A new technique for assessing glottal dynamics in speech and singing by means of optical-flow computation

Gustavo Andrade-Miranda¹, Nathalie Henrich Bernardoni², Juan Ignacio Godino-Llorente¹

¹Centro de Tecnología Biomédica, Universidad Politécnica de Madrid. 
²Univ. Grenoble Alpes / CNRS, GIPSA-lab, Grenoble, France

gxandrade@ics.upm.es, Nathalie.Henrich@gipsa-lab.fr, igodino@ics.upm.es

Abstract

The use of high-speed videendoscopy (HSV) in combination with image-processing techniques is the most promising approach to investigate vocal-folds vibration and laryngeal dynamics in speech and singing. The current challenge is to provide facilitative and informative playbacks for clinical and research purposes. We present three new facilitative playbacks using an optical-flow framework (OF). Optical-flow techniques are widely used in the field of computer vision for tracking unidentified moving objects in video sequences. The application of OF computation to HSV images is investigated. The advantages, drawbacks, and the complementarity to existing methods are discussed. The method has been tested on a database of 60 HSV sequences which covers different voice qualities for spoken and sung vowels. The new data representations have been compared with commonly-used facilitative playbacks. They provide additional information on the temporal dynamics of glottal vibratory movements during glottal closing and opening phases. 

Index Terms: Optical Flow, High-speed videendoscopy, motion field, vocal folds, glottal dynamics, laryngeal movement.

1. Introduction

The advent of high-speed videendoscopy (HSV) has revolutionized laryngeal imaging. It has exponentially increased our understanding of glottal dynamics during the phonation process. HSV is the only technique with the ability to capture the true intra-cycle vocal-folds vibratory behavior. Nowadays, due to the fast-growth of imaging technology, it is possible to find high-speed cameras with frame rates up to 10,000 fps. They allow the vocal assessment of male and female phonation in most of the clinical scenarios, such as phonation at normal pitch and loudness, onset and offset, high and low pitch in modal register, breathy and pressed phonation, falsetto register and pitch glides [1]. Yet, HSV is more than a slow-motion movie for visualizing the true movement of anatomic structures. With appropriate image-processing techniques, the time-varying data can be condensed into a few static images reducing the spatial-temporal dimensionality. The clinician or the researcher can follow the dynamics of anatomic features of interest in a more intuitive way, revealing contents which are often hidden to human eyes through the HSV movie playback. The literature reports interesting proposals to represent the HSV in a more intuitive manner. These representations improve the accuracy of quantification and increase the reliability of visual ratings. Such tools are termed facilitative playbacks. The most widespread and successful playbacks used either by clinicians or researchers are: Digital Kymograms(DKG) [2], Mucosal Wave playback(MW) [3], Mucosal Wave Kymogram playback(MWK) [3], Phonovibrogram(PVG) [4] and Glottovibrogram [5].

The DKG playback allows to select a line $y_i$ along the anteroposterior axis and replot the HSV sequences $f(x,y,t)$ as a function $f_{y_i}(x,t)$. The DKG playback provides a good understanding of the spatial and topological pattern of the vocal folds along time. The information obtained by DKG was found useful for demonstrating the change of the dynamic characteristics while viewing damaged tissues, such as lesions, scars, and discoloration of the vocal folds. The MK playback modifies the HSV sequence into a series of frames in which the pixel intensity encodes both the motion of the mucosal edges and the direction of the motion. The MK playback corresponds to viewing kymographic frames of the mucosal wave along the anteroposterior axis. Both methods provide information related with the mucosal edges propagation during the opening and closing phases. One of the most successful methods used for evaluating the vocal folds patterns is the PVG playback. It requires the segmentation of glottal area on HSV sequence [6]. The glottal axis is then identified as the main orientation of the glottal area by computing a linear regression line. Lastly, the left and right vocal-fold edges are splitted by finding the points of intersection among the glottal boundary and the regression line. The PVG synthesizes the variation of distance between vocal-fold edges and glottal axis along time. The PVG geometrical patterns are an unique representation of the underlying vocal-folds dynamics. The GVG presents an alternative way to depict the exact physiological behavior of the vocal folds by computing the distance between left and right vocal-fold edges along the anteroposterior glottal axis.

Despite the great advances that have been reached for condensing the data coming from HSV, many of the previous methods rely on glottal-area segmentation, which is not a trivial task. The obtained results are highly correlated with the quality of the HSV data, including image contrast and clarity of glottal edges. For that reason, new methods for data visualization are needed to overcome the drawbacks of existing ones. One main goal would be to provide simultaneously features that would integrate the time dynamics, such as velocity, acceleration, instants of maximum and minimum velocity, and vocal-fold pixel displacement during phonation.

