

Analyzing of the vocal fold dynamics using laryngeal videos

Gustavo Andrade-Miranda

▶ **To cite this version:**

Gustavo Andrade-Miranda. Analyzing of the vocal fold dynamics using laryngeal videos. Signal and Image Processing. Universidad Politécnica de Madrid, 2017. English. ⟨NNT : ⟩. ⟨tel-01585708⟩

HAL Id: tel-01585708

https://theses.hal.science/tel-01585708v1

Submitted on 12 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN



Analyzing of the vocal fold dynamics using laryngeal videos

TESIS DOCTORAL

Gustavo Xavier Andrade Miranda

Fecha de Sustentación:
8 de Junio de 2017

CENTRO DE TECNOLOGÍA BIOMÉDICA

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

Analyzing of the vocal fold dynamics using laryngeal videos

TESIS DOCTORAL

Autor:

Gustavo Xavier Andrade Miranda

Universidad Politécnica de Madrid

Director:

Juan Ignacio Godino Llorente

Doctor Ingeniero en Informática

Catedrático de Universidad

Universidad Politécnica de Madrid

Co-Director:

Nathalie Henrich Bernardoni

Doctora en Acústica Musical

Investigadora del CNRS

Departamento de Humanidades y Ciencias Sociales

Fecha de Sustentación:
8 de Junio de 2017

TESIS DOCTORAL

Analyzing of the vocal fold dynamics using laryngeal videos

Autor: Gustavo Xavier Andrade Miranda
Director: Juan Ignacio Godino Llorente
Co-Director: Nathalie Henrich Bernardoni

Tribunal nombrado por el Mgfco. y Excmo. Sr. Rector de la Universidad Politécnica de Madrid, el día ..... de ..... de 2017.

Presidente: D. Philippe H. DeJonckere .....
Vocal: Dña. María Arrate Muñoz Barrutia .....
Vocal: D. Jose María Martínez Sanchez .....
Vocal: D. Norberto Malpica González de Vega .....
Secretario: D. Luis Alfonso Hernández Gómez .....
Suplentes: D. Fernando Cruz Roldán .....
Suplentes: D. Juan Carlos González de Sande .....
Revisores Remotos: D. Daryush D. Mehta .....
Revisores Remotos: D. Jan G. Švec .....

Realizado el acto de defensa y lectura de la Tesis el día ..... de ..... de 2017 en .....

Calificación: .....

EL PRESIDENTE LOS VOCALES

EL SECRETARIO

A mi Familia y
a cada una de las
personas que han
formado parte de
esta gran aventura.

Acknowledgements

"On ne voit bien qu'avec le coeur.
L'essentiel est invisible pour les yeux"
Le Petit Prince

Quisiera comenzar agradeciendo al gobierno de Ecuador por su programa de becas, el cual ayuda a fomentar el talento humano, brindándonos la oportunidad de forjarnos profesionalmente en universidades de renombre internacional. Agradezco también al Ministerio de Economía y Competitividad de España que a través del proyecto TEC2012-38630-C04-01 ha cofinanciado mi tesis doctoral.

Mi imperecedera gratitud a mi director en Madrid Juan Ignacio Godino Llorente, quien con sus conocimientos y orientación ha permitido llevar adelante este nuevo reto en mi vida profesional. Un especial agradecimiento a el Sr. Oswaldo, Sra. Digna y a toda la familia Ramos en Madrid que siempre me han hecho sentir como un miembro mas de su familia. Agradezco también a todas esas personas que he conocido en Madrid y con las cuales he compartido inolvidables momentos.

Durante mi estancia en Grenoble una ciudad que será siempre parte importante en mi corazón encontré gente amable y magnífica como mi directora de tesis, Nathalie Henrich Bernardoni una apasionada de la voz. Mi eterna gratitud hacia ella por estar siempre dispuesta a nuevas ideas y por transmitirme su pasión a la investigación. Gracias a mis amigos de Grenoble, Remi que me mostró que no todo en la vida es comer "gateaux" y con el cual conocí la belleza de las montañas de Rhone Alpes en los inigualables randonnée. Francois por su alegría y por su música, sobre todo por la "flaca". A todos mis amigos del laboratorio de DPC en GIPSA: Angelique, Andrew, Annalisa, Xavier, Melaine, Alexandre y Jean Francois. También quiero agradecerles a mis dos colloc Simon y Anne, que mas que colloc han sido grandes amigos y los primeros en soportar mi francés. Un agradecimiento muy especial a mi Hortense con la que he compartido momentos maravillosos y que me ha apoyado siempre con su cariño y afecto.

Por último y no por eso menos importante quiero agradecerles a mis padres, que a pesar de la distancia siempre están presentes para transmitirme todo su amor, brindarme su apoyo incondicional y recordarme que cada día debo esforzarme, trabajar mas, hacer camino al andar y paso a paso avanzar en busca de mis mas anhelados propósitos.

Abstract

"It is not knowledge, but the act of learning,
not possession but the act of getting there,
which grants the greatest enjoyment"

Carl Friedrich Gauss

The voice is the most crucial tool allowing communication between human beings, therefore a healthy voice is important to people's daily life, especially for the professional voice users. It is imperative to find techniques to provide comprehensive information about the voice production mechanism and, more specifically, to examine the vocal folds vibratory behavior by Laryngeal High-Speed Videoendoscopy (LHSV). Thus, the present work aims to contribute to the analysis of the vocal folds vibratory function by proposing new and more robust tools based on image processing techniques.

Due to the vast amount of data that has to be evaluated both quantitatively and qualitatively, a dimensionality reduction of the spatial-temporal data is necessary by condensing the information into a few static representations that synthesize the vocal folds motion. Most of the milestones achieved until now are thanks to the segmentation and tracking of the glottal gap which is not a trivial task due to factors as noise in the image, variability in illumination, variability of the gray levels presented in the glottal gap, fuzziness, blurring edges, movements of the camera and/or patient.

In that respect, two algorithms to tackle the problem of the glottal gap segmentation are proposed. The first one, named Glottal Segmentation Based on Watershed Transform and Active Contours (**SnW**), identifies a Region of Interest (ROI) that is automatically updated, and combines Deformable Models and Watershed Transform for the final delineation of the glottal gap. Thanks to the ROI implementation, the proposal resists to the camera shiftings. The second one, called Glottal Segmentation Based on Background Subtraction and Inpainting (**InP**), presents a quasi-automatic framework to segment accurately the glottal gap introducing several techniques not explored before in the state of the art. The method takes advantage of the possibility of a minimal user intervention in cases where the automatic computation fails. In addition, a set of guidelines to measure the accuracy and efficiency of the segmentation algorithms are proposed. These guidelines are divided

into three groups according to their nature: analytical, subjective, and objective. The results obtained suggest that a more reliable delimitation of the glottal gap is obtained with **InP**, achieving an average improvement of 13% with respect to others techniques in the state of the art, and 18% with respect to **SnW**. Additionally, the results show that the set of validation guidelines proposed can be used to standardize the criteria of accuracy and efficiency of the segmentation algorithms.

Lastly, the application of Optical Flow (OF) is investigated in order to solve the problems related to segmentation. Three new playbacks are proposed to understand the dynamical information of the vocal folds. 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 advantages, drawbacks and the complementarity to existing methods are discussed. The new playbacks are tested on a database of 60 LHSV sequences which covers different voice qualities for spoken and sung vowels. The new data representations have been compared with commonly used facilitative playbacks. Results show that they provide additional information on the temporal dynamics of glottal vibratory movements during glottal closing and opening phases.

Resumen

"No es el conocimiento, sino el acto de aprendizaje, y no la posesión, sino el acto de llegar allí, lo que concede el mayor disfrute"

Carl Friedrich Gauss

La voz es una herramienta esencial en la que se fundamenta la comunicación de los seres humanos por este motivo tener una voz saludable es importante para el diario vivir de las personas, y más aún si esta es utilizada como una herramienta profesional de trabajo. Por tal motivo, es imperioso encontrar nuevas y mejores técnicas para comprender los mecanismos usados para la producción de la voz y sobretodo para entender el comportamiento vibratorio de los pliegues vocales utilizando Videos Laríngeos de Alta Velocidad (Laryngeal High-Speed Videoendoscopy (LHSV)). A partir de los antecedentes anteriormente mencionados, el presente trabajo tiene como objetivo contribuir al análisis de la función vibratoria de los pliegues vocales mediante la implementación de nuevas y más robustas herramientas basadas en el uso de técnicas de procesado de imágenes.

Debido a la gran cantidad de información que debe ser evaluada tanto cualitativa como cuantitativamente es necesario sintetizar esta información espacio-temporal en pocas representaciones estáticas que reflejen inequívocamente el movimiento de los pliegues vocales. Hasta el momento la mayoría de los hitos han sido alcanzados gracias al uso de la segmentación y del seguimiento de la abertura glotal. Dichas tareas no son fáciles debido a factores como ruido en las imágenes, variación en la iluminación, diferentes niveles de grises presentes en la abertura glotal, borrosidad de las imágenes, borrosidad de los contornos de la abertura glotal, movimiento de la cámara y/o de los pacientes.

Con la finalidad de solucionar los problemas citados anteriormente se presentan dos algoritmos para segmentar la abertura glotal. El primero, recibe el nombre de Segmentación Glotal Basada en Transformación Divisoria y Contornos Activos (Glottal Segmentation Based on Watershed Transform and Active Contours (**SnW**)), la cual identifica una Región de Interés (Region of Interest (ROI)) que se actualiza automáticamente. Este método combina el uso de Modelos Deformables (Deformable Models) y la Transformación Divisoria (Watershed Transform) para realizar la delimitación final de la abertura glotal. Gracias a la implementación del

ROI, **SnW** es robusto a los movimientos de la cámara. El segundo método recibe el nombre de Segmentación Glotal Basada en Sustracción de Fondo e Restauración de Imagen (Glottal Segmentation Based on Background Subtraction and Inpainting (**InP**)), en el que se presenta un algoritmo cuasi-automático para segmentar con precisión la abertura glotal mediante la introducción de técnicas que no habían sido exploradas antes en la literatura. La metodología propuesta en **InP** permite que el usuario realice una intervención mínima en los casos donde la segmentación automática falla. Adicionalmente se propone el uso de un conjunto de directrices para poder evaluar la precisión y eficiencia de las segmentaciones glotales. Estas directrices se dividen en tres grupos: analíticas, subjetivas y objetivas. Los resultados obtenidos a partir de estas directrices sugieren que el método más confiable para la segmentación de la abertura glotal es **InP**, logrando una mejora de un 13% con respecto a otras técnicas en la cuestión del arte y 18% con respecto a **SnW**. También quedo demostrado que el conjunto de directrices pueden ser usadas para estandarizar los criterios de precisión y eficiencia en la evaluación de los algoritmos de segmentación glotal.

Por último, se investigó el uso del Flujo Óptico (Optical Flow (OF)) para resolver los problemas relacionados con la segmentación glotal. A partir del OF tres nuevas representaciones son presentadas para comprender la dinámica de los pliegues vocales. Dos de ellas analizan la dinámica global, Flujo Óptico del Glotovibrograma (Optical Flow Glottovibrogram (OFGVG)) y el Flujo Óptico de la Forma de Onda Glotal (Glottal Optical Flow Waveform (GOFW)); el restante recibe el nombre de Flujo Óptico del Quimograma (Optical Flow Kymogram (OFKG)) y analiza las dinámicas locales de los pliegues vocales. Las ventajas, inconvenientes y como complementan a los métodos ya existentes son discutidos. Las nuevas representaciones fueron evaluadas utilizando una base de datos compuesta por 60 LHSV, la misma que incluye diferentes calidades de voz tanto en voz hablada como en voz cantada. La nuevas representaciones basadas en OF fueron comparadas con las obtenidas mediante segmentación, mostrando que proporcionan información adicional sobre la dinámica temporal de los movimientos vibratorios glotales durante las fases de cierre y apertura glotal.

