Australasian Physical & Engineering Sciences in Medicine (2019) 42:91–109
https://doi.org/10.1007/s13246-018-0717-9

SCIENTIFIC PAPER

Optical flow based waveform for the assessment of the vocal fold vibrations

Heyfa Ammar[1,2]

Received: 20 April 2018 / Accepted: 12 December 2018 / Published online: 19 December 2018

Abstract

Assessing vocal fold (VF) vibrations is important for the diagnosis of several diseases, and is made possible through the analysis of videoendoscopy recordings. However, the visual analysis of these recordings is hard due to the high acquisition rate. For this reason, it is commonly used to extract the laryngeal activity information from the recordings and represent it in a way suitable to be visually analyzed. Waveforms, images and playbacks are examples of representations reported in the literature. The main limitation of some of them is the lack of precisely locating the pathology within the VFs. Whereas others require the segmentation of the glottis in all the images of the video which is a complex and hard task given the high amount of images in the video and the necessity for the user intervention. To overcome these problems, the present study proposes a new waveform that maps the local vibrations of the VFs without the need for segmenting all the images of the video. Instead, the segmentation is restricted to only one image per vibratory cycle. Then, a new optical flow based technique is proposed to deduce the cycle-to-cycle dynamics of the VFs. The ability of the proposed approach to provide a reliable visual assessment is experimentally evaluated using different types of phonation and different vocal pathologies.

Keywords Vocal fold vibrations · Waveform · Optical flow · Pathology · Cycle-to-cycle analysis

Introduction

Assessing the vocal fold (VF) vibrations is important for the diagnosis of several diseases such as neurologic diseases (the Parkinson [1], the multiple sclerosis [2]) and vocal diseases (laryngeal paralysis, nodules, polyps [3]). In healthy phonation, the vibrations are a periodic closing and opening of the right and left VFs characterized by their phase and amplitude symmetry with respect to the glottal axis (Fig. 1). In the same way, vibratory aperiodicity and/or asymmetries can reveal a pathology. The acquisition of the true intra-cycle vibratory behavior during phonation is made possible through the high speed videoendoscopy (HSV) recordings [4] allowing to capture the VF vibrations at a rate of 4000 images per second and above. However, due to the high acquisition rate, it is hard to visually analyze the laryngeal activity from the HSV recordings. For this reason, it is commonly used to map the spatial-temporal content of the HSV into a compact representation appropriate for a visual analysis by clinicians [5].

Several 1D and 2D representations are proposed in the literature and can be classified into global and local ones. The global representations map the VF dynamics along the whole glottal length and provide a general analysis of the glottal behavior during vibratory cycles. For instance, given that the VF opening/closing is related to the temporal variation of the glottal area, one solution is to consider the glottal area waveform (GAW) that reflects the opening and the closing statuses of the glottis [6]. Although this waveform allows to deduce some important quantification measures as the fundamental frequency of the vibrations, the opening/closing times and the opening/closing speed, it is not possible to separately characterize the movements of the two VFs and precisely locate a possible pathology occurring in one or both of them. Two-dimensional representations are also proposed in the literature [7, 8]. The idea is to calculate

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13246-018-0717-9) contains supplementary material, which is available to authorized users.

✉ Heyfa Ammar
hammar@kau.edu.sa; heyfa.amar@gmail.com

[1] King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

[2] University of Tunis El Manar, Tunis, Tunisia

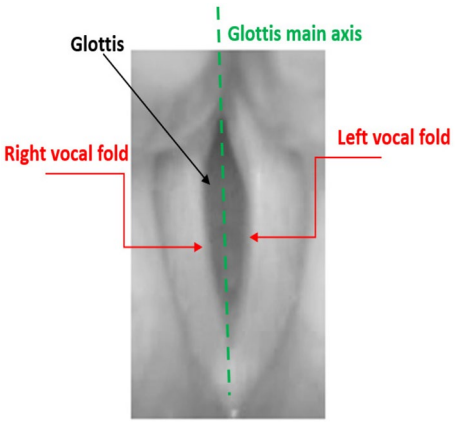


Fig. 1 Illustration of a maximum opening of the glottis

the distances between the glottal axis and some equidistant points located on the glottal edges within each image of the HSV. Then, these distances computed at different instants are mapped into an image named phonovibrogram (PVG), where the x-axis and the y-axis represent respectively the temporal dimension and the point position on the glottal edge [7]. The intensity of each point in the PVG reflects its distance relatively to the glottal axis. Despite the well characterization of the vibrations by the geometric shape of the PVG which facilitates the visual assessment of the periodicity and the symmetry properties, this approach is sensitive to the accuracy of the glottal axis computation. To alleviate this problem, a representation called glottovibrogram (GVG) is proposed in [8] and replaces the computation of the distances between the main glottal axis and the points on the glottal edges in the right/left sides, by the computation of the distances between these points. A main limitation of the aforementioned techniques is that they require the segmentation of all the processed images in the HSV. This is a complex task and requires the user intervention.

More focused analysis of the laryngeal dynamics is possible by the means of local representations. For instance, the digital kymography (DKG) locally represents the vibrations viewed on a given line perpendicular to the glottal main axis during multiple vibratory cycles [9, 10]. The vertical axis of the digital kymogram indicates the positions of the pixels on the selected line, while the horizontal axis corresponds to the temporal dimension. The value of each point in DKG is its intensity in the image. This representation makes possible the visualization of the pixel trajectories along the vibratory cycles. Although the DKG has the advantage of providing easy ways to assess the VF vibrations [5, 11], it is sensitive to the endoscope movement. To overcome the problems arising from the segmentation procedure, modified versions of the aforementioned representations are recently proposed in [12]. The VF displacements are represented by the means of the motion vectors estimated using an optical flow (OF) technique, within a rectangular window that contains the glottal region. In this context, the GVG technique is modified giving rise to a new one called optical-flow GVG (OFGVG) deduced by averaging the motion vectors in each row of the window. In the same way, a modified version of DKG called optical flow kymogram (OFKG) represents the horizontal components of the motion vectors and their directions along a given line perpendicular to the glottal axis. Although these OF based playbacks have the advantage of not requiring a segmentation step, they appear blurred due to the effect of the mucosal wave as mentioned in [12]. They are also sensitive to the position of the glottal region within the selected window. In fact, the VF movements together with the movements of the neighboring tissues are involved in the OF based playbacks although these tissues could have different motion amplitudes and orientations [13, 14], which may compromise the analysis. In addition, the behavior of these modified representations in pathological phonation are not explored [12]. To alleviate the aforementioned limitations, our rationale is to locate some points of interest on the glottal edges and perform a cycle-to-cycle tracking of them based on their motions. This allows a local and precise examination of the vibratory behavior of the VFs without involving the movements of the neighboring areas. Therefore, a localization of the glottal boundaries is first required through a segmentation approach. To alleviate the complexity of the segmentation task, we propose the segmentation of one image per vibratory cycle rather than segmenting all the images. Then, a technique that appropriately traces the movements of the VFs along the HSV sequence is necessary in order to represent their oscillatory behavior. This step gives rise to a new waveform that facilitates the clinical assessment of the vibrations. Its advantage is threefold. First, the proposed representation is one-dimensional. Second, it highly reduces the amount of data to be segmented. Third, it can precisely locate a possible anomaly within the VF. Two contributions are proposed in this study:

1. A cycle detection technique is proposed. It allows the automatic identification of the vibratory cycles within the HSV recording.
2. A new waveform reflecting the local oscillatory behavior of the VFs at each point, is proposed. The waveform allows the analysis of the VF dynamics at a preselected level of the glottis, based on the OF estimation and without requiring the segmentation of all the images. Moreover, it is robust to the glottal axis shift and to the endoscope movement.

Both healthy and pathological phonation are explored. The localization of a possible pathology can be identified: in

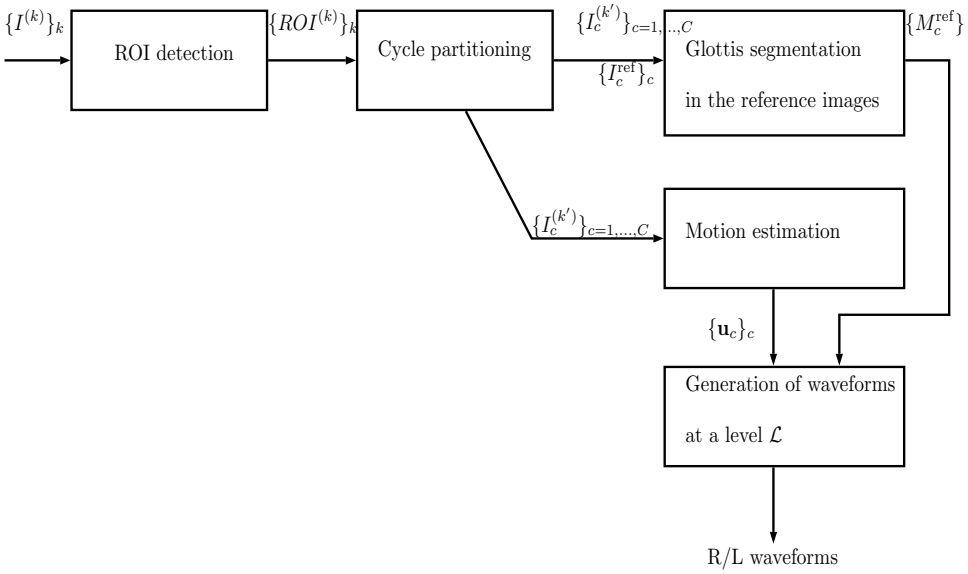


Fig. 2 Block diagram of the proposed approach. R/L refers to right/left

the right or in the left VF, and in which level from the posterior to the anterior commissures.

The remainder of this paper is organized as follows. First, the databases used in all the experiments of the present study are described in "Introduction". Second, the proposed approach is detailed in "Methods". Third, "Results" describes the evaluation results obtained in comparison to similar techniques in one hand, and the use of other modalities as the EGG signals on the other hand. Finally, the discussion is presented in "Discussion" and some conclusions are drawn in "Conclusion and future work".

Databases

Two databases of HSV recordings are used in the present work and correspond respectively to healthy and pathological VFs. The first database is provided by E. Bianco and G. Degottex-IRCAM [15, 16] and contains about 48 videos along with the corresponding EGG and audio signals. A HSV camera *ENDOCAM 5562* is used which allows the acquisition of 4000 color images per second at a spatial resolution of 256 × 256 pixels. Healthy speakers and mostly singers were recorded in a duration of about 2 s. The available EGG signals and acoustic waveforms are exploited in this work in order to evaluate the proposed partitioning technique of the HSV into cycles. The EGG signals were acquired using an EGG Portable Electroglottograph type EG 90 (from F. J. Electronics) and were sampled in a perfect synchronization with the acoustic waveforms at a frequency of 44150 Hz using the Wolf's software [15]. The EGG signals and the videos were synchronized with an uncertainty of at most three images. The HSV sequences include various laryngeal mechanisms [16]. Mechanism 1 known as *modal voice* or *chest voice*, Mechanism 2 known as *falsetto voice* and Mechanism 0 known as *vocal fry* or *creaky voice*. In Mechanism 1, the fundamental frequency is about 120 Hz for a male and 180 Hz for a female. It can reach the double in Mechanism 2 which is used by children and often by females. The Mechanism 0 is frequent in speech and is especially used at the end of English sentences. It is characterized by irregular movements of the VFs [16].

The second database contains videos of pathological phonation, used in order to evaluate the capability of the proposed technique in diagnosing unhealthy voices. It is publicly available online[1] and contains HSVs of several disorders as laryngeal paralysis, cancer, polyps and more.

Methods

The main goal of the present study is to deduce a 1D mapping of the HSV recordings that facilitates the visual interpretation of the vibrations without requiring the segmentation of all the images. The proposed approach is mainly composed of five modules as shown in the block diagram of Fig. 2:

1. Detection of the region of interest (ROI): after converting the colored images of the HSV to gray-level ones, the processing is restricted to a ROI that includes the glottal area. This step allows to obtain reliable results while reducing the computations.