In this paper, we propose a new approach for visualizing HSV data in a compact manner, which does not rely on segmentation of glottal area. It is based on computing Optical Flow (OF) for a HSV sequence. This method depicts the vocal-folds displacement by a motion field. The paper is organized as follows. Section 2 develops the methodology implemented, presenting three new representations. Section 3 details the parameterization of OF computation and the different options that
could be used. Additionally, the new representations are compared with state of art techniques such as DKG and GVG playbacks. Finally, Section 4 presents some conclusions.

2. Methodology

2.1. Principles of Optical Flow computation

Optical Flow is the pattern of apparent motion of image objects between consecutive frames caused by the movement of either the object itself or the camera. It is a 2D vector field in which some kind of velocity can be associated to each pixel in the frame or, equivalently, a displacement that represents the distance that a pixel has moved with respect to the previous frame. The most basic assumption made in OF computation is that the image brightness remains constant [7][8]. In a short time interval, an object may change its position, but the reflectivity and illumination remain constant. Brightness constancy means that the pixels in one tracked patch look the same along time. Mathematically, is is represented as:

\[ f(x + \Delta x, y + \Delta y, t + \Delta t) \approx f(x, y, t) \]  

where \( f(x, y, t) \) is the intensity of the image at position \((x, y)\) and at time \(t\). \(\Delta x, \Delta y\) are changes in the position, and \(\Delta t\) is the change at time. Applying a Taylor series expansion to the left hand side of equation (1), the optical flow constraint equation can be rewritten as follows:

\[ \nabla f \cdot v + I_x = 0 \]  

where \(\nabla f\) is the spatial gradient, \(v = (u, v) = (\Delta x, \Delta y)\) is the OF vector, and \(I_x\) is the temporal gradient. Another constraint that can be used to compute the OF is the motion tensor. The idea is that a segment of a video is a stack of images in which gray-value structures have certain orientations. The orientation in the \(xy\) - subspace is an indicator for the orientation of the structure in the space. In contrast, the orientation of the structure in the \(xt\) - subspace or \(yt\) - subspace relates to the image velocities \(u\) and \(v\), respectively. Thus, estimating the orientation of the structure in these two subspaces or a combination thereof allows estimating the OF. The work in [9] defines the motion tensor \(T\) using the minimum residuals between the input signal \(I\) and basis vectors computed with a least squares approach.

There are two main strategies for solving the OF problem: Sparse and Dense. The sparse optical flow finds the displacement only on a subset of features that have been specified beforehand; these features have certain desirable properties such as corners, dominant gradient orientation, subpixel corner locations among others. In the other hand, dense OF finds out the vector displacement of all pixels in the image, requiring a more expensive computational burden but providing more interesting information about the movement in the sequence.

2.2. Why Optical Flow in Laryngeal HSV?

The purpose of laryngeal HSV analysis is to characterize the motion of vocal folds by identifying their movements from one frame to the followings. However, this task requires to isolate the glottis and track it along time. OF computation allows the possibility to track unidentified objects solely based on its motion, with no need of additional segmentation techniques. The laryngeal HSV meets the optical flow main assumption (brightness constancy), which states that the rate of change in intensity along the motion trajectory is zero. This statement becomes more valid as frame rate increases. OF computation is adversely affected by temporal aliasing due to the object motion. This effect can be neglected when frame rate increases. These facts are more evident when the OF is computed using gradient-based methods, since the partial derivatives with respect to the temporal coordinates are calculated as brightness gradients, and the accuracy of motion estimation decreases when the object in a scene or the camera moves at high speed, because temporal partial derivatives are sensitive to and are degraded by a large image displacement between frames.

Actually the field of OF computation is making steady progress evidenced by the increasing accuracy of current methods on the Middlebury optical flow benchmark [10]. The OF can be used in a variety of situations, including time-to-collision calculations, segmentation, structure of objects, movement parameters, among many others.

2.3. Optical Flow implementation in HSV

The implementation used in this work is based on [9], since the state of the art has proved its reliability for computing dense OF. In order to obtain more accurate information, the OF had been computed only inside a region of interest (ROI). Such region includes only the vocal folds, since our attention is focused on the movements produced by glottal area and not by the surrounding tissues. The ROI was detected automatically and it refreshes after \(N\) frames based on [11].