Résumé

"Il n'est pas la connaissance, mais l'acte d'apprendre, pas la possession, mais l'acte d'y acceder, qui délivrent le plus grand plaisir"

Carl Friedrich Gauss

La voix est l'outil essentiel de la communication entre les êtres humains. C'est ainsi qu'avoir une voix en bonne santé est important dans la vie de tous les jours et plus encore si on l'utilise comme outil de travail. Par conséquent, il est impératif de trouver de nouvelles techniques plus performantes pour comprendre les mécanismes impliqués dans la production de la voix et surtout pour saisir le comportement vibratoire des plis vocaux grâce aux Vidéos Haute Vitesse du Larynx (Laryngeal High-Speed Videoendoscopy (LHSV)). Les études décrites ci-après ont pour objectif de contribuer à l'analyse de la fonction vibratoire des plis vocaux grâce à l'implémentation d'outils plus fiables utilisant des techniques de traitement des images.

La masse des informations à traiter tant sur le plan qualitatif que quantitatif est telle qu'il est nécessaire de synthétiser ces informations spatio-temporelles en quelques représentations statiques reflétant avec précision le mouvement des plis vocaux. Jusqu'à présent, la majorité des avancées dans ce domaine ont été réalisées grâce à la segmentation et au suivi de l'ouverture glottale. Ce type de travail n'est pas aisé notamment à cause de facteurs tels que le bruit sur les images, la variation lumineuse, les différents niveaux de gris représentant l'ouverture glottale, le flou des images, le flou des contours de l'ouverture glottale, le mouvement de la caméra vidéo et/ou des patients.

Afin de résoudre les problèmes précédemment cités, on a utilisé deux algorithmes pour segmenter l'ouverture glottale. Le premier algorithme, appelé segmentation glottale basée sur la technique de ligne de partage des eaux et contours actifs (Glottal Segmentation Based on Watershed Transform and Active Contours (**SnW**)) identifie une région d'intérêt (Region of Interest (ROI)) qui s'actualise automatiquement. Cette méthode combine l'utilisation de modèles déformables (Deformable Models) et de segmentation par ligne de partage des eaux (Watershed Transform) pour délimiter l'ouverture glottale. Grâce à l'implémentation

d'une ROI, cette méthode n'est pas sensibles aux mouvements de la caméra vidéo. Le deuxième algorithme, appelé segmentation glottale basée sur la soustraction des bruits et la reconstruction d'images (Glottal Segmentation Based on Background Subtraction and Inpainting (**InP**)), s'effectue semi-automatiquement pour segmenter avec précision l'ouverture glottale en utilisant différentes techniques encore jamais utilisées. La méthodologie proposée avec l'InP permet à l'utilisateur de réaliser des interventions minimes dans les cas où la segmentation automatique aurait échoué.

De plus, il sera exposé un ensemble de directives pour mesurer la précision et l'efficacité des algorithmes. Ces directives se divisent en trois groupes : analytiques, subjectives et objectives. Les résultats obtenus à partir de ces directives suggèrent que l'algorithme le plus fiable pour la segmentation de l'ouverture glottale est l'InP étant plus précis de 13% par rapport à autres et de 18% par rapport à SnW. Il est également démontré que ces directives peuvent être utilisées pour standardiser les critères de précision et d'efficacité pour l'évaluation des algorithmes de segmentation glottale.

Enfin, dans cette étude, sont présentées les recherches concernant l'usage du flux optique (Optical Flow (OF)) pour résoudre les questions liées à la segmentation glottale. L'OF permet trois nouvelles représentations pour comprendre la dynamique des plis vocaux. Deux d'entre elles analysent la dynamique glottale: le flux optique vibrogramme (Optical Flow Glottovibrogram (OFGVG)) et le flux optique en onde (Glottal Optical Flow Waveform (GOFW)). Le troisième, appelé flux optique Quimogramme (Optical Flow Kymogram (OFKG)) analyse les dynamiques locales des plis vocaux. On présentera les avantages et inconvénients ainsi que la contribution de ces représentations aux méthodes existantes. Ces nouvelles représentations ont été évaluées à l'aide d'une base de données de 60 LHSV qui inclue différentes qualités de voix parlées et chantées. Les représentations basées sur l'OF ont été comparées avec les représentations obtenues grâce aux méthodes de segmentation, démontrant qu'elles apportent des informations supplémentaires sur la dynamique temporelle des mouvements vibratoires de la glotte pendant les phases de fermeture ou d'ouverture glottales.

Contents

Acknowledgements **ix**

Abstract **xi**

Resumen **xiii**

Résumé **xv**

Introduction **xxxv**

I State-of-Art in Voice Production and Laryngeal Imaging **1**

1 Principles of Voice Production **3**
1.1 Voice Production 3
1.1.1 Air Pressure System 4
1.1.2 Phonatory System 4
1.1.3 Articulatory System 11
1.2 Vocal Registers: Definition and Historical Facts 11
1.3 Laryngeal Vibratory Mechanisms 12
1.3.1 Mechanism M0 14
1.3.2 Mechanism M1 14
1.3.3 Mechanism M2 14
1.3.4 Mechanism M3 15
1.4 Discussion 15

2 Laryngeal Imaging **17**
2.1 Laryngeal Imaging Notation and Terminology 17
2.1.1 Imaging Notation 17
2.1.2 Basic Terms and Concepts 18
2.2 Methods for Direct Observation of Vocal Folds 21
2.2.1 Laryngeal Videostroboscopy 22

2.2.2 Laryngeal High-Speed Videoendoscopy 24
2.3 Voice Disorders 25
2.3.1 Normal Behavior of the Vocal Folds 26
2.3.2 Organic Disorders 28
2.3.3 Functionals Disorders 29
2.4 Clinical Applications of the LHSV 30
2.5 Discussion 33

3 Facilitative Playback Techniques **35**
3.1 Importance of Synthesizing LHSV Information 35
3.2 Local-Dynamics Playbacks 37
3.2.1 VKG and DKG Playbacks 37
3.2.2 VFT Playback 38
3.2.3 MKG Playback 40
3.3 Global-Dynamics Playbacks 41
3.3.1 GAW Playback 41
3.3.2 PVG Playbacks 42
3.3.3 GVG Playbacks 45
3.4 Discussion 47

II State-of-Art in Image Processing and Glottal Segmentation **49**

4 Image and Video Processing Techniques **51**
4.1 Review of General Image Segmentation Methods 51
4.1.1 Thresholding 51
4.1.2 Edge-Based 52
4.1.3 Region-Based 53
4.1.4 Classification-Based 56
4.1.5 Graph-Based 57
4.1.6 Deformable Models 58
4.2 Review of Motion Estimation Techniques 60
4.2.1 Phase Correlation 61
4.2.2 Block Matching 62
4.2.3 Pel-Recursive 62
4.2.4 Optical Flow 63
4.2.5 Feature-Based Methods 69
4.3 Review of Inpainting Techniques 70
4.4 Discussion 71

5 Glottal Segmentation Techniques **73**
5.1 Overview 73

5.2 Image Enhancement 75
5.3 Region of Interest 77
5.4 Glottal Gap delimitation 79
5.5 Discussion 84

III Contribution to the laryngeal High-Speed Video Processing 87

6 Contribution to the Glottal Gap Segmentation 89
6.1 Database Description 89
6.2 Glottal Segmentation Based on Watershed and Active Contours . . 90
6.2.1 Image Enhancement 91
6.2.2 ROI Localization 93
6.2.3 First Region Merging 102
6.2.4 Correlation Regions Merging 106
6.2.5 Post-Processing: Localizing Region-Based Active Contours 107
6.2.6 Evaluation of the ROI Performance 108
6.2.7 Drawbacks of SnW Technique 110
6.3 Glottal Segmentation by Background Modeling and Inpainting . . 112
6.3.1 Image Enhancement 112
6.3.2 ROI Localization 116
6.3.3 Glottal Gap Delimitation 116
6.3.4 User Intervention: Including a Manual ROI 117
6.4 Accuracy Assessment of the Vocal Folds Segmentation 121
6.4.1 Analytical Methods: Assessing the Efficiency of the Vocal Folds Segmentations 122
6.4.2 Subjective Evaluation: Accuracy of the Vocal folds Detection by Playbacks Analysis 122
6.4.3 Objective Supervised Evaluation: Accuracy of the Vocal Folds Detection via Ground-Truth Comparison 124
6.5 Results 127
6.5.1 Analytical Assessment 129
6.5.2 Subjective Assessment 129
6.5.3 Objective Supervised Assessment 133
6.6 Discussion 134

7 Synthesizing the Vocal Folds Motion by Optical Flow 137
7.1 Optical Flow in LHSV 137
7.2 Database Description 139
7.3 Image Processing Implementation 140
7.4 New Playbacks for Visualizing Glottal Dynamics 140
7.4.1 Local Dynamics Along One Line: Optical Flow Kymogram 141

7.4.2 Global Dynamics Along the Whole Vocal Folds Length: Optical Flow Glottovibrogram 141
7.4.3 Global Velocity: Glottal Optical Flow Waveform 144
7.4.4 Definition of the Vocal Folds Displacements Trajectories . 144
7.5 Reliability Assessment of Optical Flow Playbacks 145
7.6 Results 146
7.6.1 Comparison Among Segmentation and OF Displacement Trajectories 146
7.6.2 Comparison of OF Playbacks with Traditional Ones . . . 148
7.6.3 Global Dynamics Evaluation for the Whole Database . . . 149
7.7 Discussion 155

IV Conclusions and Future Works 157

8 Conclusions and Future Works 159
8.1 Conclusions 159
8.2 Future Works 163

V Appendices 165

A Playbacks Computed Using InP 167

B OFGVG Playback Thumbnails 175

C Scientific Production 181

References 185

List of Figures

1.1 General scheme of the voice production apparatus. Adapted from (Freepik, 2016). 4

1.2 Anterolateral, front, rear, side and top view of the laryngeal apparatus with its respective anatomical structures. Adapted from (Laver, 2009). 5

1.3 Intrinsic muscles of the larynx. Adapted from (Hixon et al., 2008). 7

1.4 Schematic representation of the abduction and adduction of the vocal folds. Adapted from (Henrich, 2001). 7

1.5 Coronal section through the free edge of the vocal folds and thyroarytenoid muscles. 9

1.6 Schematic representation of the *Myoelastic-Aerodynamic* theory, with a coronal and superior visualization of the vocal folds. 10

1.7 Schematic representation of the vocal tract with its differents parts. Adapted from (Gilles, 2010). 11

1.8 Illustration of the time frequency analysis of the four laryngeal vibratory mechanisms during the production of an ascending glissando sung by a soprano. Adapted from (Henrich, 2006). 13

1.9 Illustration of the superior and coronal glottal configuration associated with the mechanism M1. Adapted from (Henrich, 2001). . . 14

1.10 Illustration of the superior and coronal glottal configuration associated with the mechanism M2. Adapted from (Henrich, 2001). . . 15

2.1 Laryngeal image sequence with its respective notation. Left side: laryngeal image sequence $I(\mathbf{x},t)$; right side: single image $I(x,y)$. . 18

2.2 Representation of one glottal cycle $G_{C_o}(\mathbf{x},t)$ with its respective phases: closed-state; opening phase; open-state; closing phase. . . 21

2.3 Illustration of the stroboscopic sampling. Adapted from (Kendall and Leonard, 2010). (1) and (3) are the rigid and flexible endoscope, respectively; (2) are the vocal folds; (4) represents the real vibratory pattern; (5) illustrates the strobe light; (6) is the estimated version of the vibratory pattern. 23

