Figure 6.23: GVG and PVG playbacks corresponding to the 10 HSVs presented in Table 6.7.

ployed in the graphic, **InP** with blue, SrG with red and **SnW** with black. The points represent the mean value of the three subjective tasks (quality, readability and shape), while the vertical bars indicate the standard deviation. **InP** provides the highest rank in 7 out of 10 patients. Meanwhile, the remaining three patients have been ranked best with SrG. Table A.1 in appendix A depicts the playbacks of the entire DB2 using **InP**.
6.5. Results

Figure 6.24: Overlapping between vocal fold trajectories and VKG in the medial axis using InP and SrG. Four glottal cycles are shown for each video sequence.

Figure 6.25: Segmentation subjective assessment of 10 patients on a 5-point scale.

6.5.3 Objective Supervised Assessment

The percentage accuracy improvements with respect to InP are summarized in Table 6.8 and are calculated from the first and last row of Table 6.3. The symbol
\( \mu \) represents the average accuracy obtained from 760 images analyzed, and \( \epsilon_{close} \) rates how many times an image is ranked with 0. The accuracy improvement with respect to \( \mu \) is computed as the percentage difference between InP and SrG, and SnW respectively. Meanwhile, The accuracy improvement with respect to \( \epsilon_{close} \) is computed as the percentage difference between InP and SrG, and SnW respectively. From Table 6.8 is observed that InP outperforms SrG and SnW, obtaining accuracy improvements up to 18\% in \( \mu \) and 25\% in \( \epsilon_{close} \).

<table>
<thead>
<tr>
<th>InP</th>
<th>Accuracy Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DICE</td>
</tr>
<tr>
<td>SrG</td>
<td>13</td>
</tr>
<tr>
<td>SnW</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 6.8: Comparison of the accuracy improvements of InP with respect to SrG and SnW.

By way of illustration, Figure 6.26 depicts 6 frames belonging to different LHSV with their respective good metrics pairwise trials: InP vs. MaN, SrG vs. MaN, and SnW vs. MaN. For a better visualization of the results, the MaN segmentation is showed in red; the segmentation obtained with InP, SrG and SnW are colored with green; and the intersection between manual (MaN) and automatic segmentation (InP, SrG and SnW) are depicted in yellow.

In Figure 6.26a the best result is obtained with InP, DICE ranks 0.49 meanwhile Pratt ranks 0.52. For the second frame (Figure 6.26b), the best performance is obtained again with InP (DICE=0.60 and Pratt=0.73) and the worst with SnW.

In Figure 6.26c, SnW and InP have similar results with values over 0.9, demonstrating an accurate segmentation. In the fourth frame, SnW is the only method able to segment correctly the anterior part of the glottis having rankings of 0.84 and 0.89 for DICE and Pratt, respectively.

In Figure 6.26e SrG and SnW the glottis is considered as closed, therefore the good metrics are ranked with zero. Contrariwise, InP presents a high accuracy to segment the glottal gap obtaining metrics of 0.79 and 0.93 for DICE and Pratt, respectively.

In Figure 6.26f, SrG presents over-segmentation since the effect of the glottal splitting is replaced by one unique gap. InP detects correctly the glottal splitting, however exists some pixels in the posterior part that are segmented wrongly (over-segmentation). On the other hand, SnW has the closest approximation to the segmentation expected (MaN).

### 6.6 Discussion

The lack of reliable glottal segmentation algorithms with minimal user interaction and of standard criteria to assess them limit the clinical acceptance of high-speed
6.6. Discussion

<table>
<thead>
<tr>
<th>Frame</th>
<th>MaN vs SnW</th>
<th>MaN vs InP</th>
<th>MaN vs SrG</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DICE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs Snw</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs InP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs SrG</td>
</tr>
<tr>
<td>a)</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs Snw</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs InP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs SrG</td>
</tr>
<tr>
<td>b)</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs Snw</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs InP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs SrG</td>
</tr>
<tr>
<td>c)</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs Snw</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs InP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs SrG</td>
</tr>
<tr>
<td>d)</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs Snw</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs InP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs SrG</td>
</tr>
<tr>
<td>e)</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs Snw</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs InP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MaN vs SrG</td>
</tr>
<tr>
<td>f)</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
</tbody>
</table>

Figure 6.26: Objective comparison between MaN and InP, SrG and SnW using the good metrics. First column: frames to be evaluated; second column: visual overlapping between MAN and SnW method; third column: visual overlapping between MaN and InP method; fourth column: visual overlapping between MaN and SrG methods; fifth column: summary of the good metrics results.

Techniques. For that reason, an attempt has been made to provide novel frameworks to automatically-or with minimal interaction-segment the glottal gap. The accuracy and efficiency of the glottal segmentations algorithm InP and SnW are...
compared against the SrG segmentation using an exhaustive analysis: one analytical assessment, three subjective tasks, and 18 objective metrics.

In the analytical assessment a direct comparison among the algorithms is not feasible, since they are implemented in different programming languages. However, based on the computation times obtained, a straight deduction is that the three algorithms analyzed are suitable to be used in a clinical environment.

