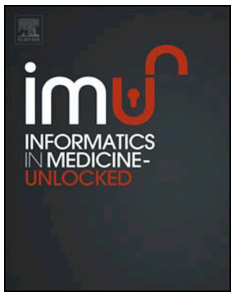

Fitting a biomechanical model of the folds to high-speed video data through bayesian estimation

Carlo Drioli, Gian Luca Foresti

PII: S2352-9148(19)30405-8

DOI: https://doi.org/10.1016/j.imu.2020.100373

Reference: IMU 100373

To appear in: *Informatics in Medicine Unlocked*

Received Date: 3 January 2020

Revised Date: 10 June 2020

Accepted Date: 10 June 2020

Please cite this article as: Drioli C, Foresti GL, Fitting a biomechanical model of the folds to high-speed video data through bayesian estimation, *Informatics in Medicine Unlocked* (2020), doi: https://doi.org/10.1016/j.imu.2020.100373.

Fitting a biomechanical model of the folds to high-speed video data through Bayesian estimation

Carlo Drioli[a,*], Gian Luca Foresti[a]

[a]*Department of Mathematics, Computer Science and Physics, University of Udine, Udine 33100, Italy*

Abstract

High-speed video recording of the vocal folds during sustained phonation has become a widespread diagnostic tool, and the development of imaging techniques able to perform automated tracking and analysis of relevant glottal cues, such as folds edge position or glottal area, is an active research field. In this paper, a vocal folds vibration analysis method based on the processing of visual data through a biomechanical model of the laryngeal dynamics is proposed. The procedure relies on a Bayesian non-stationary estimation of the biomechanical model parameters and state, to fit the folds edge position extracted from the high-speed video endoscopic data. This finely tuned dynamical model is then used as a state transition model in a Bayesian setting, and it allows to obtain a physiologically motivated estimation of upper and lower vocal folds edge position. Based on model prediction, an hypothesis on the lower fold position can be made even in complete fold occlusion conditions occurring during the end of the closed phase and the beginning of the open phase of the glottal cycle. To demonstrate the suitability of the procedure, the method is assessed on a set of audiovisual recordings featuring high-speed video endoscopic data from healthy subjects producing sustained voiced phonation with different laryngeal settings.

Keywords: Vocal folds oscillation analysis, Biomechanical modeling, Image segmentation, Voice quality

*Corresponding author: Tel.: +39-0432-558448; fax: +39-0432-558499;
Email addresses: carlo.drioli@uniud.it (Carlo Drioli), gianluca.foresti@uniud.it (Gian

1. Introduction

In the last decades, visual data recording and analysis techiques have gained a central role in the understanding of phonation and in medical applications such as larynx examination and pathology diagnosis. Among these, laryngeal videostroboscopy, high-speed videolaringoscopy, and videokymography (i.e., the high-speed line scanning of vocal fold vibration) are widely used today for clinical investigation.

Laryngeal videostroboscopy is commonly used in clinical examinations as a tool for visualizing healthy and pathological vocal fold dynamics, although it is known to perform effectively only in case of periodically vibrating vocal folds (Wendler, 1992), while high-speed videolaryngoscopy is a more effective means to visualize asymmetric and nonperiodic vocal fold vibration (Popolo, 2018). In 1996, Švec and Schutte introduced the videokymography (Švec and Schutte, 1996), a low-cost, high-speed imaging method for the examination of the vocal folds, which provides an effective way of visualizing regular and irregular vibration patterns, and whose usefulness in phonation investigation and diagnosis of voice pathologies has been documented in (Schutte et al., 1998; Švec et al., 2007).

With respect to all the aforementioned data acquisition methods, digital image processing algorithms provide the tools for essential preliminary segmentation steps, including vocal folds boundary detection and motion tracking (Turkmen et al., 2015; Osma-Ruiz et al., 2008; Yan et al., 2006), and to subsequently perform the recognition and analysis of time patterns of the visual cues related to the vocal fold edge oscillations (Wittenberg et al., 1997; Tigges et al., 1999; Lohscheller et al., 2008; Chen et al., 2014). Specific video processing issues such as calibration, lighting conditions, image brightness impact are discussed in (Wurzbacher et al., 2008; Kuo et al., 2014; Popolo, 2018), and investigations of voice disorders based on vocal folds edge tracking in high-speed video data are reported in (Lohscheller et al., 2007; Stiglmayr et al., 2008). The importance of image-based methods and their role as surgical guidance and decision making tools in laryngeal surgery is discussed in (Verikas et al., 2007; Schoob et al., 2017; Lin et al., 2018).

In addition to assessment methods based on the sole visual recordings, the possibility of gathering information from both acoustic and visual data (possibly synchronized) has recently been investigated. In (Larsson et al., 2000), vocal fold vibrations were analyzed through high-speed videokymography, allowing the estimation of glottal edge displacement and glottal area variations, and put in relation to acoustic cues computed on the voice emission. Also, some attention has been dedicated to the use of biomechanical models, originally developed to represent the acoustic emission during phonation, paired to the video analytics related to high-speed endoscopic data. In (Döllinger et al., 2002; Schwarz et al., 2008; Döllinger et al., 2017; Murtola et al., 2018), the parameters of a lumped-element biomechanical model of the vocal folds are adapted to replicate the fold vibrations as captured in digital high-speed recordings, and in (Hadwin et al., 2019) the fitting process is applied to a 2D finite element model of the folds; in (Pinheiro et al., 2012), a two-mass model is used to reproduce the glot-

Luca Foresti)

tal area evolution estimated from high-speed endoscopic video of the oscillating folds; in (Drioli and Foresti, 2015a) and (Drioli and Foresti, 2015b), we have approached the vocal fold dynamics modeling and the voice emission at lips through the use of both videokymographic and acoustic data; in (Schwarz et al., 2006), classification of a specific voice pathology (i.e., unilateral vocal fold paralysis) is addressed by an inversion procedure which tunes the parameters of a biomechanical model of the vocal folds to reproduce the irregular vocal fold oscillations, and in (Díaz-Cádiz et al., 2019) an impact model fitted to video data by vocal fold edge tracking is used to predict the contact force during the collision of the folds.

A relevant number of investigations were also dedicated to the estimation of the model parameters using a probabilistic framework. Bayesian estimation of the parameters of a lumped mass model from real and acoustic observations was investigated in (Cataldo et al., 2009, 2013), in the hypothesis of parameter stationarity. In (Hadwin et al., 2016; Hadwin and Peterson, 2017), the same Bayesian estimation method was extended to non-stationary parameter estimation based on particle filtering and on extended Kalman filtering, in which the observation data was the glottal area waveform simulated from a dynamical glottal model. Finally, in (Deng et al., 2019) the Bayesian framework is also investigated with respect to simulated videondoscopic data, focusing however on the effect that different video measurement parameters such as frame rate, resolution and viewing angle, have on the model parameter estimation.

It is worth noting that biomechanical models of the vocal folds during sustained phonation where first designed in the 1970's with the aim of understanding and representing phonation from an acoustical point of view. At that time, such models were obviously not intended to represent visual patterns from high-speed recordings. A considerable research effort was dedicated to replicate the flow induced vocal folds oscillation during voiced sound production, and the underlying mechanical and aerodynamic phenomena have been investigated by accurate numerical modelling of the folds vibration in (Ishizaka and Flanagan, 1972; Koizumi et al., 1987; Titze, 1988; Lucero, 1993). The studies on voice source have been topical for the understanding of the principles of flow-induced oscillatory phenomena, as well as for studying and understanding vocal fold pathologies (e.g., (Ishizaka and Isshiki, 1976; Neubauer et al., 2001; Tao and Jiang, 2008)). Physical models have been employed for speaker recognition and speech synthesis as well, although today their use for these purposes seems to be marginal. On the other hand, they now appear to be interesting tools for the processing and the automatic interpretation of laryngeal visual data, since today high-speed digital video recording facilities with sufficiently high time and image resolution are becoming more and more accessible.

In this paper, a high-speed endoscopic video analysis method is proposed, which is based on the fitting of a biomechanical model to real endoscopic visual data. With respect to other video analysis methods specifically designed for the processing of high-speed endoscopic data of the folds, the one presented here investigates the possibility of using a biomechanical model in a Bayesian setting to fit the position of the edge of the folds during the glottal open phase, and to use in turn tuned model to predict the observation in the next analysis window. It is argued that the method can be also used to further infer the position of vocal fold edge position in those intervals of the glottal cycle in which no observation data is available due to visual occlusion, although

the assessment of this feature will be the object of future investigation. The fitting algorithm relies on a biomechanical model whose parameters are adapted so that his time evolution is coherent with the folds edge position estimated from the high-speed video endoscopic data. The dynamical model is then used for the Bayesian inference as a state transition model, with a dual role: on one side, it models the folds edge motion to compute the likelyhood of their position in given portions of the glottal cycle; on the other side, its parameters are finely tuned to maximize the likelyhood of the visual observations. The method is assessed on a set of recordings featuring high-speed video endoscopic data from healthy subjects uttering sustained vowels. It is shown that the use of a biomechanical model of the folds as a state transition model permits to accurately fit the upper and lower vocal fold edges during the intervals in which both are visible, and to infer their position in complete fold occlusion conditions occurring during the end of the closed phase and the beginning of the open phase of the glottal cycle.