2.4 Illustrations of the LHSV sampling effect for two different frame rates. Adapted from (Kendall and Leonard, 2010). 24
2.5 Common glottal configuration: (a) complete closure; (b) posterior chink; (c) anterior chink; (d) spindle closure; (e) irregular closure; (f) incomplete closure; (g) hourglass. 27
2.6 Illustration of the Organic disorders. Adapted from (BTP, 2014). . 29

3.1 Schematical drawing of the successive phases of a glottal cycle in three views. First row: Frontal section of the vocal folds. Second row: Laryngoscopy (superior view of the vocal folds). Third row: High-Speed Digital Kymograghy at the line y_j. Adapted from (Švec and Šram, 2002). 38
3.2 Illustration of the VFT playback. First row: the image sequence of one glottal cycle $G_{C_o}(\mathbf{x},t)$. Second row: vocal folds trajectories $\delta_{seg}^{l,r}(pc,t)$, DKG and DKG+VFT playbacks of five glottal cycles. . 39
3.3 Illustration of the MKG playback. First row: six images of a particular glottal cycle G_{C_o}. $\mathbf{KG}(t_k)$ is a horizontal line at time t_k and position y_j. Second row: MKG playback, the green tonalities represent the opening phase and the red tonalities represents the closing phase. Adapted from (Deliyski et al., 2008). 40
3.4 Illustration of the GAW playback. First row: six images of a particular glottal cycle G_{C_o}. Second row: GAW playback normalised within the interval [0,1] where 0 represents the minimum area and 1 the maximum area. 42
3.5 Interpolation procedures of the intermediate images within the interval $[t_O, t_{O+1}]$. First row: six images of a particular glottal cycle G_{C_o} explaining the interpolation procedure. Second row: GAW playback normalised within the interval [0,1]. 44
3.6 Schematic representation of the PVG playback. (1) Segmentation; (2) Resampling of the extracted vocal folds edges; (3) Computation of vocal folds deflections $\delta_{seg}^{l,r}(pc,t_k)$; (4) Splitting of the glottal axis; (5) Virtual turning of the left fold; (6) Color coding of the vocal fold deflections; (7) $I_{PVG}(x,y)$ representing a LHSV with five glottal cycles. Adapted from (Lohscheller and Eysholdt, 2008b). . 45
3.7 Schematic representation of the GVG playback. (1) Segmentation; (2) Resampling of the extracted vocal folds edges; (3) Computation of vocal fold deflections $\delta_{GVG}(pc,t_k)$; (4) $I_{GVG}(x,y)$ representing a LHSV with five glottal cycles. 46

4.1 Illustration of the Otsu's thresholding method. First row: coin image thresholded automatically with a value of 126. Second row: laryngeal image thresholded automatically with a value of 163. . . 52

4.2 A laryngeal image segmented by different edge-based techniques. (a) Original image; (b) Roberts detector; (c) Prewitt detector; (d) Sobel detector; (e) LoG detector; (f) Canny detector; (g) Hough Transform with Sobel detector; (h) Hough Transform with LoG detector; (i) Hough Transform with Canny detector. 54

4.3 Laryngeal image segmented by region growing: (a) and (c) are the same image but with a different seed pixel (circles in blue); (b) and (d) are the respectively region growing segmented images (white regions). 55

4.4 Watershed transformation applied to laryngeal images: (a) and (d) original frames; (b) and (e) watershed transform computed on the gradient of the images; (c) and (f) post-processing step using morphological controlled marked and Just Noticeable Difference (JND), respectively. 56

4.5 Classification-based segmentation applied to laryngeal images: (a) original frame; (b) K-means with two classes and color features; (c) K-means with ten classes and color features; (d) K-means with five classes and color-spatial features. 58

4.6 Interactive Graph-based segmentation applied to laryngeal images: (a) and (d) original image; (b) and (e) interactive user-drawn markers; red background and blue foreground; (c) and (f) final segmentation using graph-cut method. 59

4.7 Active contours applied to laryngeal images: (a) original image; (b) snake initialization; (c) deformable model based on Chan and Vese (2001) with 300 iterations; (d) deformable model based on Lankton and Tannenbaum (2008) with 300 iterations; (e) Gradient vector flow method proposed by Xu and Prince (1998); (f) active contours based on Andrade-Miranda et al. (2013) with automatic initialization. 61

4.8 Illustration of the block matching algorithm. 62

4.9 The visualization of flow fields. Left side: color code visualization, and right side: arrow visualization. Adapted from (Liu et al., 2011). 65

4.10 Arrows and color code visualization. (a) Two laryngeal images taken during the opening phase of the vocal folds at time t_k and t_{k+3}; (b) OF based on Horn and Schunck (1981); (c) improved mathematical formulation of Horn and Schunck (1981) using (Bruhn et al., 2006); (d) OF based on Drulea and Nedevschi (2013). . . . 66

4.11 Arrows and color code visualization using LK-OF computation . . 67

4.12 Arrows and color code visualization using MT-OF computation . . 68

4.13 Arrows and color code visualization using TVL1-OF computation 69

4.14 Diffusion-based inpainting applied to laryngeal images: (a) and (c) are the same image but with a different region to be inpainted (rectangles in blue); (b) and (d) are the results of appling a Diffusion-based inpainting technique. 71

5.1 Different laryngeal images during phonation showing different illumination conditions, orientation, depth, occlusion, among others features are showed. 74

5.2 Graphic Representation of the three common steps followed to segment the glottal gap. 76

5.3 Visual representation of the different enhancement methods for three different LHSV. First row: original image; second row: anisotropic with FFT-filter (Mendez et al., 2009); third row: CLAHE (Karakozoglou et al., 2012); fourth row: nonlinear transformation with $\beta = 200$ (Skalski et al., 2008). 78

5.4 Automatic ROI detection of six different LHSV computed based on (Andrade-Miranda and Godino-Llorente, 2014; Andrade-Miranda et al., 2015b). The image in (f) illustrates a minor problem in the detection of the ROI in the posterior commissure. 79

6.1 Graphic Representation of the different steps followed to segment the glottal gap: image enhancement, ROI detection, first region merging, correlation merging and post-processing. In this case, to differentiate from an arbitrary segmentation $I_{seg}(\mathbf{x},t)$, the final glottal delineation is denoted as $SnW(\mathbf{x},t)$. 91

6.2 Comparison of the pre-processing algorithms. The objective evaluations applied to 110 HSDI images extracted from the 22 videos (DB1): (a) PSNR graph; (b) EOR graph; (c) MSOR graph. 94

6.3 Visual representation of the non linear transformation using different values of β. 95

6.4 ROI localization. Upper panel: procedure to obtain $\mathbf{TIV}_c$; bottom panel: procedure to compute $\mathbf{TIV}_r$ and the final ROI. 96

6.5 $\mathbf{TIV}_r$ for N_{ROI}=100 frames. The LHSV sequence has a glottal chink and the glottis is splitted in two parts, illustrating one of the most demanding cases. The frame is rotated in horizontal position for a better visualization. 98

6.6 Effect of the transversal motion in $\mathbf{TIV}_c$. The importance of recomputing the ROI is illustrated plotting different $\mathbf{TIV}_c$ without gaussian fitting for different values of N_{ROI}: N_{ROI}=30, N_{ROI}=1000, N_{ROI}=2000, and N_{ROI}=2975. 100

6.7 Evaluation of the effect of different N_{ROI} settings. Graphical representation of the varaition of N_{ROI} for three different LHSV sequences $\mathbf{TIV}_c$ for $N_{ROI} = 30$(blue line), $N_{ROI} = 100$(red line), $N_{ROI} = 300$(green line), $N_{ROI} = 600$(black line). 101
6.8 Visibility threshold of the human visual system as a function of the grey level in the image. 104
6.9 Illustration of the first region merging. First row: image enhanced; second row: watershed transform after thresholding the magnitude of the gradient; third row: the watershed transform after the JND cost fucntion. 105
6.10 Merging steps. (a) Standard template found empirically based on manual segmentation, $T(\mathbf{x})$; (b) and (c) show from left to right: two different frames of two different sequences, similitude matrix, first region merging, cross-correlation overlapping and correlation region merging. 107
6.11 Complete methodology representation. From left to right: enhanced image $I_{NLT}(\mathbf{x},t_k)$; segmentation obtained after watershed and first region merging $I_{JND}(\mathbf{x},t_k)$; second region merging $I_{cor}(\mathbf{x},t_k)$; final delimitation of the glottis after 100 iterations $SnW(\mathbf{x},t_k)$. . . . 109
6.12 ROI detection using two approaches. a) ROI detection according to (Karakozoglou et al., 2012); b) final ROI obtained using the approach based on intensity variation. 110
6.13 Examples of the results obtained using a ROI based on intensity variation with their respective $\mathbf{TIV}_c$ and $\mathbf{TIV}_r$ plots. 111
6.14 Graphic representation of the steps followed to segment the glottal gap. 113
6.15 Contrast enhancement procedure for two different LHSV. From left to right: original LHSV; color equalization; image free of specularity; and image after bilateral filtering 115
6.16 Complete framework of the glottal gap delimitation. 116
6.17 Glottal gap segmentation of 9 LHSVs in the instants of time t_k= 1, 3, 5, 7, 9, 11, 13, 15 using **InP** method. 118
6.18 Graphical representation of the manual intervention process. Top left panel: frame with the maximal opening; bottom left panel: representation of the vocal folds reflections; medium panel: user interaction and manual ROI; top right panel: segmentation results; bottom right panel: detail of the slide bar to manually modify the threshold. 120
6.19 Four HSVs with their respective GVG, PVG, VKG and GAW playbacks. Automatic and manual ROIs are represented by a green rectangle. a) Error due to the glottal gap orientation; b) paralysis of both vocal folds; c) unilateral paralysis of the left fold; d) partial paralysis of the right fold. 121

6.20 LHSV of a patient after a carcinoma surgery with its respective playbacks: GVG, PVG, VKG, GAW and VFT. a) Segmented with **InP**; b) segmented with SrG. (1) Vibratory pattern; (2) errors in the anterior or posterior part of the glottis; (3) playbacks discontinuities; (4) main glottal axis crossing; (5) glottal area waveform; (6) opening state length. 123
6.21 VKG and VKT playbacks of a patient after a carcinoma surgery. First row: vocal folds displacement trajectories; second row: videokymogram; third row: VKG and VKT overlapping. The dashed lines in white show the correct delimitation of the VKG. 125
6.22 Segmented frames corresponding to a dysphonic voice (P5) and to a normal voice production (P7). First row: manual segmentation (MaN); second row: segmentation based on inpainting (**InP**); third row: segmentation based on region growing (SrG); fourth row: segmentation based on snakes and watershed (**SnW**). 131
6.23 GVG and PVG playbacks corresponding to the 10 HSVs presented in Table 6.7. 132
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. 133
6.25 Segmentation subjective assessment of 10 patients on a 5-point scale. 133
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. 135

7.1 Illustration of a synthetic motion field $\mathcal{W}(\mathbf{x},t)$ located among the posterior ($\mathbf{p}(t+1)$) and anterior ($\mathbf{a}(t+1)$) part of the vocal folds during two consecutive instants of time, t_k and t_{k+1}. 138
7.2 Fluctuation of u along one line for a complete glottal cycle 139
7.3 Graphical representation of the procedure followed to compute the new playbacks. 141
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). 142
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 $U(\mathbf{x},t_k)$. Third row: schematic view of one OFGVG cycle. Dark regions indicate no velocity ($u(\mathbf{x}_{ij},t_k)=0$). 143

