Enhanced videokymographic data analysis based on vocal folds dynamics modeling

Carlo Drioli, Gian Luca Foresti

Department of Mathematics and Computer Science
University of Udine
Via delle Scienze 206,
33100 Udine, Italy
carlo.drioli@uniud.it, gianluca.foresti@uniud.it

10.21437/Interspeech.2015-489

Abstract

The automatic analysis of temporal patterns of vocal folds motion and the tracking of glottal cues such as folds edge position or glottal area, has recently become a topic of interest in the field of laryngeal video imaging. We discuss here the use of a numerically simulated model of the folds motion within a video analysis context, for the analysis of videokymographic data and glottal cues segmentation. The proposed algorithm exploits both visual and acoustic data related to the glottal excitation, to estimate the parameters of the model. The trained model is then used to enhance the analysis and segmentation of visual glottal cues, i.e. the folds edge displacement and glottal area. Objective measures are reported of the accuracy with which the visual glottal cues and the acoustic voice emission are represented by the model. The method is illustrated and assessed on data from different subjects.

Index Terms: Biomechanical Glottal Modeling, Videokymography, Voice Data Analysis, Model Inversion, Video Analysis.

1. Introduction

In the field of acoustic phonetics and speech physiology, video data acquisition and processing became in the last decades an essential tool for medical practical applications such as larynx examination and pathology diagnosis. Visual analysis techniques that are widely used, especially for clinical investigation, include Laryngeal (video) stroboscopy, high-speed videolaringoscopy, videokymography (high-speed line scanning of vocal fold vibrations). The acquisition of visual information about voice production requires that an endoscope is inserted in the mouth or in the nasal cavity to reach the vocal folds. Digital image processing algorithms can provide time patterns of visual cues related to the oscillations of the vocal fold edges for further analysis (vocal folds boundary detection and tracking) [13, 25]. Recently, a video processing based analysis scheme relying on the computation of a set of spatiotemporal geometric features from the glottal area has been proven useful in quantifying and differentiating normal and disordered vocal fold vibrations in adults and in children [26, 27].

Vocal fold vibration consists of a back-and-forth movement, which can be induced and sustained over time, and whose source of energy is a steady stream of air flowing through the glottis[2]. This phenomenon is called flow-induced oscillation. Since the 1970's, a large number of studies addressed the acoustic characterization of the glottal air flow during voiced phonation by accurate modeling of the folds vibration phenomenon [3, 4, 5, 6]. Among these, the lumped-element model proposed in 1972 by Ishizaka & Flanagan [3], in which the folds are represented by two coupled mass-spring oscillating systems, is most representative. To date, the main achievement of the studies on voice source dynamics has been to assist us in understanding the principles of flow-induced oscillatory phenomena and the causes underlying vocal fold pathologies, e.g. [7, 8].

Despite of the wide number of investigations dedicated to the analysis of acoustic data on one side, and of videoendoscopic data on the other, effective analysis schemes exploiting both modalities have been rarely addressed to date. An example is [14], in which vocal fold vibrations were analyzed using a high-speed camera, and related to sound characteristics. Analysis included automatic glottal edge detection and calculation of glottal area variations, as well as kymography.

In a recently published paper [1], we have discussed an approach to phonation modeling that relies on both acoustic and videokymographic data analysis. The information gathered from the audiovisual analysis is used to accurately fit a source-plus-vocal tract model, in which the voice source is represented by a dynamical model of the vocal folds. The videokymographic data in particular is used to improve the parameterization of the source model, by controlling the principal glottal sub-cycle features such as open/closed interval durations.

In this paper, we discuss how the vocal folds modeling based on audiovisual recordings can be exploited to improve the accuracy and robustness of the video analysis. It is shown how the acoustic data related to the glottal excitation, can be used to estimate the parameters of the model and to enhance the video-based glottal cues extraction.

2. Method

The proposed voice modeling method is based on the joint analysis of audio voice recordings and videokymographic data with the aim of extracting relevant glottal cues related to phonation quality. The acoustic pressure recorded at lips is first inverse filtered to remove the effect of vocal tract resonances and provide an estimation of the acoustic glottal source; the videokymographic data, which represent the motion of the vocal folds, is used to gather information on the fold edge displacement, on closure and opening glottal instants and on the duration of closed and open phases.