With respect to previous literature dealing with Bayesian parameter estimation we highlight the following differences: 1. in (Hadwin et al., 2016), the model is only assessed on simulated visual data, whereas here the fitting process is designed to deal with real HSV data. Using real data implies that visual artifacts must be taken into account that cannot be modelled as additive state and observation gaussian noise, such as the time-varying glottis-camera alignment offset introduced by small movements of the endoscope; 2. in (Cataldo et al., 2009, 2013; Hadwin et al., 2016; Hadwin and Peterson, 2017), both the state and the parameters of the model are estimated at sampling rate. In our proposal, instead, the model parameter estimates are updated at each other glottal cycle, whereas the state estimates are computed at sample rate and used to compute the likelyhood of the visual observations. This leads to higher computational efficiency without significantly reducing the parameter estimation effectiveness, as the physiological parameters taken into consideration do not change considerably within one cycle; 3. in the model used in this study, each fold is governed by his own state variables to allow the simulation of L-R asymmetric oscillations, whereas a symmetric model is used in (Hadwin et al., 2016; Deng et al., 2019).

2. Proposed method

The video analysis procedure under investigation is aimed at exploiting the motion of the vocal folds from a high-speed video sequence $I(x, y, t)$ in which the vocal fold vibration is captured from a top-view position. Fold motion is defined as the time-varying distance of the vocal folds edge from the glottal axis, taken at the half way from anterior to posterior glottal endings (the glottis being the opening between the opposing vocal folds). Figure 1 illustrates the schematic representation of a laryngeal high-speed videoendoscopy recording, the videokymogram (VKG) corresponding to the dashed line reported in the upper figure, and the interpretation of the data by means of a two-mass lumped model of the vocal cord edges.

Figure 2 illustrates the interpretation of the VKG patterns with relation to the upper and lower edges of the vocal fold displacement. The figure refers to two glottal cycles. The romboid regions correspond to time intervals in which both the lower and the upper edge of either folds are deflected, allowing air to pass through the open glottis. Time

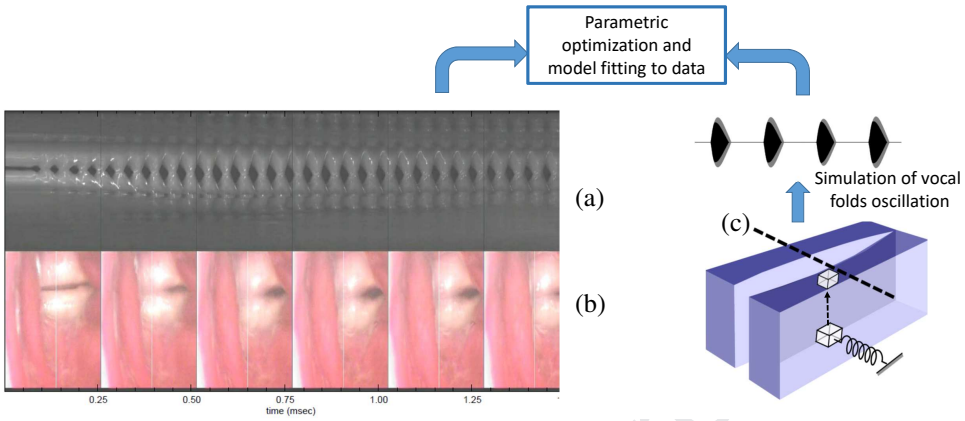

Figure 1: (a) the videokymogram related to the laryngeal high-speed videoendoscopy recording (b) (the two are not synchronized), and (c) the interpretation of the data by means of a two-mass lumped model of the vocal cord edges.

intervals between two romboidal-shaped regions corresponds to the closed phase of the glottal cycle, either because both upper and lower edges are in the closed position, or because just one edge is closed while the other is opening or closing.

Figure 3 shows the interpretation of a fragment of videokymographic data, corresponding to five glottal cycles, in terms of vocal folds edge analysis. It can be seen how actual recorded data is characterized by clearly distinguishable romboidal-shaped regions related to the open phase, but it provide barely visible information concerning the upper folds edge position during the closing interval and no information at all concerning the lower folds edge position during the opening interval (due to camera occlusion). Moreover, the VKG data is often characterized by asymmetries with respect to the L/R direction.

2.1. Pitch-synchronous joint parameter and state estimation of the model

We define the objective of fitting the model to the visual observation as the joint model parameter estimation and model state estimation as follows. Based on the partial visual information on the vocal folds edge available in a glottal cycle and on past estimations, estimate the new set of model parameters and fold edge position (i.e., the model state) within the whole glottal cycle. The process is then repeated for each other glottal cycle. Let us restrict the analysis on a given position along the glottal axis. In other terms, we interpret the video sequence as a videokymography defined at a given point along the glottal axis, and we formulate the analysis problem as the fitting of the model to the fold edge in this restricted region. If $\mathbf{z}_{1:k}$ is the set of observations up to time instant k, $\mathbf{x}_k$ is the state of the fold edge at time instant k, and θ_k is the set of parameters at time k, then we are interested in the computation of the posterior probability $p(\mathbf{x}_k, \theta_k|\mathbf{z}_{1:k})$. This probability can be recursively computed as

$$p(\mathbf{x}_k, \theta_k|\mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k|\mathbf{x}_k, \theta_k)p(\mathbf{x}_k|\theta_k, \mathbf{z}_{1:k-1})p(\theta_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \quad (1)$$

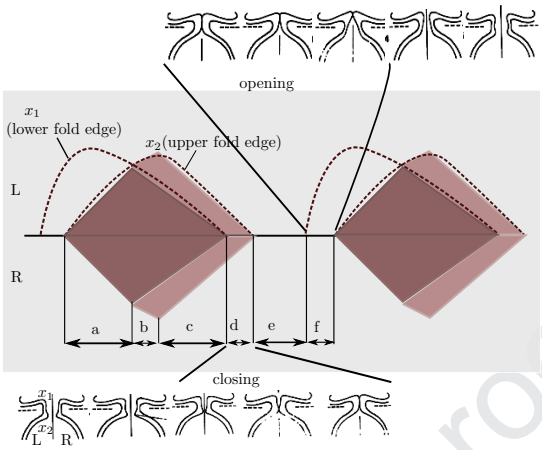

Figure 2: Scheme of the VKG patterns interpretation: (a) opening phase; (b,c) closing phase; (d) lower edge has closed, while upper edge is closing; (e) both lower and upper edges are closed; (f) lower edge is opening while upper edge is still closed (since video recording is from above, lower edge displacement is occluded in this interval).

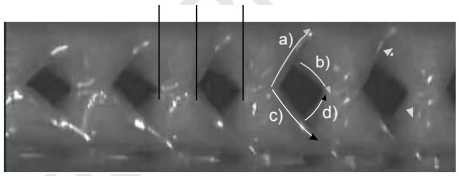

Figure 3: Vocal fold edge classification in a videokymography analysis fragment: (a) left upper edge, opening phase; (b) left lower edge, closing phase; (c) right upper edge, opening phase; (d) right lower edge, closing phase.

where $p(\mathbf{z}_k|\mathbf{x}_k, \theta_k)$ is the likelyhood probability, $p(\mathbf{x}_k|\theta_k, \mathbf{z}_{1:k-1})$ is the state prior, $p(\theta_k|\mathbf{z}_{1:k-1})$ is the parameter set prior, and $p(\mathbf{z}_k|\mathbf{z}_{1:k-1})$ is the marginal likelyhood. Since it is $p(\mathbf{x}_k|\theta_k, \mathbf{z}_{1:k-1})p(\theta_k|\mathbf{z}_{1:k-1}) = p(\mathbf{x}_k, \theta_k|\mathbf{z}_{1:k-1})$, joint parameter and state estimation can be achieved through augmentation of the state space by the parameter vector (Särkkä, 2013). Assuming that the posterior pdf is available at time $k-1$, the prior (or prediction) pdf can be computed as

$$p(\mathbf{x}_k, \theta_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k, \theta_k|\mathbf{x}_{k-1}, \theta_{k-1})p(\mathbf{x}_{k-1}, \theta_{k-1}|\mathbf{z}_{1:k-1})d\mathbf{x}_{k-1}, \theta_{k-1} \tag{2}$$

Note that the temporal prior pdf $p(\mathbf{x}_k, \theta_k|\mathbf{x}_{k-1}, \theta_{k-1})$ provides an estimate of the update of the state and parameters at time k, given the state and parameters at time $k-1$, in other words it models the dynamics of the process under observation. In general, there are various way to model such probability, depending on the type of problem to solve. A common solution, which only loosely makes use of the knowledge of the underlying dynamics, is to adopt a possibly low order linear dynamical system with Gaussian noise (Arulampalam et al., 2002). Other more specialized choises make use of some amount

of knowledge about the process, e.g. physical Newtonian simulation can be used in probabilistic motion prediction and tracking of objects or persons in a scene (Vondrak et al., 2008). In the specific case discussed here, we propose to adopt a biomechanical numerical model of the vocal folds as state transition model, and assume that the state vector $\mathbf{x}_k$ is the displacement of the vocal fold as predicted by the numerical simulation of the model. For the update of the parameters, a pitch-synchronous random walk model is assumed, i.e.

$$\theta_{Tk} = \theta_{Tk-1} + \phi_k \tag{3}$$

where $\phi_k \in N(0, W_\phi)$ satisfies a Gaussian distribution with zero mean and covariance matrix W_ϕ. The parameters are thus assumed constant during a glottal cycle. Note that the optimization process of the parameters can be very sensitive to the initial hypothesis and to the variance of the parameter. An advantage of using a physically informed model in the process is that often a starting hypothesis can be done on a physiological basis (see, e.g. (Drioli, 2005) for a discussion on the empirical tuning of these parameters.

