2 M_{bl,r}^2\right]^2} \quad (6)$$

with the frequency relative to the resonance point given as

$$\varpi = \frac{\omega^2 - \omega_r^2}{\omega} \quad (7)$$

and the two selected points in the *power spectral density of the glottal* source, corresponding to the peak and the second harmonic are given as

$$\begin{aligned} T_r &= T(\omega = \omega_r) = \frac{1}{R_{bl,r}^2} \\ T_2 &= T(\omega = 2\omega_r) \end{aligned} \quad (8)$$

The evaluation methodology must produce first a very accurate estimation for the value of pitch, which is used to evaluate ω_r. This leads to the determination of the losses from (8) and to the mass and stiffness from (4) and (5), respectively.

The specific interest on the biomechanical stiffness of the vocal fold is due to the fact that this parameter is directly and strongly related with the neuron firing rate acting on the laryngeal muscles, and retains emotional stress in marks of hypo- and hyper-tension, as well as tremor [12]. Important correlates quantifying emotional stress are thus vocal fold stress, and tremor, expressed in frequency and rms amplitude relative to its average value. This study will be oriented to give results on biomechanical tension, tremor being left for future research.

III. RESULTS

The exploitation of the database for the present study consisted in selecting fillers and phonation lengthened segments found in running speech from interviews of male (M1-20) and female (F1-20) speakers. Segments were 500 ms long to grant enough number of phonation cycles (an average between 50 and 100), in order to grant statistical stability in the estimates. Results of vocal fold stiffness estimates (average biomechanical tension in N/m along all the phonation cycles produced in the segment) from the *pro* and *con* conditions are given in Table 1.

Table 1. Average body stiffness in pro and con situations for male and female speakers in the experiment.

Speaker	Body St. *pro* (N/m)	Body St. *con* (N/m)
M1	8.557	9.488
M2	7.894	9.095
M3	12.401	13.637
M4	14.595	18.001
M5	13.695	12.969
M6	9.831	9.572
M7	13.262	14.558
M8	9.684	9.840
M9	7.893	8.015
M10	14.492	13.097
M11	9.627	10.819
M12	11.117	10.220
M13	9.658	9.579
M14	9.802	9.150
M15	9.266	9.994
M16	9.221	9.684
M17	7.958	10.806
M18	11.884	12.679
M19	9.281	10.087
M20	9.841	9.053
F1	17.615	18.388
F2	23.300	20.820
F3	23.288	21.480
F4	19.635	15.117
F5	17.809	12.821
F6	14.724	13.307
F7	13.954	12.477
F8	21.715	21.674
F9	18.422	22.493
F10	33.220	23.297
F11	19.682	15.307
F12	18.838	18.554
F13	21.222	21.354
F14	30.758	19.522
F15	19.700	17.281
F16	14.354	21.277
F17	16.725	16.059
F18	17.794	16.488
F19	21.136	16.248
F20	22.799	20.228

The statistical analysis of the results can be carried out as a t-Student paired test, where the null hypothesis (H0) is formulated as that the difference between average estimates for each speaker's body stiffness in self agreement (*pro*) and disagreement (*con*) may be modelled by the same statistical distribution, or in other words, that there is no significant difference between the estimates taken under both conditions below the threshold of 0.05. Table 2 illustrates the results of the paired test. The average of differences between *pro* and *con* is given in the row labelled *aver*. It may be seen that the average deviation is positive for males (0.519 N/m) whereas it is negative for females (-2.125 N/m). This would indicate a

relaxation of the vocal fold stiffness in contradictory situations as the average tendency in females, where the results indicate an increment in the same parameter for males.

Table 2. Statistics of t-Student tests.