An example of the analysis data used in this investigation is shown in Figure 1. It reproduces a videokymography (VKG), i.e. a high-speed line scanning of vocal fold vibrations at a given point along the vocal folds length [18, 19]. Looking at the VKG from left to right, each vertical line of the image provides infor-

mation about the displacement of the vocal folds' edges at a new time instant. The video frame has an x-axis resolution of 512 pixels, which at a frame rate of $Fps_v = 25$ frames per second corresponds to 12800 lines recorded per second.

The available acoustic pressure at lips is recorded with a 44.1 kHz sampling rate and 16 bit resolution.

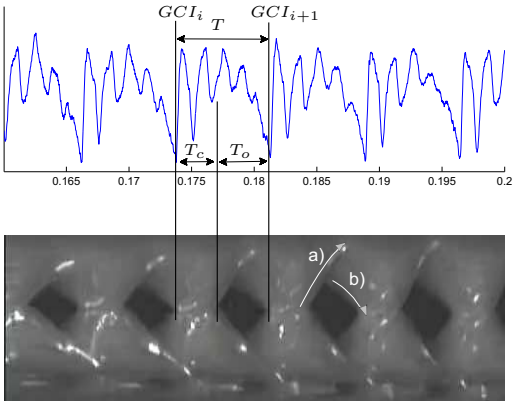

Figure 1: The audio visual data used in this investigation. The acoustic pressure recorded at lips (upper plot) is used to gather information on the vocal tract formants and to provide an estimation of the glottal source by inverse filtering; the video kymograph data (lower plot) provides information on the folds edge displacement and on the duration of closed and open glottal phases. Sub-cycle timing details are highlighted: GCI are glottal closure instants, T is the glottal cycle period, T_c and T_o are the closed and open phase intervals. Arrows indicate motion pattern of fold upper edge (a) and lower edge (b).

The voice modeling scheme is the well established feed-forward source-filter model, in which the lip pressure signal measured by the microphone is given by $y(t) = A(z)\dot{u}_g(t)$, where $A(z)$ is an all-pole filter representing the vocal tract resonances, and $\dot{u}_g(t)$ is the first derivative of $u_g(t)$, the excitation glottal pulse waveform. The voice source model used to represent u_g relies on the mass-spring paradigm adopted, among others, by the well known Ishizaka-Flanagan one-mass and two-mass models. The details of the glottal excitation model can be found in [15].

The lower edge of the folds is represented by a single mass-spring system k, r, m and the propagation of the displacement x along the thickness Th of the fold is represented by a propagation line of length τ. Let x_1 be the displacement of the fold at glottis entrance, and x_2 the displacement at the exit. An impact model reproduces the impact distortions on the fold displacement and adds an offset x_0 (the resting position of the folds). The driving pressure P_m acting on the folds is computed from the lung pressure P_l, the flow u_g and the lower glottal area A_1, using Bernoulli's law: $P_m = P_l - \frac{1}{2}\rho\frac{u_g}{A_1}$ (ρ being the air density). A flow model $\mathcal{F}$ converts the glottis area given by the fold displacements into the airflow at the entrance of the vocal tract. The glottal area is computed as the minimum cross-sectional area between the area at lower vocal fold edge, $A_1 = L \cdot x_1$, and the area at upper vocal fold edge, $A_2 = L \cdot x_2$. The flow is then assumed proportional to the glottal area, i.e. $u_g = \mathcal{F}(x_1, x_2) = k_g \min(x_1, x_2)$ (where the lung pressure P_l is included in k_g). The propagation line of length τ reproduces the vertical phase difference of the vibration of the cord edges, which is essential for the production of self-sustained oscillations without a vocal tract load. We thus assume that $x_2(k) = \xi x_1(k - \tau)$, where ξ takes into account that the amplitude of the fold edge displacement might be non-uniform along the vertical axis (thickness) of the glottis. The pressure lung, P_l, has a role in determining the onset and offset of the oscillation. In our simulations, it is kept constant during the system evolution and is omitted for simplicity in what follows. The mass-spring system k, r, m is modeled as a second-order resonant filter, characterized by a resonance frequency $f_0 = \frac{1}{2\pi}\sqrt{k/m}$.