2.2. Biomechanical numerical model

The biomechanical model of the vocal folds is a lumped-elements representation in which the lower edge of each fold is modeled by a single mass-spring-damper system with stiffness k, damping r and mass m, and the phase difference of the vibration between the lower and the upper edge, which is essential for the modeling of self-sustained oscillations, is modeled by a delay of the displacement induced by its propagation along the cover of the fold (Titze, 1988; Drioli, 2005). The scheme of the model is illustrated in Fig. 4. Let us call $x_{1,l}$ the displacement of the left fold at the entrance of the glottis (lower edge), and $x_{2,l}$ the displacement at the exit (upper edge). The displacements of the right fold are named accordingly $x_{1,r}$ and $x_{2,r}$. The distortions on the folds during the mutual impact is represented by an impact model f_X, and the offsets $x_{0,l}$ and $x_{0,r}$ represent the resting positions of the folds. The driving pressure P_m acting on the folds is computed from the flow U_g and the lower glottal area $a_1 = Lx_{1,l} + Lx_{1,r}$ using Bernoulli's law (L is the length of the folds). The total glottal area is then computed as the minimum cross-sectional area between the area $a_1 = Lx_{1,l} + Lx_{1,r}$ at lower vocal fold edge and the area $a_2 = Lx_{2,l} + Lx_{2,r}$ at upper vocal fold edge, and the flow U_g is finally assumed proportional to the total glottal area (this flow model is referred to as f_U in the following). The propagation of the displacement x along the thickness T of the folds is represented by a propagation delay line of length $\tau = TF_s/c_f$ samples, where F_s is the sampling rate and c_f is the propagation velocity on the cover of the fold (in what follows we assume that right and left folds have equal thickness T and length L). The propagation line is an approximation of the fold edge displacement along its vertical axis (thickness), and models the vibration phase differences between the lower and the upper edges of the cords, which is an essential cue of the glottal cycle. Moreover, to also account for the fact that the amplitude of the fold edge displacement might be non-uniform along the vertical axis, we assume, for the left fold, that $x_{2,l}(k) = \xi_l x_{1,l}(k - \tau)$, where ξ_l is a gain factor (in other terms, $x_{2,l}$ is obtained from $x_{1,l}$ through a filter with transfer function $h_{\xi,l}(z) = \xi_l z^{-\tau}$). Similarly, the gain factor for

the right fold displacement along the vertical axis is called ξ_r. The whole system can be described by the following set of continuous-time equations.

$$\begin{cases} m_\alpha \ddot{x}_\alpha(t) + r_\alpha \dot{x}_\alpha(t) + k_\alpha x_\alpha(t) = F_m(t) \\ F_m(t) = P_m(t) \cdot S_m \\ P_m(t) = P_l - \frac{1}{2}\rho \frac{U_g(t)^2}{(Lx_{1,l}(t)+Lx_{1,r}(t))^2} \\ x_{1,\alpha}(t) = f_X(x_\alpha(t), x_{01,\alpha}) \\ \qquad = \begin{cases} x(t) + x_{01,\alpha} & \text{if } x(t) + x_{01,\alpha} > 0 \\ 0 & \text{otherwise} \end{cases} \\ x_{2,\alpha}(t) = \xi_\alpha f_X(x_\alpha(t) - \frac{T}{c_f}\dot{x}_\alpha(t), x_{02,\alpha}) \\ U_g(t) = f_U(P_l, a_1(t), a_2(t)) \\ \qquad = \sqrt{\frac{2P_l}{\rho}} \min\{a_1(t), a_2(t)\} \end{cases} \tag{4}$$

where ρ is the air density, and S_m and L are respectively the fold surface and length. We used the index α to distinguish between the left ($\alpha = l$) and the right ($\alpha = r$) portion of the vocal fold.

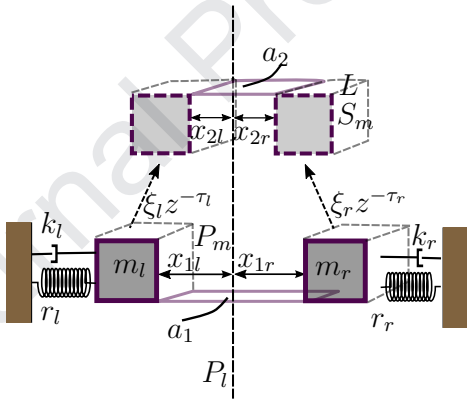

Figure 4: The biomechanical model of the vocal folds.

The discretization of the equations in (4) leads to a discrete-time system that can be numerically solved to obtain an estimate of the glottal flow $U_g(kT_s)$ and of the folds displacements $x_1(kT_s)$ and $x_2(kT_s)$ at discrete time k, with $T_s = 1/F_s$ being the sampling interval (Drioli, 2005).

The biomechanical model is able to effectively reproduce the self sustained oscillations of the vocal folds and can be used as a glottal waveform generator. The natural frequency of a mass-spring system is $f_0 = 1/2\pi\sqrt{k/m}$, thus its parameters k and m can be tuned accordingly when a given oscillation period of the model is desired. However, note that the resulting (closed loop) observed vibration frequency may happen to be different from f_0, due to coupling of the vocal folds via the airstream. The dynamical scheme, however, can also serve as a signal predictor at instants $k, k+1, ...$, given that it fits well the data observed at previous instants $1, \cdots, k-1$. Thus, to correctly pre-

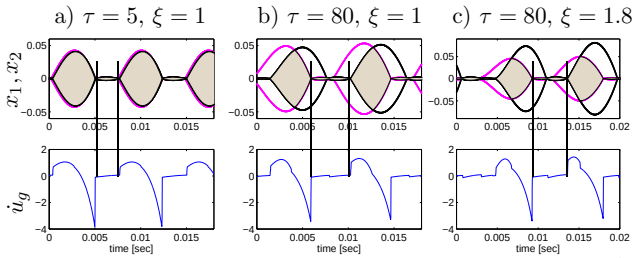

Figure 5: Simulations of the vocal folds vibration through the illustrated glottal model, for different values of the phase delay parameter τ (in samples): folds edge displacements (upper plots), and glottal source (lower plots). The output of a left-right symmetric model is shown here (i.e., $\tau = \tau_l = \tau_r$, and $\xi = \xi_l = \xi_r$). The plots show how the parameters τ and ξ can be put in relation with 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. The areas depicted in grey correspond to open phase intervals.

dict the state of the system at time k, the model needs to be previously tuned to behave coherently with the observed data. We don't further insist here on the model parameters tuning, since this topic has been extensively discussed in (Drioli, 2005; Drioli and Calanca, 2014). Figure 5 shows the numerical simulation of the vocal folds vibrations through the model discussed, for different values of the parameters τ and ξ.

A closed-form solution of eq. 1 and eq. 2 is in general not feasible, and a numerical approximation is often sought instead. We propose here the use of a Particle Filtering scheme (PF), with a Sequential Importance Resampling algorithm (SIR) to represent the posterior (Arulampalam et al., 2002; Vermaak et al., 2002; Vondrak et al., 2008; Dore et al., 2010). The underline principle is to form a weighted particle representation of the posterior distribution, as $p(\mathbf{x}_k, \theta_k|\mathbf{z}_{1:k}) \approx \sum_i w^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)})$, where $\{(w_k^{(i)}, x_k^{(i)}), i = 1, \ldots, N\}$ is the set of particles and of the corresponding weigths at instant k. A scheme of the Bayesian tracking algorithm is reported in Algorithm 1.

The biomechanical model is involved in the prediction step, where each particle can be considered as an independent instance of the model simulation. In the following, we will include in the estimation proces three model parameters for each fold, i.e. the natural frequency f_α, the vertical phase delay τ_α, and the upper-to-lower edge amplitude ratio ξ_α. Hence the parameter vector is $\theta = \{f_l, f_r, \tau_l, \tau_r, \xi_l, \xi_r\}$.