On the other hand, the subjective assessments reveal that InP outperforms the other segmentation algorithms. Figure 6.22 illustrates the glottal segmentation of (P5) and (P7). For these particular cases, SrG and InP segmentations are slightly different. For instance, InP and SrG differ in the number of pixels assigned as glottis in the first and third frame of (P5). Contrariwise, SnW has the worst performance, showing over-segmentation in some frames (second frame (P5)), and wrongly detecting the instants of total closure (first and third frame of (P7)). The PVG and GVG playbacks of Figure 6.23 provide a good reference of the whole recording without the need of a frame by frame visual inspection, and let us infer that SnW was the most affected by the under and over-segmentation, either in the anterior or the posterior part of the glottis. Contrariwise, SrG and InP deal better with these issues but there are some cases when both algorithms disagree. For instance, in (P4) and (P8), SrG playbacks show a completely closed behavior, whereas InP playbacks deployed a common pattern related with the presence of a glottal chink. Another important aspect in the subjective assessment is concerned with the shape of the vocal folds trajectories. Figure 6.24 depicts the overlapping between VKG and VFT. The shapes obtained either with InP or SrG differ especially during the closing and opening phases which can be observed clearly in (P9).

Lastly, the results of the objective assessment are illustrated in Figure 6.26 for six different frames. In Figure 6.26a the glottis is divided in two parts. SnW method assumes a complete closure of the glottal gap, ranking all the metrics with 0. Meanwhile, InP detects almost correctly the upper part of the glottis, but fails to detect completely the inferior part. SrG has a similar performance than InP but is not able to segment the bottom part. The metrics for both segmentations show this slight difference. Figure 6.26b also splits the glottis. In this case SnW fails again to detect the glottal gap introducing an erroneous object. InP ranks well but there is a small artifact incorrectly segmented that is penalized with DICE but not using Pratt. Meanwhile, SrG has the same problem than in Figure 6.26a: only one region is segmented. In Figure 6.26f, good rankings were obtained using SnW and InP for the two metrics. Contrariwise, SrG presents problems of over-segmentation.

Based on the subjective and objective assessments, the comparative study concludes that the best results in average are obtained using InP, achieving an average accuracy improvement in the segmentation up to 13% with respect to the SrG and 18% with respect to SnW.
Chapter 7

Synthesizing the Vocal Folds Motion by Optical Flow

“Logic will get you from A to B. Imagination will take you everywhere”
Albert Einstein

SUMMARY: In this chapter three new playbacks are proposed to synthesize the dynamical information of the vocal folds based on Optical Flow (OF) computation. Two of them, called Optical Flow Glottovibrogram (OFGVG) and Glottal Optical Flow Waveform (GOFW), analyze the global dynamics; and the remaining one, called Optical Flow Kymogram (OFKG), analyzes the local dynamics. The reliability of the proposed playbacks is evaluated by comparison with traditional representations such as DKG, GAW, and GVG. Results show a great correlation in the shape of the vibratory pattern, allowing also the identification of the most important instants of time, such as closed-state and maximal opening. In addition, the playbacks based on OF computation provide complementary information to the common spatio-temporal representations.

7.1 Optical Flow in LHSV

The purpose of LHSV analysis is to characterize the motion of the 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. Advantageously, Optical Flow (OF) computation allows the possibility to track unidentified objects solely based on its motion, with no need of additional segmentation techniques.
The LHSV sequences present challenging scenarios such as complex reflectance phenomena that appears due to intrinsic mucosal surface properties, motion discontinuities due to the mucosal wave dynamics and occlusion in the glottal-area region. On the other hand, the OF accuracy is improved by the high frame rate of LHSV; it reduces the temporal aliasing not only for areas with large displacements but also for areas with small displacements and high spatial frequencies. Additionally, the BBC assumption becomes even more valid with high frame rates (Lim et al., 2005). In two consecutive frames the OF should describe precisely the vocal-folds motion pattern. The direction of the motion field is expected to be inwards during the closing phase and outwards during glottal opening. In order to illustrate this idea, Figure 7.1 presents a synthetic representation of the vocal folds motion among the posterior \( p(t) \) and anterior \( a(t) \) part of the glottal main axis for two consecutive frames during the opening phase.

![Figure 7.1: Illustration of a synthetic motion field \( W(x, t) \) located among the posterior \( p(t+1) \) and anterior \( a(t+1) \) part of the vocal folds during two consecutive instants of time, \( t_k \) and \( t_{k+1} \).](image)

Additionally, the fluctuations over time of the motion field \( W(x, t) \) have to reflect the glottal dynamics solely. In order to prove this fact, the magnitude changes of \( U(x, t) \) are analyzed for one line \( pc = 50\% \) in a complete glottal cycle (see Figure 7.2). As expected, the flow is concentrated in the glottal region since it is the region with 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.
Despite its suitability to the problem under study, the use of OF for assessing the vocal folds dynamics has been recently introduced in (Andrade-Miranda et al., 2015c,a). Nevertheless, the authors in (Saadah et al., 1996) had used motion estimation techniques to describe the vocal folds deformation but only around glottal edges.

Currently, the field of OF computation is making steady progress evidenced by the increasing accuracy of current methods on the Middlebury OF benchmark (Baker et al., 2011a). The OF can be used in a variety of situations, including time-to-collision calculations, segmentation, structure of objects, movement parameters, among many others.