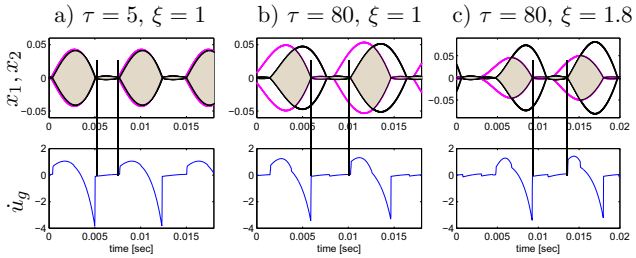

Figure 2: A simulation of the glottal model, for different values of the phase delay parameter τ (in samples): folds edge displacements (upper plots), and glottal source (lower plots). The plots show how the parameter τ and ξ directly affect the closed phase interval of the glottal flow cycle, i.e. the interval in which x_1 or x_2 is in the closed position. Grayed areas correspond to open phase intervals.

This model is able to provide stable oscillatory behaviour in a wide range of parametric configurations of interest and to be suited for applications in which automatic fitting to recorded speech data is involved [15, 16, 17]. Moreover, with respect to traditional multi-mass based glottal models, it has the property that the the phase delay parameter τ directly affects the closed/open phase ratio of the glottal flow waveform, as shown in Figure 2.

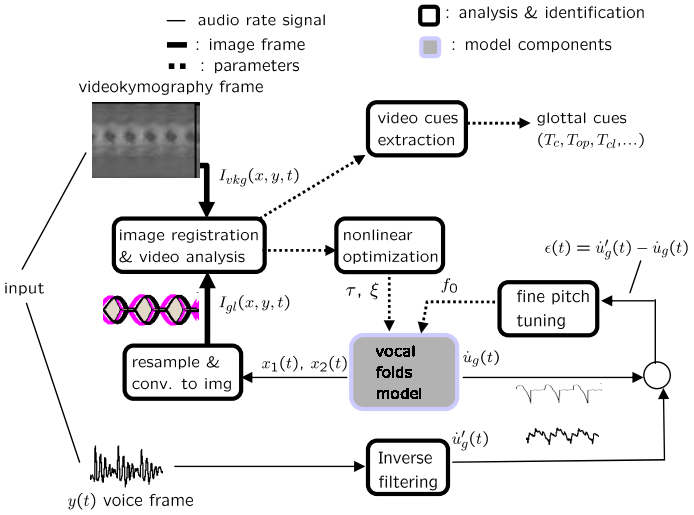

Figure 3: Scheme of the model parameter identification and video analysis procedure.

The glottal source model is fitted to time-varying recorded speech data, by a pitch-synchronous parameter identification

procedure which performs a vocal tract identification and cancellation, followed by a source parameters identification. The vocal folds edge displacement predicted by the model is then used to track the lower and upper edge motion of the vocal folds observed in the VKG video data. The procedure is summarized in Figure 3, and operates on a frame by frame basis through the following steps:

1. the vocal folds model is first tuned to oscillate at the same fundamental frequency of the acoustic signal. This step is accomplished by estimating the fundamental frequency f_0 through a pitch detection algorithm, and by tuning the mass-spring parameters k, r, m of the fold model.

2. the remaining parameters of the model (τ and ξ) are tuned so that the modeled glottal area evolution (grayed areas in Fig. 2) maximally overlap. This step requires that the vocal folds displacement pattern predicted by the model and the videokymographic image frame are aligned through a normalized 2D cross-correlation.

3. The inferior and superior vocal fold edges are tracked on the videokymography frame, and the GCI (glottal closure instants) and the other closed/open phase durations of the glottal cycle are estimated from the videokymography through a video analysis routine.

In the procedure sketched above, the fitting procedure in Step 2 and Step 3 is used to accurately tune also those parameters of the model that principally affect the open phase to close phase duration ratio, i.e. τ and ξ (other parameters are held constant during the simulations). To this purpose, a Levenberg-Marquardt nonlinear least square optimization is used, which searches for the best τ and ξ parameter that minimizes a cost function proportional to the distances between the target and the reproduced glottal area evolution.

