SIViP
DOI 10.1007/s11760-013-0597-0

ORIGINAL PAPER

Accurate glottal model parametrization by integrating audio and high-speed endoscopic video data

Carlo Drioli · Gian Luca Foresti

Received: 17 July 2013 / Accepted: 10 December 2013

Abstract The aim of this paper is to evaluate the effectiveness of using video data for voice source parametrization in the representation of voice production through physical modeling. Laryngeal imaging techniques can be effectively used to obtain vocal fold video sequences and to derive time patterns of relevant glottal cues, such as folds edge position or glottal area. In many physically based numerical models of the vocal folds, these parameters are estimated from the inverse filtered glottal flow waveform, obtained from audio recordings of the sound pressure at lips. However, this model inversion process is often problematic and affected by accuracy and robustness issues. It is here discussed how video analysis of the fold vibration might be effectively coupled to the parametric estimation algorithms based on voice recordings, to improve accuracy and robustness of model inversion.

Keywords Physical glottal modeling · Videokymography · Voice data analysis · Model inversion · Video analysis

1 Introduction

The glottal flow waveform has a fundamental role in the characterization of a speaker's voice. There is experimental evidence that flow waveforms obtained by inverse filtering actual voice recordings are characterized by a wide variety of different shapes and cues. The waveform of the glottal volume velocity is influenced by a number of factors, e.g., the sex and the age of the speaker, the vocal fold health, the style of phonation. Physiological parameters used to control the glottal cycle characteristics include the subglottal pressure, the laryngeal muscles tension, and the resting position of the vocal folds [1]. 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. This phenomenon is called flow-induced oscillation. In the early 1950s and 1960s, the vocal fold oscillation was explained with the myoelastic-aerodynamic theory. According to these theories, Bernoulli forces (negative pressure) cause the vocal folds to be sucked together, creating a closed airspace below the glottis. Continued air pressure from the lungs builds up underneath the closed folds. Once this pressure becomes high enough, the folds are blown outward, thus opening the glottis and releasing a single "puff" of air. Since the 1970s, 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 [2–5]. Among these, the lumped-element model proposed in 1972 by Ishizaka and Flanagan [2], 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., [6,7]. The potentialities of employing source model tracking in conjunction with vocal tract analysis in voice modeling and disorder diagnosis [8] are interesting, yet poorly investigated if compared with other non-dynamical representations of the glottal source [9–11].

On the other hand, 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)

C. Drioli (✉) · G. L. Foresti
Department of Mathematics and Computer Science,
University of Udine, Via delle Scienze 206, 33100 Udine, Italy
e-mail: carlo.drioli@uniud.it

G. L. Foresti
e-mail: gianluca.foresti@uniud.it

stroboscopy, high-speed videolaryngoscopy, and 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 fold boundary detection and tracking) [12,13]. 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 [14,15].

Despite the wide number of investigations dedicated in the analysis of acoustic data on one side and of video endoscopic data on the other, effective analysis schemes exploiting both modalities have been rarely addressed to date. An example is [16], 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 this paper, we illustrate 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 parametrization of the source model, by controlling the principal glottal sub-cycle features such as open/closed interval durations. A pilot experiment is presented in which the method is used on a dataset featuring two different subjects uttering a sustained vowel.

The paper is organized as follows: in Sect. 2, the numerical model of the voice source and the parametrization algorithm, addressing the fitting of visual and acoustic data, is presented. In Sect. 3, the proposed method is assessed on a dataset consisting of a videokymographic plus acoustic recordings of sustained phonation, and the results are discussed. In Sect. 4, the conclusions are presented.

2 Method

The proposed voice modeling method is based on the joint analysis of audio and video data with the aim of inverting a physiologically inspired model representing the dynamics of the vocal folds and the vocal tract resonances. The acoustic pressure recorded at lips is used to gather information on the vocal tract formants and to provide an estimation of the glottal source by inverse filtering; the videokymographic data, providing accurate information on the closure and opening glottal instants and on the duration of closed and open phases, are used to improve the accuracy in the fitting of the glottal model to the acoustic data.

In our modeling scheme, the lip pressure signal measured by the microphone is given by

$$y(t) = -\sum_{k=1}^{N} a_k y(t-k) + \dot{u}_g(t) \qquad (1)$$

where $a_1, \ldots, a_N$ are the auto regressive (AR) coefficients of an all-pole model of the vocal tract, 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, illustrated in Fig. 1, can be found elsewhere [17], and here we only briefly recall the essential components.