Another fact to consider before computing the OF is to eliminate features that can produce erroneous flows vectors. The first step is to remove the artefacts produced on the images by the complex reflectance phenomena due to intrinsic surface properties. These artifacts appear in the endoscopic images because light source and viewing direction are almost identical; thereby, wet mucosa surfaces perpendicular to the viewing direction produce white flashing spots. A common approach to mitigate these effects is by isolating the information about material properties in a scene. One of such properties is the surface reflectance which is described often by the bi-directional reflectance distribution function (BDRF) [12]. Various specular removal techniques have been proposed in the state of art; they differ in the information used and in how this information is treated. One simple way to isolate the specular reflection effects when the illuminant color is known and the reflectance of surfaces can be represented by a dichromatic model is by linearly transforming the RGB color space by using rotation axes. For instance, one of the axes becomes aligned with the direction of the effective RGB source vector \(s\); this transformation defines a new color space, which is referred as \(SUV\) color space [13]. The \(SUV\) color space separates the diffuse and specular reflection effects where the \(S\) channel encodes the entire specular component and an unknown fraction of the diffuse component, and the remaining two channels \((U\) and \(V\)) are independent of specular invariants. Finally in order to solve the discontinuities and to regularize the glottal surface a bilateral filter is applied to each frame.

The OF computed afterward is represented by the coordinates \(u\) and \(v\), where \(u\) is the projection of the velocity on the \(x\) axis and \(v\) on the \(y\) axis. Additionally, it is possible to compute the magnitude of each vector and its angle. The next step is to synthesize this spatio-temporal information obtained between consecutive frames in a 2D representation, in which the information of the behavior of the vocal folds movement is readable. In order to achieve this task, three representations have been elaborated.
2.3.1. Optical-Flow Kymogram

The Optical-Flow Kymogram playback (OFKG) uses the same principle that DKG to compact the information. However, the information used to comprise the data is taken from the displacements of the OF in \( x \), which is the \( u \) component of the flow vector. When the OF moves to the right, the direction angle of displacement ranges from \( [-\pi/2, \pi/2] \) and is coded in red, as shown in Fig.1a. Contrariwise, the direction angle of displacement ranges from \( [\pi/2, 3\pi/2] \) for a movement to the left and is coded in blue. The OFKG playback is illustrated in Fig.1b for a phonatory sequence of six glottal cycles.

![Figure 1: (a) Color representation of the Optical-Flow during the closing phase (b) Optical-Flow kymogram Playback (blue represents the left displacement and red the right displacements of \( u \))](image)

2.3.2. Optical-Flow Glottovibrogram

The Optical-Flow Glottovibrogram (OFGVG) represents the velocity of glottal movement per cycle (see Fig.2). It is obtained by averaging each row of the \( u \) component of the flow and represents it as a column vector. This procedure is repeated along time for each new frame. The resulting plot is illustrated in Fig.3. The aim of the OFGVG playback is to complement the spatio-temporal information provided by the common techniques (GVG, PVG) in adding velocity information for each displacement of the vocal folds. Velocity comes close to zero at glottal maximum opening and during the closed phase.

![Figure 2: Optical-Flow Glottovibrogram playback](image)

2.3.3. Glottal Optical-Flow Waveform

The Glottal Optical-Flow Waveform (GOFW) is a 1D representation of the velocity that is derived from the glottal area waveform (GAW). The GAW is the sum of the number of pixels that belongs to the glottal area, each of these sums are represented as a point in the \( y \) axis where the \( x \) axis represents the time. GOFW is based on the same principle of the GAW, but here the total magnitude of the velocity is computed over the ROI for each instant of time. Graphically the GOFW represents the change of velocity along time at the same instant in which it is quantified the total velocity variation. The Fig.3 summarizes the complete framework and the three playbacks obtained.

![Figure 3: Outline for the elaboration of new playbacks](image)

2.4. Database

The image data used during this work were taken from the UKE highspeed database, kind courtesy of Nathalie Henrich, which was recorded at the University Medical Center Hamburg-Eppendorf (UKE) in Hamburg, Germany, by the team of Pr. Hess (Frank Müller and Anna-Katharina Licht) [5]. It consists of synchronized highspeed, audio and EGG recordings of two subjects, one singer and one speaker. For the highspeed recordings, a rigid endoscope (Wolf 90 E 60491) equipped with a continuous source of light (Wolf 5131) driven by optic fiber was used. The system was equipped with a grayscale charged coupled device (CCD) with a spatial resolution of 256x256 pixels and sampling frequency of either 2000 or 4000 fps. The speech sequences consisted in sustained vowels produced with several voice qualities (relaxed, pressed, breathy, creaky) and pitch glides. The singing sequences consisted in vowels sung at different pitches, different loudness, and exploring the four laryngeal mechanisms [14].