[1] At www.entusa.com.

2. HSV partitioning into cycles: the HSV sequence is partitioned into vibratory cycles to allow the cycle-to-cycle analysis of the VF dynamics.
3. OF estimation: a new approach is proposed aiming to account for the oscillatory behavior of the VF during the OF estimation process.
4. Glottal segmentation: the local interpretation of the VF movements requires the segmentation of the glottis in order to determine the points of interest located on the VF edges at a level $\mathcal{L}$ from the posterior commissure to the anterior commissure.
5. Waveform generation: the trajectories of the points of interest located respectively on the right (R) and left (L) VFs are traced giving rise to R/L waveforms.

ROI detection

We believe that the best images suitable to give a precise localization of the ROI are those containing the maximal opening of the VFs. Such images are characterized by a local minimum value of the sum of their pixel intensities as the glottis is the darkest region in the image. Once the images corresponding to the minima are selected, the ROI is automatically determined in each of them following the technique proposed in [8]. First, for each image of maximum glottal opening, the edges are detected using a Sobel filter. Then, small related regions are connected using a morphological closing. Second, a connected component analysis [17] is performed in order to detect the regions of the image. The region R_g with the largest area is considered as the one that contains the glottis, and the bounding box surrounding it is the ROI. As the ROI contains the information needed to analyze the VF vibrations, all the processing is performed within this region, starting from the cycle partitioning of the HSV sequence described below.

HSV partitioning into cycles

Given a set of K successive images $I^{(k)}$, $k = 1, \ldots, K$ obtained by sampling the high speed video into K images, the aim is to label each image to a given vibratory cycle c giving rise to C subsets $\{I_c^{(k')}\}_{c=1,\ldots,C}$ of images. The key idea of the technique we propose is motivated by the relationship between the VF movements and the temporal variation of the sum of the pixel intensities $S^{(k)}$ defined by:

$$S^{(k)} = \sum_{(i,j)\in ROI} I^{(k)}(i, j). \tag{1}$$

As the glottis opens, the sum of the pixel intensities decreases because of the darkness of the glottis. In the same way, the sum of the pixel intensities increases as the glottis closes. Figure 3 illustrates the temporal variation of

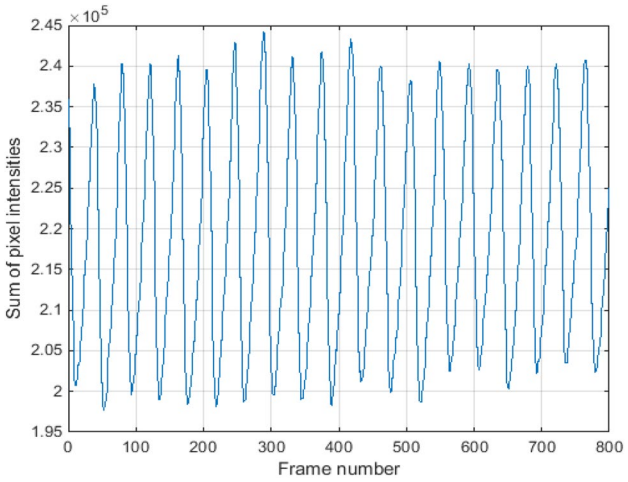


Fig. 3 Temporal variation of the sum of the pixel intensities related to healthy vibrations

$\mathcal{S} = \{S^{(k)}\}_{k=1,\ldots,K}$ related to a healthy phonation and well highlights the periodic aspect of the vibrations. Each period corresponds to the duration of a vibratory cycle; it is then possible to partition the sequence into cycles by estimating the fundamental period T of $\mathcal{S}$ using the non linear curve fit model defined by [18]:

$$\begin{aligned}\mathcal{S} = s_0 &+ a_1 \sin\left(\frac{2\pi}{T}k\right) + b_1 \cos\left(\frac{2\pi}{T}k\right) \\ &+ a_2 \sin\left(\frac{4\pi}{T}k\right) + b_2 \cos\left(\frac{4\pi}{T}k\right)\end{aligned} \tag{2}$$

where s_0 is the direct component corresponding to the average value of $\mathcal{S}$ over time, a_1, b_1, a_2 and b_2 are coefficients.

The maxima of $\mathcal{S}$ correspond to the glottal maximal closings. Each cycle starts by a closed state of the glottis. Then, the glottis opens along an opening stage that terminates when the image I_c^{ref} of maximal opening is reached. The image I_c^{ref} is designated by *reference image* of the cycle c. Next, a closing phase is performed until reaching the next maximum. Each vibratory cycle $c \in \{1, \ldots, C\}$ is composed of a set $\{I_c^{(k')}\}_{k' \in \{-n_1^c, \ldots, -1, 0, 1, \ldots, n_2^c\}}$ of $K_c = n_1^c + n_2^c + 1$ images, where n_1^c and n_2^c are the number of images that respectively precede and follow the reference image within the cycle c; k' refers to the signed sequential number of the image $I_c^{(k')}$ in the cycle c, relatively to the reference image $I_c^{\text{ref}} = I_c^{(0)}$. That is, the image that immediately precedes the reference image is denoted by $I_c^{(-1)}$, the one that precedes $I_c^{(-1)}$ is $I_c^{(-2)}$, and so on. In the same way, the image that immediately follows the reference image in the cycle c is denoted by $I_c^{(1)}$ and so on. A pseudo-code of this step is depicted by Algorithm 1.

Algorithm 1 Pseudo code for HSV partitioning into cycles

Input: ROI, $\{I^{(k)}\}_{k=1,\ldots,K}$, F_s (the sampling frequency of the video).
Output: $\{I_c^{(k')}\}_{k' \in \{-n_1^c,\ldots,-1,0,1,\ldots,n_2^c\}}$.
$[nrows,ncols]$=size(ROI)
for $k = 1, \ldots, K$ **do**
  $S^{(k)} = \sum_{(i,j)\in ROI} I^{(k)}(i,j)$
end for
$T = CurveFitModel(\{S^{(k)}\}_k, 'Fourier2')$
$K_c = T \times F_s$
$c = 1$ ▷ Initialization of the sequential number of the cycle.
for $k = 1, \ldots, K - K_c$ by step of K_c
  $ref = \min(\{S^{(i)}\}), \forall i = k, \ldots, k + K_c - 1$ ▷ Determine the position ref of the reference image

  $I_c^{(0)} = I^{ref}$
  $n_1^c = ref - k$
  $n_2^c = c \times K_c - ref$
  $l = 0$
  for $k' = -n_1^c, \ldots, 0$ **do**
    $I_c^{(k')} = I^{(k+l)}$
    $l = l + 1$
  end for
  for $k' = 1, \ldots, n_2^c$ **do**
    $I_c^{(k')} = I^{(k+l+1)}$
  end for
  $c = c + 1$
end for do

Algorithm 2 Pseudo code for motion estimation within a single cycle c

Input: $\{I_c^{(k')}\}_{k' \in \{-n_1^c,\ldots,n_2^c\}}$, n_1^c, n_2^c
Output: $\{\mathbf{u}_c^{(k',k'-1)}\}_{k' \in \{-n_1^c+1,\ldots,0\}}$, $\{\mathbf{u}_c^{(k',k'+1)}\}_{k' \in \{0,\ldots,n_2^c-1\}}$
for $k' = 0, \ldots, n_2^c - 1$ **do**
  $ROI^{(k')} = ROI(I_c^{(k')})$
  $ROI^{(k'+1)} = ROI(I_c^{(k'+1)})$
  $[\mathbf{u}_c^{(k',k'+1)}] = ComputeOF(ROI^{(k')}, ROI^{(k'+1)})$ ▷ Motion estimation in the forward direction
end for
for $k' = 0, \ldots, -n_1^c + 1$ by step of -1 **do**
  $ROI^{(k')} = ROI(I_c^{(k')})$
  $ROI^{(k'-1)} = ROI(I_c^{(k'-1)})$
  $[\mathbf{u}_c^{(k',k'-1)}] = ComputeOF(ROI^{(k')}, ROI^{(k'-1)})$ ▷ Motion estimation in the backward direction
end for

Once the partitioning into cycles is performed, the vibratory behavior of the VFs is analyzed within each cycle based on their movements.

Motion estimation of the VF

In order to analyze the vibratory behavior of the VFs, the motion field called OF is estimated within the ROI according to a particular way that aims to be compliant to the specificities of the laryngeal dynamics. More precisely, in order to highlight the oscillatory behavior of the VF movements, the motion field is estimated in the backward or the forward directions depending on the position of the considered images relatively to the reference one. As detailed in Algorithm 2, the OF estimation between $I_c^{(k')}$ and $I_c^{(k'+1)}$ is performed in the forward direction when

$0 < k' \leq n_2^c$; which means that the images under consideration temporally follow the reference image. In the same manner, the estimations between $I_c^{(k')}$ and $I_c^{(k'-1)}$, $\forall -n_1^c \leq k' < 0$ are performed in the backward direction. The general principles of the OF estimation technique are described below. For the sake of simplicity, only the estimation in the forward direction is detailed. The backward estimation is easily derived.

Many efforts are invested to estimate the OF since the work of Horn and Schunk [19]. The idea is to minimize a functional energy involving two basic assumptions: data conservation and spatial coherence. The data conservation means that the objects of the video continue to be present in the scene during a certain period of time and with the same brightness level. This statement is valid in laryngeal HSV as the frame rate is high. The spatial coherence implies that the pixels within a certain local neighborhood have the same motion characteristics. Taking into consideration these assumptions, the motion field $\mathbf{u}_c^{(k',k'+1)}$ between two images $I_c^{(k')}$ and $I_c^{(k'+1)}$ within a region $\mathcal{R} \subset$ ROI is estimated by minimizing a functional energy classically defined by:

$$\begin{aligned} E(u,v) = \sum_{(i,j)\in\mathcal{R}} & \left(\rho_D\left(I_c^{(k')}(i,j)\right) - I_c^{(k'+1)}\left(i+u_{(i,j)}, j+v_{(i,j)}\right)\right) \\ & +\lambda\left[\rho_S\left(u_{(i,j)} - u_{(i+1,j)}\right) + \rho_S\left(u_{(i,j)} - u_{(i,j+1)}\right)\right. \\ & \left.+\rho_S\left(v_{(i,j)} - v_{(i+1,j)}\right) + \rho_S\left(v_{(i,j)} - v_{(i,j+1)}\right)\right] \end{aligned} \quad (3)$$

where u is the horizontal component of the motion vector $\mathbf{u}_c^{(k',k'+1)}$ and v is its vertical component; ρ_D and ρ_S are penalty functions related respectively to the data conservation and the spatial coherence terms and λ controls the relative importance of the two terms. In order to reduce the possibility that some artifacts be present in the waveform to be generated, robust functions [20] are selected to be used as penalty functions because of their robustness to the noise and to a possible presence of multiple motions within $\mathcal{R}$. A possible presence of noise in the images may be due to the acquisition conditions of the video recordings.