2.3. Likelyhood

A likelihood function should provide a reliable measure of how well an image observation $I(x, y, k)$ is explained by a particular hypothesis. If we suppose that a set of video features $\mathbf{f}(I(x, y, k))$ related to the folds edge can computed from the image frame, then we can define the likelyhood $p(\mathbf{z}_k|\mathbf{x}_k)$ at discrete instant k as

$$p(z_k|\mathbf{x}_k) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{|\mathbf{f}(I(x, y, k)) - \mathbf{x}_k|^2}{2\sigma^2}) \tag{5}$$

The issue then is to compute a set of features that can be related to the observable state of the folds model, i.e. the lower and upper edge of both left and right vocal folds. This information is only partially contained in the video endoscopic imaging data, as

Algorithm 1 Bayesian estimation algorithm

Initialization: Draw N samples $\{\mathbf{x}_0^{(i)}, \theta^{(i)}\}$ from the prior $p(\mathbf{x}_0, \theta_0)$ and set $\mathbf{w}_0^{(i)} = 1/N$
while $k < K$ **do**
  Prediction: Draw N new samples $\mathbf{x}_k^{(i)}$ from the temporal prior $p(\mathbf{x}_k^{(i)}|\mathbf{x}_{0:k-1}^{(i)}, \theta_{Tk-1}^{(i)})$
  Draw new samples $\theta_{Tk}^{(i)}$ through the random walk model $\theta_{Tk}^{(i)} = \theta_{Tk-1}^{(i)} + \phi_k$
  Likelyhood: Compute $p(\mathbf{z}_k^{(i)}|\mathbf{x}_k^{(i)}, \theta^{(i)})$
  Update: Calculate new weights according to $\mathbf{w}_k^{(i)} \propto \mathbf{w}_{k-1}^{(i)} \frac{p(\mathbf{z}_k^{(i)}|\mathbf{x}_k^{(i)}) p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{p(\mathbf{x}_k^{(i)}|\mathbf{x}_{0:k-1}^{(i)}, \mathbf{y}_{1:k}^{(i)})}$
  Resample: Generate a new set of N samples $\{\mathbf{x}^{(i)}, \theta^{(i)}\}$ by sampling each $\{\mathbf{x}^{(l)}, \theta^{(l)}\}$ with probability $\mathbf{w}_k^{(l)}$
  $k \leftarrow k + 1$
end while

discussed before. In the next sections we will illustrate how to design an ad-hoc likelyhood function that effectively uses the incomplete information available. The use of such function within the particle filter framework will allow to fit the folds displacement in the regions where features can be computed from the available information, and to provide an estimation of the position based on the prediction of the model in those time intervals in which information is missing.

3. Video data processing and folds-specific likelyhood function

The computation of the likelyhood related to the visual data under observation requires to go through a number of subsequent analysis steps, which include a preliminary video processing stage, the extraction of the visual cues related to the target motion, and the computation of the likelyhood function of choice. In what follows, these steps as well as the design of a problem-specific likelyhood function will be described in detail.

3.1. Preliminary video processing

Each input image $I(x, y, t)$ might contain one or more glottal cycles, in each of which an open glottis interval can be distinguished as a rhomboid-shaped convex area. The pixels of the image are thus classified as belonging to either a rhomboid-shaped convex area, i.e. an open phase, or to the time interval beetween two convex areas, i.e. a closed phase interval.

The video analysis first aims at detecting all open phase pixels in each frame of the video temporal sequence by a change detection method based on the fast Euler number (FEN) (Snidaro and Foresti, 2003). This procedure returns a binary image $B(x, y, t)$ in which open phase pixels are set to 1 and background pixels are set to 0.

Since noise may still affect the binary image $B(x, y, t)$, a further processing step based on a morphological focus of attention mechanism is performed (Foresti and Regazzoni, 2000), which operates in two steps: first, a statistical erosion is applied to the binary image $B(x, y, t)$, $B' = B \ominus_{\beta_1} S$, with S being a square structuring element and β_1 being a parameter which regulates the statistics of the operators (Foresti and Regazzoni, 2000; Maragos et al., 1996); secondly, a statistical dilation is applied to the set B', $B'' = B' \oplus_{\beta_2} S'$, with S' being a cross structuring element and $\beta_2 > \beta_1$. Finally, the contours cnt_i of the open phase regions is detected as the pixels where the vertical gradient assumes maximum values (upper semi-contour, corresponding to the left cord) and minimum values (lower semi-contour, corresponding to the right cord). The resulting denoised video frame and the open phase regions contours are shown in Fig. 6, lower panel. Finally, glottal opening and closing instants t_{GO}'s and t_{GC}'s can be estimated as the leftmost pixel and rightmost pixel of each countour curve, and closed/open phase durations can be estimated as $T_{c,i} = t_{GO,i+1} - t_{GC,i}$ and $T_{o,i} = t_{GC,i} - t_{GO,i}$ respectively.

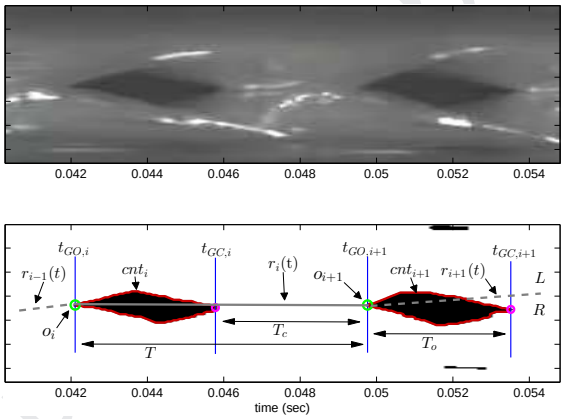

Figure 6: Image data preprocessing, glottal cycle segmentation and piecewise linear trend computation. Upper panel: video frame analysis; lower panel: video frame after thresholding, denoising and open phase regions contouring. Lower panel also illustrates the picewise linear trend identification,

At the same time, information on DC component, modeled with a piecewice linear trend function, is also computed. If we call $\mathbf{o}_i$ and $\mathbf{c}_i$ the leftmost corner and rightmost corner of each countour curve at opening and closing instants $t_{GO,i}$ and $t_{GC,i}$ respectively, the linear trend segment $r_i(t)$ is computed pitch-synchronously as

$$r_i(t) = \mathbf{o}_i + \frac{\mathbf{o}_{i+1} - \mathbf{o}_i}{t_{GO,i+1} - t_{GO,i}} \cdot (t - t_{GO,i}), \tag{6}$$

for $t_{GO,i} < t < t_{GO,i+1}$. The linear trend information is then used to vertically align the fold displacement prediction provided by the dynamical model, and the visual information:

$$\begin{aligned} x_{1(2),r}(t) &= x'_{1(2),r}(t) + r_i(t),\ t \in [t_{GO,i}, t_{GO,i+1}) \\ x_{1(2),l}(t) &= r_i(t) - x'_{1(2),l}(t),\ t \in [t_{GO,i}, t_{GO,i+1}) \end{aligned} \tag{7}$$

where the notation $x'_{1(2),\alpha}(t)$, $\alpha = \{l, r\}$, has been used here to indicate the folds position estimates provided by the dynamical model, which are originally not affected by any DC component. The glottal opening and closing instants, closed/open phase duration, and piecewice linear trend, are shown in Fig. 6, lower panel, for an analysis window corresponding to approximately 16 msec.

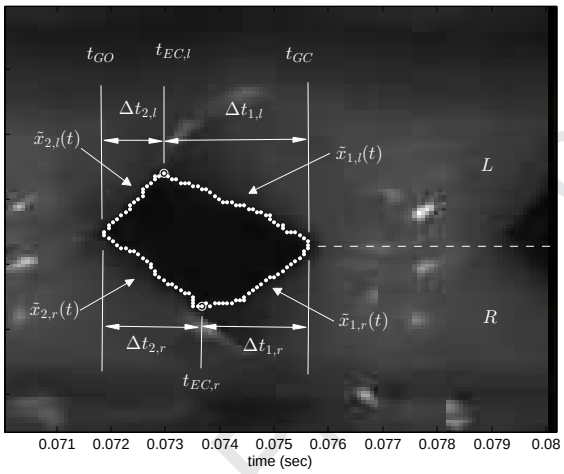

Figure 7: Computation of the visual cues related to folds displacement during the open phase.

3.2. Feature extraction

Finally, visual cues extraction is required to gather information concerning folds displacement during the open phase. Figure 7 illustrates how the different contour sections related to the romboidal-shaped region are related to the left fold opening lower edge displacement provided by the model ($x_{2,l}(t)$), to the left fold closing upper edge ($x_{1,l}(t)$), to the right fold opening lower edge ($x_{2,r}(t)$), and to the right fold closing upper edge ($x_{1,r}(t)$).