7.2 Database Description

The Database3 (DB3) was acquired by means of a Wolf high-speed cinematographic system and it is composed for 60 high-speed sequences. The laryngeal HSVs were sampled at either 2000 or 4000 fps with a spatial resolution of 256 × 256 pixels. The recording took place at the University Medical Center Hamburg-Eppendorf (UKE) in Germany (Karakozoglou et al., 2012) and two male subjects (one speaker, one singer) participated in the experiment. The sequences include different phonatory tasks: sustained sounds with specific voice qualities (creaky, normal, breathy, pressed), pitch glides, sung vowels at different pitches and loud-
ness. Additionally, they cover a huge variety of vocal folds vibratory movements, including symmetrical and asymmetrical left-right movements, transients, aperiodicities, and antero-posterior modes of vibration.

To ensure the processing of sustained phonation only, the processed sequences were chosen approximately at the middle of phonation. They all comprise of 501 frames, which correspond to roughly 125 msec of sustained phonation.

7.3 Image Processing Implementation

In order to obtain a more accurate information, reduce computational burden and mitigate the effect produced by noisy regions, the OF has been computed only inside a ROI. Such region is detected automatically based on the procedure presented in section 6.2.2.

The OF techniques used for the implementation of the new playbacks are Total Variation L1 Optical-Flow (TVL1-OF), Motion Tensor Optical-Flow (MT-OF) and Lukas Kanade Optical-Flow (LK-OF). The principal reason for this selection is to explore the performance of different kinds of OF implementations since these methods use different strategies to deal with the complex reflectance phenomena and motion discontinuities. Other algorithms were also explored in this work (Brox et al., 2004; Drulea and Nedevschi, 2013; Horn and Schunck, 1981; Bruhn et al., 2006) but due to the computational burden needed to process a whole video and the similarities in the computation of the flow field with the aforementioned, they were not included in the OF-based playback evaluation.

Although TVL1-OF and LK-OF are based on the BBC assumption, they differ in the approach followed to compute OF, being TVL1-OF global and LK-OF local. Meanwhile, MT-OF does not have a direct connection with the BBC since the flow field is computed by orientation tensors. The implementation provided in the C++ OpenCV library was adopted for TVL1-OF and MT-OF flow computation. Since LK-OF is one of the fastest algorithms to compute OF, it was programmed in Matlab.

The implementation procedure is shown graphically on Figure 7.3 and the computation of each playback is explained below.

7.4 New Playbacks for Visualizing Glottal Dynamics

Three facilitative playbacks are proposed: Optical Flow Kymogram (OFKG) which depicts local dynamics along one line, Optical Flow Glottovibrogram (OFGVG) that represents global dynamics along the whole vocal folds length, and Glottal Optical Flow Waveform (GOFW) which plots the glottal velocity. They are described next:

---

1 A detail study of the database can be found in (Henrich, 2006; Roubeau et al., 2009).
7.4. New Playbacks for Visualizing Glottal Dynamics

Figure 7.3: Graphical representation of the procedure followed to compute the new playbacks.

7.4.1 Local Dynamics Along One Line: Optical Flow Kymogram

The OFKG playback shows the direction and magnitude of the vocal folds motion in a single line. It follows the same idea as DKG to compact the LHSV information. However, the information used to synthesize the data comes from the displacements produced in the x-axis at each time \( t_k \) \((U(x,t_k))\). For rightwise displacements, the direction angle ranges from \([-\pi/2, \pi/2]\) and is coded with red intensities. Conversely, the angle for leftwise displacements ranges from \([\pi/2, 3\pi/2]\) and is coded with blue tonalities. The OFKG playback is depicted in Figure 7.4 for a sequence of six glottal cycles. Algorithm 2 explains in detail the procedure followed to obtain the OFKG playback.

7.4.2 Global Dynamics Along the Whole Vocal Folds Length: Optical Flow Glottovibrogram

The OFGVG playback represents the global dynamics of the vocal folds by plotting the glottal velocity movement per cycle. The OFGVG playback has the goal to complement the spatiotemporal information provided by common techniques (GVG, PVG), adding velocity information of the vocal folds cycles. It is obtained
Figure 7.4: Schematic view of OFKG playback for the line represented in yellow, which is located in the median part of the vocal folds; The new local playback distinguishes the direction of motion (rightwise: red; leftwise: blue).

Algorithm 2: Pseudocode for OFKG playback

```plaintext
input : ROI, I(x, t), Line
output: OFKG

foreach k in I(x, t_k) do
    I(x, t_k) ← ROI(I(x, t_k))
    I(x, t_{k+1}) ← ROI(I(x, t_{k+1}))
    [U(x, t_k), V(x, t_k)] ← computeOpticalFlow(I(x, t_k), I(x, t_{k+1}))

foreach u(x_i, Line, t_k) in U(x, t_k) do
    if θ(u(x_i, Line, t_k), v(x_i, Line, t_k)) ∈ [−π/2, π/2] then
        OFKG(t_k, x_i) ← colorCode(|u(x_i, Line, t_k)|, blue)
    else
        OFKG(t_k, x_i) ← colorCode(|u(x_i, Line, t_k)|, red)
end
end
```

by averaging each row of \( U(x, t_k) \) and representing it as a column vector. This procedure is repeated along time for each new frame. Algorithm 3 presents the pseu-
docode for computing OFGVG and the third row of Figure 7.5 shows its graphic representation.