Commonly, the OF estimation technique is performed at different resolution levels. More precisely, at the full resolution level $j = 0$, the images $I_c^{(k')}$ and $I_c^{(k'+1)}$ are spatially filtered and downsampled giving rise to the respective sub-images $I_c^{(k')j}$ and $I_c^{(k'+1)j}$ at the resolution level $j = 1$. This process is recursively repeated until a pyramid of a preselected number J of resolution levels is constructed. Then, starting from the coarsest resolution level $j = J$, the motion field $\mathbf{u}_j^{(k',k'+1)}$ is estimated between $I_c^{(k')j}$ and $I_c^{(k'+1)j}$ and projected to the finer resolution level $(j-1)$ of the pyramid after appropriately scaling it. At that finer level, one of the two sub-images is warped toward the second sub-image with respect to the estimated flow $\mathbf{u}_j^{(k',k'+1)}$ and according to the technique suggested in [21]. This warped image is used to estimate the motion at the resolution level $(j-1)$. Then, in order to alleviate the effect of the presence of possible artifacts like the bright spots that appear in the laryngeal recordings due to the acquisition conditions, a 5×5 median filter is applied to the motion field [22], giving rise to $d\mathbf{u}_{j-1}$. The final motion vector at this level $(j-1)$ is then $(\mathbf{u}_{j-1} + d\mathbf{u}_{j-1})$. This process is repeated until the computation of the flow $\mathbf{u}_c^{(k',k'+1)}$ at the full resolution level $j = 0$. It is worth pointing out that a one-level pyramidal decomposition ($J = 1$) is sufficient to estimate the VF movements as the motion magnitudes between two consecutive images are small enough.

As mentioned above, the vibratory behavior of the VFs is analyzed at a chosen level $\mathcal{L}$ of the glottis. This requires the localization of the glottis in the image using a segmentation technique. In order to limit the segmentation errors and reduce the user intervention, the segmentation is only applied to the reference image of each cycle.

Glottis segmentation in the reference image

The segmentation of the glottis in $I_c^{\text{ref}}, \forall c \in \{1,\ldots,C\}$ is performed according to the technique proposed in [8] and is based on an active contour model: an initial contour ζ_0 located within the ROI iteratively deforms to produce a contour that ideally superimposes the boundaries of the glottis after a number of iterations. The contour deformation is the consequence of the minimization of an energy functional until convergence or until reaching a predefined number of iterations. Two main steps are necessary and are briefly described in what follows: the contour initialization and the choice of the energy functional.

As the glottal shape is mostly elliptic when the glottis is in its maximum opening, it is convenient to choose an ellipse as an initial curve for the active contour model. The center of the ellipse corresponds to the center of the region R_g. The major and minor axe lengths of the initial contour are proportional to the dimensions of R_g, and its orientation is the same as the one of R_g. At each iteration t, the curve $\zeta(t)$ evolves by taking a position and a topology depending on the value of an energy E_c. The energy has to be minimized, and is evaluated on all the N points p_i of the curve. In the present work, the Chan-Vese model is used [23]. It assumes that the region to be segmented has a constant intensity. However, as the glottal region presents some inhomogeneities, the Chan-Vese model is applied to small local regions defined by balls $\mathcal{B}_i$ respectively centered at the points p_i, $i = 1,\ldots,N$ of the curve, and having a radius r [24]. Given two points of spatial positions respectively $\mathbf{x}$ and $\mathbf{y}$ in the domain of the image, the ball $\mathcal{B}_i$ is defined by :

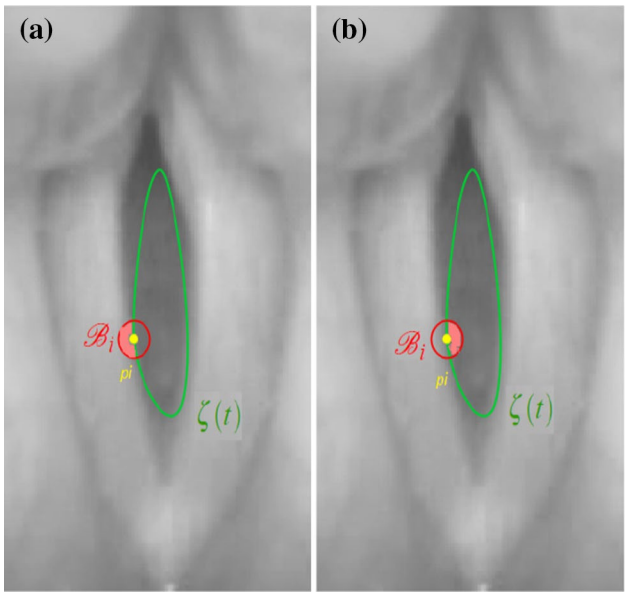


Fig. 4 Local energy minimization for the active contour model. Local regions are defined by balls $\mathcal{B}_i$ (in red) centered at each point p_i (in yellow) of the curve $\zeta(t)$ (in green). The energy to be minimized is related to the mean of the intensities **a** in the ball and outside the curve, in one hand and **b** in the ball and inside the curve on the other hand

$$\mathcal{B}_i(x, y) = \begin{cases} 1 \text{ if} \|\mathbf{x} - \mathbf{y}\| < r \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

Let m_i^{in} (resp. m_i^{out}) be the mean value of the intensities of the pixels that are in the ball $\mathcal{B}_i$ and inside (resp. outside) the curve $\zeta(t)$ (Fig. 4). The energy to be minimized at each iteration t is thus defined by:

$$\begin{aligned} E_c\left(m_i^{\text{in}}, m_i^{\text{out}}, \zeta(t)\right) = & \sum_{(x_i, y_i) \in \mathcal{B}_i} \lambda_1 \left(I_c^{\text{ref}}(x_i, y_i) - m_i^{\text{in}}\right)^2 b\left(x_i, y_i\right) \\ & + \lambda_2 \left(I_c^{\text{ref}}\left(x_i, y_i\right) - m_i^{\text{out}}\right)^2 \left(1 - b\left(x_i, y_i\right)\right) \\ & + \mu \,\text{length}\left(\zeta(t)\right) \end{aligned} \tag{5}$$

where λ_1, λ_2 and μ control the contribution of each term, and $b\,(x_i, y_i)$ is a binary function that takes 1 if the pixel of coordinates (x_i, y_i) in the ball $\mathcal{B}_i$ is inside the curve $\zeta(t)$ at the iteration t, and 0 if it is outside the curve. The model is solved using a level set formulation [23] and the algorithm runs until 200 iterations are reached. This segmentation step gives rise to a binary map M_c^{ref} for each cycle $c = 1, \ldots, C$, where the detected glottal region is identified by pixels of value 1. It is worth pointing out that the segmentation is carried out only within the selected ROI. It is also to be noted that the image of maximal opening (the reference image in the cycle) is chosen to be segmented as the VFs are more visible in this image.

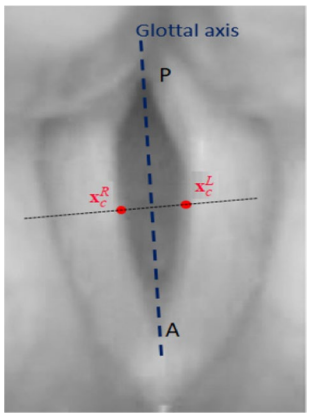


Fig. 5 Points of interest $\mathbf{x}_c^R$ and $\mathbf{x}_c^L$ (in red) located at $\mathcal{L} = 50\%$ from the posterior commissure (P) to the anterior commissure (A)

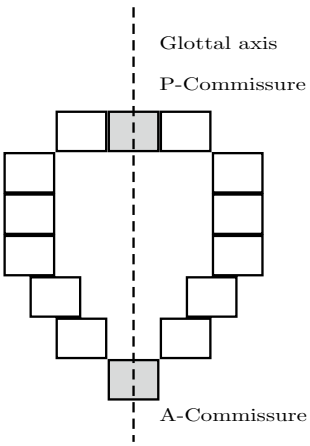


Fig. 6 Illustration of the way to locate the posterior (P) and the anterior (A) commissures

The proposed OF based waveform for oscillatory behavior tracing

Given the detected glottis in the reference image of each cycle, we aim to analyze the cycle-to-cycle vibrations by individually describing the oscillations at some pairs of points $\{(\mathbf{x}_c^R, \mathbf{x}_c^L)\}$ on the glottal edges, where $\mathbf{x}_c^R$ (resp. $\mathbf{x}_c^L$) is the spatial position of a point on the right (resp. left) glottal edge located at a level $\mathcal{L}$ from the posterior commissure to the anterior commissure as shown in Fig. 5.

Given the contour of the glottal region in M_c^{ref}, the P-commissure (resp. A-commissure) is approximately determined as the median point in the top (resp. in the bottom) of the glottal contour as illustrated in Fig. 6. The segment that binds P and A corresponds to the glottal axis. It splits the glottal contour into the right and the left edges. It is worth pointing out that when the glottis is not vertical or does not have a slight deviation from the vertical direction, a rotation

of the image is necessary. The glottal level of interest $\mathcal{L}$ is determined by selecting the point o_c located at the level $\mathcal{L}$ of the glottal axis. The intersection point between the perpendicular to the glottal axis on o_c and the right (resp. left) contour is $\mathbf{x}_c^R$ (resp. $\mathbf{x}_c^L$). In what remains, these points are denoted by $\mathbf{x}_c^\delta$, where $\delta \in \{R, L\}$. The aim is to quantify the displacements of $\mathbf{x}_c^\delta$ between the pairs of successive images of the cycle based on the estimated flow field. Starting from the reference image I_c^{ref}, the point $\mathbf{x}_c^\delta$ of coordinates $(x^{\delta(0)}, y^{\delta(0)})$ in I_c^{ref} is tracked along the cycle, in the backward and forward directions with respect to the reference image. Hence, given the motion vector $\mathbf{u}_c^{(\text{ref},l)}(\mathbf{x}_c^\delta)$, $l \in \{1, -1\}$ of $\mathbf{x}_c^\delta$ between the reference image and $I_c^{(l)}$, the spatial position $\mathbf{x}_c^{(\delta,l)} = (x_c^{(\delta,l)}, y_c^{(\delta,l)})$ of $\mathbf{x}_c^\delta$ in $I_c^{(l)}$ taking into account the estimated displacement $\mathbf{u}_c^\delta$ $(u_c^{(\delta,0,l)}, v_c^{(\delta,0,l)})$ is deduced by:

$$\begin{aligned} x_c^{(\delta,l)} &= x_c^{(\delta,0)} + u_c^{(\delta,0,l)}, \\ y_c^{(\delta,l)} &= y_c^{(\delta,0)} + v_c^{(\delta,0,l)}. \end{aligned} \tag{6}$$

More generally, given the spatial position $(x_{c,\mathcal{L}}^{(\delta,k')}, y_{c,\mathcal{L}}^{(\delta,k')})$ of $\mathbf{x}_{c,\mathcal{L}}^{(\delta,k')}$ in the image $I_c^{(k')}$ at the level $\mathcal{L}$ of the glottis, the spatial position $(x_{c,\mathcal{L}}^{(\delta,k'+l)}, y_{c,\mathcal{L}}^{(\delta,k'+l)})$ of this point in $I_c^{(k'+l)}$, $l \in \{-1, 1\}$ taking into consideration its displacement $(u_c^{(\delta,k',k'+l)}, v_c^{(\delta,k',k'+l)})$ between $I_c^{(k')}$ and $I_c^{(k'+l)}$ is deduced by:

$$\begin{aligned} x_{c,\mathcal{L}}^{(\delta,l,k'+l)} &= x_{c,\mathcal{L}}^{(\delta,k')} + u_c^{(\delta,k',k'+l)}, \\ y_{c,\mathcal{L}}^{(\delta,k'+l)} &= y_{c,\mathcal{L}}^{(\delta,k')} + v_c^{(\delta,k',k'+l)}. \end{aligned} \tag{7}$$

The displacements from an image to another one with respect to the reference image are stacked into the vector $\mathbf{V}_{c,\mathcal{L}}^\delta$ defined by :

$$V_{c,\mathcal{L}}^\delta = \left[d_{c,\mathcal{L}}^{(\delta,-n_1^c)}, \ldots, d_{c,\mathcal{L}}^{(\delta,-1)}, 0, d_{c,\mathcal{L}}^{(\delta,1)}, d_{c,\mathcal{L}}^{(\delta,2)}, \ldots, d_{c,\mathcal{L}}^{(\delta,n_2^c)} \right] \tag{8}$$

where:

$$\begin{aligned} d_{c,\mathcal{L}}^{(\delta,k')} &= \sqrt{\left(x_{c,\mathcal{L}}^{(\delta,k')} - x_{c,\mathcal{L}}^{(\delta,0)}\right)^2 + \left(y_{c,\mathcal{L}}^{(\delta,k')} - y_{c,\mathcal{L}}^{(\delta,0)}\right)^2}, \\ &\forall k' \in \left[-n_1^c, \ldots, n_2^c\right]. \end{aligned} \tag{9}$$

Ideally, for healthy VF, the vector $\mathbf{V}_{c,\mathcal{L}}^\delta$ is symmetric with respect to the position of the reference image, reflecting the oscillatory behavior of the points $\mathbf{x}_{c,\mathcal{L}}^\delta$ selected on the VF edge. A pseudo code detailing the computation of this vector is given by Algorithm 3.