Statistic	Males	Females
aver.	0.519	-2.125
std. err.	1.208	4.079
norm. s. e.	0.382	1.290
t-statist.	1.36	1.65
p-value	0.0348	0.0155
Reject H0	YES	YES

Another parameter of interest is the normalized standard error for the sample sizes, which is almost three times larger for the female set. Based on these values the results of the t-test are given as *t-statist.*, and their associated *p-value*. It may be seen that in both cases the null hypothesis (H0) has to be rejected for a significance interval limited to 0.05, certifying that the *pro-con* methodology is statistically significant for both male and female sets. It may be seen that male results are less disperse than those from the female set, and this is indicated by a larger Pearson's coefficient (0.87 vs 0.56). The null hypothesis (H0) is signaled by the red-dash line corresponding to $<\xi_{pbi}>=<\xi_{cbi}>$, where $<\xi_{pbi}>$ is the average body stiffness (b) under *pro* conditions for a given speaker i, and $<\xi_{cbi}>$ is the corresponding value under *con* conditions. A normalized standard error has been used to trace two delimitation lines (melba-dash) around the H0 line in both sets. Samples within the two standard error delimitation lines are considered low sensitive to *pro-con* test. Samples over (in dark red) are increased stiffness-sensitive, and samples below (in green) are reduced stiffness-sensitive.

These results are better visualized in the templates given in Fig. 3 and Fig. 4.

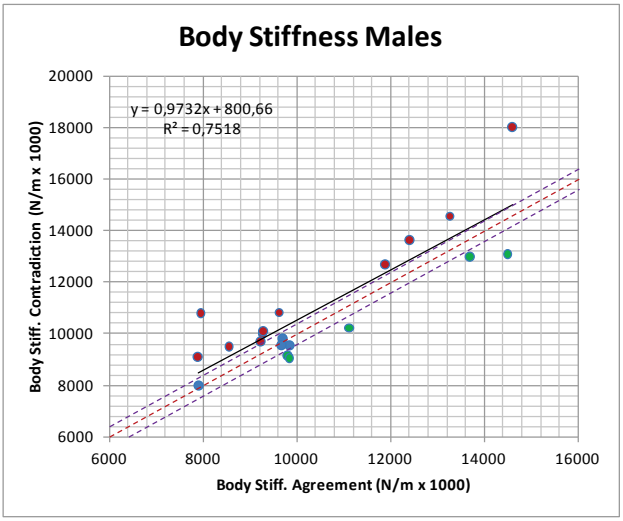

Fig. 3 Comparison of average vocal fold body stiffness estimates in *con* vs *pro* conditions for the male set. The regression line expression is given and depicted as a black solid line.

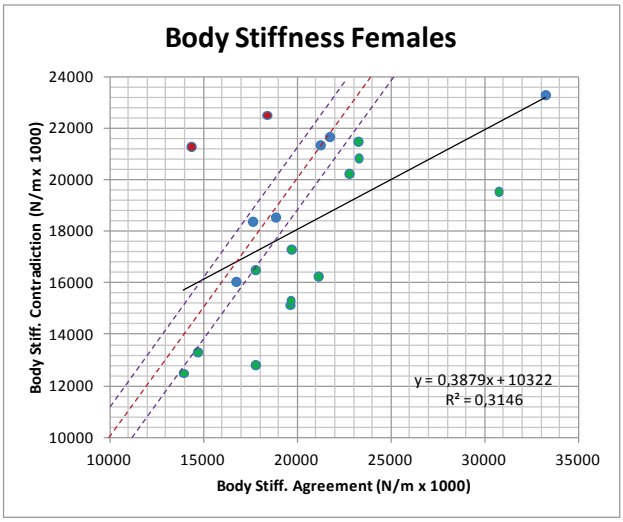

Fig. 4 Comparison of average vocal fold body stiffness estimates in *con* vs *pro* conditions for the female set. The regression line expression is given and depicted as a black solid line.

IV. Implications to CogInfoCom Methodology

The idea of applying the technologies of information and communications to disentangle and better understand natural cognitive structures and organisms, as well as the reverse i.e. using bioinspiration and neuromorphism to help artificial intelligent systems to become more adaptive to complex problem solving is a very attractive and intriguing field since long. A good formal starting point has been defined by the work of [13].