Algorithm 3: Pseudocode for OFGVG playback

\begin{verbatim}
input : ROI, I(x,t)
output: OFGVG
foreach k in I(x,t_k) do
  I(x,t_k) ← ROI(I(x,t_k))
  I(x,t_k+1) ← ROI(I(x,t_k+1))
  [U(x,t_k), V(x,t_k)] ← computeOpticalFlow(I(x,t_k), I(x,t_k+1))
  foreach Row in U(x,t_k) do
    OFGVG(t_k, y_j) ← \( \sum_{i=1}^{n} |U(x_i,y_j,t_k)| \) \( \forall j \in m \)
  end
end
\end{verbatim}

Figure 7.5: First row: frames representation of one glottal cycle. Second row: schematic view of GOFW. Each point in the playback (dark circles) is obtained by averaging the absolute magnitude of \( U(x,t_k) \). Third row: schematic view of one OFGVG cycle. Dark regions indicate no velocity (\( u(x_j,t_k) = 0 \)).
7.4.3 Global Velocity: Glottal Optical Flow Waveform

The GOFW playback is a 1D representation of the glottal velocity. It is computed following the same criteria of GAW but averaging the absolute magnitude of $U(x,t_k)$. Additionally, overlapping this information with GAW highlights the velocity variation in each instant of the glottal cycles. The second row of Figure 7.5 explains schematically how the GOFW is computed, showing the different velocity instants (black circles). Algorithm 4 summarizes the procedure to obtain the GOFW playback.

Algorithm 4: Pseudocode for GOFW playback

```
input : ROI, I(x,t)
output: GOFW
foreach k in I(x,t_k) do
   I(x,t_k) ← ROI(I(x,t_k))
   I(x,t_{k+1}) ← ROI(I(x,t_{k+1}))
   [U(x,t_k), V(x,t_k)] ← computeOpticalFlow(I(x,t_k), I(x,t_{k+1}))
   \[ \sum_{i=1}^{n} \sum_{j=1}^{m} |U(x_i,y_j,t_k)| \]
   GOFW(t_k) ← \[
end
```

7.4.4 Definition of the Vocal Folds Displacements Trajectories

The Vocal Folds Displacement Trajectories (VFDT) follow the same framework introduced in VFT (section 3.2.2) with the difference that the accuracy of the displacement is measured rather than the distance between vocal-folds edges and glottal axis.

Firstly, a trajectory line $L(t_k)$ at time $t_k$, which intersects perpendicularly with glottal main axis $G(t_k)$ in a predefined point $g_{pc}(t_k)$ is defined and updated every image using eq 3.2. Following, the intersection between the vocal folds edges $C^{lr}(t_k)$ and trajectory line $L(t_k)$ is computed, \{c^{pc}_{l,k} : c^{pc}_{r,k} ∈ L(t_k) \land c^{pc}_{l,k} ∈ C^{lr}(t_k)\}. Then, the displacement trajectories $\hat{\delta}^{lr}_{OFW}(pc,t_k)$ at $t_k$ and position $pc$ is defined by eq 7.1 as:

$$\hat{\delta}^{lr}_{OFW}(pc,t_k) = W(c^{lr}_{pc}(t_k)) \tag{7.1}$$

In view of the aforementioned, two additional trajectories can be derived from eq 7.1: $\hat{\delta}^{lr}_{OFU}(pc,t_k) = U(c^{lr}_{pc}(t_k))$ and $\hat{\delta}^{lr}_{OFV}(pc,t_k) = V(c^{lr}_{pc}(t_k))$. However, as the glottal edges have a motion pattern mainly perpendicular to the glottal axis, $\hat{\delta}^{lr}_{OFV}(pc,t_k)$ is negligible. Hence $\hat{\delta}^{lr}_{OFW}(pc,t_k)$ reflects primarily the fluctuations along $t_k$ produced by $\hat{\delta}^{lr}_{OFU}(pc,t_k)$. From now, both terms are used indistinctly and denoted for
simplicity as $\hat{\delta}_{OF}^{l,r}(pc,t_k)$. The graphical procedure followed to plot $\hat{\delta}_{OF}^{l,r}(pc,t)$ is described in Figure 7.6 and expressed in vector notation in eq 7.2.

$$\hat{\delta}_{OF}^{l,r}(pc,t) = [\hat{\delta}_{OF}^{l,r}(pc,t_1) \hat{\delta}_{OF}^{l,r}(pc,t_2) \ldots \hat{\delta}_{OF}^{l,r}(pc,t_k) \ldots \hat{\delta}_{OF}^{l,r}(pc,t_N)]$$  (7.2)

where $\hat{\delta}_{OF}^{l,r}(pc,t_k)$ is positive when the glottal edges are moving from right to left, and contrariwise, negative when the edges are moving from left to right.

Figure 7.6: Schematic procedure to compute $\hat{\delta}_{OF}^{l,r}(pc,t_k)$ during the opening phase.