The concatenation of all the vectors $\{\mathbf{V}_{c,\mathcal{L}}^\delta\}_{c\in[1,\ldots,C]}$ gives rise to a waveform that traces the local oscillatory movements of the VF at the selected level $\mathcal{L}$ during the whole duration of phonation. Figure 7 shows the right and the left waveforms at a level of 50% from the posterior to the anterior commissures corresponding to a glissando phonation along 14 cycles (200 images).

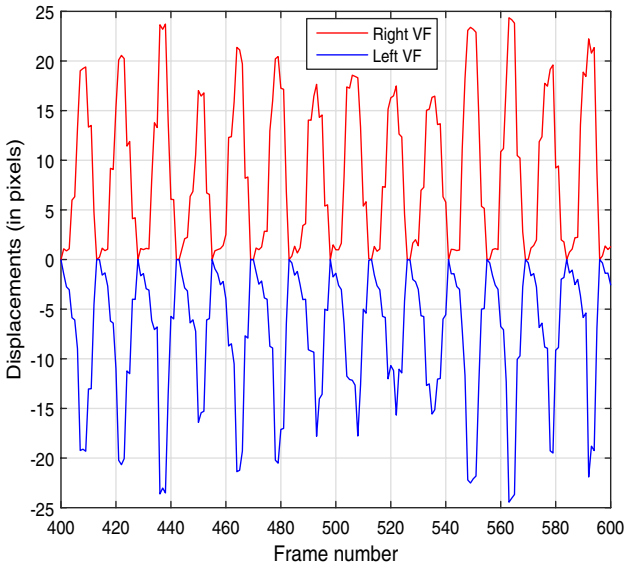


Fig. 7 Illustration of the waveforms generated for the right (in red) and the left (in blue) VF. The waveforms are based on the displacements (deduced from the OF field) of the right and the left vocal folds with respect to their positions at the maximum opening of the glottis. The amplitudes of the waveform are eventually not negative but represented as such for more clarity

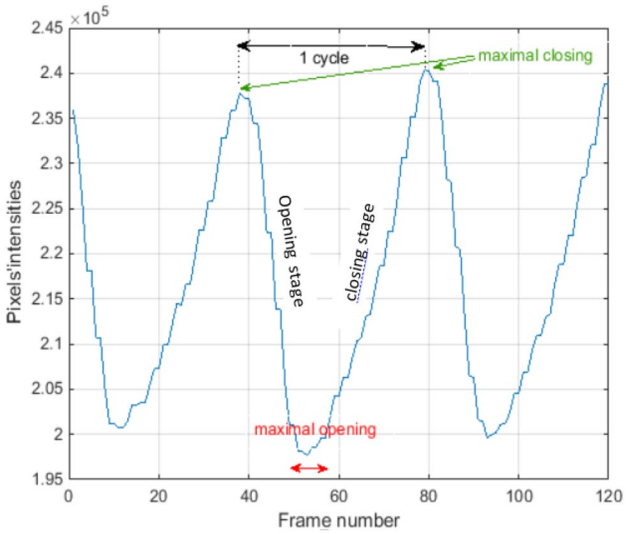


Fig. 8 Details of a single cycle within the waveform

Visually, the movements of the two VF are well synchronized and have almost the same amplitudes. For the sake of clarity, the displacements of the left VF are represented in the figure in the negative part of the y-axis. Unlike the traditional representations, the minima of the proposed OF based waveform correspond to the maximal opening of the glottis whereas the maxima in the waveform correspond to the maximal closing as illustrated in Fig. 8. This is an immediate

Algorithm 3 Waveform generation for one single cycle

Input: $\{I_c^{(k')}\}_{k'=-n_1^c,\ldots,n_2^c}$, M_c^{ref}, $\mathbf{u}_c$, $\{\mathbf{x}_c^{\delta,k'}\}_{k'=-n_1^c,\ldots,n_2^c}$
Output: $\mathbf{V}_{c,\mathcal{L}}^{\delta}$
$d_{c,\mathcal{L}}^{(\delta,0)} = 0$
for $k' = 0,\ldots,n_2^c - 1$ **do**
$x_{c,\mathcal{L}}^{(\delta,k'+1)} = x_{c,\mathcal{L}}^{(\delta,k')} + u_c^{(\delta,k',k'+1)}$
$y_{c,\mathcal{L}}^{(\delta,k'+1)} = y_{c,\mathcal{L}}^{(\delta,k')} + v_c^{(\delta,k',k'+1)}$
$d_{c,\mathcal{L}}^{(\delta,k'+1)} = \sqrt{(x_{c,\mathcal{L}}^{(\delta,k')} - x_{c,\mathcal{L}}^{(\delta,0)})^2 + (y_{c,\mathcal{L}}^{(\delta,k')} - y_{c,\mathcal{L}}^{(\delta,0)})^2}$
end for
for $k' = 0,\ldots,-n_1^c + 1$ by a step of -1 **do**
$x_{c,\mathcal{L}}^{(\delta,k'-1)} = x_{c,\mathcal{L}}^{(\delta,k')} + u_c^{(\delta,k',k'-1)}$
$y_{c,\mathcal{L}}^{(\delta,k'-1)} = y_{c,\mathcal{L}}^{(\delta,k')} + v_c^{(\delta,k',k'-1)}$
$d_{c,\mathcal{L}}^{(\delta,k'-1)} = \sqrt{(x_{c,\mathcal{L}}^{(\delta,k')} - x_{c,\mathcal{L}}^{(\delta,0)})^2 + (y_{c,\mathcal{L}}^{(\delta,k')} - y_{c,\mathcal{L}}^{(\delta,0)})^2}$
end for
$i = 1$
for $k' = -n_1^c,\ldots,n_2^c$ **do**
$\mathbf{V}_{c,\mathcal{L}}^{\delta}(i) = d_{c,\mathcal{L}}^{(\delta,k'-1)}$
$i = i + 1$
end for

consequence of considering the image of maximal opening as a temporal starting point from which the displacements of the VF are calculated.

It is worth noting that the proposed method is robust to the glottal axis shift and to the endoscope movement. The right and left displacements are compared relatively to each other without the need of determining the axis position in each image: a shift in the position of the glottal region leads to the perception of a larger or smaller displacement than in reality and impacts similarly both VFs but does not influence their relative displacements.

Results

The experimental assessment of the proposed technique is performed in three stages aiming to evaluate the proposed approach.

First, the partitioning of the HSV sequence into cycles is evaluated and compared to manually identified cycles. Its reliability is also verified by comparing the estimated fundamental period with the one estimated based on the analysis of the EGG signals by the method 3/7 [25]. Tests are conducted on healthy and pathological phonation and the results are shown and discussed in this section.

Second, the reliability of the proposed technique to trace the VF trajectories is evaluated on healthy phonation. Given the difficulty of creating a ground truth to evaluate the estimation of the OF field within the glottal region [12, 26], the assessment is conducted similarly to the one adopted in [12]: the obtained VF trajectories are compared to those resulting from the segmentation of all the images of the HSV followed by the calculation of the displacements of the points through the sequence.

Third, the proposed waveform is assessed on healthy and pathological phonation using (1) a visual interpretation of the waveform. (2) A comparison between the proposed waveform and DKG in terms of reliability of their interpretation.

Evaluation of the cycle detection

Figure 9 depicts the cycle partitioning associated to excerpts taken from four HSV recordings of various non-pathological phonation: relaxed, glissando, a sustained sound moving from a breathy to a tense phonation, and crescendo. In all these excerpts, the temporal variation of the sum of the pixel intensities within the ROI well depicts the cyclic behavior of the VF vibrations. The evaluation of the proposed cycle partitioning technique is performed by comparing the estimated fundamental period of S to the fundamental period estimated on the associated EGG signals based on the technique proposed in [25, 27]. The results depicted in Table 1 show a high accuracy of the proposed technique.

In the pathological phonation, the cycle partitioning appears to be less intuitive, which is in accordance with the pathological character of the phonation. Figure 10 shows the temporal variation of the pixel intensities related to pathological VFs suffering respectively from paralysis and cancer of the larynx, where the periodicity of the obtained signal is

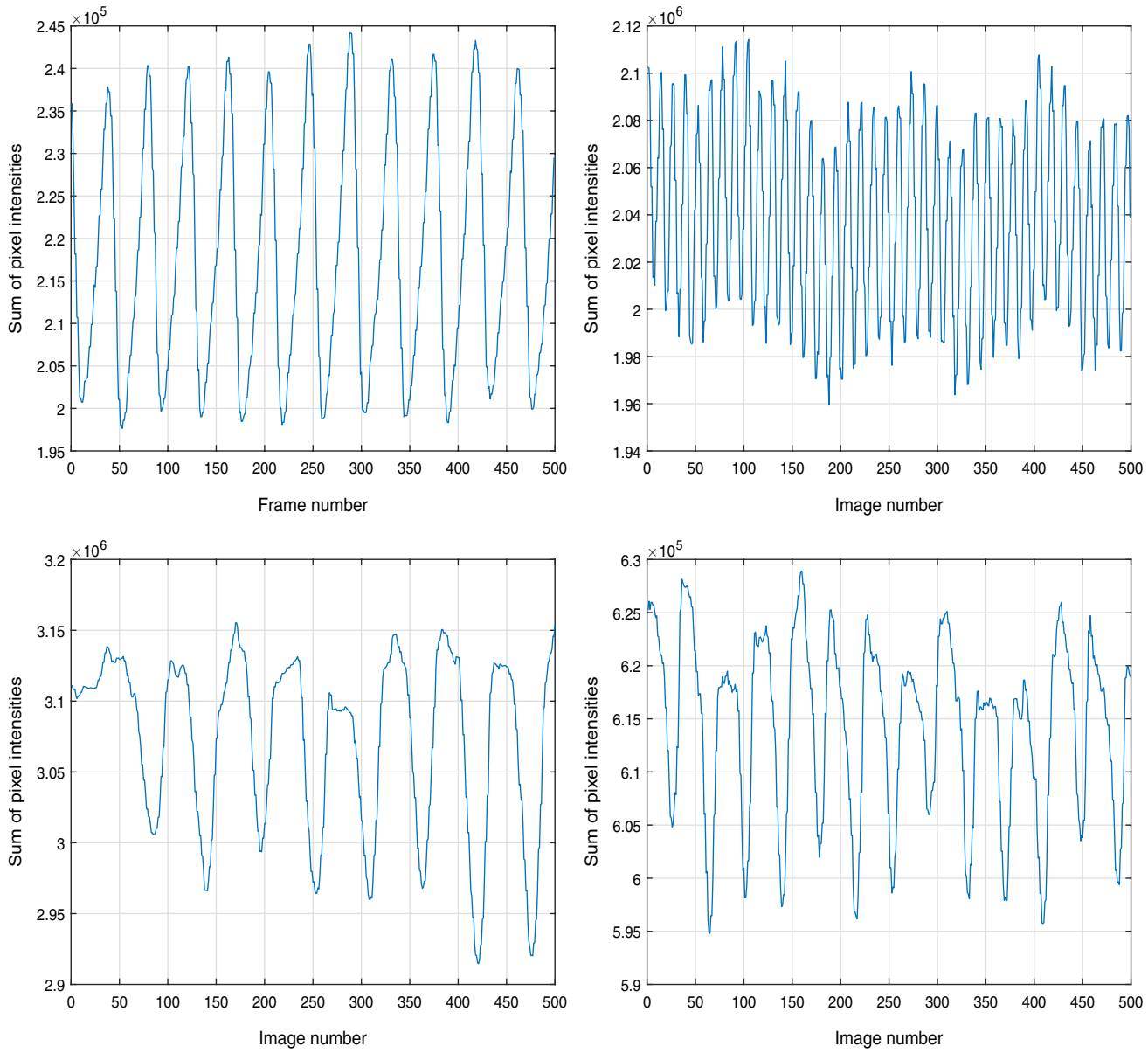


Fig. 9 Temporal variation of the sum of pixel intensities in healthy vocal folds. From top to down and from left to right: relaxed phonation, glissando, sustained sound, crescendo