This idea is not alien to the field of Speech Production and Perception. Just to present a clear paradigm well known in the field of Automatic Speech Recognition, where a lot of bio-inspired knowledge has been integrated, and a lot more will be in the next years, the pioneer paper by Hermansky is to be cited [14]. Since then, many researchers have oriented their work to better understand natural cognitive systems to extract useful knowledge to be integrated in Speech Production and Perception [15]. In this line our research group has carried out a series of studies to explore the field of Speech Perception under the bioinspired point of view [16]. The basic idea is to develop simple yet multimodal functional units replicating the basic macroscopic functionality of the neurons in the Auditory System to emulate the detection of simple phonetic patterns roughly described as “basic vowels and consonants”. The functions of the real receptive fields known to be resident in the Human Auditory System were thus emulated using these “neuromorphic units” [17]. Following the descending paths in these studies, the next steps are oriented to better understand and describe Speech Production combining the indirect estimation of biomechanical and neuromotor pathways from the speech motor cortex to the groups of muscles active in the larynx, tongue, velopharynx and face, with the direct reconstruction of their emulated counterparts. The specific emphasis is to be placed in their application to neuromotor and cognitive diseases (e.g. Parkinson and Alzheimer) as well as to behavioural diseases. An especially sensitive target is autism. In all these cases the “emotional temperature” of the patient is a very important clue [18]. One of the most lively and semantic correlates in estimating the emotional temperature of the patient is

phonation stress revealed from biomechanical vocal fold correlates, thus closing the loop Perception-Processing-Production of Speech.

The big issue now is how to embed the knowledge provided by these correlates into a rich semantic framework for clinicians, neurologists, psychologists, or other care providing professionals. This is not a simple and easy task. Indeed it may be the most problematic one. The basic difficulty comes from the nature of the correlates being detected, because these are given as streams of data more or less aligned to the speech frame, but lacking interpretation by themselves. A visually semantic interface should be designed to present and exploit such information by the user, and this professional needs to be a properly trained expert. Fig. 5 (end page after References) gives a first approximation to a possible user interface exploiting this knowledge for specific patient or speaker monitoring tasks. This same approach may be extended to many other applications, as in reactive emotional monitoring for use in neuromarketing, experimental psychology, or in the field of speech forensic, among others.

V. Discussion and conclusions

Through the present work a methodology to monitor stress in phonation detectable in *pro* vs *con* statements has been presented. A very interesting result is apparently the dominant tendency is gender dependent, the fact that vocal fold stiffness relaxation in *con* situations for female voice being specially relevant. It must be said that this fact was very intriguing at first sight. Later discussions convinced us that it may be due to the strong dimorphism of adult male and female larynges, corresponding to a much larger stiffness in the case of the female vocal fold, thus explaining the higher fundamental tone observed in female voice modal phonation. It seems that for female voice, instead of increasing biomechanical tension further, contradictory stress conveys a reduction of this parameter, corresponding with a decrease of tone in almost a fifth. After arriving at this conclusion the authors have had the opportunity to verify this situation in many occasions from real life experience, far from artificially induced laboratory situations. Stress in female voice is perceived in most of the cases with a decrease of almost a fifth (practically from an average 220 Hz to around 180 Hz) associated to an increase in spontaneous tremor around 5-8 Hz. The analysis of tremor is out of the scope of the present paper being currently further investigated, as well as other study lines to be accomplished in the near future.

The hot issue is that of political "correctness" regarding self-agreed messages regarding sensitive issues. This is what Arciuli labels as "low-stakes" or "high-stakes". It seems highly reasonable that people tend to express opinions or statements in a very different way if these imply further consequence for self-esteem, image or civil status (risk of receiving a guilty declaration, for instance). Therefore, the spoken message needs to be fabricated accordingly, and this leaves important clues in phonation stress, which can be estimated form biomechanical clues in voicing. A number of speakers in an interview in which their self-image may be damaged by issuing an opinion contrary to that one admitted by the majority may be suspects of expressing *pro* opinions under *con* conditions, and the reverse, thus introducing a bias in the results of these tests. Better test strategies and protocols to cope with this possibility are necessarily to be designed to establish a baseline for contrasting results from real tests.

Acknowledgments

This work has been funded by grant TEC2012-38630-C04-04 from Plan Nacional de I+D+i, Ministry of Economic Affairs and Competitiveness of Spain.