7.5 Reliability Assessment of Optical Flow Playbacks

Due to the high amount of data in LHSV and the complexity of the vocal folds motion, it is difficult to create a ground-truth to evaluate the OF performance (Baker et al., 2011b). Therefore, it is necessary to find alternative options to assess the reliability of the new playbacks. An intuitive way to evaluate the accuracy of the OF playbacks is to compare against those obtained using glottal segmentation algorithms, since both results should be related. This premise comes from the fact that these two techniques represent the motion originated in the vocal folds, with the difference that in glottal segmentation the motion is reflected only on the glottal edges, while in the OF procedures the entire vocal folds region is analyzed.
Therefore, DB3 was segmented automatically, having as a result: well-segmented videos and videos with minor errors in the segmentation. In this way, the benefits of the OF playbacks are explored when the segmentation is not 100% reliable.

Three assessments are carried out. Firstly, the VFDT obtained by OF are correlated with the one obtained via segmentation, which are defined in eq 7.3 for a particular time $t_k$.

$$\delta_{\text{seg}}^l(p_c;k) = c_{p_c}^l(t_{k+1}) - c_{p_c}^l(t_k) \quad (7.3)$$

Since we are using three different OF methods, $\delta_{\text{seg}}^l(p_c,t)$ is compared with each of them. The OF displacement trajectories are renamed as: $\delta_{\text{TVL1}}^l(p_c,t)$, $\delta_{\text{MT}}^l(p_c,t)$ and $\delta_{\text{LK}}^l(p_c,t)$ for TVL1-OF, MT-OF and LK-OF respectively. All the displacement trajectories are computed in the medial glottal axis position $p_c = 50\%$.

The second assessment tries to find out the similarities of traditional playbacks with OF playbacks by visually analyzing their common features and quantifying their resemblance through two metrics: Structural Similarity Index (SSIM) (Wang et al., 2004) and Normalize Correlation Coefficient (CC).

The last assessment explains the contributions of the glottal contour and the contribution of the mucosal wave in the OFGVG playback. First, the motion field generated only by the points belonging to $C_{l,r}^l(t)$ is computed. Following, the OFGVG of such points is subtracted to the OFGVG obtained from the whole image. Lastly, the contribution to the OFGVG playback of each of them is explained.

### 7.6 Results

#### 7.6.1 Comparison Among Segmentation and OF Displacement Trajectories

The correlation between the segmentation trajectory and OF-based trajectories is depicted in Figure 7.7a (CC($\delta_{\text{seg}}^l, \delta_{\text{TVL1}}^l$), CC($\delta_{\text{seg}}^l, \delta_{\text{LK}}^l$) and CC($\delta_{\text{seg}}^l, \delta_{\text{MT}}^l$)). Each point of the graphic corresponds to the correlation of one LHSV sequence. The greatest correlation is 0.98 which belongs to CC($\delta_{\text{seg}}^l, \delta_{\text{TVL1}}^l$). Meanwhile, CC($\delta_{\text{seg}}^l, \delta_{\text{LK}}^l$) and CC($\delta_{\text{seg}}^l, \delta_{\text{MT}}^l$) do not exceed the value of 0.93. Additionally, 62% of the trajectories computed via TVL1-OF presented a correlation greater or equal than 0.8 while only 23% and 8% of the trajectories reached this value using LK-OF and MT-OF respectively. On the other hand, there are 8 CC($\delta_{\text{seg}}^l, \delta_{\text{TVL1}}^l$) with values below to 0.5, representing 13% of the videos in the database.
In order to understand the differences between $\hat{\delta}_{TVL1}^{t,r}$ and $\hat{\delta}_{seg}^{t,r}$, a breathy and creaky phonation are analyzed visually (see Figure 7.7b, 7.7c). The trajectories computed via TVL1-OF are smoother than the ones obtained via segmentation but the shape and the amplitude of both are comparable. Additionally, during a short period of time (regions enclosed by dashed lines in black at see Figure 7.7b) $\hat{\delta}_{TVL1}^{t,r}$ presents some fluctuations originated from a vibration of the vocal folds, while $\hat{\delta}_{seg}^{t,r}$ does not show any motion. The close up of one frame belonging to these regions is shown on the right hand side of the displacement trajectories. From them, it is observed that the segmentation does not delineate correctly the glottal area causing an erroneous estimation of the trajectory displacements for $\delta_{seg}^{t,r}(pc,t)$.
7.6.2 Comparison of OF Playbacks with Traditional Ones

Global Dynamics Along the Whole Vocal Folds Length: Derivative of Glottovibrogram and Optical Flow Glottovibrogram

Five playbacks are depicted in Figure 7.8 for three phonation cases: GVG and its derivative $|d_t, GVG|$ and three OFGVG. Similarities between $|d_t, GVG|$ and the OFGVG playbacks can be noticed, especially in shape appearance. In pressed phonation there is a long closed-state that can be observed along the five playbacks, taking place at the same time for all of them. Glide up phonation has a posterior glottal chink that produces a constant tonality of gray at the top part of the GVG plot. In contrast, this is perceived as a no-motion region in the $|d_t, GVG|$ and in the OFGVGs, so it is depicted in black for those playbacks. In the glide down sequence the vocal folds open as two separate regions until it gets fused in a short period of time. This effect can be observed easily in the GVG (dashed circle in red) and in its derivative. However, due to the blurring effect induced by the presence of mucus, it is not obviously readable in the OFGVG.