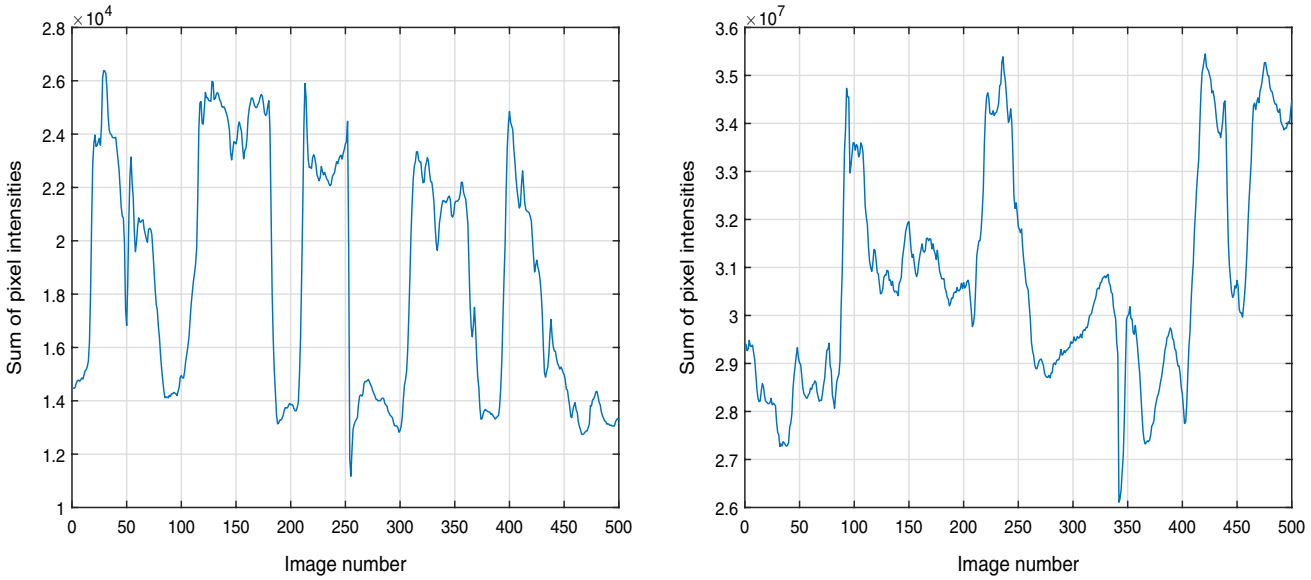
Table 1 Evaluation of the fundamental period estimates (and hence, of the cycle partitioning technique)

	EGG (ms)	Proposed technique (ms)
Voice 1	6.5	6.5
Voice 2	1.9	2.1
Voice 3	8.6	8.3
Voice 4	5.5	5.9

altered. It is to be noted that the EGG signals associated to HSV of pathological VF are not available.

Comparisons with segmentation based trajectories

In order to evaluate the reliability of the proposed approach, the VF waveforms are compared to the VF trajectories obtained through the segmentation of all the images. The displacement $d_{\text{seg}}^{(k',k'+1)}(\mathbf{x}_c^{\delta})$ of a point $\mathbf{x}_{c,\mathcal{L}}^{\delta}$ ($\delta \in \{R, L\}$) between its position $\mathbf{x}_{c,\mathcal{L}}^{\delta(k')}$ in the image $I_c^{(k')}$ and its position $\mathbf{x}_{c,\mathcal{L}}^{\delta(k'+1)}$ in the image $I_c^{(k'+1)}$ is defined by :



(a) $T = 16.1$ ms **(b)** $T = 16.6$ ms

Fig. 10 Temporal variation of the sum of pixel intensities in pathological vocal folds. **a** Paralysis, **b** cancer of the larynx

$$d_{\text{seg}}^{(k',k'+1)}(x_{c,\mathcal{L}}^{\delta}) = \left\| \mathbf{x}_{c,\mathcal{L}}^{\delta(k'+1)} - \mathbf{x}_{c,\mathcal{L}}^{\delta(k')} \right\|_{L^2}. \tag{10}$$

The tracing of the trajectories in each cycle, based on the calculated displacements between each pair of images, is performed in backward and forward directions starting from the reference image. It is worth noting that the segmentation is automatically performed on the reference images according to the technique described above, and manually executed on the remaining images of the cycles in order to minimize the segmentation errors. By this way, we aim to compare the proposed waveforms to the full segmentation based algorithm when this latter gives as accurate results as possible. For this purpose, the Pearson's correlation coefficient between the OF based- and the segmentation based-trajectories is calculated for each cycle of a sequence. The correlation results aim to provide an idea about how well both trajectories have a similar behavior. Two HSV sequences of about ten cycles related to healthy VF are used. The first sequence contains about 45 images/cycle, while the second one contains about 12 images/cycle. A HSV sequence related to a paralysis is also used to evaluate the accuracy of the OF based waveforms in presence of disorders. It is composed of about four cycles, each cycle contains about 170 images. The first column of Fig. 11 depicts the correlation coefficients between the OF based waveforms and the full segmentation based trajectories of the right VF (the curve in red) and those of the left VF (the curve in blue). The mean value of the correlation ranges between 0.73 and 0.94 for the right VF, and between 0.75 and 0.92 for the left VF. Note that the trajectories that provide the highest correlation in each sequence are depicted in the second column, while those that provide the lowest correlation are depicted in the third column. A cycle by cycle examination of the obtained values shows a compliance between the proposed technique and the fully segmented sequence although some differences naturally exist between the two types of trajectories. In fact, very small vibrations are hardly perceived by the segmentation procedure which explains the small variations of at most one pixel observed in the OF based waveforms, against a null displacement in the segmentation based trajectories. This behavior is illustrated in Fig. 11b, c between the images 8 and 40 and in Fig. 11f between the images 6 and 10. The same holds for the trajectories related to the pathological phonation given in Fig. 11h, i of the same figure. This reduces the correlation between the two trajectories even though they globally have the same aspect. Larger differences in the pixel displacements between the OF based- and the segmentation based- approaches are noticed in some situations during the closing stage of the VF, as illustrated in Fig. 11. This is due to the difficulty of spatially locating in the image the glottal edges during the segmentation step when the VF are completely closed (even when the segmentation is carried out manually for the evaluation purposes as stated above). This explains rarely obtained low correlations as shown in Fig. 11d at the 6th cycle of the right VF vibrations. Figure 11g shows a high correlation in the case of pathological VFs suffering from a paralysis of the right cord.

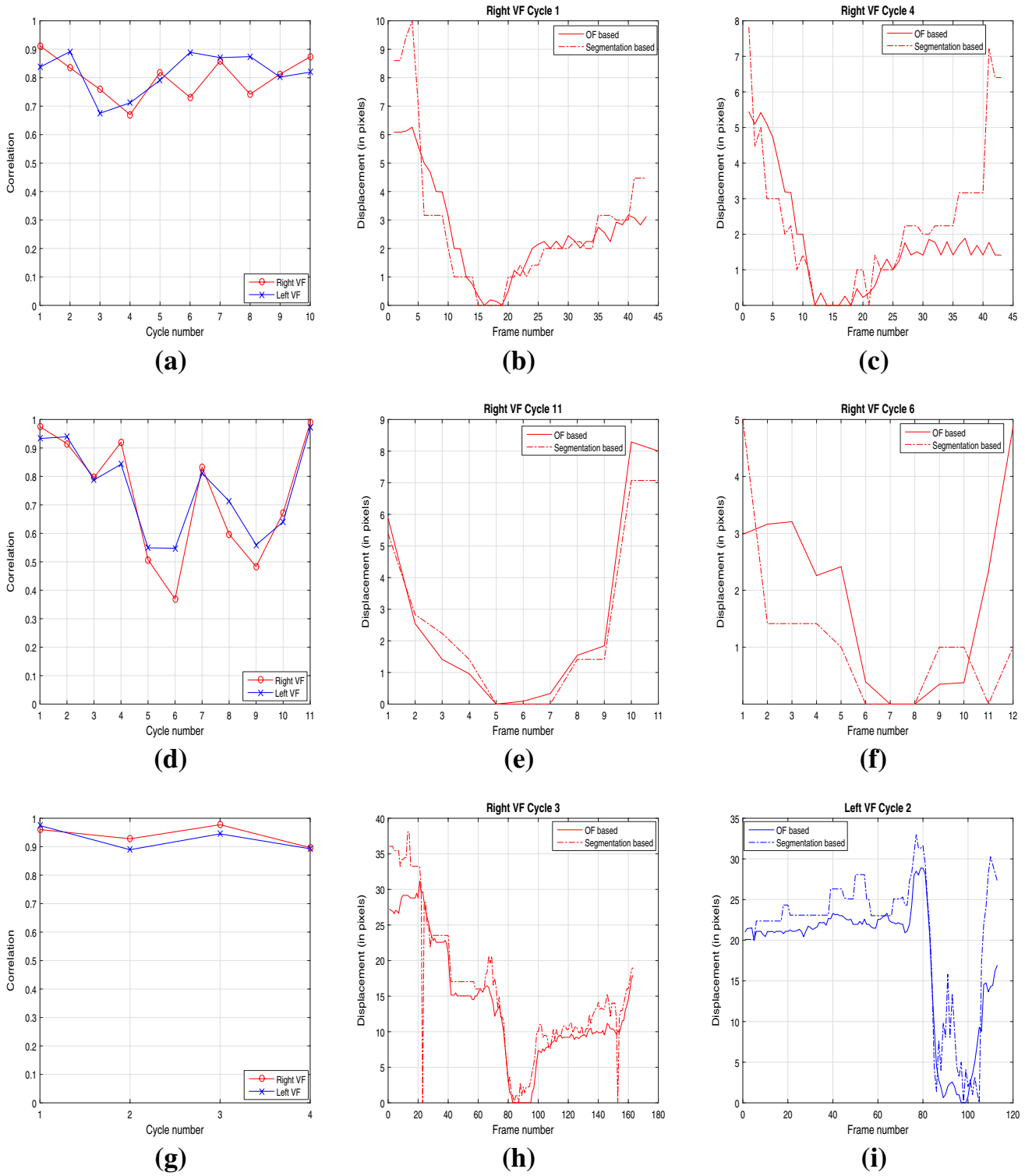
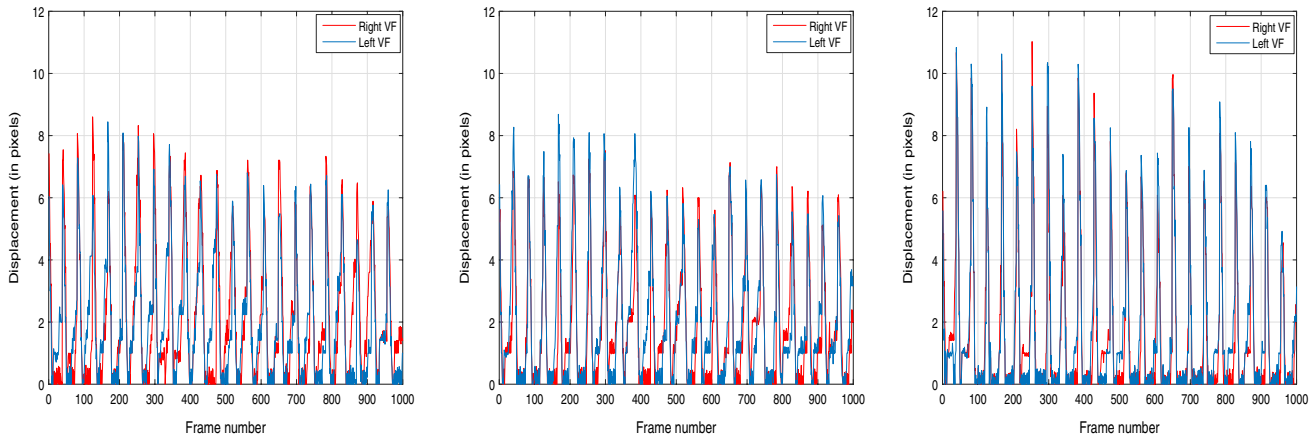
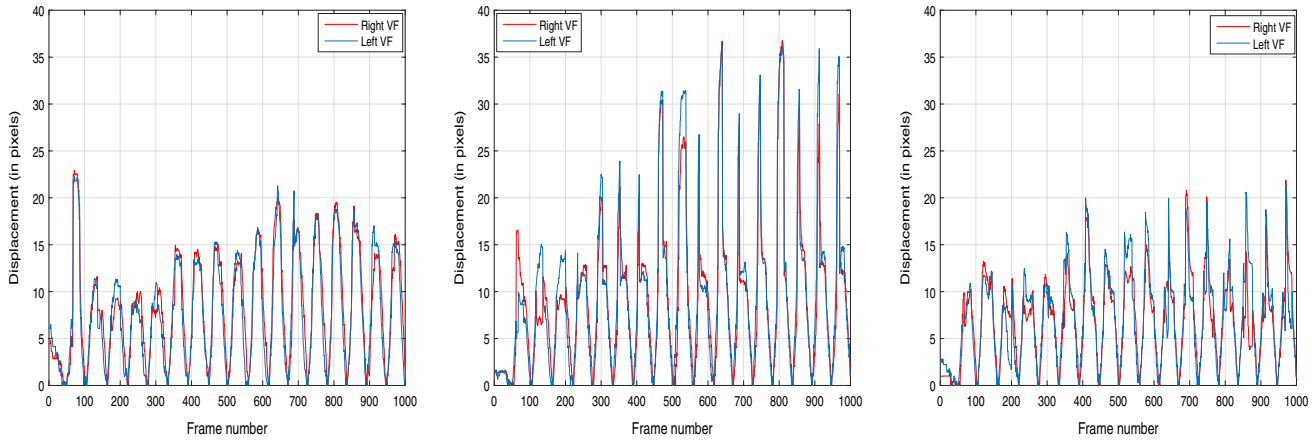