References

[1] G. Zhou, J. H. Hansen, J. F. Kaiser, “Nonlinear feature based classification of speech under stress”, *IEEE Trans. on Speech and Audio Processing,* Vol. 9 (3), 2001, pp. 201-216.

[2] E. Moore, M. A. Clements, J. W. Peifer, L. Weisser, “Critical analysis of the impact of glottal features in the classification of clinical depression in speech”, *IEEE Trans. on Biomedical Engineering,* Vol. 55 (1), 2008, pp.96-107.

[3] M. El Ayadi, M. S. Kamel and F. Karray. "Survey on speech emotion recognition: Features, classification schemes and databases", *Pattern Recognition*, Vol. 44, 2011, pp. 572-587.

[4] H. Lou, et al, "StressSense: Detecting Stress in Unconstrained Acoustic Environments using Smartphones" *Proceedings of the 2012 ACM Conference on Ubiquitous Computing,*pp. 351-360.

[5] J. Arciuli, G. Villar, D. Mallard, “Lies, lies and more lies”, *Proceedings of the 31st Annual Conference of the Cognitive Science Society (CogSci 2009)*, pp. 2329-2334.

[6] J. Arciuli, D. Mallard, G. Villar, "“Um, I can tell you’r lying”: Linguistic markers of deception versus truth-telling in speech", *App. Psycholing.*, Vol. 31, 2010, pp. 397-411.

[7] G. Villar, J. Arciuli, D. Mallard, "Use of “um” in the deceptive speech of a convicted murderer", *App. Psycholing.*, Vol. 22, 2012, pp. 83-95.

[8] E. Bartolomé, "Contribución al estudio de las alteraciones de la fonación en habla contradictoria frente a espontánea", MSc Thesis, Universidad Internacional Menéndez Pelayo, July 2012 (in Spanish).

[9] V. Rodellar, D. Palacios, E. Bartolomé and P. Gómez, "Vocal fold stiffness estimates for emotion description in speech", *International Conference on Bio-inspired Systems and Signal Processing, BIOSIGNALS'2013*, pp. 112-119.

[10] V. Rodellar-Biarge, D. Palacios-Alonso, V. Nieto-Lluis and P. Gómez-Vilda, “Speech parameter selection for emotional stress characterization in women”, *IEEE 3rd International Work Conference on Bioinspired Intelligence IWOBI 2014*, pp. 21-24.

[11] P. Gómez, R. Fernández, V. Rodellar, V. Nieto, A. Álvarez, L. M. Mazaira, R. Martínez, and J. I. Godino, “Glottal Source Biometrical Signature for Voice Pathology Detection”, *Speech Comm.*, vol. 51, 2009, pp. 759-781.

[12] P. Gómez-Vilda, V. Nieto-Lluis, V. Rodellar-Biarge, A. Álvarez-Marquina, L. M. Mazaira-Fernández, R. Martínez-Olalla, C. Muñoz-Mulas, Mario Fernández-Fernández, Carlos Ramírez-Calvo, “Estimating Tremor in Vocal Fold Biomechanics for Neurological Disease Characterization”, *Proc. of the 18th Int. Conf. on Digital Signal Processing DSP2013*, M1C2.

[13] P. Baranyi, A. Csapó, “Definition and Synergies of Cognitive Infocommunications”, *Acta Pol. Hungarica*, vol. 9, No. 1, 2012, pp. 67-83.

[14] H. Hermansky, “Should Recognizers Have Ears?”, *ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Uncknown Communication Channels*, Pont-à-Mousson, France, 17-18 April, 1997, pp. 1-10.

[15] B. J. Kröger, J. Kannampuzha, E. Kaufmann, “Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception”, EPJ Nonlinear Biomedical Physics, vol. 2, No. 2, 2014, http://www.epjnonlinearbiomedphys.com/content/2/1/2.

[16] P. Gómez, et al., “Bio-inspired broad-class phonetic labeling”, Proc. of 1st Workshop on Cognitive Information Processing, June 9-10, Santorini, Greece, 2008, pp. 221-226.