7.6 Schematic procedure to compute $\hat{\delta}^{l,r}_{OF}(pc,t_k)$ during the opening phase. 145

7.7 First row: correlation between OF trajectories and segmentation trajectory for each sequence. Second and third row: $\hat{\delta}^{l,r}_{seg}$ and $\hat{\delta}^{l,r}_{TVL1}$ are compared for two phonatory tasks: breathy and creaky (sequences selected on first panel). The left panel shows four frames of each LHSV with their respective segmentation and trajectories. The right panel shows the close up of two frames with segmentation errors corresponding to the interval in dashed lines. 147

7.8 Illustration of GVG, $|d_x\text{GVG}|$, OFGVG-LK, OFGVG-MT and OFGVG-TVL1 playbacks for three different phonatory tasks (pressed, glide up and glide down). 149

7.9 GAW vs GOFW representation for a pressed and glissando task. First row: GAW and $|d\text{GAW}|$; second row: GAW and GOFW-LK; third row: GAW and GOFW-MT; fourth row: GAW and GOFW-TVL1. 150

7.10 Correlation and SSIM obtained by comparing $|d_x\text{GVG}|$ with each OFGVG. The horizontal axis represents the video sequence in the DB3 database, and the vertical axis the value of the metrics. . . . 151

7.11 Upper panel: nine segmented frames, the rectangle dotted with red correspond to the space between the margin of the ROI and to the area with a glottal chink; middle panel: $|d_x\text{GVG}|$ playback; lower panel: OFGVG with a vertical length that depends on the ROI size. The effect caused by the mucosal wave motion and the vibratory shape pattern for three consecutive cycles are marked with dotted and continuous red lines respectively. 152

7.12 Upper panel: nine segmented frames, the areas dotted with red correspond to the posterior glottal chink; middle panel: $|d_x\text{GVG}|$ 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 $|d_x\text{GVG}|$. The vibratory shape pattern for three consecutive cycles is marked with a continuous red line. 153

7.13 $|d_x\text{GVG}|$ and OFGVG 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. 154

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. 154

7.15 Illustration of the mucosal wave contribution in an OFGVG playback. First row: OFGVG of the whole vocal folds surface motion (OFGVG$_{OF}$), OFGVG of the vocal folds edges (OFGVG$_{seg}$) and mucosal wave propagation (MW). Second row: overlapping of both contributions (in red, the edges motion, and in blue, mucosal wave surface motion) 155

List of Tables

1.1 Classification of registers depending on the laryngeal mechanisms involved (table extracted and modified from (Roubeau et al., 2009)). 13

3.1 Summary of the main studies carried out to synthesize the vocal folds vibratory pattern. 36

5.1 Summary of the main studies carried out from glottal segmentation 80

6.1 Summary of the results reported in Figure. 6.2 for the different image enhancement techniques used. 93

6.2 Overview of the findings in the 16 LHSV that required manual intervention. 121

6.3 Summary of the 18 metrics with the selection guidelines. Each row corresponds to one of the metrics. The first seven columns correspond to the properties evaluated to be part of the *good metrics* set. A check (✓) denotes that the metric is recommended for the corresponding property; a cross (X) denotes that the metric is not recommended; and empty cells denote neutrality. The last three columns are the average values of each metric for the three assessments. 128

6.4 Pearson's Correlation coefficients among the *good metrics*. Correlations correspond to MaN vs. **InP**, MaN vs. SrG and MaN vs. **SnW** trials. 128

6.5 Segmentation times of the three algorithms (in fps) of 400 high-speed images. 129

6.6 Subjective assessments used to evaluate the segmentation performance (in a 0-5 point scale). 130

6.7 Summary of the clinical information for a subset of 10 HSV taken from the database DB2. 130

6.8 Comparison of the accuracy improvements of **InP** with respect to SrG and **SnW**. 134

A.1 Used recordings from the database DB2. The LHSV sequences indices are shown, as well as the glottal segmentation at time t_k=1, 5, 10, 15, 20. The last column from top to down presents the corresponding GVG, PVG, DKG and GAW playbacks. 173

B.1 Used recordings from the database DB3. The LHSV sequences indices are shown, as well as the used laryngeal mechanisms and the glottal vibration pattern. The last column from top to down presents a small portion of the corresponding GVG, $|d_x\text{GVG}|$ and OFGVG-TVL1 thumbnail. 180

List of Acronyms

AHE	Adaptive Histogram Equalization
BBC	Brightness Constancy Constraint
CC	Normalize Correlation Coefficient
CLAHE	Contrast Limited Adaptive Histogram Equalization
CLHE	Contrast Limited Histogram Equalization
CQ	Closed Quotient
DB1	Database1
DB2	Database2
DB3	Database3
DFTA	Discrete Fourier Transform Analysis
DKG	High-Speed Digital Kymograghy
DTW	Dynamic Time Warping Analysis
EFA	Eigenfolds Analysis
EGG	Electroglottography
EOF	Empirical Orthogonal Eigenfunctions Analysis
EOR	Edge Overlapping Ratio
fps	Frames per Second
GAW	Glottal Area Waveform
GMM	Gaussian Mixture Model
GOFW	Glottal Optical Flow Waveform
GTG	Glottaltopogram
GVG	Glottovibrogram
HE	Histogram Equalization
HTA	Hilbert Transform Analysis

InP	Glottal Segmentation Based on Background Subtraction and Inpainting
JND	Just Noticeable Difference
LGT	Laryngotopography
LHSV	Laryngeal High-Speed Videoendoscopy
LK-OF	Lukas Kanade Optical-Flow
LVS	Laryngeal Videostroboscopy
M0	Mechanism M0
M1	Mechanism M1
M2	Mechanism M2
M3	Mechanism M3
MAE	Mean Absolute Error
MaN	Manual Segmentation
MKG	Mucosal Wave Kymography
MSE	Mean Square Error
MSOR	Mean Segment Overlapping Ratio
MT-OF	Motion Tensor Optical-Flow
MW	Mucosal Wave
NDA	Nonlinear Dynamic Analysis
OF	Optical Flow
OFGVG	Optical Flow Glottovibrogram
OFKG	Optical Flow Kymogram
OQ	Open Quotient
PCA	Principal Component Analysis
PGAW	Phasegram Analysis
PSNR	Peak-Signal-to-Noise-Ratio
PVG	Phonovibrogram
PVG-wavegram	Phonovibrographic Wavegram
RGB	RGB Color Space
ROI	Region of Interest
SnW	Glottal Segmentation Based on Watershed Transform and Active Contours
SQ	Speed Quotient
SrG	Seed Region Growing
SSIM	Structural Similarity Index

$\mathbf{TIV}_c$	Total Intensity Variation in Columns
$\mathbf{TIV}_r$	Total Intensity Variation in Rows
TVL1-OF	Total Variation L1 Optical-Flow
VF	Vocal Folds
VFDT	Vocal Folds Displacement Trajectories
VFT	Vocal Folds Trajectories
VKG	Videokymography
VP	Vibration Profiles
WDA	Waveform Decomposition Analysis

Introduction

"I never did anything worth doing by accident, nor did any of my inventions come by accident; they came by work"

Thomas A. Edison

The voice is the most basic tool that supports the usual method of communication of the human being, with which the culture is transmitted and feelings and emotions are expressed. Therefore, a healthy voice is very important to people's daily life, especially for the professional voice users. However, due to the misuse and overuse of the voice, changes in the laryngeal structures and vocal folds may lead to voice disorders. The consequences of such disorders have a different impact depending on the population affected. For instance, an auto mechanic who loses one or two notes at the top of his range is likely to have his personal and professional life unaffected. On the other hand, a singer with the same symptoms may be totally disabled. This negative impact will affect its ability to work, on their overall sense of well-being, and sometimes on their very sense of self.

For this reason, a clinical voice assessment is an important component to the diagnosis of voice disorders, and for planning the appropriate treatment strategies. According to the American Academy of Otolaryngology-Head and Neck-Surgery, the basic protocol to evaluate a patient with a voice disorder has to include a rigorous clinical history, physical examination, and visualization of the larynx via laryngoscopy. However, compared with physical examination, only the laryngoscopy allows the etiology determination of a voice disorder. Hence, the examination of the vibratory characteristics of the vocal folds function has taken a great importance.

For clinicians, an essential part of a thorough examination is the use of Laryngeal Videostroboscopy (LVS). It has been used to examine subtle abnormalities along the vibratory margin of the vocal folds such as small cysts, scars from previous injury, and to detect subtle problems such as mild inflammation, subtle swelling of the vocal folds, white patches, or excessive mucus. However, significant vibration details might be overlooked while using the LVS due to its low recording frame rate (e.g., around 30 Frames per Second (fps)) in the presence of voice disorders that results in irregular vocal folds vibration or short phonation duration. In this situation, Laryngeal High-Speed Videoendoscopy (LHSV), with its significantly

higher capturing frame rate addresses the limitations of LVS, making it helpful to investigate the vocal folds vibratory features.

Nowadays, due to the fast-growth of imaging technology, it is possible to find high-speed cameras with frame rates up to 10000 fps, so LHSV is currently regarded as a superior method to LVS for the assessment of vocal folds vibration. First, LHSV is applicable to the assessment of unstable phonations such as transient, subharmonic, or aperiodic phonations, and thus is more useful for investigating vocal pathology. Second, LHSV allows 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 which provides greater validity for assessment of intracycle and intercycle vibratory characteristics compared with LVS. Third, LHSV data can be analyzed by a wider variety of methods than LVS data, enabling more interpretable and extensive evaluation on both a qualitative and quantitative basis.

The use of LHSV 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 objective information of the time-varying data so the clinician or the researcher can follow the dynamics of anatomical features of interest in a more intuitive way, revealing contents which are often hidden to human eyes. For this reason, the literature reports different representations, usually called facilitative playbacks extracted from LHSV, able to identify objectively the presence of organic voice disorders, classify functional voice disorders, categorize vibratory patterns, and to discriminate early stages of malignant and precancerous vocal folds lesions, among others. Most of these milestones have been achieved thanks to the segmentation and tracking of the glottal gap which is not a trivial task due to noise in the image, variability in illumination, variability of the gray levels presented in the glottal gap, fuzziness, blurring edges, movements of the camera and/or patient.

Despite the great progress obtained up to date, the total adoption of the LHSV into routine clinical practice requires additional development. Therefore, the researchers need new methods for data visualization to overcome the drawbacks of existing ones, providing simultaneously features that would integrate the time dynamics, such as: velocity, acceleration, instants of maximum and minimum velocity, vocal folds displacements during phonation and motion analysis. In this way, the LHSV can prove that it is capable of determining the nature and extent of voice disorders.

Contribution

The present work aims to contribute to the analysis of the vocal folds function by proposing the following objectives:

1. Presenting a detailed description of the concepts and definitions associated with the voice production as well as the laryngeal vibratory mechanisms used to define the different configurations of the glottal vibrator that allow the production of the entire frequency range of the human voice.

2. Presenting a detailed description of the laryngeal imaging techniques, pointing out to their concepts, notation, advantages, limitations, and the importance for clinical applications.

3. Presenting a detailed review of the concepts and definitions associated with the main techniques of image segmentation.

4. Presenting a detailed review of the literature devoted to solving the problem of the glottal gap segmentation by mentioning the different algorithms proposed until now.

5. Implementation of a complete framework to segment and track the glottal gap accurately. The algorithm has to consider a minor user intervention in cases when the segmentation is not as expected.

6. Proposing a set of guidelines to measure the accuracy and efficiency of the glottal gap segmentation algorithms. The guidelines have to include an analytical, subjective, and objective assessment to provide robust criteria to decide which is the most appropriate method to delineate the glottal gap.

7. Presenting a detailed review of the concepts and definitions associated with the main techniques of image segmentation and motion estimation. Within this objective, a special attention is given to the Optical Flow computation since it allows the possibility to track unidentified objects solely based on its motion.

8. Finding out new methods to synthesize the vibratory pattern of the vocal folds to overcome the drawbacks of existing ones, providing information not only for those points belonging to the glottal edges but also those regions that originated such movements.