Fig. 11 Correlation between OF based waveforms (in red) and segmentation based trajectories (in blue). First and second rows: healthy VF. Third row: paralysis



(a) $\mathcal{L} = 25\%$ **(b)** $\mathcal{L} = 50\%$ **(c)** $\mathcal{L} = 75\%$

Fig. 12 OF based waveforms of the VF at the **a** posterior, **b** mid-glottis, **c** anterior levels. relaxed voice (1000 images)



(a) $\mathcal{L} = 25\%$ **(b)** $\mathcal{L} = 50\%$ **(c)** $\mathcal{L} = 75\%$

Fig. 13 OF based trajectories of the VF at **a** 25%, **b** 50%, **c** 75%. (Sample USC2008.02.050.ouv.renf.50.frames of the database, 130 cycles, 7200 images)

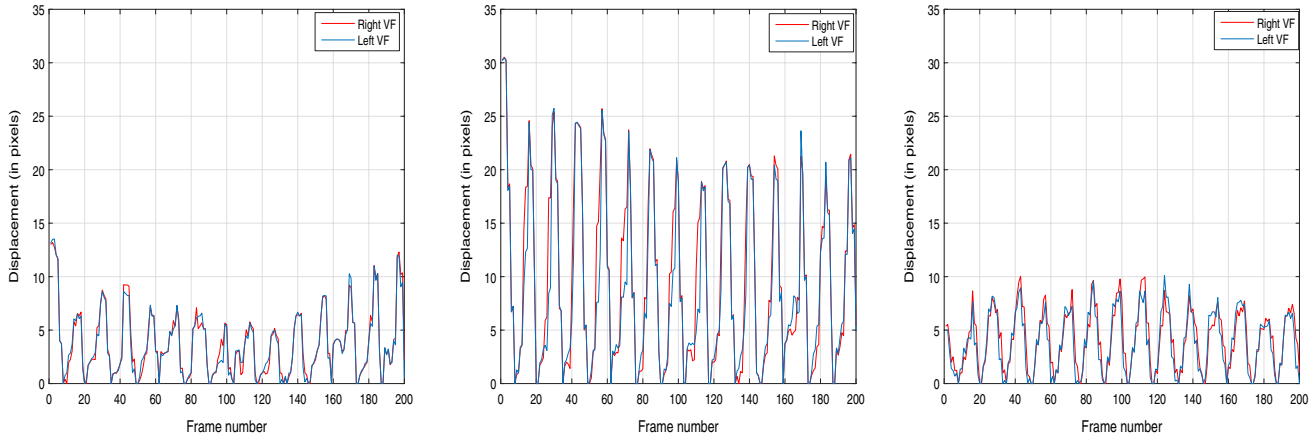
Visual assessment of the proposed waveform

The interpretation of the proposed waveform is possible visually and is evaluated on healthy and pathological VF.

Healthy vocal folds

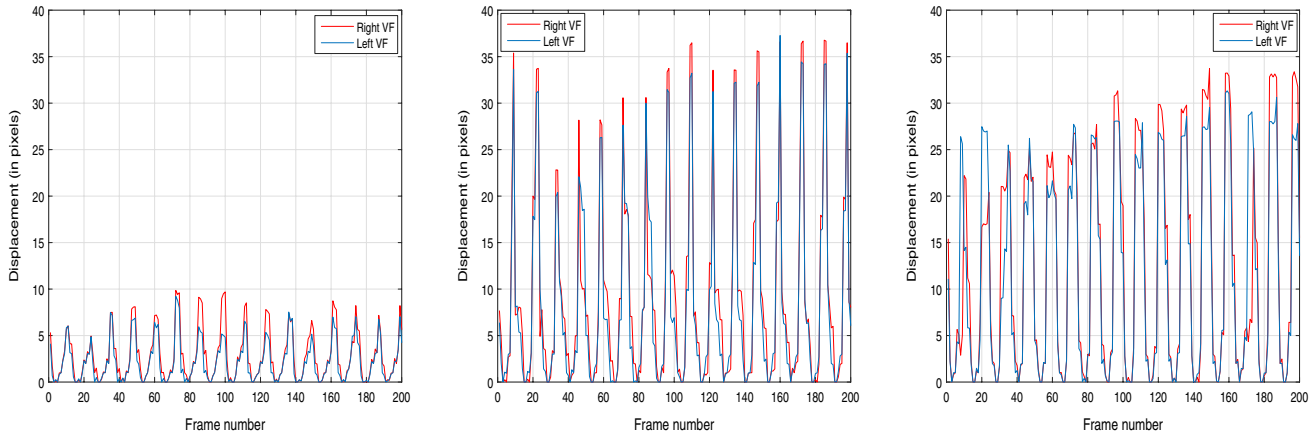
Tests are conducted using several laryngeal mechanisms on healthy VFs in order to evaluate the proposed waveform in terms of reliability of the visual assessment. Ideally, the waveforms of the left and right VFs should be as superimposed as possible ensuring phase and amplitude symmetries. In addition, the same pattern should be observed among many cycles.

Figures 12, 13, 14 and 15 show the waveforms of the left and right VFs at the glottal levels 25%, 50% and 75% from the posterior to the anterior commissures related to four different types of phonation. The waveforms of the complete videos are given in the Online Resource as a supplementary material. Visually, the vibrations are phase-symmetric, a slight difference in amplitude and a periodic behavior along time are observed which indicate a healthy behavior of the phonation. Furthermore, Fig. 16 illustrates the similarities between the proposed waveform and DKG of the VF vibrations. In fact, the first column of the figure presents the vibrations of a relaxed voice. The DKG representation shows periodic and symmetric vibrations; the same interpretation holds based on the visualization of the generated waveform.



(a) $\mathcal{L} = 25\%$ **(b)** $\mathcal{L} = 50\%$ **(c)** $\mathcal{L} = 75\%$

Fig. 14 OF based waveforms of the VF at **a** 25%, **b** 50%, **c** 75%. (Sample USC2008.02.065.glissendo.65.frames, 200 images)



(a) $\mathcal{L} = 25\%$ **(b)** $\mathcal{L} = 50\%$ **(c)** $\mathcal{L} = 75\%$

Fig. 15 OF based waveforms of the VF at **a** 25%, **b** 50%, **c** 75%. (Sample USC2008.02.004.M1.4.frames, 51cycles, 655 images)

The second column represents the vibrations related to a glissando phonation and both DKG and the proposed waveform show the same behavior: periodic and symmetric vibrations with a progressive increase in the amplitudes. Furthermore, the waveform allows a direct quantification of the vibration amplitudes and allows a better visualization of a possible variation of them during phonation. Whereas the quantification of the vibrations through DKG requires the segmentation of this latter [11].

Pathological vocal folds

Three HSVs related to different VF disorders are used to evaluate the efficiency and the reliability of the proposed waveform in assisting clinicians in the diagnosis. The first video is about a right true vocal cord paralysis where the cause of the paralysis is unknown or idiopathic. The video is sampled into 940 images corresponding to nine cycles according to the proposed cycle detection approach. The waveforms of the right and the left VF are depicted in Fig. 17 and clearly show dissimilarities between both VF dynamics.

The patient of the second video suffers from a cancer in the larynx involving the left VF. The waveforms of the two VFs depicted in Fig. 18 visually show a dissimilarity between them. Moreover, a periodic pattern seems to be absent in the waveforms.