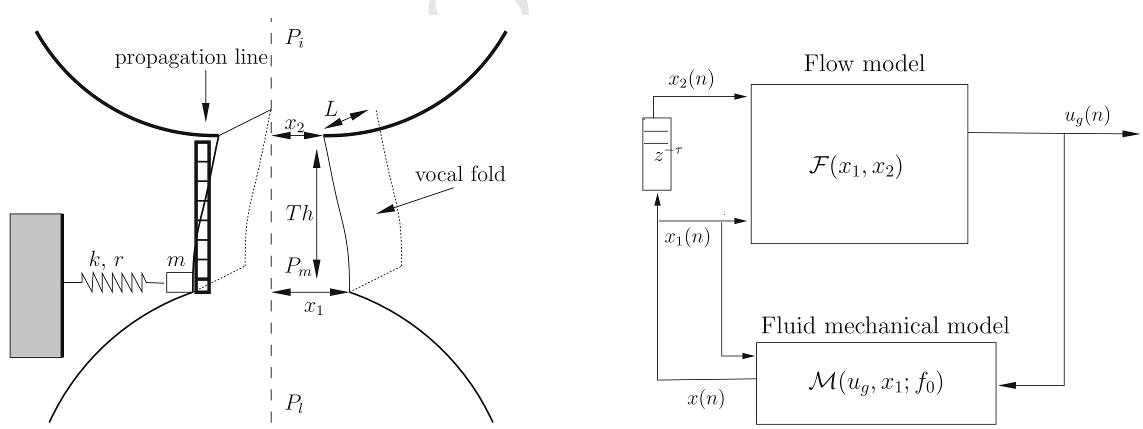

Fig. 1 Scheme of the low-dimensional voice source used as glottal waveform generator (note that the vocal tract model is not represented here). *Left* representation of the vocal folds in terms of a mass-spring system; phase delay between lower and upper edges of the fold are modeled through the propagation of the fold displacement along the thickness of the fold. *Right* the discrete counterpart of the mass-spring model

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). In Fig. 1, the vocal folds and the Bernoulli term are enclosed in the fluid mechanical component $\mathcal{M}$. A flow model $\mathcal{F}$ converts the glottis area given by the fold displacements into the airflow at the entrance of the vocal tract. In its simplest form, the glottis 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. 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 $f_0 = \frac{1}{2\pi}\sqrt{k/m}$.

In previous investigations, this model has shown to provide stable oscillatory behavior in a wide range of parametric configurations of interest [18], and to be suited for applications in which automatic fitting to recorded speech data is involved [17,19]. Moreover, with respect to traditional multi-mass-based glottal models, it has the property that the phase delay parameter τ directly affects the closed/open-phase ratio of the glottal flow waveform, as shown in Fig. 2. This is of particular interest here, since the method that we propose relies especially on the optimization of τ in order to match the closed/open-phase ratio measured from the visual data.

An example of the analysis data used in this investigation is shown in Fig. 3. It reproduces a videokymography, i.e., a high-speed line scanning of vocal fold vibrations in a given point along the vocal folds length [20,21]. Given the video

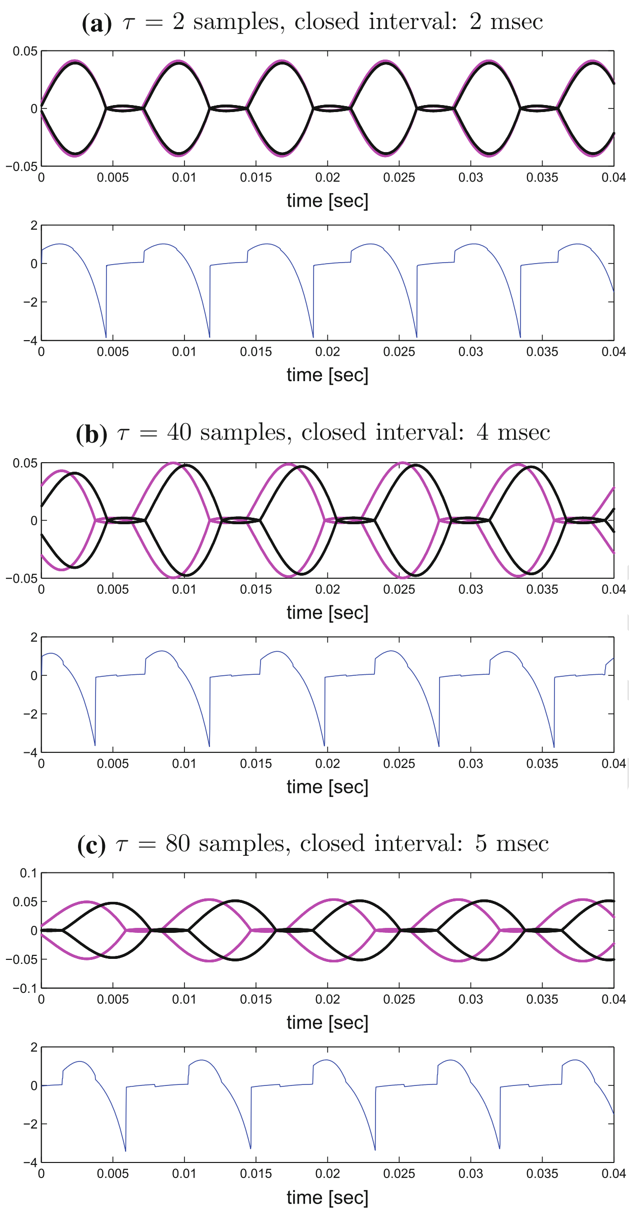

Fig. 2 A simulation of the glottal model, for different values of the phase delay parameter τ (in samples): folds edge displacements (*upper plots*), and glottal source (*lower plots*). The plots show how the phase delay parameter τ directly affects the closed-phase interval of the glottal flow cycle, i.e., the interval in which x_1 or x_2 is in the closed position