2.5. Playbacks computation

The algorithm was developed in C++ using the OpenCV library. In addition to OFKG, OFGVG and GOFW playbacks, the GVG playback was also computed by means of an adaptive thresholding segmentation algorithm. The threshold was manually tuned during the video sequence, fixing the problems with the segmentation in any instant of time. All the different playbacks were simultaneously generated during image processing, allowing a real time visualization. Different trials were performed in order to check the reliability and applicability of the proposed method. Firstly, the different parameters of OF computation were tested and combined in order to verify how their variability may affect the new playbacks. Secondly, the variation in magnitude of the \( u \) component was tracked for one fixed line through a complete cycle in order to understand the fluctuation of OF density. Finally, the new playbacks were contrasted with GVG ones for different phonatory tasks.

3. Results and Discussion

3.1. How do optical-flow parameters affect OFKG playback?

Three main parameters are of interest for the OF computation: the number of image pyramids, the averaging window size, and the frame rate. The image pyramids [15] increase the temporal resolution of the frames, allowing to track objects with fast movements. In HSV sequences, the movement between two consecutive frames is small enough to use only one level, and thus it is not necessary to implement pyramids for the case of laryngeal HSV. The window size allows robustness to image noise, and it gives better chances for fast-motion detection. However, increasing too much the window size yields to a blurred motion field. As illustrated in Fig.4, the wider the window size, the smoother the black spot caused by the zero veloci-
ties. In such case, beginning and ending of opening and closing phases are still preserved. Fig. 4 shows also the impact of frame rate. At high frame rate, velocity may not vary too much between two consecutive frames. For that reason, the flow can be smoothed by a computation each three/four frames without affecting the shape and main information extracted.

![Digital kymogram and Optical Flow](image)

**Figure 4:** First Row: Digital kymogram, Second Row: OFKG between two consecutive frames and window size=5, Third Row: OFKG between two consecutive frames and window size=15, Fourth Row: OFKG each 4 frames and window size=5, Fifth Row: OFKG each 4 frames and window size=15

### 3.2. Is glottal dynamics reflected on optical-flow variations?

The fluctuation over time of the motion field are a good indicator of the algorithm stability. Do they reflect glottal dynamics solely? To answer this question, the magnitude changes of $u$ were analyzed for one line (without considering the ROI boundaries) for a complete glottal cycle (see Fig.5). As expected, the flow is concentrated in the glottal region, since it is the region with the strongest movements. Another remarkable feature is the valley formed between two peaks. The valley can be understood as the region inside the glottis, in which motion field is zero. The two peaks can be interpreted as the pixels along the selected line with maximal positive and negative displacements.

![Fluctuation of u along one line for a complete glottal cycle](image)

**Figure 5:** Fluctuation of $u$ along one line for a complete glottal cycle

### 3.3. Comparison between GVG and OFGVG playbacks

The aim of OFGVG playback is to provide information related to the velocity of glottal movement along time. The results obtained for six different phonatory sequences are shown in Fig.6. OFGVG playback evidences the variation of glottal moving-edge velocity per glottal cycle. The greater the intensity in OFGVG, the higher the velocity. Conversely, if OFGVG intensity comes close to black, velocity tends to zero. Such information can be complemented by GOFW, and it may be combined with GVG playback for better visualization of the temporal-spatial relation (see an illustration in Fig.2).

![Playbacks GVG, OFGVG and GOFW for the cases of Breathy, Creaky, Pressed and M0, M1, M2](image)

**Figure 6:** Playbacks GVG, OFGVG and GOFW for the cases of (a) Breathy, Creaky, Pressed and (b) M0, M1, M2

### 4. Discussion and Conclusions

Highspeed videendoscopy is the most promising technique for direct investigation of glottal dynamics in speech and singing. We present here a novel approach to synthesize dynamical information from HSV recordings in a compact way which does not depend on prior glottal segmentation. The glottis is treated as an unidentified object, and attention is focused on the motion field produced by vocal-folds vibration. Dense optical flow is computed along consecutive frames to extract dynamical information related to the pattern of glottal displacement. It is sensitive to averaging window size and frame rate. Three new playbacks are proposed to visualize the computed optical flow: OFKG, OFGVG and GOFW playbacks. They provide complementary information to the common spatio-temporal representations. Their applicability to assess glottal dynamics has been illustrated on several phonatory sequences with varying voice qualities.

### 5. Acknowledgements

This work has been funded by the Spanish Ministry of Economy and Competitivity under grant TEC2012-38630-C04-01.
6. References