The third video is about a left true vocal cord paralysis from injury to the vagus nerve during carotid endarterectomy surgery. More similarity is observed in the

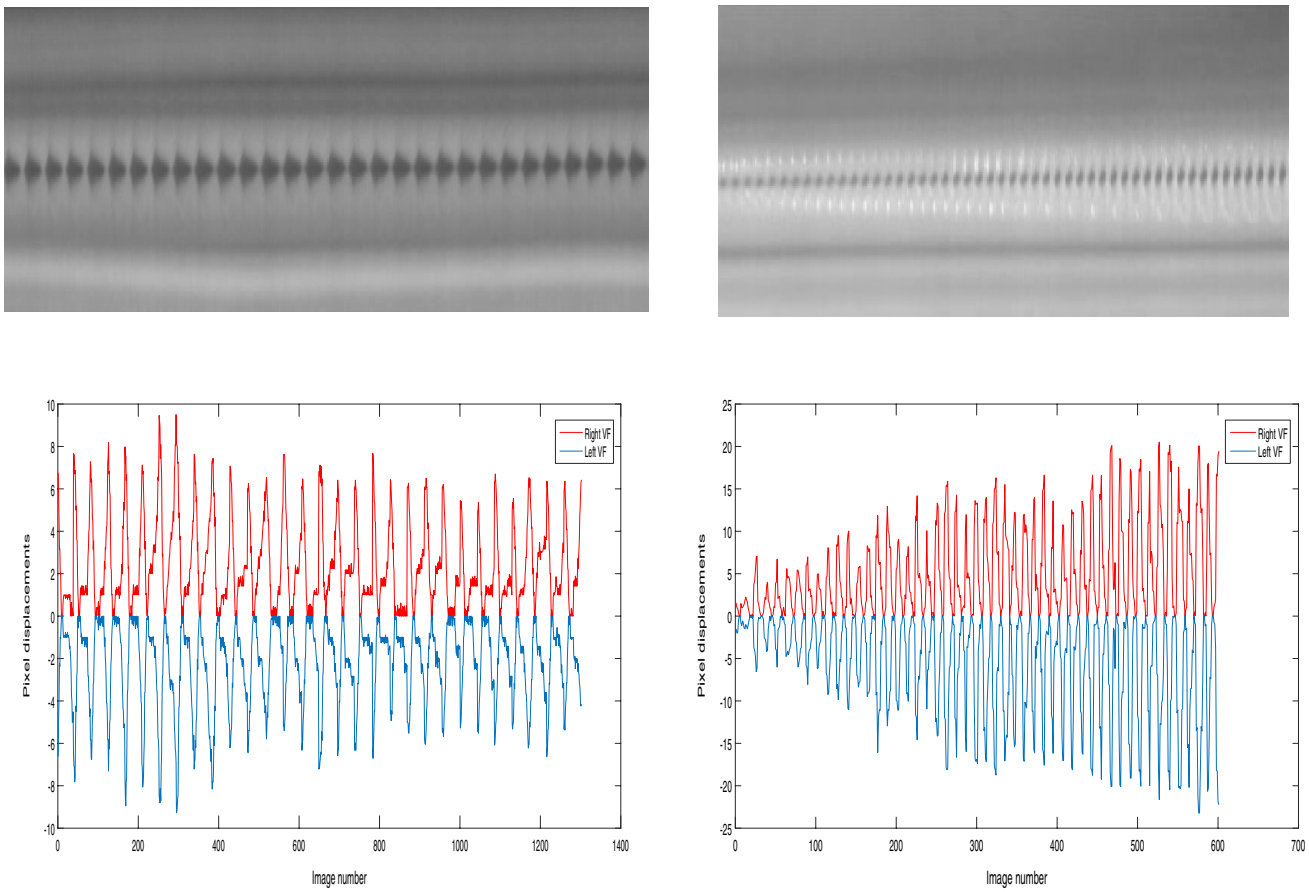
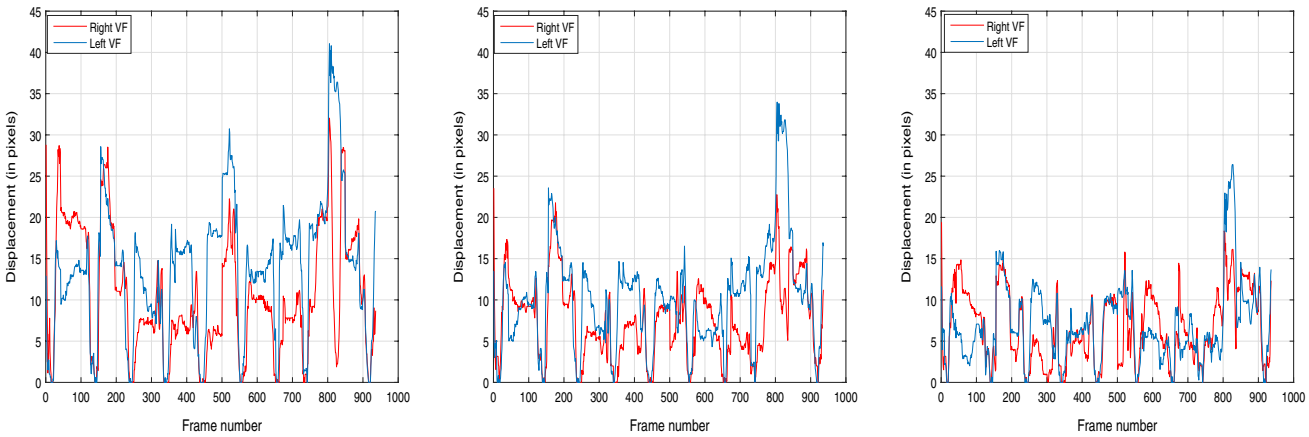
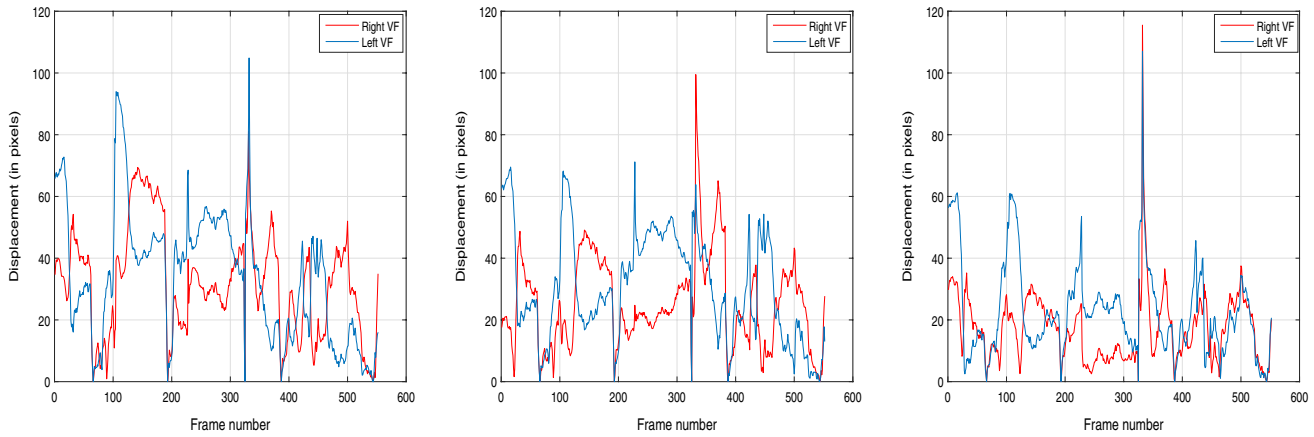


Fig. 16 Comparison between DKG and the proposed waveform computed on healthy phonation. First column: relaxed phonation. Second column: glissando phonation



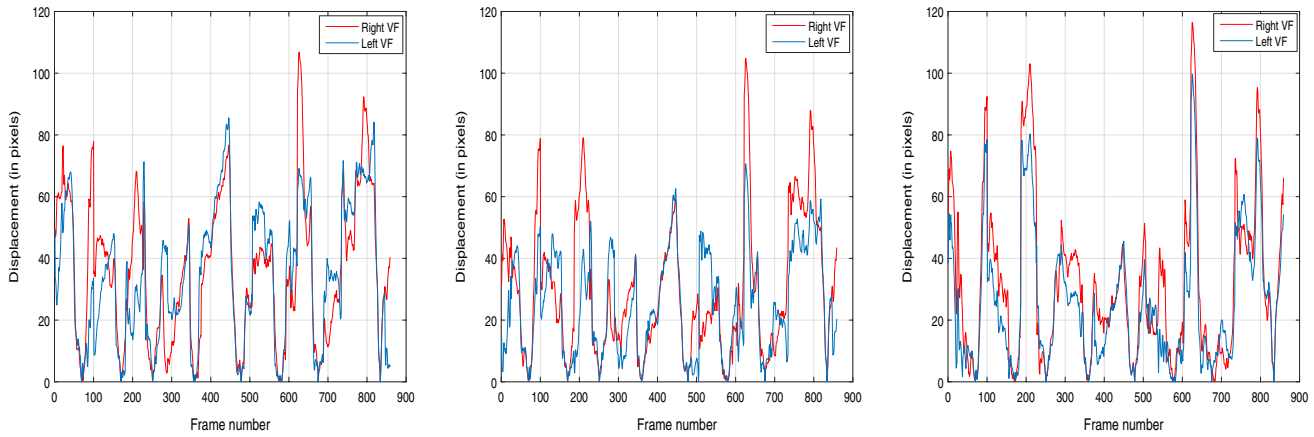
(a) $\mathcal{L} = 25\%$ **(b)** $\mathcal{L} = 50\%$ **(c)** $\mathcal{L} = 75\%$

Fig. 17 OF based waveforms of the VFs at the **a** posterior, **b** mid-glottis, **c** anterior level. paralysis (nine cycles, 940 images)



(a) $\mathcal{L} = 25\%$ **(b)** $\mathcal{L} = 50\%$ **(c)** $\mathcal{L} = 75\%$

Fig. 18 OF based waveforms of the VF at the **a** posterior, **b** mid-glottis, **c** anterior level. Cancer of the left true vocal cord, five cycles, 552 images



(a) $\mathcal{L} = 25\%$ **(b)** $\mathcal{L} = 50\%$ **(c)** $\mathcal{L} = 75\%$

Fig. 19 OF based waveforms of the VF at the **a** posterior, **b** mid-glottis, **c** anterior level. Uncompensated vocal fold paralysis (eight cycles, 860 images)

trajectories of the two VF compared to the previous two videos (Fig. 19), but still the magnitudes of both VF vibrations are noticeably different in addition to the absence of a periodic pattern.

Figure 20 illustrates the DKG representations of two pathological vibrations as well as the resulting waveforms. The pathological behavior is well depicted using both techniques. The first column of Fig. 20 depicts the DKG and the waveforms of VF suffering from cancer. The DKG seems to be affected by an endoscope movement which makes difficult the interpretation of the vibration amplitudes. The second column shows the DKG and the waveforms related to VF suffering from a paralysis in the right VF. This paralysis cannot be perceived through the DKG depicted in the figure; whereas by visualizing the waveforms, low magnitudes of the right VF with respect to the left VF magnitudes is easily noticed.

Discussion

The present study proposes a new waveform that allows a visual assessment of the VF vibrations. Two contributions are proposed whose efficiency and accuracy are experimentally proved. The first contribution consists of automatically identifying the vibratory cycles within the HSV recording.