Additionally, two peculiarities are observed in the OFGVGs representation of Figure 7.8. Firstly, the playbacks do not show gray tonalities in the middle part of the glottal cycle (open-state), which means no motion of the vocal folds (velocity close to 0). Secondly, the presence of mucus is depicted as gray regions that produce a blurring effect (bottom panels in Figure 7.8). Lastly, for all the phonatory tasks a certain degree of noise is found when the OF is computed via LK-OF and MT-OF. Contrariwise, OFGVG based on TVL1-OF is more readable and its shape pattern resembles are closer to $|d_t, GVG|$.

Glottal Velocity: Derivative of the Glottal Area Waveform and Glottal Optical Flow Waveform

Since GOFW computes an absolute velocity, it is possible to obtain a similar representation by differentiating GAW and computing its absolute value ($|d_t GAW|$). The GOFW provides valuable information about the total velocity of the vocal folds motion for each instant of time. Additionally, if $|d_t GAW|$ is overlapped with the GAW (as shown in Figure 7.9), it is feasible to analyze the velocity variation with respect to the glottal cycles.

Figure 7.9 shows that in the open-state the velocity decreases, creating a valley in the $|d_t GAW|$ and in the GOFW playbacks. Additionally, it shows that the maximum velocities take place in the same instants of time but with different amplitude values depending on the OF techniques. A velocity variation can be seen in all $|d_t GAW|$ playbacks since in some glottal cycles the maximum occurs during the opening, in others during the closing phase, and sometimes both amplitudes are similar. This fact can be clearly observed in Figure 7.9 for the pressed voice quality where the amplitude of the peaks oscillates around different values. Contrariwise, GOFW always has its maximum velocity during the opening phase, but the ampli-
7.6. Results

Figure 7.8: Illustration of GVG, $|d_x GVG|$, OFGVG-LK, OFGVG-MT and OFGVG-TVL1 playbacks for three different phonatory tasks (pressed, glide up and glide down).

tude values are different depending on the OF used. In the glissando task, $|dGAW|$ and GOFW have a discrepancy with respect to maximal velocity occurrence. In $|dGAW|$, it occurs during the closing-state, while in GOFW, during the opening. Among GOFW playbacks, the main dissimilarity relies on the peak amplitude. For instance, in the glissando phonation, GOFW-LK and GOFW-MT maximum fluctuates between opening and closing states. In contrast, GOFW-TVL1 always has maximum velocity during the opening phase.

7.6.3 Global Dynamics Evaluation for the Whole Database

The GVG playback is a compact way to assess the entire vocal folds dynamics. Therefore, it is important to compare objectively its resemblance with the OFGVG playbacks. To accomplish this task, correlation and SSIM are used to measure the similitude between $|d_x GVG|$ and OFGVGs. The correlation between $|d_x GVG|$ and OFGVGs is depicted in the first row of Figure 7.10. Each point corresponds to the correlation of one HSV sequence. The best correlations are obtained using OFGVG-TVL1. The average correlation achieved in this case is 0.47. Meanwhile, OFGVG-LK and OFGVG-MT correlate in 0.38 and 0.37 respectively. The maximum correlation for OFGVG-TVL1 is 0.76 in the LHSV #7. Contrariwise, the
lowest value occurs for LHSV #54 with a metric of 0.02. Only 45% of the videos have a correlation greater than 0.5. Using SSIM, the metrics obtained are 0.22, 0.16 and 0.18 for TVL1-OF, LK-OF and MT-OF respectively.

Figure 7.11 and Figure 7.12 show two examples where the vibratory patterns are more distinctly represented in the OFGVG-TVL1 than in the GVG. Figure 7.11 presents an example with a glottal chink in the posterior part, so the motion only appears at the anterior part of the vocal folds. Nevertheless, $|d_x\text{GVG}|$ indicates a vibratory pattern in the posterior part of the vocal folds edges due to an imprecise contour detection. Contrariwise, OFGVG synthesizes the motion of the anterior part and includes the vibration of the mucosal wave as blurring gray tonalities during the closed-phase. Figure 7.12 shows an LHSV sequence also with a glottal chink in the posterior part. Here the length of the glottal edges detected by segmentation does not completely reach the anterior part of the vocal folds, affecting the legibility of the GVG. For instance, a close look to the frame $t_{13}$ and $t_{32}$ shows that there is no left glottal edge defined for the anterior part (red edge). So the distance between the glottal edges is different to zero in spite of the glottis is closed, producing vertical gray lines in the $|d_x\text{GVG}|$ playback. In contrast, the vibratory pattern of OFGVG is more readable and remains similar for all the glottal cycles. Lastly, its tolerance to highly asymmetrical vocal folds vibration is illustrated in
Figure 7.13 during a glissando with a transition between two laryngeal mechanisms. Here, OFGVG and \(|d,GVG|\) playbacks have features in common such as cycle shape and time of occurrence of mechanism transition. Table B.1 in Appendix B depicts the GVG, \(|d,GVG|\) and OFGVG playbacks using TVL1-OF of the entire DB3.