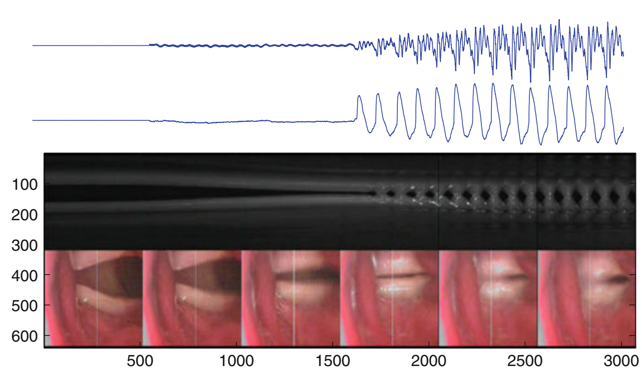

Fig. 3 The audio visual data used in this investigation. The acoustic pressure recorded at lips (*upper plot*) is used to gather information on the vocal tract formants and to provide an estimation of the glottal source by inverse filtering; the video kymograph data (*lower plot*) provides accurate information on the closure and opening glottal instants and on the duration of closed and open phases, which is used in turn to accurately fit the glottal model to the acoustic data

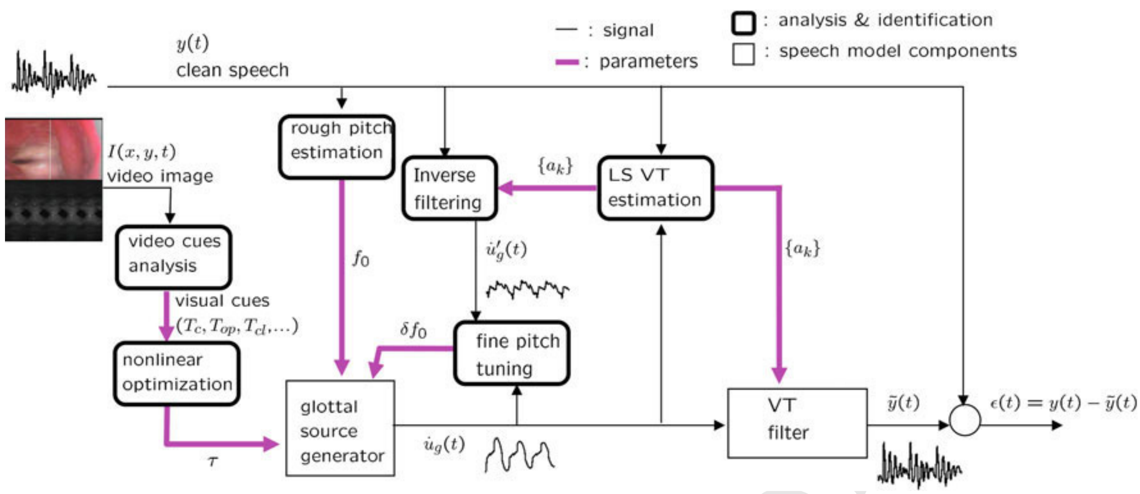

Fig. 4 The pitch-synchronous parameter identification procedure performing a joint source-vocal tract identification

frame rate Fps_v and the image resolution Xres_v along the time axis, the time interval Tpix corresponding to an image pixel can be computed as $\mathrm{Tpix} = (1/\mathrm{Fps}_v)\mathrm{Xres}_v$. In the example shown, the image has an x-axis resolution of 512 pixel at 25 frames per second, resulting in a pixel time Tpix = 0.0781 ms. The available acoustic pressure at lips is recorded with a 44.1 kHz sampling rate and 16 bit resolution.

The voice model (glottal source plus vocal tract) is fitted to time-varying recorded speech data, by a pitch-synchronous parameter identification procedure which performs a joint source-vocal tract identification. The procedure is summarized in Fig. 4 and operates through the following steps:

1. a fixed length running analysis window is shifted by a variable hop size equal to the period length.
2. for the audiovisual analysis frame under investigation, whose length corresponds to around three periods of speech, a traditional LPC analysis is performed on the audio signal to obtain a rough estimate of the vocal tract model parameters a_k, which also represent its principal resonances called formants.
3. the fundamental frequency is estimated through an audio pitch detector (and the analysis of the videokymography); the GCI (glottal closure instants) and the closed/open-phase durations of the glottal cycle are estimated from the videokymography through video analysis routines.
4. the cues computed in the previous step are used to synchronize and tune the mass-spring system representing the folds (through the mass-spring system resonance frequency f_0 and the folds edges delay parameter τ), and the glottal model is used to generate a glottal pulse.
5. a least-square fitting procedure, based on QR factorization, is used to solve the estimation problem which provides the final parameters a_k of the vocal tract filter, given its time aligned input (the glottal source) and output (the target speech signal) time series.