The time support intervals for the four classes are defined respectively as:

$$\begin{cases} \Delta t_{2,l} = \{t : t_{GO} < t < t_{EC,l}\} \\ \Delta t_{1,l} = \{t : t_{EC,l} < t < t_{GC}\} \\ \Delta t_{2,r} = \{t : t_{GO} < t < t_{EC,r}\} \\ \Delta t_{1,r} = \{t : t_{EC,r} < t < t_{GC}\} \end{cases} \tag{8}$$

where $t_{EC,l}$ and $t_{EC,r}$ (upper-lower "edge crossing") are defined as the instants at which the left and light lower edge displacements become smaller than the respective upper edge displacements, and are computed by estimating the maximum value of the left and right contour respectively. The following visual data cues are then defined:

$$\begin{cases} \tilde{x}_{1,l}(t) = cnt_l(t),\ t \in \Delta t_{1,l} \\ \tilde{x}_{2,l}(t) = cnt_l(t),\ t \in \Delta t_{2,l} \\ \tilde{x}_{1,r}(t) = cnt_r(t),\ t \in \Delta t_{1,r} \\ \tilde{x}_{2,r}(t) = cnt_r(t),\ t \in \Delta t_{2,r} \end{cases} \tag{9}$$

where $cnt_l(t)$ and $cnt_r(t)$ are the open phase region contours related to the left and right fold, respectively.

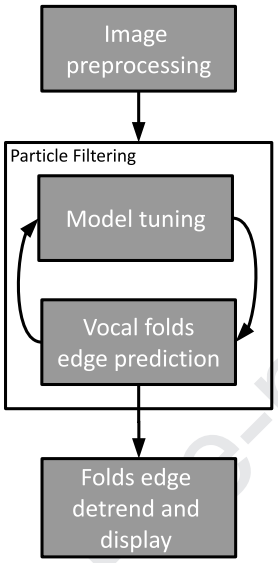

Figure 8: Method workflow.

3.3. Edge-based likelyhood

Once that each point of the open phase contour has been assigned to one of the four classes defined above, the likelyhood function for the particle filter relying on the glottal model can now be defined as

$$L = p(\mathbf{z}_k|\mathbf{x}_k) = L_{x_{1,l}} + L_{x_{2,l}} + L_{x_{1,r}} + L_{x_{2,r}} \tag{10}$$

where

$$\begin{aligned}
L_{x_{1,l}} &= \frac{1}{\sqrt{2\pi\sigma}} \int_{t\in\Delta t_1^L} \exp^{(-\frac{|\tilde{x}_{1,l}(t)-x_{1,l}(t)|^2}{2\sigma^2})}\, dt \\
L_{x_{2,l}} &= \frac{1}{\sqrt{2\pi\sigma}} \int_{t\in\Delta t_2^L} \exp^{(-\frac{|\tilde{x}_{2,l}(t)-x_{2,l}(t)|^2}{2\sigma^2})}\, dt \\
L_{x_{1,r}} &= \frac{1}{\sqrt{2\pi\sigma}} \int_{t\in\Delta t_1^R} \exp^{(-\frac{|\tilde{x}_{1,r}(t)-x_{1,r}(t)|^2}{2\sigma^2})}\, dt \\
L_{x_{2,r}} &= \frac{1}{\sqrt{2\pi\sigma}} \int_{t\in\Delta t_2^R} \exp^{(-\frac{|\tilde{x}_{2,r}(t)-x_{2,r}(t)|^2}{2\sigma^2})}\, dt
\end{aligned} \tag{11}$$

The method workflow, including the image pre-processing stage, the model tuning to the vocal fold edges, and the final detrending and display stage, is illustrated in Fig. 8.

Figures 9 and 10 show an example of vocal fold edge fitting obtained through particle filtering with the glottal model prediction procedure. Figures 9 evidentiates the evolution of L/R and x_1/x_2 asymmetries relates to four subsequent frames of the onset region of the phonation, and the adaptation of the model-driven particles to the data.

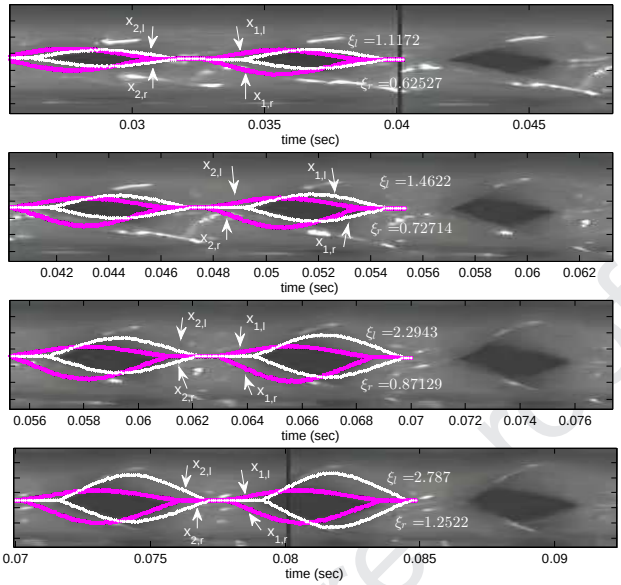

Figure 9: VKG video analysis and vocal fold edge fitting: asymmetries L vs R and x_1 vs x_2. The scattered plots superimposed to the VKG image represent the evolution of particles related to $x_{1,r}$ (magenta, upper portion), $x_{1,l}$ (magenta, lower portion), $x_{2,r}$ (white, upper portion), and $x_{2,l}$ (white, lower portion).

Figure 10 shows a wider picture of the folds edge fitting, which evidentiates the adaptation of the model to the slowly varying trend induced by the relative shifts of the endoscope with respect to the oscillating folds during the recording. Figure 11 shows the Left and Right fold parameter optimization during the fitting process (the maximum likelyhood parameter values and the credible intervals (CI) are shown). It can be noticed that the natural frequencies of the folds are approximately constant above 140 Hz, and that the vertical phase delays around 10 samples, which reflects the fact that the pitch and the ratio between the closed and open phase are rather stable in the analysis interval. On the other hand, in the Left fold the amplitude ratio reaches a factor of 2 whereas in the Right fold the ratio is around 1 on average, reflecting the asymmetry of the vocal cord oscillation already noted in this visual data fragment.

4. Experimental results

In this section, the vocal folds displacement reconstruction and model fitting procedure is assessed on a videokimographic dataset obtained from real high-speed video endoscopic recordings.

In order to provide a measure of the performance of the automatic fitting process, we define the error of the edge displacement fitting related to the open phase, as follows. Let first define the partial edge fitting errors as

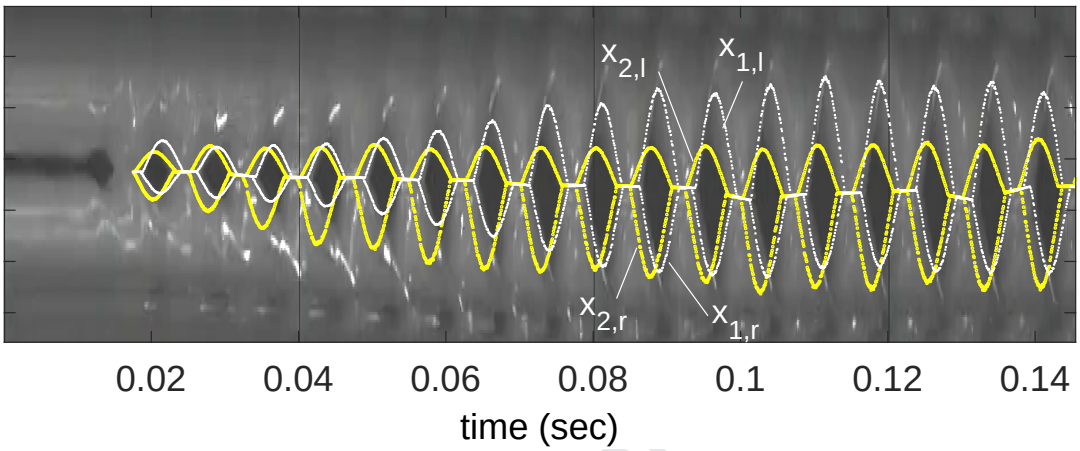

Figure 10: VKG video analysis and vocal fold edge fitting: an analysis windows of approximately 140 msec, showing the particle filtering fit to the observation (yellow and white scatter plot refer to x_1 and x_2 estimates respectively).