Thesis Structure

This thesis report is structured in eight chapters and two appendices. This section provides a global view of the document organization to make easy its reading and understanding. A summary of the contents of the eight chapters of this thesis is detailed as follows:

Chapter 1: Principles of Voice Production
This chapter presents a brief review of the concepts and definitions related to the voice production. Additionally, a description of the anatomy and physiology of the larynx is presented. Lastly, the concept of vocal register, laryngeal vibratory mechanisms and some particular phonatory situations are introduced.

Chapter 2: Laryngeal Imaging
This chapter addresses the most important aspects of the vibratory behavior of the vocal folds from an image-based point of view. First, a detailed description of the laryngeal imaging techniques is presented. Later on, the vibratory behaviour of the vocal folds is studied. Lastly, a review of the clinical applications of the laryngeal imaging until now is presented.

Chapter 3: Facilitative Playback Techniques
This chapter introduces the concept of facilitative playbacks to better visualize the features of the vocal folds dynamic and highlights the importance of synthetizing the LHSV information. Later on, the most widespread playbacks are presented and divided into two groups based on how the vocal folds motion is assessed.

Chapter 4: Image and Video Processing Techniques
This chapter presents a brief review of the basic concepts and definitions related to the most relevant techniques for image and video processing. The examples of the different segmentation and motion estimation algorithms are focused on solving the problem of the glottal gap delimitation.

Chapter 5: Glottal Segmentation Techniques
This chapter reviews the literature devoted to solve the problem of the glottal gap segmentation dividing the different approaches into three main stages: Image Enhancement, identification of the Region of Interest (ROI), and Glottal Gap Delimitation.

Chapter 6: Contribution to the Glottal Gap Segmentation
This chapter proposes two algorithms to tackle the problem of the glottal gap segmentation. The first one, named as Glottal Segmentation Based on Watershed Transform and Active Contours (**SnW**), uses traditional image segmentation methods but adding the temporal information of the videos. The second one, named Glottal Segmentation Based on Background Subtraction and Inpainting (**InP**), presents a quasi-automatic framework and introduces several techniques

never explored previously in the state of the art. Lastly, a set of validation guidelines are proposed in order to standardize the criteria of accuracy and efficiency of the segmentation algorithms.

Chapter 7: Synthetizing the Vocal Folds Motion by Optical Flow
This chapter introduces three new playbacks 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.

Chapter 8: Conclusions and Future Works
Finally, this chapter presents the conclusions and futures works.

Part I

State-of-Art in Voice Production and Laryngeal Imaging

Chapter 1

Principles of Voice Production

"The human voice is the first and most natural musical instrument, also the most emotional"

Klaus Schulze

SUMMARY: This chapter presents a brief review of the concepts and definitions relating to the voice production. The different organs with their specific roles in voice production are described. Additionally, a description of the anatomy and physiology of the larynx is presented. This description is very important for the interpretation of laryngeal imaging in the evaluation of patients with voice disorders. Lastly, the concept of vocal register, laryngeal vibratory mechanisms and some particular phonatory situations are introduced.

1.1 Voice Production

The voice production is a complex process that includes several structural and functional components. The voice mechanism is composed of three systems where each of them uses different organs and has specific roles in the voice production (Howard and Murphy, 2008). The first is the source of air which is the power supply for the voice (air pressure system). This interaction between air and structures makes that a set of components start to vibrate (phonatory system), producing acoustic waves. These waves propagate and are radiated towards the external medium (articulatory system). Figure 1.1 depicts a general scheme of the voice production apparatus, including the location of each system.

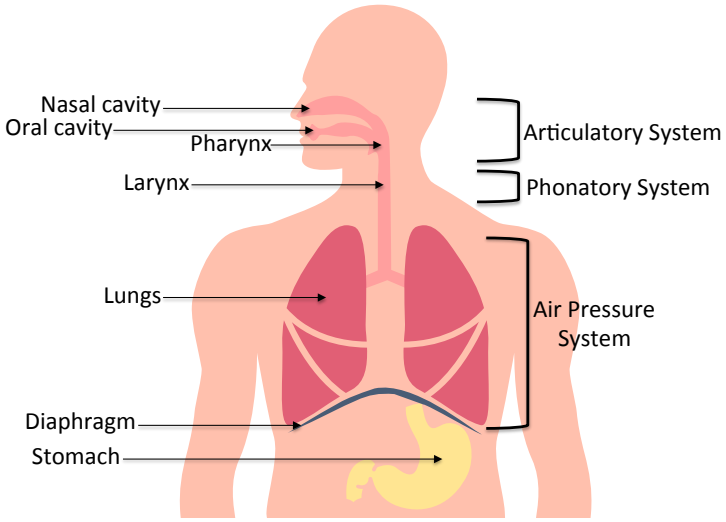


Figure 1.1: General scheme of the voice production apparatus. Adapted from (Freepik, 2016).

1.1.1 Air Pressure System

The air pressure system provides and regulates the airflow and it is composed by the diaphragm, trachea, chest muscles, ribs, abdominal muscles and the lungs. The voice production begins with the inspiration (inhalation), then the air goes through the mouth and nose, passes down the trachea, and is inhaled into the lungs. For air to be inhaled into the lungs, the ribcage needs to expand and the dome-like diaphragm which forms the base of the chest needs to flatten downwards. Once the air has been inhaled into the lungs and they reach capacity, the elastic tissue of the lung recoils and the air is exhaled or breathed out. The exhaled air then returns up through the trachea and then through the larynx where it encounters the closing Vocal Folds (VF) (Godino-Llorente, 2002).

1.1.2 Phonatory System

In the phonatory system, the aerodynamic and mechanical energy coming from the air pressure system is converted to acoustic energy by the fluid-structure interaction between the air and the movable walls of the larynx (vocal folds, vestibular folds or false folds) (Henrich, 2015).

1.1.2.1 Anatomy of the Laryngeal Apparatus

The larynx is a structure supported by a cartilage framework and is located in the anterior portion of the neck, just below to the hyoid bone and above the trachea (see Figure 1.1). It is composed of three large, unpaired cartilages (thyroid, epiglottis, cricoid); three pairs of smaller cartilages (arytenoids, corniculate, cuneiform); two

pairs of laryngeal joints that articulate the cartilages (cricothyroid, cricoarytenoid); two nerves (recurrent laryngeal nerve, superior laryngeal nerve); and two groups of muscles (intrinsic muscles, extrinsic muscles) (Kendall and Leonard, 2010; Hixon et al., 2008; Simpson and Rosen, 2008). Figure 1.2 shows the anterolateral view of the laryngeal apparatus with its respective anatomical structures.

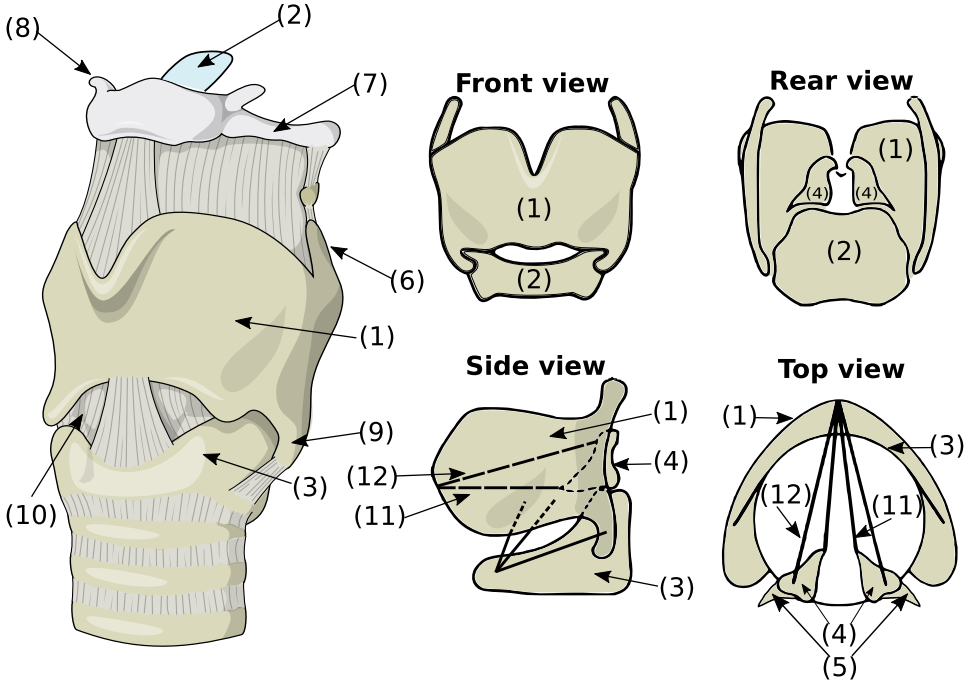


Figure 1.2: Anterolateral, front, rear, side and top view of the laryngeal apparatus with its respective anatomical structures. Adapted from (Laver, 2009).

The thyroid cartilage is the largest of the laryngeal cartilages. It is shaped like a shield with a right and left lamina fusing in the midline, forming the prominence known as Adam's apple. The back edges of the thyroid lamina extend upward into two long horns (superior horn) and downward into two short (inferior horn). The upper ones are coupled to the hyoid bone meanwhile, the lower horns have areas where other structures join. The epiglottis is a leaf-shaped cartilage that is positioned behind the hyoid bone and root of the tongue. This cartilage moves down to form a lid over the glottis and protects the larynx from aspiration of foods or liquids being swallowed. The cricoid cartilage has a ring-shaped structure located above the trachea and sits inside the posterior aspect of the thyroid cartilage. It is the only complete ring of cartilage around the trachea. Sitting on the superior surface of the posterior cricoid lamina are the paired arytenoid cartilages. The arytenoid

cartilages are pyramidal and articulate with the cricoid cartilage through a joint that allows the arytenoids to both swivel and slide relative to the cricoid cartilage. The arytenoid cartilages form the part of the larynx to which the vocal ligaments and vocal folds are attached. Over the top of each arytenoid exists a small cone-shaped cartilage called corniculate cartilage which serves to prolong the arytenoids posteriorly and medially. Lastly, the cuneiform cartilages have a club-shaped that lies anterior to the corniculate cartilages in the aryepiglottic folds.

The laryngeal joints allow the interconnection between the cricoid and thyroid cartilages (cricothyroid joints), and the interconnection between the cricoid and arytenoid cartilages (cricoarytenoid joints). The cricothyroid joints allow the anteroposterior sliding and rotation of the inferior cornu upon the cricoid cartilage, meanwhile the cricoarytenoid joints permit motion in a sliding, rocking, and twisting fashion of the arytenoid cartilages.

Regarding nervation, there are two nerves coming from the brain to the larynx which control the movement of the larynx: the recurrent laryngeal nerve is responsible for the opening of the vocal folds (as in breathing and coughing), and the closing of them during voice use and swallowing; and the superior laryngeal nerve is in charge of adjusting the tension of the vocal folds for high notes during singing.

The extrinsic muscles provide laryngeal stabilization, vertical mobility, and indirectly may affect vocal folds position. On the other hand, the intrinsic muscles act directly on the vocal folds by controlling the adduction[1] and abduction[2] length, tension, shape, position and vibratory motion. They can be subdivided into three major vocal fold adductors (thyroarytenoid, lateral cricoarytenoid, interarytenoid), one abductor (posterior cricoarytenoid), and one tensor muscle (cricothyroid). Figure 1.3 depicts the intrinsic muscles of larynx from the top, rear and side view.