In the procedure sketched above, the cues provided by the video analysis procedure in Step 3 are used in Step 4 to accurately tune those parameters of the model that principally affect the open-phase to close-phase duration ratio, i.e., principally, the phase parameter τ (the vocal fold resting position x_0 and the lung pressure P_l may also affect the glottal cycle, however, the focus in this paper will be on the phase delay control, and the 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 parameter that minimizes a cost function proportional to the distances between target and reproduced open-phase/closed-phase duration ratio.

2.1 Video features extraction

Several cues of the glottal waveform can be extracted from videokymographic data in order to estimate voice source parameters. Glottal opening and closing instants are clearly identified as the left and right corners of the rhomboid-shaped convex regions, denoting the open phase of the glottal cycle. Closed- and open-phase time localization and duration are the principal parameters that will be used here to tune the model fitting. The skewness of the rhomboid-shaped regions is potentially interesting as well, since it relates to the degree of left–right asymmetry in the vocal folds oscillation. Here, we will adopt a symmetrical model of the folds oscillation and will not take left–right asymmetries into consideration.

The input image $I(x, y, t)$ is considered as a set of pixels that belong to one of two regions: rhomboid-shaped convex areas or background. Convex area pixels are those which

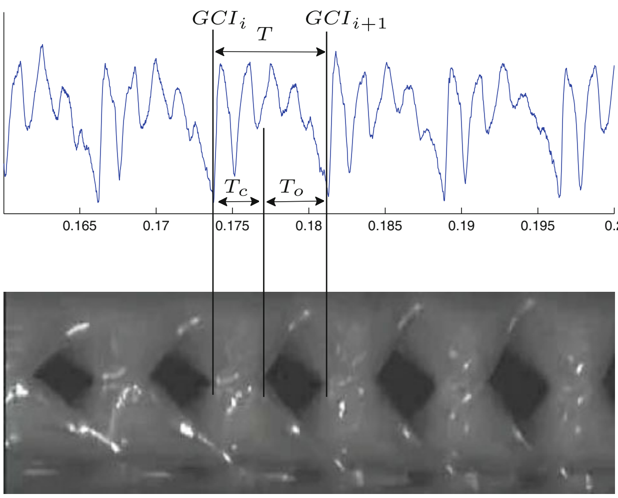

Fig. 5 A frame showing sub-cycle timing details. GCI are glottal closure instants, T is the glottal cycle period, T_c and T_o are the closed- and open-phase intervals

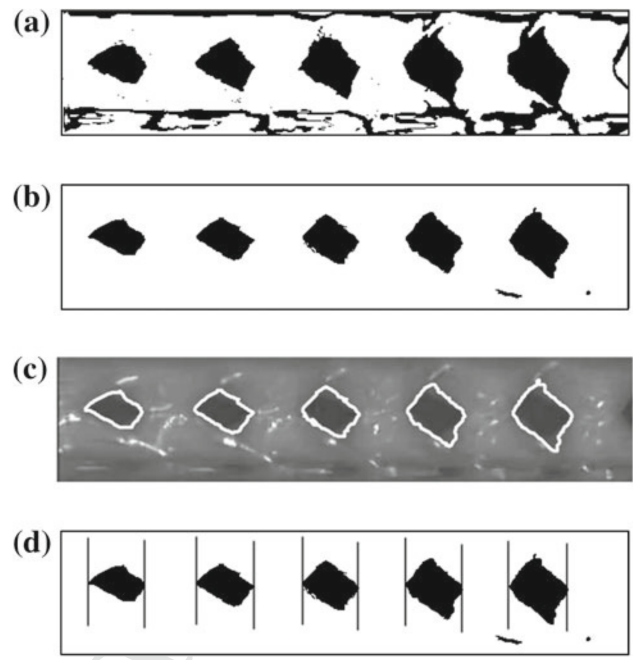

Fig. 6 A frame showing the image processing steps for sub-cycle cues analysis: thresholding for convex regions-background separation (**a**), denoising (**b**), contour detection (**c**), computation of opening and closing instants (**d**)