The target area evolution in the VKG image $I_{vkg}(x, y, t)$ is the sequence of rhomboid-shaped convex areas (open phase) separated by time intervals corresponding to closed phase segments. The segmentation of these areas of interest in each video frame has been previously addressed by an image analysis procedure which included a FEN based thresholding technique, a denoising step and a fast active contour algorithm, with the aim to obtain a binary image in which the open-phase intervals are clearly represented by uniform and compact rhomboid-shaped regions [1]. Since we now have a reliable indication, provided by the predicted folds edge trajectories, on where the open phase region should lie in the image, the region detection algorithm is now performed only on a neighborhood of the predicted open phase regions (the grayed regions in Fig. 4), resulting in a procedure which is much less sensitive to thresholding and denoising inaccuracies.

Glottal opening and closing instants, GOI's and GCI's, are computed respectively as the leftmost pixel and rightmost pixel of each countour curve (Figure 4), and the closed/open phase durations are computed as $T_{c,i} = GOI_{i+1} - GCI_i$, and $T_{o,i} = GCI_i - GOI_i$.

Closed and open phase time localization and duration are the principal parameters measured by the procedure. The skewness of the rhomboid-shaped regions is potentially interesting as well, since it relates to the degree of left-right asimmetry in the vocal folds oscillation. Here, however, we will adopt a symmetric model of the folds oscillation, and will not take left-right asymmetries into consideration.

3. Results and discussion

The video analysis procedure discussed was assessed on a dataset of sustained phonations from two healthy subjects. The subjects, both males, uttered a sustained vowel (/a/ for S1, and /i/ for S2) for approximately 7 seconds, subject S1 with a fundamental frequency of 130.0 Hz, and subject S2 with a fundamental frquency of 178.6 Hz. The procedure was applied on a total of 30 frames for each subject. In the vocal folds model adaptation algorithm sketched in Sec. 2, part of the parameters are adapted to the acoustic pressure radiated at lips, whereas part of the parameters are tuned using the visual information related to the glottal area function evoution in time. Specifically, the visual related adaptation step is performed using a Levenberg-Marquardt gradient descent optimization method, targeted at reproducing the same closed and open glottis intervals as mesaured from the videokymography frames. The cost function used here in the gradient descent algorithm, referred to a frame of data, is defined as:

$$F(\tau, \xi) = \alpha_1 (T_c^M(\tau, \xi) - T_c^V)^2 + \alpha_2 ||(\mathbf{u_g} - \tilde{\mathbf{u}}_\mathbf{g}(\tau, \xi)||_{L_2} \quad (1)$$

where T_c^M and T_c^V are the closed interval durations from the model and from the video analysis repectively, $\mathbf{u_g} = [u_g(n_i), ..., u_g(n_i + N_{fr})]$ and $\tilde{\mathbf{u}}_\mathbf{g} = [\tilde{u}_g(n_i), ..., \tilde{u}_g(n_i + N_{fr})]$ are the target and reproduced glottal pulse waveforms respectively. The parameters α_1 and α_2 allow to weight the importance of the glottal time parameter term over the acoustic waveform term, and are set both to 0.5 in our experiments. Figure 4 show the result of the adaptation of the folds model with respect to the target waveforms and area parameters.

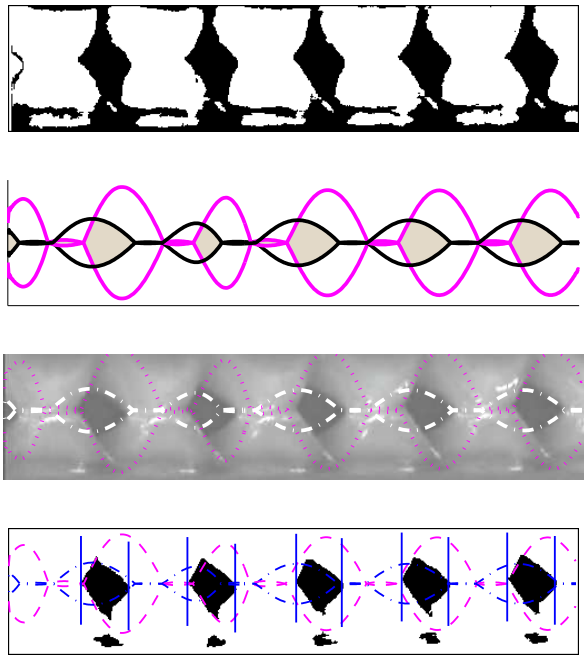

Figure 4: An analysis frame from subject S1 showing the adaptation of the folds model with respect to glottal area time intervals measured from videokymograph image.