1.1.2.2 Anatomy of the Vocal Folds

The VF are composed of twin infoldings of mucous membrane attached between the midline of the thyroid cartilage and the anterior aspect of the arytenoid cartilages. On the one hand, the contraction of the posterior cricoarytenoid muscles is the origin of the abduction of the vocal folds. The gap created by the vocal folds is referred to as glottis or glottal opening. On the other hand, when the arytenoids are closed by the lateral cricoarytenoid and interarytenoid muscles contraction, the vocal folds are brought to the midline resulting in glottal closure. Figure 1.4 depicts a schematic representation of the abduction and adduction of the vocal folds due to the action of the intrinsic muscles.

The vocal folds consist of three layers that work together to permit the vocal folds vibration (Simpson and Rosen, 2008). The first and second layer are called cover. The vibration of this layer results in glottal opening and closing, creating

[1] Adduction is the action to close the vocal folds.
[2] Abduction is the action to open the vocal folds.

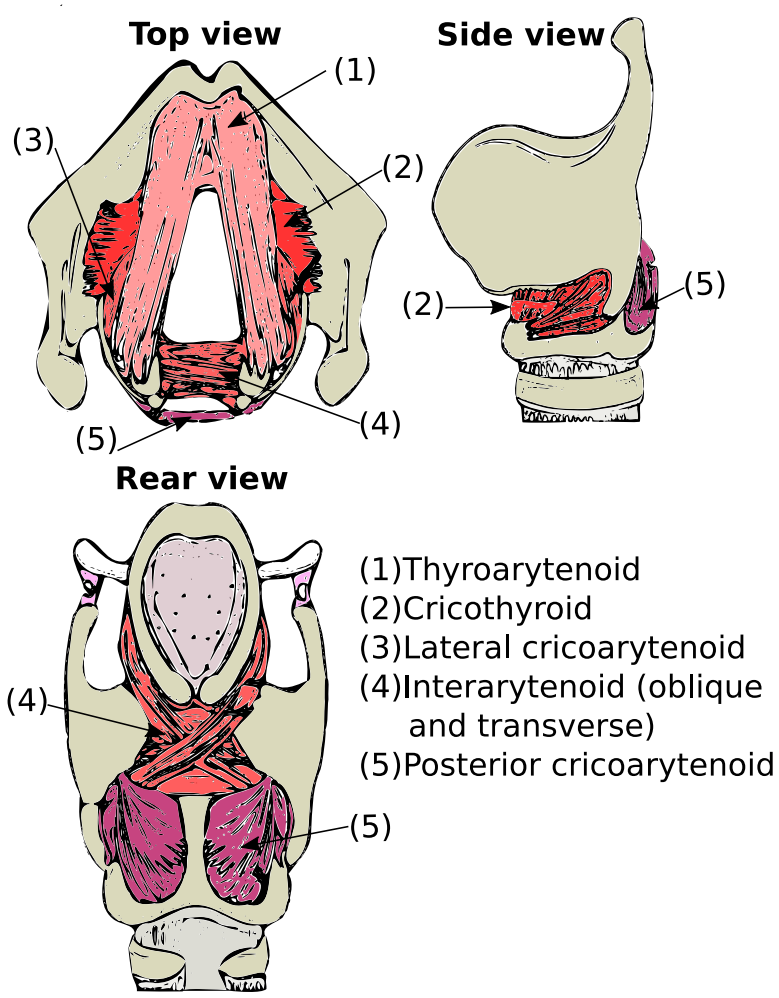


Figure 1.3: Intrinsic muscles of the larynx. Adapted from (Hixon et al., 2008).

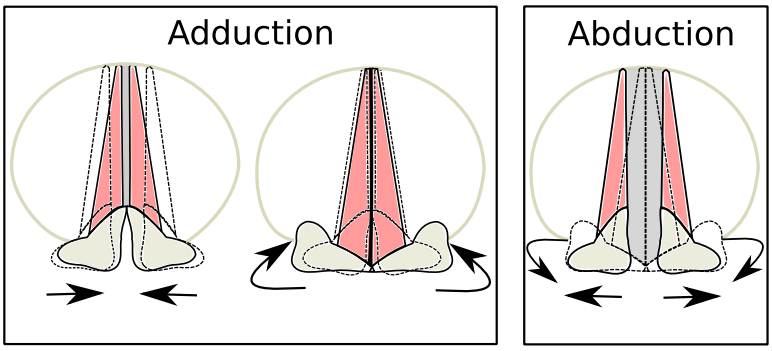


Figure 1.4: Schematic representation of the abduction and adduction of the vocal folds. Adapted from (Henrich, 2001).

a series of intermittent pulses, the sounds source. The first layer is composed of the epithelium and superficial lamina propia. The second layer is the vocal ligament and is composed of the intermediate lamina propia and deep lamina propia. Lastly, in the most inner part, the "body" of the vocal folds is found. The body is composed of thyroarytenoid muscle. At the same time, the thyroarytenoid consists of two distinct parts (Hixon et al., 2008), called the external thyroarytenoid or thyromuscularis and the internal thyroarytenoid or vocalis. The body regulates the resistance to the airflow which affects the vocal folds tension (Kendall and Leonard, 2010). Figure 1.5 depicts a coronal section of the vocal folds histology, demonstrating their structural layers and the subdivision of the thyroarytenoid muscles.

The vocal folds have a length around 13-17 mm for women and 17-24 mm for men and they can elongate 3-4 mm approximately (Titze, 1993). The human vocal folds open and close repeatedly in ranges between a few Hz and hundred times per second in the spoken voice and between *mi1* ($\sim$80 Hz) and *mi5* ($\sim$1320 Hz) in the singing voice (Sundberg, 1996).

In summary, the cartilages support and house the vocal folds; the contraction of the intrinsic laryngeal muscles moves the cartilages relative to one another in order to open and close the glottis; this movement modifies the length and mechanical properties of the vocal folds tissue; and the vibration of the vocal folds hundreds of time per second produces a sound source[3].

1.1.2.3 Physiology of the Laryngeal Apparatus During Phonation

As it was mentioned previously, the intrinsic laryngeal muscles rule the abduction and adduction instants but also determines the length, mass, stiffness, and tension of the vocal folds. These biomechanical parameters have a close relation with the vibratory characteristics of the vocal folds, and thus with the nature of the sounds produced.

Currently, the most accepted theory of the laryngeal vibration is the *Myoelastic-Aerodynamic* theory proposed by Van den Berg (1958). This theory maintains that vocal folds oscillation is determined by an interaction between aerodynamic stresses applied to the free surfaces of the vocal folds and myoelastic restoring forces generated within the tissues. This biomechanical system is self-oscillating. In other words, the frequency of the mechanical vibration is not determined by periodic neural impulses or any other periodic input imposed mechanically or aerodynamically upon the system (Titze, 1980). Despite that the *Myoelastic-Aerodynamic* theory is a closer representation of the motion of the vocal folds, some refinements are proposed constantly in order to make the model capture the complexity of human phonation (Titze, 1993). In the following, a brief description of the *Myoelastic-Aerodynamic* theory is presented.

[3]It is worth to mention that this sound source is considered as "*voiced*" in speech. Meanwhile, "*unvoiced sounds*" are produced with abducted vocal folds (Sundberg, 1996).

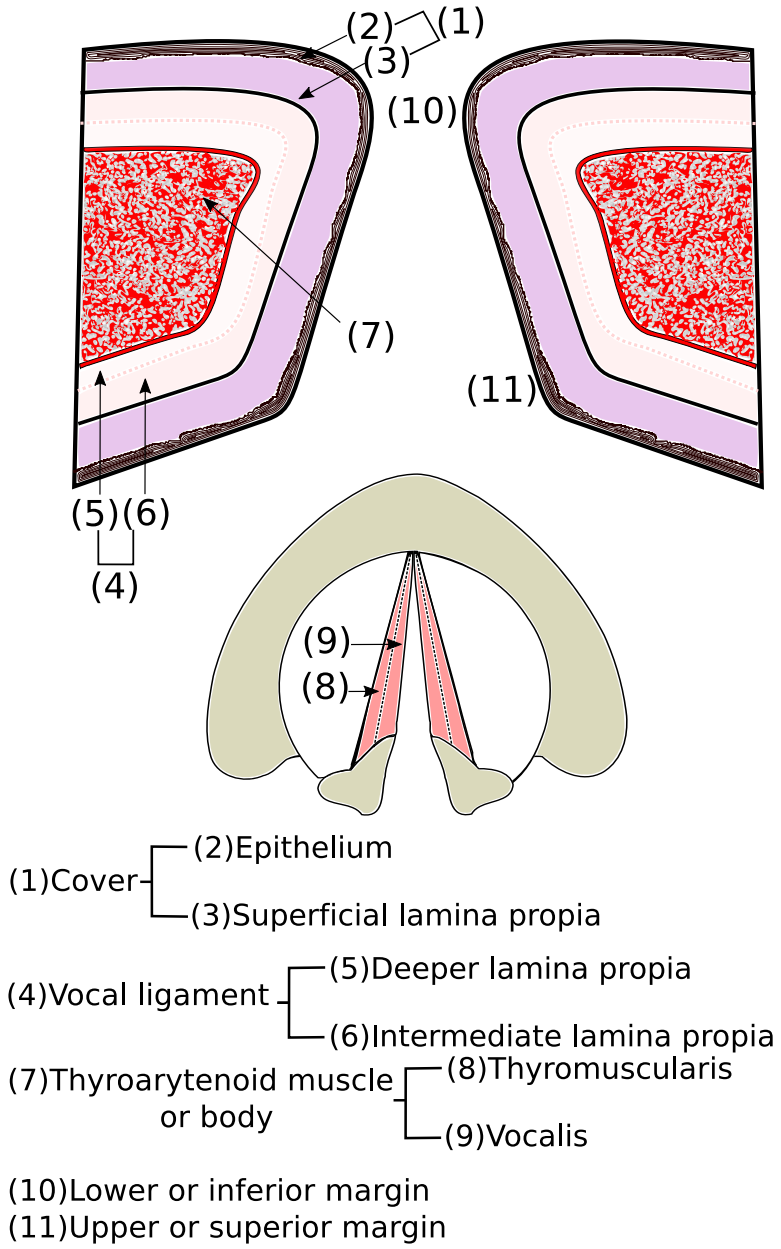


Figure 1.5: Coronal section through the free edge of the vocal folds and thyroarytenoid muscles.

The process begins with inhalation and subsequent glottal closure (Figure 1.6 (1)), then the glottal closure creates a resistance to the pass of air that comes from the lungs. Eventually, this pressure (subglottic pressure) overcomes the closing forces that maintain together the vocal folds and gradually produces the separation of the folds. The lower (or inferior) margins of the folds open first (Figure 1.6 (2))

until reaching the upper (or superior margins) surface of the folds (Figure 1.6 (3)). Once the vocal folds are completely open (Figure 1.6 (4)), the subglottic pressure begins to decrease and the inferior margin of the folds becomes re-approximated due both to their elastic properties, and to an aerodynamic sucking Bernoulli effect (Van den Berg et al., 1957) (Figure 1.6 (5), (6), (7), (8)). The aforementioned process is known as "*vibratory cycle*" and is repeated hundreds of times per second (for instance, the motion of 400 times per second of the vocal folds will produce a sound of fundamental frequency $f_o = 400$ Hz). The generated acoustic waves are propagated to the articulatory system.