belong to a region associated with the open phase, while background pixels between two convex areas are associated with the closed phase (see Fig. 5). The first step of the proposed method is to detect figure pixels in each frame of the temporal sequence. There is a wide variety of techniques that could be used for the identification of whether a pixel is part of the figure or the background. For example, a model of the average shape of the figure can be built and an attempt to fit this model to locations within the image can be done. However, model-based identification schemes are computationally intensive and may not be able to complete the detection in real-time. In order to satisfy the real-time constraint and to reach a high level of accuracy, the change detection method based on the fast Euler number (FEN) has been applied [22]. Such a method consists in thresholding the difference image at g different levels, computing the Euler number for each binarization, and choosing the "optimal" threshold value that better separates signal from noise. At the end of this process, a binary image $B(x, y, t)$ is obtained where figure pixels are set to 1 and background pixels are set to 0. The output of this step can be seen in Fig. 6a. However, noise points may appear in $B(x, y, t)$, due to wrong illumination conditions or errors of the FEN method. In practice, isolated points represent noise points, while compact regions of black pixels represent possible regions associated with the open phase. In order to reduce noise and obtain a binary image characterized by uniform and compact regions, a morphological focus of attention mechanism is used [23]. First, a statistical erosion is applied to the binary image $B(x, y, t)$, $B' = B \ominus_{\beta_1} S$, where S is a 3×3 square structuring (SE) element and β_1 is a parameter which regulates statistical operators [23,24]. Then, a statistical dilation is applied to the set B', $B'' = B' \oplus_{\beta_2} S'$, where S' is a cross SE and $\beta_2 > \beta_1$. The resulting denoised video frame is shown in Fig. 6b. Finally, a fast active contour algorithm [25] is applied to detect the contours of the open-phase regions (see Fig. 6d), and each region is approximated with an elliptical shape. Glottal opening and closing instants, GOI's and GCI's, are computed, respectively, as the leftmost pixel and rightmost pixel of each contour curve (Fig. 6d), and the closed/open-phase durations are computed as $T_{\mathrm{c,i}} = \mathrm{GOI}_{\mathrm{i+1}} - \mathrm{GCI}_{\mathrm{i}}$, and $T_{\mathrm{o,i}} = \mathrm{GCI}_{\mathrm{i}} - \mathrm{GOI}_{\mathrm{i}}$.

The procedure discussed so far has been implemented in Matlab as a semi-automatic program. It requires a certain amount of supervision, including the preliminary segmentation of the portion of data to be analyzed, the tuning of the parameters of the numerical model not involved in the adaptation procedure, and the tuning of the video analysis threshold parameters.

3 Results and discussion

In this section, the proposed fitting procedure is assessed on a dataset consisting of a videokymographic plus acoustic recordings of sustained phonation from two healthy subjects. The subjects, both males, uttered a sustained vowel (/a/ for S1, and /i/ for S2) for approximately 7 s, subject S1 with a fundamental frequency of 130.0 Hz, and subject S2 with a fundamental frequency of 178.6 Hz. The procedure was applied on a total of 30 frames for each subject, in the stationary portions of the recordings (voice onsets and offsets were discarded in this investigation).

The video analysis process aimed at measuring the principal cues that could be of interest for the parametrization of a glottal model able to represent the motion of the vocal folds and the fluid dynamics of the airflow passing through the folds and originating the glottal waveform. Some cues of the glottal waveform have been recognized to be particularly 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 [1,26]. We define here a set of glottal area time parameters which are strictly related to the ones used in the literature to characterize the air flow. 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. Also, the following derived parameters are defined: the closed quotient $CQ = T_c/T$, the *opening quotient* $OQ = T_o/T$, the *speed quotient* $SQ = T_{op}/T_{cl}$. Table 1 reports the values of time-related area function parameters computed from the video data, upon segmentation of the visual glottal area cues as illustrated in the video analysis section.

In the speech model adaptation procedure sketched in Sect. 2, part of the parameters adaptation relies on the measure of the acoustic pressure radiated at lips, whereas part of the glottal source parameters are tuned using the visual information related to the glottal area function evolution 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 measured from the videokymography frames. The cost function used here in the gradient descent algorithm, referred to a frame of data, is defined as:

$$F(\tau, f_0) = \alpha_1 (T_c^M(\tau, f_0) - T_c^V)^2 + \alpha_2 ||(\mathbf{y} - \tilde{\mathbf{y}}(\tau, f_0))||_{L_2} \quad (2)$$

where T_c^M and T_c^V are the closed interval durations from the model and from the video analysis respectively, $\mathbf{y} = [y(n_i), \ldots, y(n_i + N_{fr})]$ and $\mathbf{y} = [\tilde{y}(n_i), \ldots, \tilde{y}(n_i + N_{fr})]$ are the target and reproduced speech waveforms, respectively. The parameters α_1 and α_2 allow to weight the importance of the glottal time parameter term over the speech waveform term and are set both to 0.5 in our experiments. The order of the AR filter representing the vocal tract filter was set to 40 (the sampling rate of the audio data being 44,100 Hz). Figures 7 and 8 show the result of the adaptation of the folds