$$\begin{aligned} E_{x1,\alpha} &= \sqrt{\frac{1}{\Delta t_{1,\alpha}} \int_{t\in\Delta t_{1,\alpha}} |\tilde{x}_{1,\alpha}(t) - x_{1,\alpha}(t)|^2 dt} \\ E_{x2,\alpha} &= \sqrt{\frac{1}{\Delta t_{2,\alpha}} \int_{t\in\Delta t_{2,\alpha}} |\tilde{x}_{2,\alpha}(t) - x_{2,\alpha}(t)|^2 dt} \end{aligned} \tag{12}$$

for $\alpha = \{l, r\}$. Time intervals $\Delta t_{1,\alpha}$ and $\Delta t_{2,\alpha}$ are defined as in Eq. 8, and target edge displacements $\tilde{x}_{1,\alpha}$ and $\tilde{x}_{2,\alpha}$ are defined as in Eq. 9. The overall root mean squared error (RMSE) on the edge fitting, normalized to the maximum excursion range $M_x = (x_{cr,l}+x_{cr,r})$, where $x_{cr,l}$ and $x_{cr,r}$ are the left and right edge displacements at lower-upper edge crossing, is then defined as

$$ETE = (E_{x1,l} + E_{x1,r} + E_{x2,l} + E_{x2,r})/M_x \tag{13}$$

The normalized errors (NE) made in the estimation of the left and right lower-upper edge crossing instants can be defined respectively as

$$\begin{aligned} ECEt_l &= |\tilde{t}_{EC,l} - t_{EC,l}|/T_o, \\ ECEt_r &= |\tilde{t}_{EC,r} - t_{EC,r}|/T_o \end{aligned} \tag{14}$$

where $\tilde{t}_{EC,l}$ and $\tilde{t}_{EC,r}$ are the left and right crossing instants estimated by the contour identification procedure on the VKG image, and $t_{EC,l}$ and $t_{EC,l}$ are the Left and Right crossing instants provided as the result of the Bayesian estimation procedure. No calibration of the the dynamical model is done with respect to the kymogram, thus the modeling is not calibrated to physical units, and the RMSE values are referred to an arbitrary normalization operated on the model output.

We also define a set of glottal cycle time parameters to characterize the glottal area during the closed phase. If T is the glottal cycle period, T_c the closed glottis phase duration, and T_o the open glottis phase duration, we define $R_{CP} = T_c/T$ as the closed

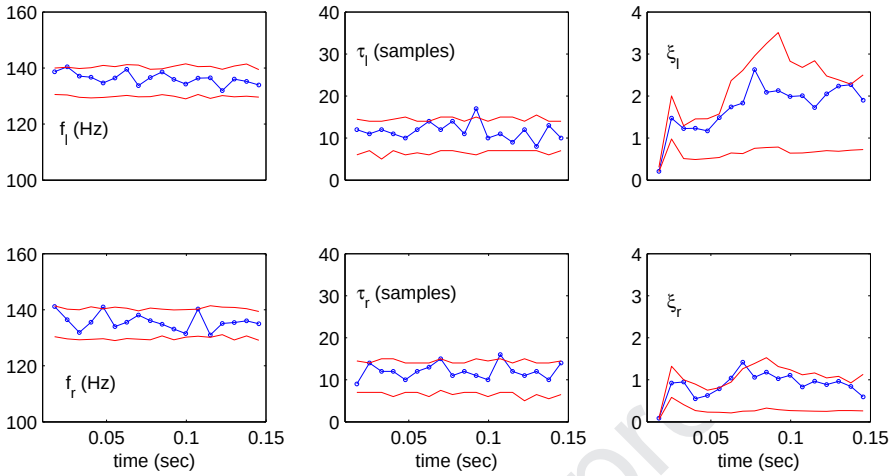

Figure 11: Maximum likelyhood parameters optimized during the fitting process, with related credible intervals. Upper plots: natural frequency, vertical phase delay, and superior-to-inferior edge amplitude ratio, for the Left cord; lower plots: same parameter for the Right cord.

phase ratio, and $R_{OP} = T_o/T$ as the open phase ratio. We further define $R_{CPd} = \Delta t_d/T_c$, $R_{CPe} = \Delta t_e/T_c$, and $R_{CPf} = \Delta t_f/T_c$ as a measure of the duration of divergent, parallel and convergent configuration of the glottis during the closed phase (refer to Fig. 2).

4.1. High-speed video dataset

The video analysis procedure discussed was tested on a selection of recordings from the database by Erkki Bianco and IRCAM (Degottex et al., 2008). The recordings, consisting in sustained phonations by different healthy subjects, are characterized by a video rate of 6665 frames per second at a resolution of 256x256 pixels per frame. Acoustic phonatory data is also provided, at a sampling rate of 44100 Hz with 16 bit resolution, which however was not used in this study. For each recording, the subjects uttered sustained voiced sounds with a different phonatory setting or a different fundamental frequency (pitch). Pitch was held constant or, in some cases, it was rised or lowered continuously.

For each fragment of high-speed video recording used in the experimental assessment, the line scan position for the kymographic data computation was manually selected, and the kymographic image was derived as the time sequence of video frame pixels corresponding to that line scan. The image preprocessing and the pitch synchronous procedure illustrated so far was applied to the resulting kymographic data, consisting in approximately 200 msec of speech for each fragment. The parameters of the glottal model where updated at every new speech period.

Table 1 reports the estimates of the area function parameters computed from the video data, based on the glottal area segmentation procedure illustrated in the video analysis section.

4.2. Performance analysis and assessment

The experimental results presented in Figs. 12 and 13 show, for each recording, a frame of the high-speed endoscopic video and the videokymograpic data computed along the scan line highlighted in the frame, the magnification of a two-periods analysis window with the model fitting to the fold edges, and an estimate of the corresponding glottal flow $U_g(t)$ as predicted by the biomechanical model.

The video data and fitting results illustrated in Figure 12 are related to two data excerpts recorded from the same subject (male, healthy) uttering a sustained vowel with same pitch (113 Hz) and with different phonatory settings (left: tense phonation, right: breathy phonation). The tense phonation case on the left is characterized by a long closed phase and a consequently short open phase, resulting in the open phase ratio $R_{OP} = 0.35$. The model-based x_1 and x_2 fitting correctly matches the glottal area evolution (romboidal shape) during the open phase. Moreover, it predicts that during the closed phase the time durations of segments d, e, and f, normalized to the total closed phase duration $T_c = 5$ msec, are respectively $R_{CPd} = 0.3$, $R_{CPe} = 0.5$, $R_{CPf} = 0.2$. If compared to the tense phonation case, the breathy phonation on the right has a much longer open phase and a shorter closed phase, resulting in an estimated $R_{OP} = 0.65$. The model-based x_1 and x_2 fitting correctly matches the glottal area evolution during the open phase, and predicts normalized intervals $R_{CPd} = 0.5$, $R_{CPe} = 0.0$, $R_{CPf} = 0.5$ for the closed phase with total duration $T_c = 2$ msec. Apart from open/closed phase matching, the model-based tracking also evidentiates a more marked R/L asymmetry for the breathy phonation sample, for which the analysis procedures provides $\{\xi_r = 1.12, \xi_l = 0.81\}$, than for the tense phonation, for which it is $\{\xi_r = 0.91, \xi_l = 0.87\}$.

The full set of performance measures and glottal parameters resulting from the fitting, averaged on the whole 200 msec time interval used in the analysis, are reported in the first two rows (S1a and S1b) of Table 1.

The video data and the fitting results illustrated in Figure 13 are related to two data excerpts recorded from a different male subject uttering a sustained vowel with modal phonation at two different pitches (left: 160 Hz, right: 135 Hz). The sample on the left is characterized by a marked folds edge L/R asymmetry which is opposite to the asymetry observed in the previous case. The related parameters for this case are is in fact $\{\xi_r = 0.88, \xi_l = 1.19\}$. The sample on the right, on the other hand, shows a rather symmetric L/R oscillatory pattern, as confirmed by the values $\{\xi_r = 0.93, \xi_l = 0.86\}$. In terms of open/closed phase ratios, the higher pitched sample on the left has equal open/closed phase durations ($R_{OP} = 0.5$), whereas in the sample on the right the closed phase is longer ($R_{OP} = 0.42$).

The full set of performance measures and glottal parameters resulting from the fitting, averaged on the whole 200 msec time interval used in the analysis, are reported in the third and fourth rows (S2a and S2b) of Table 1.

Finally, Table 1 also reports the results for three different modal phonation recordings from a female speaker: S3a, with a pitch of 526.3 Hz, S3b, with a pitch of 357.1 Hz, and S3c, with a pitch of 294.1 Hz. It can be seen that the normalized errors related to the open phase ($ECTE_l$, $ECTE_r$, and the normalized root mean squared error ETE) are around 0.10 on average (with a maximum of $0,28$ overall edge fitting error for sample S3a(f)).