The video analysis process is aimed at measuring the principal cues recognized to be relevant for the study of the perceptual influence of the voice source characteristics, and for comparing different voice qualities. Well established voice source

quantification parameters, computed from the flow and the differentiated flow, are usually defined in terms of the time intervals in which air is allowed to flow through the glottis (opening and closing intervals) or not (closed interval), and in terms of flow amplitude [24, 2]. The following set of glottal area time parameters are used here: if T is the glottal cycle period, and $F_0 = 1/T$ the fundamental frequency of oscillation, we call T_c the closed glottis interval, T_{op} the opening interval, T_{cl} the closing interval, and $T_o = T_{op} + T_{cl}$ the open interval. Table 1 reports the values of time-related area function parameters computed from the video data.

Table 1: Time-based parameters (mean values) computed from the video data for subject S1 (male, pitch: 130.0 Hz), and S2 (male, pitch: 178.6 Hz). Parameters reported are $1/T$ (pitch, in Hz), T_c/T (ratio of closed interval to period), T_{op}/T (ratio of opening interval to period), T_{cl}/T (ratio of closing interval to period).

Subj.	$1/T$	T_c/T	T_{op}/T	T_{cl}/T
S1	130.0 Hz	46%	27%	27%
S2	178.6 Hz	46%	18%	36%

4. Conclusions

The use of a non-linear dynamical model of the vocal folds to improve the video analysis and cues extraction from videokymographic data has been discussed. The glottal model adopted allows to accurately control glottal sub-cycle features such as open and closed phase durations, and was used to predict the folds edge motion and thus the glottal area time evolution. A model inversion and video analysis procedure was designed to track the folds motion in the high speed video data, and to extract glottal cues which are not directly observable from lip pressure signals. A joint audio-video parametric identification procedure allows to accurately tune the glottal numerical model, and the superposition of actual and modeled vocal folds edge displacement finally allows to estimate the glottal area and related glottal cues.

Further developments are forseen in terms of model details and tracking procedure. The model used here is intrinsically symmetrical, i.e. it is assumed that left and right folds behave is actually represented. It is often the case that the motion of the left and of the right fold is slightly asymmetrical, even in healthy subjects. An improved representation of the folds motion is possible by explicitally modelling each fold independently.

Also, the fitting of opening and closing time intervals, summing up to the open interval, has not been addressed in this paper. The ratio of these two intervals is considered to be an interesting glottal parameter (speed quotient) to characterize non modal phonation. More in general, the usefulness of the method for the analysis of non-modal voicing modes as observed in emotional speech or in some tonal languages for phonetically contrasting vowels, is to be further investigated.

5. Acknowledgements

We wish to thank Cymo B.V., Groningen, The Netherlands, for kindly providing the acoustic and videokymographic data used in this paper.

6. References

[1] C. Drioli, G. L. Foresti, Accurate glottal model parametrization by integrating audio and high speed endoscopic video data, Signal, Image and Video Processing (2014). In press. DOI: 10.1007/s11760-013-0597-0

[2] K. N. Stevens, Acoustic Phonetics, Current studies in linguistics, The MIT Press, Cambridge, Massachussetts, 1998.

[3] K. Ishizaka, J. L. Flanagan, Synthesis of voiced sounds from a two-mass model of the vocal cords, The Bell Syst. Tech. J. 51 (6) (1972) 1233–1268.

[4] T. Koizumi, S. Taniguchi, S. Hiromitsu, Two-mass models of the vocal cords for natural sounding voice synthesis, J. Acoust. Soc. Am. 82 (4) (1987) 1179–1192.

[5] I. R. Titze, The physics of small-amplitude oscillations of the vocal folds, J. Acoust. Soc. Am. 83 (4) (1988) 1536–1552.

[6] X. Pelorson, A. Hirschberg, R. R. van Hassel, A. P. J. Wijnands, Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model, J. Acoust. Soc. Am. 96 (6) (1994) 3416–3431.