Table 1 Time-based parameters (mean values and standard deviations) computed from the video data for subject S1 (male, pitch: 130.0 Hz), and S2 (male, pitch: 178.6 Hz). Parameters reported are T (period), T_c (closed interval), T_{op} (opening interval), T_{cl} (closing interval), expressed in milliseconds, and OQ (open quotient), SQ (speed quotient)

Subj.	T	T_c	T_{op}	T_{cl}	CQ	SQ
S1	7.4 (130.0 Hz)	3.4	2.0	2.0	0.46	1.0
S2	5.6 (178.6 Hz)	2.6	1.0	2.0	0.46	0.5

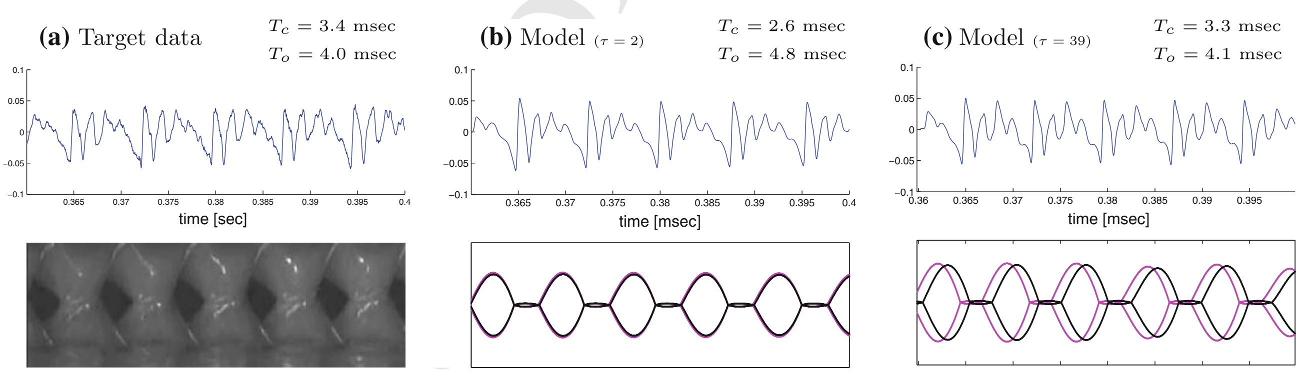

Fig. 7 An analysis frame from subject S1 showing the adaptation of the folds model with respect to glottal area time intervals measured from videokymographic image: **a** shows the acoustic pressure recorded at lips and the videokymographic data and reports the closed and open intervals (T_c and T_o, respectively) estimated by the image analysis process; **b** shows the reproduced lip pressure and the time evolution of the modeled folds (lateral displacement of lower edge and upper delayed edge), when the model is fitted to the acoustic data using a randomly chosen value for the parameter τ, affecting the edges phase difference. An arbitrary value of 2 samples was used for the parameter τ, resulting in a short closed interval and longer open interval; **c** shows the reproduced lip pressure and the time evolution of the modeled folds when the target intervals measured from video are used to tune the parameter τ (reaching the final value of 39 samples)

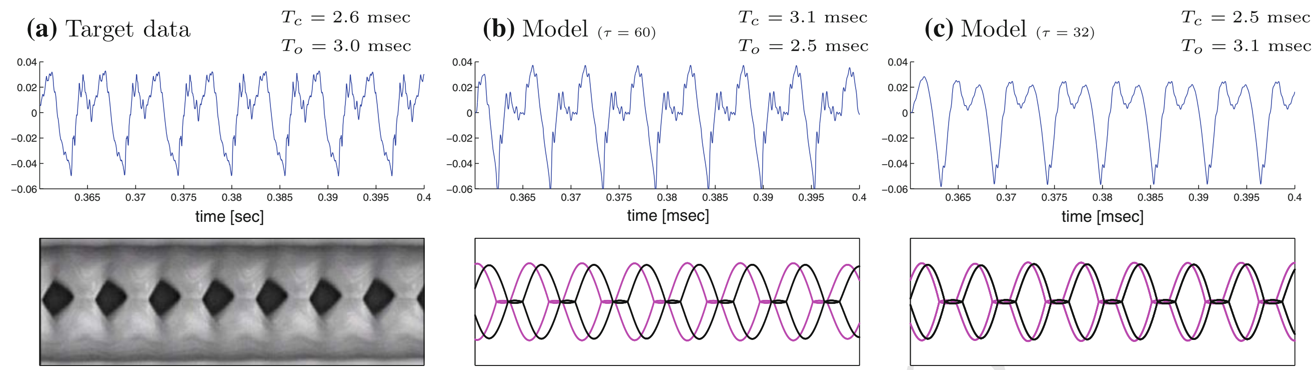

Fig. 8 An analysis frame from subject S2 [plots and parameters are as in Fig. 7. For this subject, the arbitrary value used in the audio-only procedure for τ was 60 samples, providing a wide closed phase (**b**)]. The parameter reached a value of 32 samples upon tuning (**c**)