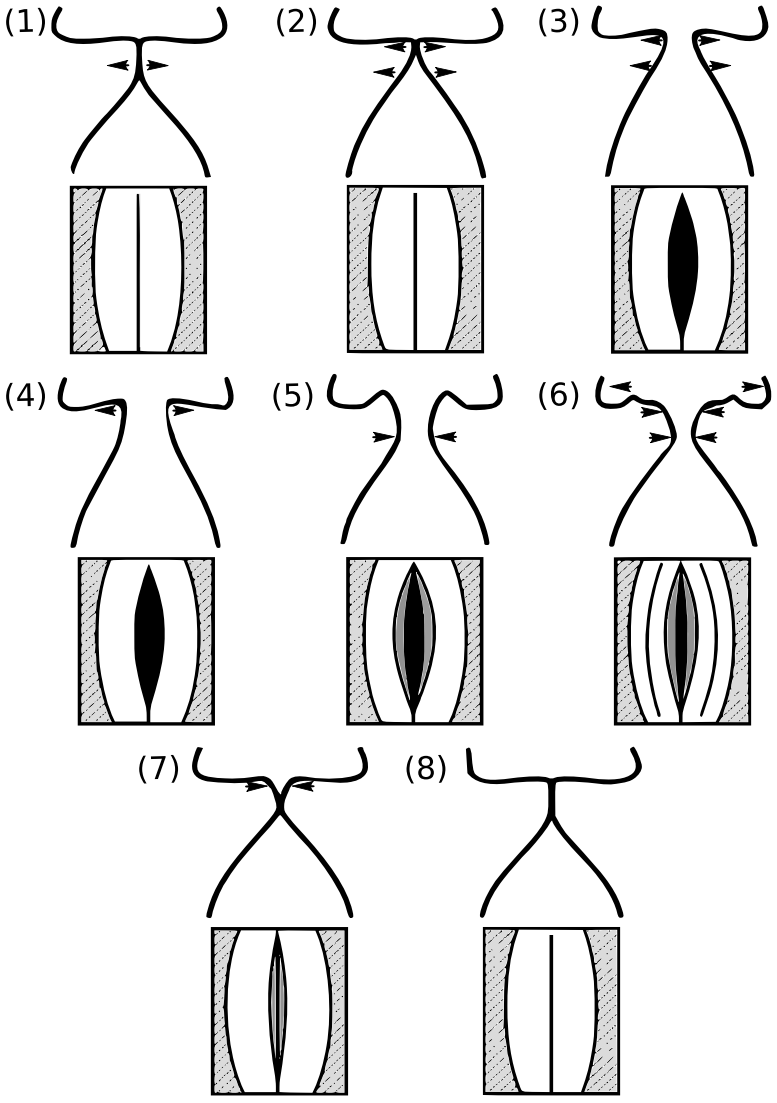


Figure 1.6: Schematic representation of the *Myoelastic-Aerodynamic* theory, with a coronal and superior visualization of the vocal folds.

1.1.3 Articulatory System

The sound waves produced by the phonatory system travel up through the vocal tract where the cavities filter the acoustic signal until it emerges from the lips and/or nostrils and radiates to the external medium (Howard and Murphy, 2008).

The vocal tract is composed of two spaces; the buccal and the nasal cavity. The buccal cavity includes the space between the glottis and the lips, and it can be altered by the motion of the tongue, jaw and lips also known as speech articulators. These articulators alter the speech by varying the height of the jaw, the position of the lips and by changing the shape of the tongue increasing the constriction with the hard palate. Contrariwise, the nasal cavity does not change its shape, so it is only the soft palate which controls the pass or not of sounds through the nose. Figure 1.7 shows the vocal tract with its respective parts.

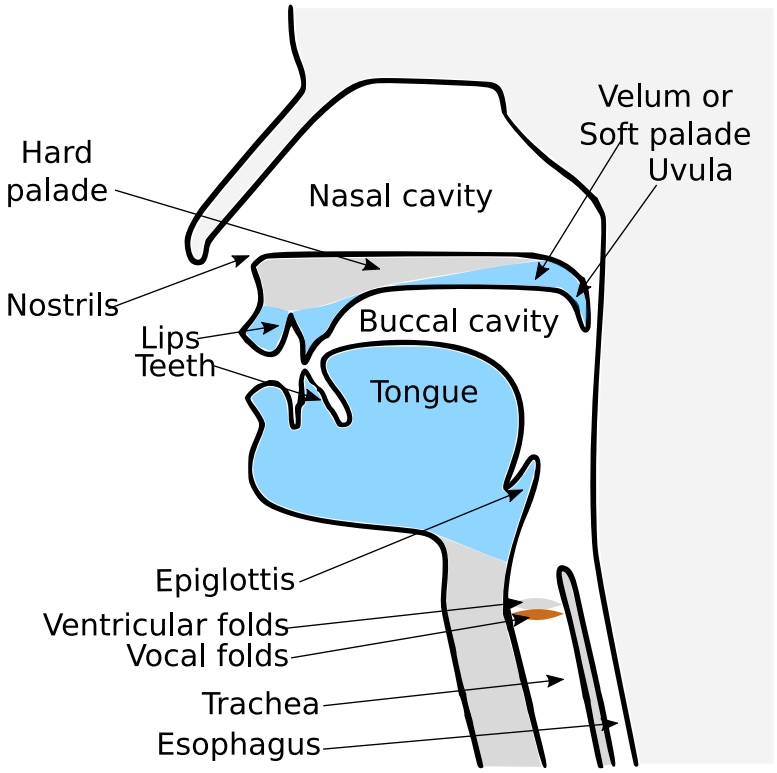


Figure 1.7: Schematic representation of the vocal tract with its differents parts. Adapted from (Gilles, 2010).

1.2 Vocal Registers: Definition and Historical Facts

The term "*vocal registers*" in human phonation arises from the need to explain the discontinuities or transition phenomena which occur voluntarily or involun-

tarily during the production of voiced sounds. However, this term has been used ambiguously creating confusion to the readers. For instance, some authors define the vocal registers related to the way the laryngeal source works (physiological), whereas others relate it to the timbre qualities of the sound produced (acoustic) (Henrich, 2006).

The first notion about registers came from the singer perspective and it was based principally on the auditory perception and proprioceptive sensations. With Manual García, a voice teacher interested in the vocal physiology mechanism, the term register took a meaning on the basis of a mechanical principle. García claims that the human voice is composed of three registers: chest, falsetto-head, and counter-bass, and he defined the term register as follows:

"*By the word register, we mean a series of consecutive and homogeneous tones going from low to high, produced by the same mechanical principle, and whose nature differs essentially from another series of tones equally consecutive and homogeneous produced by another mechanical principle. All the tones belonging to the same register are consequently of the same nature, whatever may be the modifications of timbre or of force to which one subjects them*" (García, 1847)[4].

Despite the definition of García highlights interesting facts, it has some shortcomings that are related to the lack of details about the mechanical principles.

Nowadays, the terms used to list the different registers are basted in the literature. Some of them are fry, strohbass, and pulse which are related to the perception of very low frequency. The terms heavy, thick thin and light referred to the aspect of the vocal folds; the terms normal and modal includes the range of fundamental frequencies used in speaking and singing; the term chest and head refer to the vibratory sensation at the level of the chest or head; the terms flageolet, flute, whistle and siffet refer to a high pitch of frequencies produced.

1.3 Laryngeal Vibratory Mechanisms

This work will follow the same definitions as the ones presented by (Roubeau, 1993; Henrich, 2001; Henrich et al., 2003; Henrich, 2006; Roubeau et al., 2009) in their seminars works. They characterize the different registers based on an acoustic and physiological point of view. They use the concept of *laryngeal vibratory mechanisms* to define the different configurations of the glottal vibrator that allow the production of the entire frequency range of the human voice. These mechanisms are classified based on the vibration or not of the vocalis muscle. They go from low to high and numbered from zero to three (M0, M1, M2, M3). The frequency ranges produced by two neighboring mechanisms can partially overlap each other, and the sounds produced by one and the same mechanism can present great variations in timbre and intensity (Roubeau et al., 2009). The modification of timbre and the proprioceptive sensations with which they are associated contribute to the deter-

[4]Translation of the García original paper extrated from (Henrich, 2006).

mination of the registers. Table 1.1 summarizes the classification of the different registers based on the laryngeal mechanism involved and Figure 1.8 illustrates the spectral analysis of an ascending glissando[5].

Mechanism M0	**Mechanism M1**	**Mechanism M2**	**Mechanism M3**
Fry	Modal	Falsetto	Whistle
Pulse	Normal	Head (W)	Flageolet
Strohbass	Chest	Loft	Flute
Voix de contrebasse	Heavy	Light	Sifflet
	Thick	Thin	
	Voix mixte (M)	Voix mixte (W)	
	Mixed (M)	Mixed (W)	
	Voce finta (M)		
	Head operatic (M)		

Abbreviations: M, men; W, women

Table 1.1: Classification of registers depending on the laryngeal mechanisms involved (table extracted and modified from (Roubeau et al., 2009)).

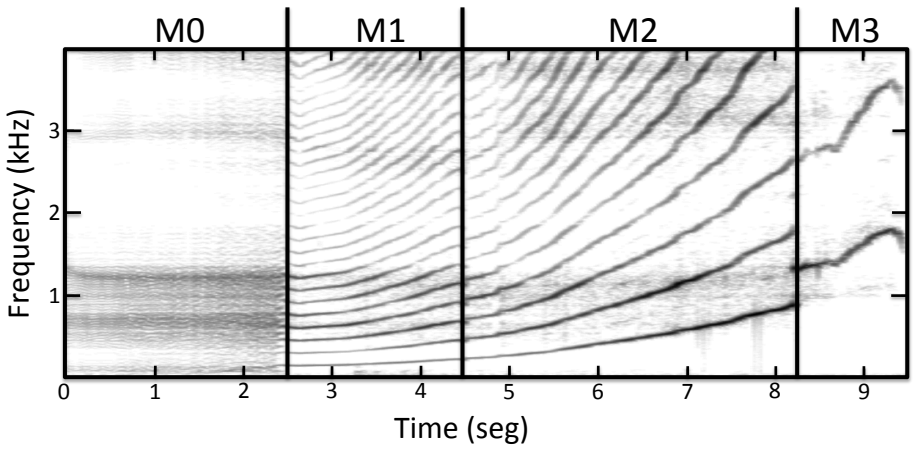


Figure 1.8: Illustration of the time frequency analysis of the four laryngeal vibratory mechanisms during the production of an ascending glissando sung by a soprano. Adapted from (Henrich, 2006).

[5]An ascending glissando is a vocal production during which the frequency progressively goes from the lowest pitch (sometimes around 20 Hz) to the highest (in some cases up to 1000, even 1500 Hz) in the vocal range (Roubeau et al., 2009).

1.3.1 Mechanism M0

In the Mechanism M0, the vocal folds are very short, thick and lax. Laryngeal Mechanism M0 is characterized by a long closed phase and a very short open phase (Roubeau et al., 2009). It is commonly observed in spoken voice and relatively little in singing (occidental lyric singing) at low frequency. It is worth to mention that there is no overlapping between mechanism M0 and M1, except in rare cases of male voices, and it can be found in men and women and in singers and non-singers.

1.3.2 Mechanism M1

In the Mechanism M1, the vocal folds are thick and they vibrate over their whole length with a vertical phase difference. Vocalis muscle participates in the vocal-folds motion and the vocal-folds microstructures are coupled with it. The vocalis muscle activity is dominant over cricothyroid and both activities increase with the pitch (Henrich, 2006). Closed-state is often longer than open-state, and this laryngeal mechanism is used by both males and females in the low to mid part of their frequency range. Figure 1.9 depicts the glottal configuration associated with the mechanism M1.