Comparison Between LK, MT and TVL1 Optical Flow Using Local Dynamics Along One Line: Digital Kymogram and Optical Flow Kymogram

OFKG is computed using TVL1-OF, LK-OF and MT-OF for three different glottal locations, each of them corresponding to a percentage of the glottal axis (\(pc_1=10\%\), \(pc_2=50\%\) and \(pc_3=90\%) as shown in Figure 7.14.

The results show that OFKG has a shape similar to DKG but blurred over the vocal folds. Such blurring effect is caused by the mucosal wave propagation. One outstanding characteristic appears during the change between opening and closing phases due to the presence of a discontinuity in the OFKG. This can be understood as an instant of time in which the velocity decreases considerably. In \(pc_1\), there is a quasi-static behavior of the vocal folds due to a glottal chink. The DKG represents the absence of motion when the shape of the glottal gap (dark region) does not
change over time. Meanwhile, OFKG is displayed with low intensity tonalities ($u(pc_1,t) \approx 0$). The lines located at $pc_2$ and $pc_3$ present a visible triangular pattern in OFKG which is a characteristic of DKG for a normal voice production. LK-OF and MT-OF computation produce, roughly speaking, the shape expected for OFKG. Yet the images are blurred, this effect is propagated to the close-state and to the inner part of the glottis. Contrariwise, OFKG-TV1 motion pattern is more readable and distinguishable.

**Contribution of the Mucosal Wave on OFGVG Playback**

The OF playbacks encode the average velocity along the vocal folds and perpendicular to the glottal axis which means that the entire mucosal wave activity is included. Contrariwise, the segmentation based techniques reveal solely the behavior of the vocal folds since only the motion of the glottal contours is computed.

In order to investigate the mucosal wave contribution on OFGVG playback, two versions of OFGVG are depicted in Figure 7.15. The first, named $OFGVG_{OF}$,
7.6. Results

Figure 7.12: Upper panel: nine segmented frames, the areas dotted with red correspond to the posterior glottal chink; middle panel: $|d_{GVG}|$ playback; lower panel: OFGVG with a vertical length that depends on the ROI size. The misleading calculation of the distance between edges is observed as gray vertical lines in $|d_{GVG}|$. The vibratory shape pattern for three consecutive cycles is marked with a continuous red line.

is computed using the whole motion field $U(x,t)$ inside a ROI. Meanwhile, the second, named OFGVG$_{seg}$, uses only the displacement vectors $U(C^r(t))$ originated by the motion of the glottal contours. Lastly, the subtraction among both playbacks is carried out, having as a result a new playback. The new playback reveals a hidden feature associated with the wave-like movement of the superficial tissues covering the musculus vocalis. This movement is referred to as residual and suggests that the MW is also identified by the OF methods.

Some remarks can be distinguished from the three playbacks. Firstly, OFGVG$_{OF}$ and OFGVG$_{seg}$ differ in the length of each glottal cycle. For instance, the glottal cycles in OFGVG$_{seg}$ are smaller than OFGVG$_{OF}$ since they do not include the motion originated by the mucosal wave propagation. Secondly, Figure 7.15 (OFGVG$_{OF}$) shows that the mucosal wave motion appears after a closed-state and before the open-state of the vocal folds (MW is depicted on blue tonalities). The mucosal wave propagation is perceived as a flashing bright highlight along the vocal fold edges due to the variation of the mucosa surface. Lastly, it is corroborated that the algorithms based on segmentation are not able to detect the mucosal wave propagation as it is observed via OFGVG$_{seg}$ playback.
Figure 7.13: $|d, GVG|$ and OFVG visualization of peculiar vocal-folds vibratory movements during glissando with a laryngeal-mechanism transition. Upper panel: 24 glottal cycles. Lower panel: 23 glottal cycles. The laryngeal mechanism transition is pointed out with red arrows and the dashed lines in red indicate different glottal cycles.

Figure 7.14: Illustration of DKG and OFKG at three different positions of the LHSV sequence. First row: VKG playback; second row: OFKG using LK-OF; third row: OFKG using MT-OF; fourth row: OFKG using TVL1-OF.
7.7 Discussion

This chapter addresses the use of OF techniques to embody the time-varying vocal folds vibratory pattern into efficient and easy to read playbacks. The aim is to find an alternative to current facilitative playbacks which requires glottis segmentation. Therefore, three new facilitative playbacks are proposed named as OFGVG, GOFW, and OFKG.

The reliability of the OF-based playbacks was assessed by comparison with the segmentation-based playbacks. The degree of similarity between the playbacks was measured objectively by computing the correlation displacements trajectories and image resemblance (CC and SSIM). In both cases, the greatest similitude was obtained with TVL1-OF. The TVL1-OF trajectories depict a smoother behavior that is originated by the denoising feature embedded in its computation. TVL1-OF is presumably the best option to compute the OF playbacks for its ability to preserve discontinuities in the flow field and its robustness against illumination changes, occlusions, and noise.

With respect to the playbacks, OFGVG provides information about the velocity
of the vocal-folds surface motion in an ROI. It allows to visualize glottal dynamics of an entire LHSV sequence in a single image, and put emphasis on moments of maximal and minimal velocities. The OFGVG playback shares similar characteristics in shape with their segmentation-based counterparts (GVG) with regards to glottal dynamics. But OFGVG is more robust to peculiar types of vibrations (Figure 7.13) when a highly asymmetrical vocal-folds vibration during a glissando with transitions is analyzed.