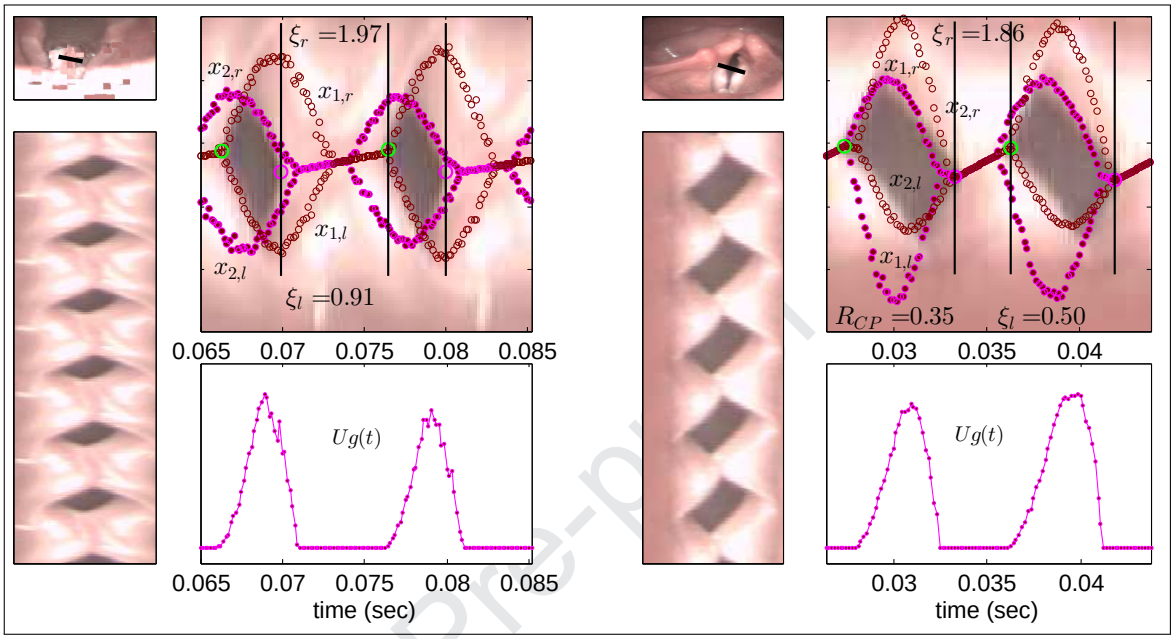

Figure 12: The image processing and fitting results for two recordings from the same subject, uttering a sustained vowel with different phonatory settings (left: tense phonation, and right: breathy phonation). Recordings from the database by E. Bianco and G. Degottex, IRCAM.

4.3. Limitations and future work

The model fitting to oberved raw video data was assessed on a small set of recordings from healthy subjects, thus the statistical assessment of the method is still limited, and future work is overseen to include assessment on a larges set of healthy and pathological voice HSV recordings. Moreover, the model does not have one-to-one correspondence to the actual values of the vocal fold masses, stiffness, and subglottal pressure, i.e. no calibration was taken into account. Nonetheless, the inspection of the modeling and parameter optimization outcome allows for objectively evaluating the biomechanical interrelationships between these variables.

Once finely tuned on the available visual data, the dynamical model can be used as a state transition model in a Bayesian motion estimation setting to obtain a physiologically motivated estimation of upper and lower vocal folds edge position, even where this information is missing in the observation due to measurement limitations, i.e., upper edge indistinguishability during divergent glottis intervals (closing and closed phases of the glottal cycle) and lower edge occlusion during convergent glottis intervals (beginning of the open phase of the glottal cycle). Based on the model prediction, an hypothesis on the lower and upper fold position can be made even in complete fold occlusion conditions occurring during the end of the closed phase and. To asses such

Table 1: Glottal parameters values for different subjects and phonation settings, calculated over 200 msec time interval for each fragment. In the *open phase cues* columns, the values listed in parentheses refer to the error with respect to the reference cue values related to the open phase.

		open phase			closed phase		
Subj.	T (msec)	$ECTE_l$ (NE)	$ECTE_r$ (NE)	ETE (NRMSE)	R_{CPd}	R_{CPe}	R_{CPf}
S1a	10.2	0.18	0.04	0.10	0.16	0.44	0.40
S1b	8.9	0.04	0.19	0.11	0.30	0.05	0.65
S2a	6.6	0.00	0.16	0.11	0.52	0.00	0.48
S2b	7.5	0.05	0.00	0.06	0.34	0.21	0.45
S3a(f)	1.9	0.14	0.14	0.28	0.57	0.00	0.43
S3b(f)	2.8	0.08	0.08	0.19	0.00	0.71	0.29
S3c(f)	3.4	0.08	0.08	0.08	0.17	0.25	0.58

perspective, however, the problem arise of obtaining a ground thruth dataset for the time intervals in which some of the fold edges are not observable.

Finally, we also recall that in this study the proposed scheme was tested with data from healty phonation, but it is potentially suitable as a tool for pathologic phonation detection and classification. It is plausible that in this context the model fitting procedure within the tracking scheme would require further improvements to deal with irregular oscillatory patterns and severe left-rigth asymmetries. This will be the subject of future research as well.

5. Conclusions

We discussed the analysis of videokymographic data with a Bayesian estimation procedure based on the prediction of the folds edges, provided by a nonlinear dynamical model of the vocal folds. The low-dimensional glottal model adopted is asymmetrical in the L/R plane and was shown to be able to accurately fit the vocal folds edge displacement information extracted from the videokymographic high-speed video data. A video processing analysis procedure was designed, that computes the likelyhood of the observed video data in terms of the predictive glottal model. A relevant characteristic of this analysis scheme is the possibility to predict the fold lower edge trajectory in the occluded intervals, where no video data is available. The application on a set of different endoscopic high-speed recordings demonstrated the suitability of the procedure. A performance analysis and assessment was conducted by computing standard glottal sub-cycle features such as open/closed phase durations and glottal area evolution. The experimental results conducted on a set of recordings featuring different types of phonations, show that the Bayesian estimation driven by the numerical glottal model provides a robust fitting to the fold motion video cues, where available, and a tool to predict glottal sub-cycle features and fold edge trajectories in those time intervals in which no useful video data is available due to poorly contrasted or too noisy image, or due to occlusion conditions. However, if on one hand it was possible to measure the performance of the procedure in the open phase regions with respect to recorded data, on the other hand it was not possible to assess, with the data at hand, the accuracy of

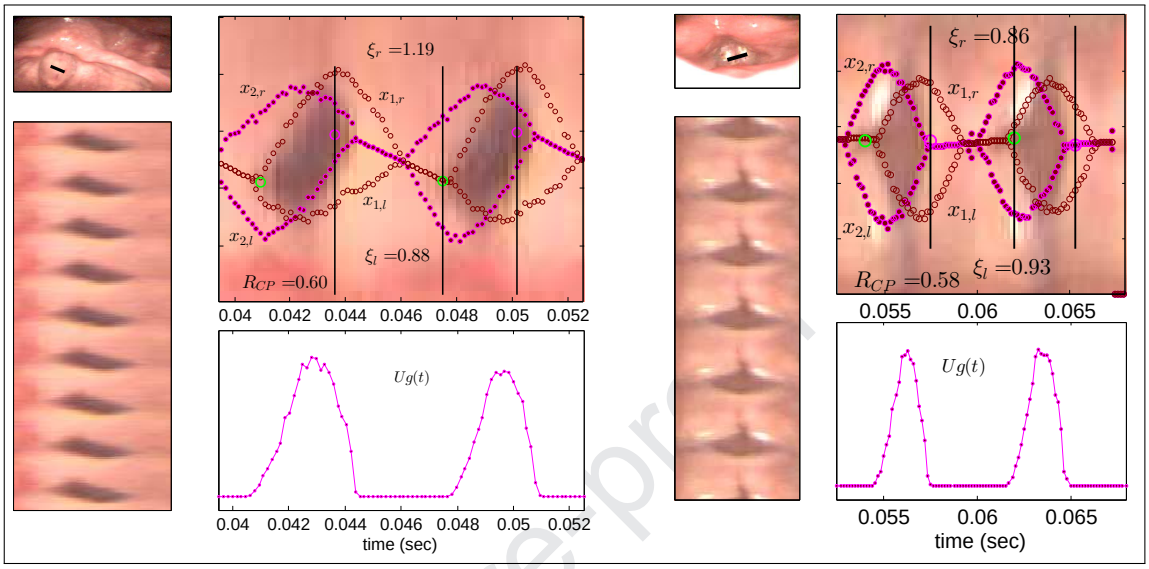

Figure 13: Fitting results for two recordings from the same subject. Left: modal phonation, pitch: 160 Hz; Right: modal phonation, pitch: 135 Hz. Recordings from the database by E. Bianco and G. Degottex, IRCAM.

the predicted cues in regions were no useful video data can be extracted or in the occluded regions. In fact, assessment related to occluded regions would only be possible if the vocal folds oscillation could be recorded not only from above the glottis, but also from below. Such dataset recording would be possible for example with an in-vitro experimental setup, and will be the subject of future research.

Acknowledgments

We wish to thank E. Bianco and G. Degottex for kindly providing the high-speed video recordings used in this paper.

Arulampalam, M.S., Maskell, S., Gordon, N., 2002. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Transactions on Signal Processing 50, 174–188.

Cataldo, E., Soize, C., Sampaio, R., 2013. Uncertainty quantification of voice signal production mechanical model and experimental updating. Mechanical Systems and Signal Processing 40, 718–726.