Table 2 Segmental SNR and IS spectral distance values for distinct modeling settings, referring to audiovisual data from subjects S1 and S2 (average values, calculated over 20 frames for each subject)

Subj.	Glottal area (rel. error)				Speech waveform	
	T (%)	T_c	T_{op} (%)	T_{cl}	SNR	IS
S1	<1	3.1 % (23 %)	18	–	1.8 (0.8)	3.8 (5.0)
S2	<1	3.8 % (19 %)	32	–	2.7 (0.6)	2.9 (8.1)

model with respect to the target waveforms and area parameters. Note that here we only addressed the matching of the open and closed glottis time intervals and did not attempted at matching the correct opening and closing intervals within each open phase. This is because, given the present design of the model, there is no direct relation that links these intervals to one prevailing parameter, as it is the case for the phase delay parameter τ and the closed phase. Most probably, all the parameters of the fluid mechanical model of the folds affect the evolution of the open phase, as well as the interaction with the vocal tract. This issue will be the object of future investigation.

Looking at Figs. 7 and 8, it can be seen that in both cases the fitting procedure which also relies on video analysis (Figs. 7c, 8c) allows to match the closed and open intervals with good approximation. To provide a measure of the acoustic reconstruction quality, two objective measures are adopted: the SNR, defined as the ratio of signal energy over the reconstruction error energy, and Itakura–Saito (IS) spectral distance, a measure of the perceptual difference between the target signal spectrum and the modeled signal spectrum. With respect to the fitting results based only on audio (Figs. 7b, 8b), in which the parameter τ was chosen arbitrarily, the quality measures computed on the speech waveform also are improved for this specific frame: SNR improves from 0.91 to 1,93 for subject S1, and from 0.60 to 2.92 for subject S2; the IS distance decreases from 4.7 to 3.68 for subject S1, and from 6.5 to 0.5 for subject S2. In Table 2, the fitting performances of the proposed model are compared in terms of glottal area time-related parameters, and in terms of SNR and IS distance. Values refers to the average of SNR and IS values, calculated on the acoustic speech signal over a total of 20 analysis frames for each subject (segmental measures). The glottal area-related parameters are expressed as relative errors given by modeled values compared with the target values computed from video: rel_err $= |(T^M - T^V)|/T^V$.

First column refers to the glottal period, and error is below 1 % in both cases; the second column shows the average improvement provided by using video data analysis if compared with audio-only analysis (values in parentheses). The third column reports the values related to the opening interval, although the fitting of opening and closing intervals is not addressed here. The value for subject S1 is around 20 % and does not improve significantly with the video-based analysis. This happens for subject S2 too, for which, however, the error value is rather high, probably because the opening interval is shorter in average than the closing interval, whereas the model shows rather symmetrical opening and closing intervals with the parametric configuration used here.

It is to be stressed that the experiments were conducted on a minimal dataset, due to the limited availability of pre-recorded audiovisual endoscopic data and to the semi-automatic nature of the procedure illustrated. Thus, the improvements documented here cannot be claimed to be statistically significant. Nonetheless, we believe that the outcome of this experiment provides interesting information on the potentials of such a data analysis setting, in which a physiologically motivated model is adapted to both acoustic and video endoscopic data.

Journal: 11760 MS: 0597 TYPESET DISK LE CP Disp.:2013/12/16 Pages: 9 Layout: Large

4 Conclusions

The use of videokymographic data to improve the audio-based parametrization of a nonlinear dynamical model of the vocal folds has been investigated. A low-dimensional glottal model, provided with features which permit to accurately control glottal sub-cycle features such as open- and closed-phase durations, was adopted. A video processing analysis procedure was designed, to extract glottal cues form the high-speed video data, which are not directly observable from lip pressure signals. The video cues were used in a joint audio–video parametric identification procedure, to obtain an accurate tuning of the glottal numerical model. This in turn provides an improved superposition of actual and modeled vocal fold edge displacement and an accurate open phase/closed phase-related glottal cues. It has finally been shown that improved glottal closed/open intervals is also beneficial to the vocal tract parameter identification, resulting in improved speech signal reconstruction error and IS spectral distance.

Given the pilot nature of this investigation and due to the scarce availability of audiovisual videokymographic recordings, the experiments were conducted on a limited amount of data. Future experiments will address the statistical significance of the method by assessing it on a larger number of subjects and on wider spectrum of variables, including gender, age, and phonatory settings.

Further developments are also foreseen in terms of model details and tracking procedure. The model used here is intrinsically symmetrical, i.e., only one fold is actually represented by a moving mass. 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 explicitly modeling each fold independently.

Also, it has been noted that 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. The possibility of accurately matching these cues by extending the proposed procedure will be further investigated.