[17] P. Gómez, J. M. Ferrández, V. Rodellar, "Simulating the phonological auditory cortex: from vowel representation spaces to categories", *Neurocomputing*, vol. 114, 2013, pp. 63-75.

[18] K. López de Ipiña et al., "On Automatic Diagnosis of Alzheimer's Disease based on Spontaneous Speech Analysis and Emotional Temperature", Cognitive Computation, vol. 5, No. 3, 2013, DOI 10.1007/s12559-013-9229-9.

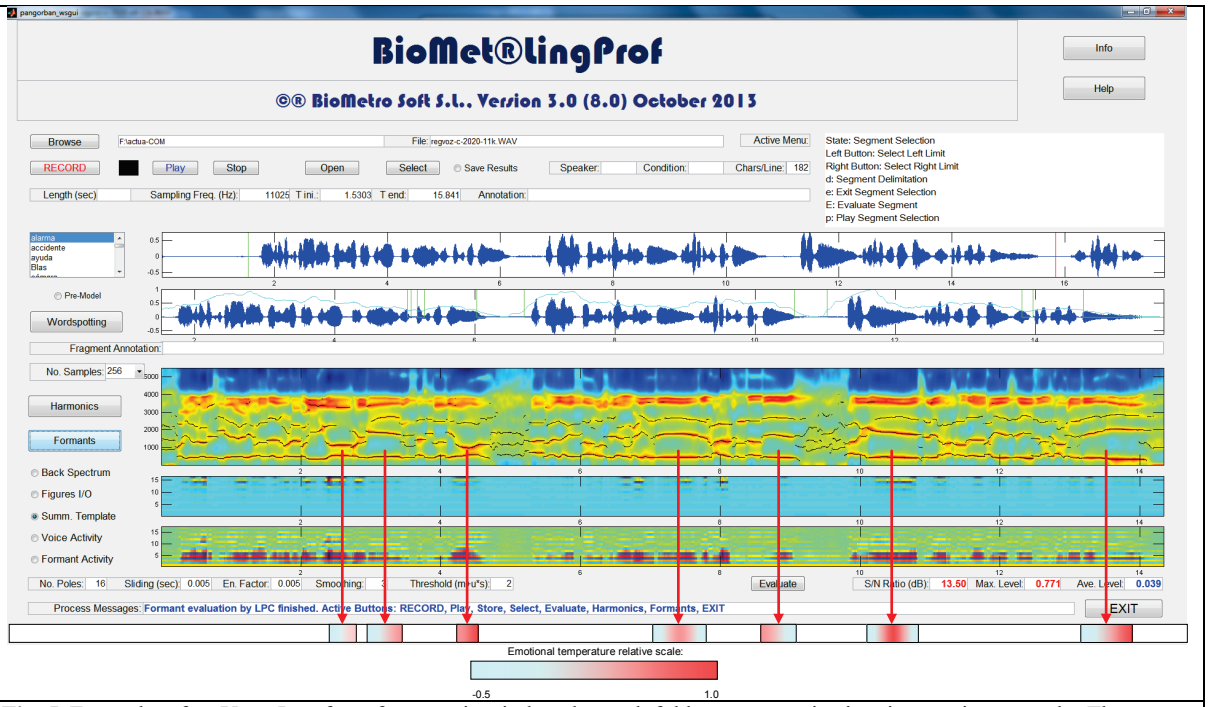

Fig. 5 Example of a User Interface for emotion-induced vocal fold stress monitoring in running speech. The upper graphical window exhibits the running speech segment being analyzed. Immediately below a specific fragment is separated. The formant spectrum from linear predictive inversion is given in the middle graphical window. The segments of stable vowels corresponding to phonated filled pauses correspond with segments where the first two formants are more stable in time (two dark lines at the bottom, from which red arrows are drawn). The fourth and fifth graphical templates give the distributions of the cepstral and LPC coefficients. The stiffness monitoring slider is at the bottom part of the interface, where emotional temperature from emotional stress is plotted blue if under, or red if above the baseline in a normalized scale from [-0.5 - 1.0].