Cataldo, E., Soize, C., Sampaio, R., Desceliers, C., 2009. Probabilistic modeling of a nonlinear dynamical system used for producing voice. Computational Mechanics 43, 265–275.

Chen, G., Kreiman, J., Alwan, A., 2014. The glottaltopogram: A method of analyzing high-speed images of the vocal folds. Computer Speech And Language 28, 1156–1169.

Degottex, G., Bianco, E., Rodet, X., 2008. Usual to particular phonatory situations studied with high-speed videoendoscopy, in: Proceedings of the 6th International Conference on Voice Physiology and Biomechanics (ICVPB), Tampere, Finland. pp. 19–26.

Deng, J.J., Hadwin, P.J., Peterson, S.D., 2019. The effect of high-speed videoendoscopy configuration on reduced-order model parameter estimates by bayesian inference. The Journal of the Acoustical Society of America 146, 1492–1502.

Díaz-Cádiz, M.E., Peterson, S.D., Galindo, G.E., Espinoza, V.M., Motie-Shirazi, M., Erath, B.D., Zañartu, M., 2019. Estimating vocal fold contact pressure from raw laryngeal high-speed videoendoscopy using a hertz contact model. Applied Sciences 9, 2384.

Döllinger, M., Gómez, P., Patel, R.R., Alexiou, C., Bohr, C., Schützenberger, A., 2017. Biomechanical simulation of vocal fold dynamics in adults based on laryngeal high-speed videoendoscopy. PLOS ONE 12, 1–26.

Döllinger, M., Hoppe, U., Hettlich, F., Lohscheller, J., Schuberth, S., Eysholdt, U., 2002. Vibration parameter extraction from endoscopic image series of the vocal folds. IEEE Trans. Biomed. Engineering 49, 773–781.

Dore, A., Soto, M., Regazzoni, C., 2010. Bayesian tracking for video analytics. Signal Processing Magazine, IEEE 27, 46–55.

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

Drioli, C., Calanca, A., 2014. Speaker adaptive voice source modeling with applications to speech coding and processing. Computer Speech & Language 28, 1195–1208.

Drioli, C., Foresti, G.L., 2015a. Accurate glottal model parametrization by integrating audio and high-speed endoscopic video data. Signal, Image and Video Processing 9, 451–1459.

Drioli, C., Foresti, G.L., 2015b. Quantitative characterization of functional voice disorders using motion analysis of highspeed video and modeling, in: Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–6.

Foresti, G.L., Regazzoni, C., 2000. A hierarchical approach to feature extraction and grouping. IEEE Transactions on Image Processing 9, 1056–1074.

Hadwin, P.J., Galindo, G.E., Daun, K.J., Zañartu, M., Erath, B.D., Cataldo, E., Peterson, S.D., 2016. Non-stationary Bayesian estimation of parameters from a body cover model of the vocal folds. The Journal of the Acoustical Society of America 139, 2683–2696.

Hadwin, P.J., Motie-Shirazi, M., Erath, B.D., Peterson, S.D., 2019. Bayesian inference of vocal fold material properties from glottal area waveforms using a 2d finite element model. Applied Sciences 9.

Hadwin, P.J., Peterson, S.D., 2017. An extended kalman filter approach to non-stationary bayesian estimation of reduced-order vocal fold model parameters. The Journal of the Acoustical Society of America 141, 2909–2920.

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

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

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

Kuo, C.F.J., Wang, H.W., Hsiao, S.W., Peng, K.C., Chou, Y.L., Lai, C.Y., Hsu, C.T.M., 2014. Development of laryngeal video stroboscope with laser marking module for dynamic glottis measurement. Computerized Medical Imaging and Graphics 38, 34 – 41.

Larsson, H., Hertegård, S., Lindestad, P., Hammarberg, B., 2000. Vocal fold vibrations: high-speed imaging, kymography, and acoustic analysis: a preliminary report. Laryngoscope 110, 2117–22.

Lin, J., Clancy, N.T., Qi, J., Hu, Y., Tatla, T., Stoyanov, D., Maier-Hein, L., Elson, D.S., 2018. Dual-modality endoscopic probe for tissue surface shape reconstruction and hyperspectral imaging enabled by deep neural networks. Medical Image Analysis 48, 162 – 176.

Lohscheller, J., Eysholdt, U., Toy, H., Döllinger, M., 2008. 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, 300–309.

Lohscheller, J., Toy, H., Rosanowski, F., Eysholdt, U., Döllinger, M., 2007. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. Medical Image Analysis 11, 400–413.

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

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

Murtola, T., Alku, P., Malinen, J., Geneid, A., 2018. Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy. Speech Communication 96, 67 – 80.

Neubauer, J., Mergell, P., Eysholdt, U., Herzel, H., 2001. Spatio-temporal analysis of irregular vocal fold oscillations: Biphonation due to desynchronization of spatial modes. J. Acoust. Soc. Am. 110, 3179–3192.

Osma-Ruiz, V., Godino-Llorente, J.I., Sáenz-Lechón, N., Fraile, R., 2008. Segmentation of the glottal space from laryngeal images using the watershed transform. Computerized Medical Imaging and Graphics 32, 193 – 201.

Pinheiro, A.P., Stewart, D.E., Maciel, C.D., Pereira, J.C., Oliveira, S., 2012. Analysis of nonlinear dynamics of vocal folds using high-speed video observation and biomechanical modeling. Digital Signal Processing 22, 304 – 313.

Popolo, P., 2018. Investigation of flexible high-speed video nasolaryngoscopy. Journal of Voice 32, 529–537.

Schoob, A., Kundrat, D., Kahrs, L.A., Ortmaier, T., 2017. Stereo vision-based tracking of soft tissue motion with application to online ablation control in laser microsurgery. Medical Image Analysis 40, 80 – 95.

Schutte, H.K., Švec, J.G., Šram, F., 1998. First results of clinical application of videokymography. Laryngoscope 108, 1206–10.

Schwarz, R., Döllinger, M., Wurzbacher, T., Eysholdt, U., Lohscheller, J., 2008. Spatio-temporal quantification of vocal fold vibrations using high-speed videoendoscopy and a biomechanical model. The Journal of the Acoustical Society of America 123, 2717–2732.

Schwarz, R., Hoppe, U., Schuster, M., Wurzbacher, T., Eysholdt, U., Lohscheller, J., 2006. Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model. IEEE Trans. Biomed. Engineering 53, 1099–1108.

Snidaro, L., Foresti, G.L., 2003. Real-time thresholding with Euler numbers. Pattern Recognition Letters 24, 1533–1544.

Särkkä, S., 2013. Bayesian Filtering and Smoothing. volume 3 of *IMS Textbooks*. Cambridge University Press.

Stiglmayr, M., Schwarz, R., Klamroth, K., Leugering, G., Lohscheller, J., 2008. Registration of pe segment contour deformations in digital high-speed videos. Medical Image Analysis 12, 318 – 334.

Tao, C., Jiang, J.J., 2008. A self-oscillating biophysical computer model of the elongated vocal fold. Computers in Biology and Medicine 38, 1211 – 1217.

Tigges, M., Wittenberg, T., Mergell, P., Eysholdt, U., 1999. Imaging of vocal fold vibration by digital multi-plane kymography. Computerized Medical Imaging and Graphics 23, 323 – 330.

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

Turkmen, H.I., Karsligil, M.E., Kocak, I., 2015. Classification of laryngeal disorders based on shape and vascular defects of vocal folds. Computers in Biology and Medicine 62, 76 – 85.

Verikas, A., Gelzinis, A., Bacauskiene, M., Valincius, D., Uloza, V., 2007. A kernel-based approach to categorizing laryngeal images. Computerized Medical Imaging and Graphics 31, 587 – 594.

Vermaak, J., Andrieu, C., Doucet, A., Godsill, S.J., 2002. Particle methods for Bayesian modeling and enhancement of speech signals. Speech and Audio Processing, IEEE Transactions on 10, 173–185.

Vondrak, M., Sigal, L., Jenkins, O.C., 2008. Physical simulation for probabilistic motion tracking., in: CVPR, IEEE Computer Society.

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

Švec, J.G., Šram, F., Schutte, H.K., 2007. Videokymography in voice disorders: What to look for? Annals of Otology Rhinology and Laryngology 116, 172–180.

Wendler, J., 1992. Stroboscopy. Journal of Voice 6, 149–154.

Wittenberg, T., Mergell, P., Tigges, M., Eysholdt, U., 1997. 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, Washington, DC, USA. pp. 1663–1666.

Wurzbacher, T., Voigt, I., Schwarz, R., Döllinger, M., Hoppe, U., Penne, J., Eysholdt, U., Lohscheller, J., 2008. Calibration of laryngeal endoscopic high-speed image sequences by an automated detection of parallel laser line projections. Medical Image Analysis 12, 300–317.

Yan, Y., Chen, X., Bless, D., 2006. Automatic tracing of vocal-fold motion from high-speed digital images. IEEE Trans. Biomed. Engineering 53, 1394–1400.