Acknowledgments We wish to thank Cymo B.V., Groningen, The Netherlands, for kindly providing the acoustic and videokymographic data used in this paper. We also wish to thank the two anonymous reviewers for their valuable comments and suggestions.

References

1. Stevens, K.N.: Acoustic Phonetics, Current Studies in Linguistics. The MIT Press, Cambridge (1998)
2. Ishizaka, K., Flanagan, J.L.: Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell Syst. Tech. J. **51**(6), 1233–1268 (1972)
3. Koizumi, T., Taniguchi, S., Hiromitsu, S.: Two-mass models of the vocal cords for natural sounding voice synthesis. J. Acoust. Soc. Am. **82**(4), 1179–1192 (1987)
4. Titze, I.R.: The physics of small-amplitude oscillations of the vocal folds. J. Acoust. Soc. Am. **83**(4), 1536–1552 (1988)
5. Pelorson, X., Hirschberg, A., van Hassel, R.R., Wijnands, A.P.J.: 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), 3416–3431 (1994)
6. Lucero, J.C.: Dynamics of the two-mass model of the vocal folds: equilibria, bifurcations and oscillation region. J. Acoust. Soc. Am. **94**, 3104–3111 (1993)
7. Ishizaka, K., Isshiki, N.: Computer simulation of pathological vocal-cord vibration. Bell Syst. Tech. J. **60**, 1193–1198 (1976)
8. Scalassara, P.R., Maciel, C.D., Guido, R.C., Pereira, J.C., Fonseca, E.S., Montagnoli, A.N., Júnior, S.B., Vieira, L.S., Sanchez, F.L.: Autoregressive decomposition and pole tracking applied to vocal fold nodule signals. Pattern Recogn. Lett. **28**(11), 1360–1367 (2007)
9. Alku, P.: Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun. **11**(2–3), 109–118 (1992)
10. Funaki, K., Miyanaga, Y., Tochinai, K.: Recursive ARMAX speech analysis based on a glottal source model with phase compensation. Signal Process. **3**, 279–295 (1999)
11. Rao, P., Barman, A.D.: Speech formant frequency estimation: evaluating a nonstationary analysis method. Signal Process. **80**(8), 1655–1667 (2000)
12. Wittenberg, T., Mergell, P., Tigges, M., Eysholdt, U.: 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)-vol. 3, ICASSP'97, pp. 1663–1666 (1997)
13. Döllinger, M.: The next step in voice assessment: high-speed digital endoscopy and objective evaluation. Curr. Bioinform. **4**(2), 101–111 (2009)
14. Lohscheller, J., Eysholdt, U., Toy, H., Döllinger, M.: Phonovibrography: mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics. IEEE Trans. Med. Imaging **27**(3), 300–309 (2008)
15. Döllinger, M., Dubrovkiy, D., Patel, R.: Spatiotemporal analysis of vocal fold vibrations between children and adults. Laryngoscope **122**(11), 2511–2518 (2012)
16. Larsson, H., Hertegård, S., Lindestad, P., Hammarberg, B.: Vocal fold vibrations: high-speed imaging, kymography, and acoustic analysis: a preliminary report. Laryngoscope **110**(12), 2117–22 (2000)
17. Drioli, C.: A flow waveform-matched low-dimensional glottal model based on physical knowledge. J. Acoust. Soc. Am. **117**(5), 3184–3195 (2005)
18. Drioli, C., Avanzini, F.: Non-modal voice synthesis by low-dimensional physical models. In: Proceedings of 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA) (2003)
19. Drioli, C., Calanca, A.: Voice processing by dynamic Glottal models with applications to speech enhancement. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), pp. 1789–1792 (2011)
20. Švec, J.G., Schutte, H.K.: Videokymography: high-speed line scanning of vocal fold vibration. J. Voice **10**(2), 201–205 (1996)

Journal: **11760** MS: **0597** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2013/12/16** Pages: **9** Layout: **Large**

21. Qiu, Q., Schutte, H.: A new generation videokymography for routine clinical vocal fold examination. Laryngoscope **116**(10), 1824–8 (2006)
22. Snidaro, L., Foresti, G.L.: Real-time thresholding with euler numbers. Pattern Recogn. Lett. **24**(9–10), 1533–1544 (2003)
23. Foresti, G., Regazzoni, C.: A hierarchical approach to feature extraction and grouping. IEEE Trans. Image Process. **9**(6), 1056–1074 (2000)
24. Maragos, P.A., Schafer, R.W., Butt, M.A. (eds.): Mathematical Morphology and Its Applications to Image and Signal Processing, Computational Imaging and Vision, 3rd edn. Kluwer, Atlanta (1996)
25. Eviatar, H., Somorjai, R.L.: A fast, simple active contour algorithm for biomedical images. Pattern Recogn. Lett. **17**(9), 969–974 (1996)
26. Backstrom, T., Alku, P., Vilkman, E.: Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range. IEEE Trans. Speech Audio Process. **10**(3), 186–192 (2002)