[7] J. C. Lucero, Dynamics of the two-mass model of the vocal folds: Equilibria, bifurcations and oscillation region, J. Acoust. Soc. Am. 94 (1993) 3104–3111.

[8] K. Ishizaka, N. Isshiki, Computer simulation of pathological vocal-cord vibration, The Bell Syst. Tech. J. 60 (1976) 1193–1198.

[9] P. R. Scalassara, C. D. Maciel, R. C. Guido, J. C. Pereira, E. S. Fonseca, A. N. Montagnoli, S. B. Júnior, L. S. Vieira, F. L. Sanchez, Autoregressive decomposition and pole tracking applied to vocal fold nodule signals, Pattern Recogn. Lett. 28 (11) (2007) 1360–1367.

[10] P. Alku, Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, Speech Commun. 11 (2-3) (1992) 109–118.

[11] K. Funaki, Y. Miyanaga, K. Tochinai, Recursive armax speech analysis based on a glottal source model with phase compensation, Signal Processing (3) (1999) 279–295.

[12] P. Rao, A. D. Barman, Speech formant frequency estimation: evaluating a nonstationary analysis method, Signal Processing 80 (8) (2000) 1655–1667.

[13] T. Wittenberg, P. Mergell, M. Tigges, U. Eysholdt, Quantitative characterization of functional voice disorders using motion analysis of highspeed video and modeling, in: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)-Volume 3, ICASSP '97, (1997), pp. 1663–1666.

[14] H. Larsson, S. Hertegrd, P. Lindestad, B. Hammarberg, Vocal fold vibrations: high-speed imaging, kymography, and acoustic analysis: a preliminary report., Laryngoscope 110 (12) (2000) 2117–22.

[15] C. Drioli, A flow waveform-matched low-dimensional glottal model based on physical knowledge, J. Acoust. Soc. Am. 117 (5) (2005) 3184–3195.

[16] C. Drioli, A. Calanca, Voice Processing by Dynamic Glottal Models with Applications to Speech Enhancement, in Proc. of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011) (2011) 1789-1792.

[17] C. Drioli, A. Calanca, Speaker adaptive voice source modeling with applications to speech coding and processing, Computer Speech and Language 28 (5) (2014) 1195-1208.

[18] J. G. Švec, H. K. Schutte, Videokymography: High-speed line scanning of vocal fold vibration, Journal of Voice 10 (2) (1996) 201–205.

[19] Q. Qiu, H. Schutte, A new generation videokymography for routine clinical vocal fold examination, Laryngoscope 116 (10) (2006) 1824–8.

[20] L. Snidaro, G. L. Foresti, Real-time thresholding with euler numbers, Pattern Recognition Letters 24 (9-10) (2003) 1533–1544.

[21] G. Foresti, C. Regazzoni, A hierarchical approach to feature extraction and grouping, Image Processing, IEEE Transactions on 9 (6) (2000) 1056–1074.

[22] P. A. Maragos, R. W. Schafer, M. A. Butt (Eds.), Mathematical morphology and its applications to image and signal processing, Computational imaging and vision, Kluwer Academic, 3rd, Atlanta, Ga., 1996

[23] H. Eviatar, R. L. Somorjai, A fast, simple active contour algorithm for biomedical images, Pattern Recognition Letters 17 (9) (1996) 969–974.

[24] T. Backstrom, P. Alku, E. Vilkman, Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range, Speech and Audio Processing, IEEE Transactions on 10 (3) (2002) 186–192.

[25] M. Döllinger, The Next Step in Voice Assessment: High-Speed Digital Endoscopy and Objective Evaluation, Current Bioinformatics 4 (2) (2009) 101–111.

[26] J. Lohscheller, U. Eysholdt, H. Toy, M. Döllinger, Phonovibrography: Mapping High-Speed Movies of Vocal Fold Vibrations Into 2-D Diagrams for Visualizing and Analyzing the Underlying Laryngeal Dynamics, Medical Imaging, IEEE Transactions on 27 (3) (2008) 300–309.

[27] M. Döllinger, D. Dubrovkiy, R. Patel, Spatiotemporal analysis of vocal fold vibrations between children and adults, Laryngoscope 122 (11) (2012) 2511–2518.