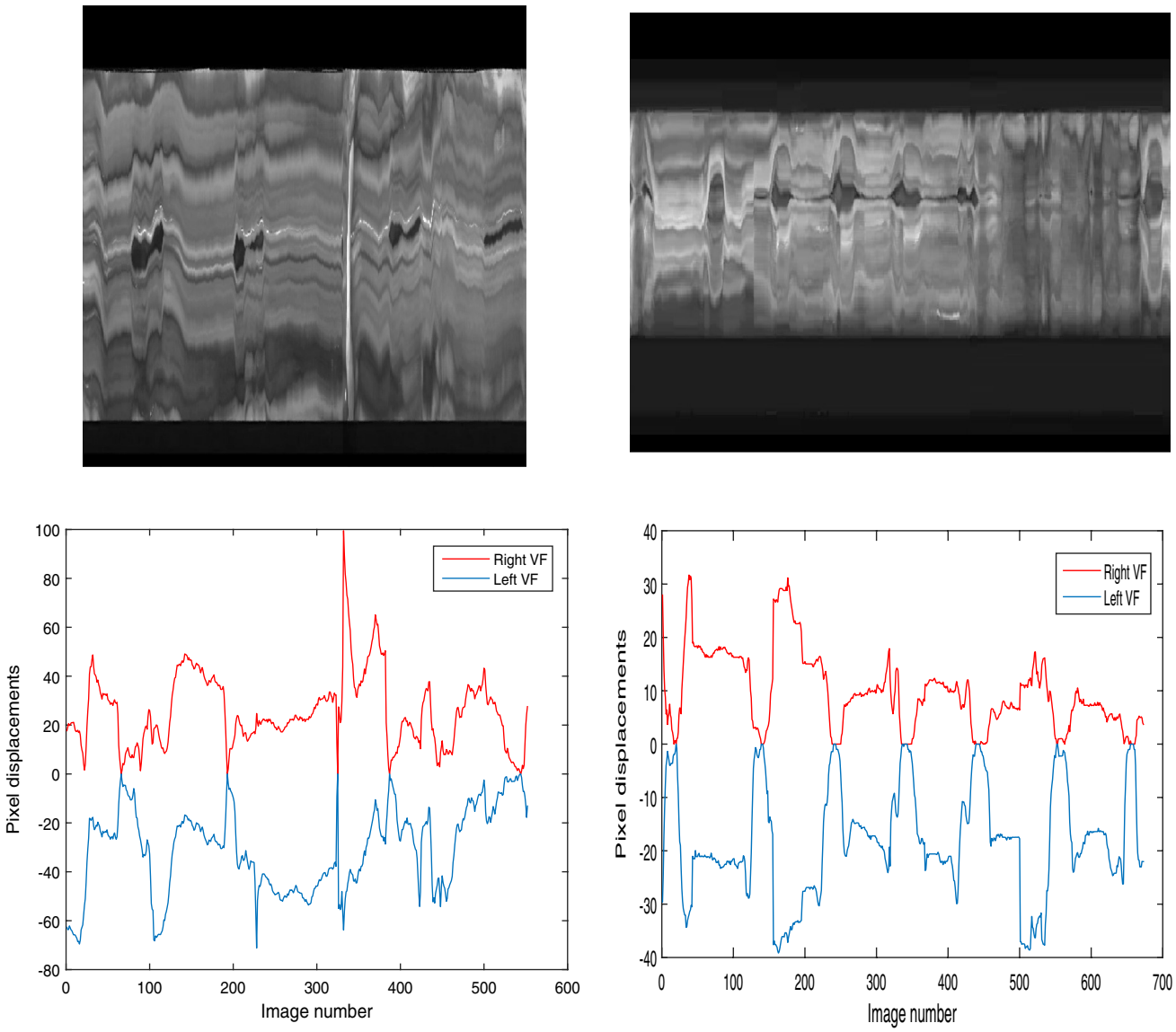


Fig. 20 Comparison between DKG and the proposed waveform computed on pathological phonation. First column: VF suffering from cancer. Second column: VF suffering from paralysis

To this aim, a new feature corresponding to the sum of pixel intensities is proposed. In healthy phonation, the temporal variation of this feature highlights the periodic aspect of the vibrations. This motivated us to identify the vibratory cycles by estimating the fundamental period of the resulting time-series using the non linear curve fit model [18]. For each HSV recording, the estimated fundamental period is compared to the one estimated using the EGG signal related to the same phonation, and found to be very close to it as shown in Table 1. Moreover, this technique can be exploited to suspect a pathological behavior of the VF. For instance, in Fig. 10, a visual analysis of the obtained curve points out the absence of a periodic pattern. In addition, the estimated fundamental frequency (around 60 Hz) is not within the range of the frequencies related to a healthy phonation, as explained in "Introduction".

The second contribution of the present work aims to map into a waveform, the VF dynamics at a selected level of the glottis. The proposed waveform overcomes some limitations of the state-of-the-art techniques. First, while the majority of the existing approaches require the segmentation of all the images of the HSV recording, the generation of the proposed waveform requires the segmentation of only one image per vibratory cycle. Moreover, the segmented image is the one containing a maximal opening of the glottis, and thus ensures a high accuracy of the segmentation result. Second, the proposed waveform is the trajectory of one VF at a given level of the glottal area. In order to comply with the oscillatory

Table 2 Recapitulation of the state-of-the-art representations of the HSV recordings and comparison with the proposed waveform

	Type	Cycle identif.	Local analysis	Robustness to axis shift	Robustness to endoscope movement	Segmented images	Distinction healthy/patho. VF
GAW [6]	W	Visual	No	Yes	Yes	All	Yes [28]
PVG [7]	P	Visual	No	No	Yes	All	Yes
GVG [8]	P	Visual	No	Yes	Yes	All	Yes
OFGVG [12]	P	Visual	No	No	No	None	Not explored
DKG [9]	P	Visual	Yes	Yes	No	All	Yes
OFKG [12]	P	Visual	Yes	No	Yes	None	Not explored
Proposed	W	Automatic	Yes	Yes	Yes	1 Image per cycle	Yes

W waveform, *P* playback, *Identif.* identification

behavior of the VF, the OF is estimated within the glottal region in the backward and the forward directions with respect to the reference image. Moreover, the OF estimation technique used in this work is robust to noise and to undesirable patterns, like the common presence of bright spots in the glottal region. A comparison between the reported techniques and the proposed one is given in Table 2. In order to evaluate the reliability of the proposed scheme, trajectories have been generated through segmenting all the images of the video. Next, the cycle-by-cycle displacements of the selected point of interest are computed with respect to the reference image. High correlations are obtained between the resulting trajectories and the proposed technique. This confirms the efficiency of the proposed waveform where the amount of segmented data is remarkably reduced while offering an easy and accurate way for a visual analysis of the vibrations. Numerous experiments are conducted on HSV recordings of healthy phonation, in order to deduce waveforms at three levels of the glottal region. The visual assessment of the vibrations using the waveforms confirms the periodicity and the symmetry aspects of the examined healthy phonation, and hence the high accuracy of the waveforms. Moreover, the comparison between the DKGs of several HSV recordings and the proposed waveforms shows a high similarity of the trajectories, as illustrated by Fig. 16. Furthermore, the proposed waveform provides a direct visual quantification of the vibrations.

Unhealthy phonation is also explored in the present work. The proposed scheme is twofold sensitive to pathological vibrations. First, the cycle partitioning technique lets suspect abnormal vibrations of the VF. Second, the pathological behavior is confirmed and precisely located on the glottal area by the visual analysis of the waveforms. Figure 20 depicts the same behavior of the vibrations obtained by DKG and the proposed technique.

Conclusion and future work

A new waveform allowing an accurate representation of the cycle-to-cycle VF dynamics is proposed in this study. Three stages are necessary to generate the proposed waveform. First, the HSV recording is partitioned into vibratory cycles. Second, the image of maximum opening in each cycle is segmented in order to locate the glottis. Third, the level of interest is selected and the corresponding points on the right and left VF are tracked in a specific way, giving rise to the waveform. The experiments are conducted on healthy and pathological VFs. They show a high reliability of the proposed system in suspecting a pathology and precisely locating it. However, one issue should be emphasized in the future in order to improve the proposed system. Although the amount of segmented data is highly reduced while obtaining accurate results, the segmentation step remains a complex task especially in the presence of suboptimal illumination conditions. A solution aiming to locate the glottal edges robustly to these conditions is to be investigated with the aim of improving the system performances.

Acknowledgements This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia (Grant No. J-552-612-38). The author, therefore, gratefully acknowledges the DSR financial support.

Compliance with ethical standards

Conflict of interest The author declares having no competing interests.

Informed consent Informed consent was obtained from all individual participants included in the study.

Research involving human participants and/or animals The video recordings related to healthy phonation as well as the EGG and audio signals used in this paper are provided by Erkki Bianco (erkki.bianco@gmail.com) & Gilles Degottex, IRCAM (Institut de Recherche et Coordination Acoustique/Musique, gilles.degottex@ircam.fr), upon permission. The recordings are part of the USC_2008 database produced

by the owners. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The recordings related to pathological phonation are publicly available at www.entusa.com.

References

1. Islam MS, Parvez I, Deng H, Goswami P (2014) Performance comparison of heterogeneous classifiers for detection of Parkinson's disease using voice disorder (dysphonia). In: International conference on informatics, electronics & vision (ICIEV), IEEE, pp 1–7
2. Hartelius L, Svensson P (1994) Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey. Folia Phoniatr Logop 46(1):9–17
3. Yamauchi A, Yokonishi H, Imagawa H, Sakakibara K-I, Nito T, Tayama N, Yamasoba T (2016) Quantification of vocal fold vibration in various laryngeal disorders using high-speed digital imaging. J Voice 30(2):205–214
4. Deliyski D, Kendall K, Leonard R (2010) Laryngeal high-speed videoendoscopy. In: Laryngeal evaluation: indirect laryngoscopy to high-speed digital imaging. Thieme, New York, pp 245–270
5. Deliyski DD, Petrushev PP, Bonilha HS, Gerlach TT, Martin-Harris B, Hillman RE (2007) Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. Folia Phoniatr Logop 60(1):33–44
6. Titze IR (1984) Parameterization of the glottal area, glottal flow, and vocal fold contact area. J Acoust Soc Am 75(2):570–580
7. Lohscheller J, Eysholdt U, Toy H, Dollinger 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(3):300–309
8. Karakozoglou S-Z, Henrich N, d'Alessandro C, Stylianou Y (2012) Automatic glottal segmentation using local-based active contours and application to glottovibrography. Speech Commun 54(5):641–654
9. Wittenberg T, Tigges M, Mergell P, Eysholdt U (2000) Functional imaging of vocal fold vibration: digital multislice high-speed kymography. J Voice 14(3):422–442
10. Hertegård S, Larsson H, Wittenberg T (2003) High-speed imaging: applications and development. Logop Phoniatr Vocol 28(3):133–139
11. Qiu Q, Schutte H, Gu L, Yu Q (2003) An automatic method to quantify the vibration properties of human vocal folds via videokymography. Folia Phoniatr Logop 55(3):128–136
12. Andrade-Miranda G, Bernardoni NH, Godino-Llorente JI (2017) Synthesizing the motion of the vocal folds using optical flow based techniques. Biomed Signal Process Control 34:25–35
13. Löfqvist A, Yoshioka H (1980) Laryngeal activity in Swedish obstruent clusters. J Acoust Soc Am 68(3):792–801
14. Löfqvist A, McGarr NS, Honda K (1984) Laryngeal muscles and articulatory control. J Acoust Soc Am 76(3):951–954
15. Degottex G, Bianco E, Rodet X (2008) Usual to particular phonatory situations studied with high-speed videoendoscopy. In: International conference on voice physiology and biomechanics, pp 19–26
16. Degottex G (2010) Glottal source and vocal-tract separation: estimation of glottal parameters, voice transformation and synthesis using a glottal model. Ph.D. thesis, Paris 6
17. Samet H, Tamminen M (1988) Efficient component labeling of images of arbitrary dimension represented by linear bintrees. IEEE Trans Pattern Anal Mach Intell 10(4):579–586
18. Jiang JJ, Chang C-I, Raviv JR, Gupta S, Banzali FM, Hanson DG (2000) Quantitative study of mucosal wave via videokymography in canine larynges. Laryngoscope 110(9):1567–1573
19. Horn BK, Schunck BG (1981) Determining optical flow. Artif Intell 17(1–3):185–203
20. Black MJ, Anandan P (1996) The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. Comput Vis Image Underst 63(1):75–104
21. Wedel A, Pock T, Zach C, Bischof H, Cremers D (2009) An improved algorithm for tv-l 1 optical flow. In: Statistical and geometrical approaches to visual motion analysis. Springer, Heidelberg, pp 23–45
22. Sun D, Roth S, Black MJ (2010) Secrets of optical flow estimation and their principles. In: IEEE conference on computer vision and pattern recognition (CVPR), IEEE, pp 2432–2439
23. Chan TF, Vese LA (2001) Active contours without edges. IEEE Trans Image Process 10(2):266–277
24. Lankton S, Tannenbaum A (2008) Localizing region-based active contours. IEEE Trans Image Process 17(11):2029–2039
25. Howard DM, Lindsey GA, Allen B (1990) Toward the quantification of vocal efficiency. J Voice 4(3):205–212
26. Baker S, Scharstein D, Lewis J, Roth S, Black MJ, Szeliski R (2011) A database and evaluation methodology for optical flow. Int J Comput Vis 92(1):1–31
27. Howard DM (1995) Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers. J Voice 9(2):163–172
28. Larsson H, Hertegård S, Lindestad P-Å, Hammarberg B (2000) Vocal fold vibrations: high-speed imaging, kymography, and acoustic analysis: a preliminary report. Laryngoscope 110(12):2117–2122