On the other hand, GOFW reveals as a valuable tool to study the total velocity of the vocal folds. It can advantageously be combined with GAW to comprehend the relationship between glottal opening and velocity. GOFW can be used along each glottal cycle to identify the instants of maximum and minimum velocity.

Lastly, OFKG represents the vocal-folds velocity motion of one line, being a complement to DKG for a better understanding of local dynamics. OFKG represents valuable information about vocal folds displacements direction, providing also a clear and comparable representation of the vocal-folds vibration, similar to the differentiated GVG while reducing errors due to glottal delineation.

The OF-based playbacks have demonstrated a great correlation in shape with the traditional playbacks, allowing the identification of the most important instants of time, such as closed-states and maximal opening, and providing complementary information to the common spatio-temporal representation. In addition, they are a good alternative when segmentation is not available, or when it is not reliable enough due to failures in the glottal-edges detection. Furthermore, the contributions of both glottal contour and vocal folds mucosal wave can be addressed. Since OF playbacks provide information about the whole vocal fold dynamics, and thus include the horizontal mucosal wave contribution of the vocal folds movement. Such information about vocal folds tissues dynamics can not be reflected using segmentation-based playbacks.
Part IV

Conclusions and Future Works
Chapter 8

Conclusions and Future Works

“The human voice is the most perfect instrument of all”
Arvo Part

**SUMMARY:** This chapter provides conclusions and future lines of research that are particularly relevant to the continuity and transferral of the results. The methodology, results, and conclusions described in this thesis, as well as the publications derived from it, have attempted to contribute to the state of the art in the understanding of the vocal folds dynamics and to help in the automatic detection of clinical disorders based on the analysis of laryngeal imaging.

8.1 Conclusions

The vibration of the vocal folds is one of the most important processes during the voice production. Therefore, the investigation and the examination of the vocal folds dynamics and mucosal wave vibration have been a subject of great interest in the past, and this interest continues today. The most extended methods to capture the vibratory movement of the vocal folds are LVS and LHSV. LHSV systems record images of the larynx at a typical rate of 4000 fps, while the rate obtained with LVS is only around 30 fps. LHSV illuminates using a continuous light whereas LVS uses a stroboscopic lamp to show the movement of the vocal folds taking advantage of the stroboscopic phenomenon. In the case of LVS, they present an important intra-video variation and do not provide a real view of the vocal folds vibratory pattern, so its use is restricted to stable and periodic vocal fold vibrations. In contrast, LHSV systems record every glottal cycle without temporal perturbation, being the only technique capable to register the true intra-cycle
vibratory behavior of the vocal folds oscillations. Despite the obvious advantages of LHSV, it has not been widely adopted in the clinic yet because of the lack of information regarding its validity and clinical relevance. Therefore, the aims of the present work, as well as the publications derived from it (see appendix C), have attempted to contribute to the state of the art in the understanding of the vocal folds dynamics and to help in the automatic detection of clinical disorders based on the analysis of laryngeal imaging.

Firstly, the problem of the glottal gap segmentation has been addressed since it is an essential operation for the correct characterization of vocal-folds vibrations. Commonly, the glottal segmentation is used as a prior step to identify different phonation features in an objective way, i.e. the periodicity and amplitude of vocal folds vibration, mucosal wave, glottal closure, closed-state, symmetry of vibration, presence of non-vibrating portions of the vocal folds (Tao et al., 2007; Lohscheller et al., 2013), etc. However, in spite of the extensive literature devoted to solving the glottal segmentation, they have some shortcomings in terms of accuracy and intervention. The lack of more accurate algorithms with minimal user supervision has limited the clinical acceptance of high-speed techniques. For this reason, two algorithms have been proposed in this thesis to tackle the problem of the glottal gap segmentation: Glottal Segmentation Based on Watershed Transform and Active Contours (SnW) and Glottal Segmentation Based on Background Subtraction and Inpainting (InP).

The SnW consists of a set of modules to pre-process, detect ROI, delineate the contours, and refine the glottis shape. In the first module, the point-wise nonlinear transformation algorithm is chosen since it presents the better trade-off between objective and subjective evaluation and also mitigates the influence of flashes which affects the performance of the ROI detection. On the other hand, the ROI detection takes advantage of the temporal intensity information of the LHSV and is adaptively updated every $N_{ROI}$ frames according to an extensive evaluation. Thanks to its adaptability, the ROI provides reliability against the camera and/or patient displacement, reduces the influence of false detections, it is robust when the glottis is divided into two or more regions and is able to manage the presence of a glottal chink when the cut-off points are chosen appropriately. The segmentation module uses the well-known Watershed Transform with two merging steps based on JND and a template correlation. The first merging step fuses regions based on the sensibility of the human visual system to the changes of luminance. The correlation merging step gives additional information about the position and shape of the glottis which lets differentiate between glottal and non-glottal regions. Finally, to refine the segmentation and solve any problem with the previous steps, a Region-Based Active Contours modeling is performed. The main novelty of SnW relies on the methods used to identify the ROI, as well as the combination of the watershed transform with a standard template for the merging process. In spite of the good performance of SnW, it has some shortcomings with respect to the empirical way in which some parameters were set up. For instance: the contrast factor $\beta$ in the