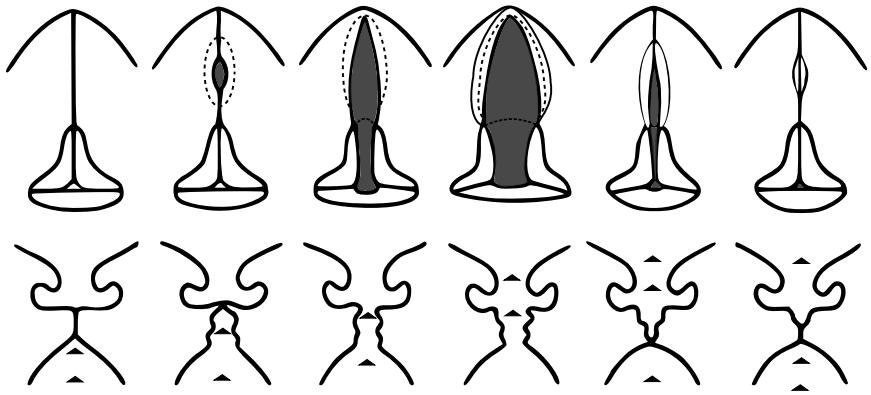


Figure 1.9: Illustration of the superior and coronal glottal configuration associated with the mechanism M1. Adapted from (Henrich, 2001).

1.3.3 Mechanism M2

In the Mechanism M2, all the vocal folds layers are stretched and the vocalis muscle does not participate any longer in the vocal-folds motion. The vocal-folds microstructures are decoupled. The open state is always longer than the closed state, lasting at least 50% of the fundamental period. There is not vertical phase difference in the glottal vibratory movement and this laryngeal mechanism is used by both males and females in the mid to high part of their frequency ranges (Roubeau

et al., 2009). Figure 1.10 depicts the glottal configuration associated with the mechanism M2.

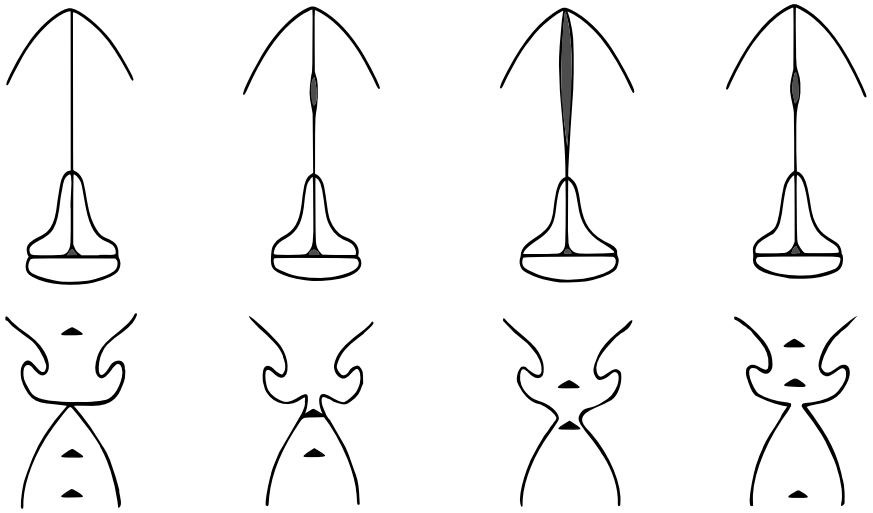


Figure 1.10: Illustration of the superior and coronal glottal configuration associated with the mechanism M2. Adapted from (Henrich, 2001).

1.3.4 Mechanism M3

In the Mechanism M3, the vocal folds are thin and very tensed (Roubeau et al., 2009). The vibratory amplitude is much reduced in comparison with mechanism M2. The highest frequencies and registers as whistle, flute or flagelot, are found in this mechanism.

1.4 Discussion

The anatomical and physiological characteristics of the phonatory system have been reviewed. The combined interaction between the air pressure system and the phonatory system is the starting point in the voice production. This interaction originates the acoustic waves that will posteriorly propagate to the articulatory system producing the voice sounds.

In the literature, different configurations of the glottal vibrator have been defined. However, each of them uses different terminology and different approaches creating a great confusion to the readers. In order to avoid this problem, the terminology presented in the seminal works (Roubeau, 1993; Henrich, 2001; Henrich et al., 2003; Roubeau et al., 2009) is followed. Its choice is not arbitrary. Contrariwise, it has been chosen because up to now is the one that better explains the laryngeal vibratory pattern based on acoustic and physiologic points of view.

Nowadays, the *Myoelastic-aerodynamic* theory proposed by Van den Berg (1958) presents the best explanation of the origin of the acoustic waves at the larynx level. Therefore, special attention has paid to develop different laryngeal imaging techniques to study and visualize the vocal folds dynamics during voice production.

Chapter 2

Laryngeal Imaging

"An image... is a message without code"

Anonym

SUMMARY: This chapter addresses the most important aspects of the vibratory behavior of the vocal folds from an image-based point of view. First, a detailed description of the laryngeal imaging techniques is presented, pointing out their respective advantages and limitations. Later on, the fascinating vibratory behaviour of the vocal folds which has been a subject of great interest along the years is studied. The understanding of this behavior is the basis to distinguish between healthy and pathological vibratory patterns. Lastly, a review of the clinical applications of the laryngeal imaging until now is presented.

2.1 Laryngeal Imaging Notation and Terminology

Before starting with the description of the laryngeal imaging techniques, some concepts and notations are reviewed. A special emphasis is made on the imaging notation since this will be the basis for future formulation along the work.

2.1.1 Imaging Notation

Let us denote a video sequence as $I(\mathbf{x},t)$, where $\mathbf{x} = (x,y) \in \mathbb{R}^2$ represents the position of the pixels and t represents the time instants of the sequence. Hence, a single frame at instant t_k, $k = \{1,2,...,N\}$, is denoted as $I(\mathbf{x},t_k)$. Therefore, the intensity of a given pixel $\mathbf{x}_{ij} = (x_i,y_j)$, $i = \{1,2,...,n\}$ and $j = \{1,2,...,m\}$, at time t_k can be defined as $I(\mathbf{x}_{ij},t_k) \in \mathbb{R}$, which represents a pixel in gray-scale.

For the case of a color image sequence, the components of the color space[1] are represented by superscripts. For instance, in the RGB space, the image sequence is denoted as $I^{R,G,B}(\mathbf{x},t)$, where R, G, and B are the 3 components of the space. Following the same criterion, a frame in time t_k is denoted as $I^{R,G,B}(\mathbf{x},t_k)$ and a given pixel is denoted as $I^{R,G,B}(\mathbf{x}_{ij},t_k) \in \mathbb{R}^3$.

For single images that do not belong to an image sequence, the notation is simplified to $I(\mathbf{x})$ or $I(x,y)$ for gray-scale images and $I^{R,G,B}(\mathbf{x})$ or $I^{R,G,B}(x,y)$ for images in the RGB space. Additionally, it is worth mentioning that the notation of an image is not restricted to use I, but also there will be cases where other symbols are introduced to denote an image. Figure 2.1 summarizes graphically the notation for a gray-scale image sequence.

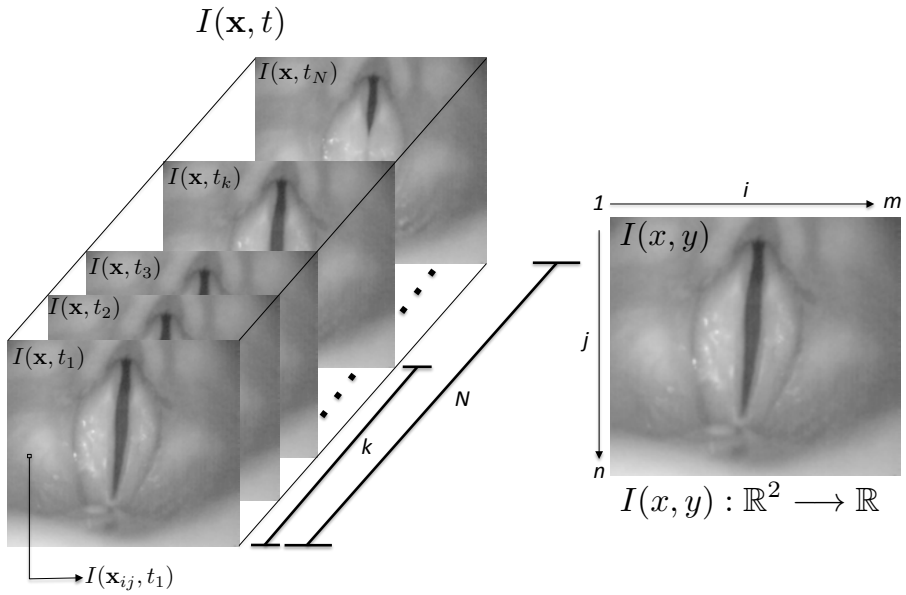


Figure 2.1: Laryngeal image sequence with its respective notation. Left side: laryngeal image sequence $I(\mathbf{x},t)$; right side: single image $I(x,y)$.

2.1.2 Basic Terms and Concepts

In the following, some terms are introduced with their respective graphical interpretation depicted in Figure 2.2:

Glottis, glottal gap or glottal opening: space between the vocal folds during the abduction. Its area at time t_k is denoted as $A_o(t_k) \in \mathbb{R}^+$.

Posterior or dorsal commissure: membranous portion of the vocal folds that is

[1]Color space is a mathematical model which simply describes the range of colors as tuples of numbers, typically as 3 or 4 values (e.g. RGB).

inserted into the vocal process of the arytenoid cartilage which is denoted as $\mathbf{p}(t_k)$ at time t_k.

Anterior or ventral commissure: membranous portion of the vocal folds that is inserted in the midline of the thryroid cartilage which is denoted as $\mathbf{a}(t_k)$ at time t_k.

False, ventricular or vestibular vocal folds: two laryngeal structures located above and in the vicinity of the vocal folds. These folds have a less differentiated layered structure than the vocal folds. They are composed of a membrane, abundant adipose tissues, and seromucous glands (Bailly et al., 2014).

Glottal main axis or posterior-anterior axis: connection line between $\mathbf{p}(t_k)$ and $\mathbf{a}(t_k)$ at time t_k that divides in two the vocal folds: left and right fold. The glottal main axis $\mathbf{G}(t_k)$ is defined within the image plane $I(\mathbf{x},t_k)$ by eq 2.1 as

$$\mathbf{G}(t_k) = [\mathbf{p}(t_k)\ \mathbf{g}_1(t_k)\ \mathbf{g}_2(t_k)\ \cdots\ \mathbf{g}_{pc}(t_k)\ \cdots\ \mathbf{a}(t_k)]^T \subset I(\mathbf{x},t_k) \tag{2.1}$$

it is worth mentioning that the total number of points in $\mathbf{G}(t_k)$ is variable and depends on the position, orientation and size of the glottis which significantly differ between frames of the same image sequence. For that reason, $\mathbf{G}(t_k)$ is often equidistantly sampled, so each point $\mathbf{g}_{pc}(t_k)$ corresponds to a percentage of the total length of $\mathbf{G}(t_k)$ where the subscript $pc \in [0,M]$ indicates the percentage.

Open-state: instant of time where the vocal folds reach the maximum aperture during one glottal cycle.

Closed-state: instant of time where the vocal folds reach the minimum aperture during one glottal cycle.

Glottal cycle or vibratory cycle: subsequence of $I(\mathbf{x},t)$ denoted as $G_{C_o}(\mathbf{x},t)$. The glottal cycle has N_{C_o} frames where $N_{C_o} \leq N$. Figure 2.2 illustrates one glottal cycle with its respective phases.

Opening phase: percentage of time when the vocal folds go from the closed-state to the open-state.

Closing phase: percentage of time when the vocal folds go from the open-state to the closed-state.

Symmetry of vibration: movement of the right and left vocal folds relative to each other. "Normally" both folds vibrate as mirror images of one another which means that they start to open and close at the same time.

Amplitude of vibration: amount of the lateral movement of the vocal folds during vibration.

Periodicity of vibration: relative length of the glottal cycle. This should be stable from cycle to cycle.

Glottal configuration: shape or contour of the glottal opening. Other terms for this feature are: vocal folds contours, glottal contours, glottal edges and vocal folds edges. The vocal folds contours are defined within the image plane $I(\mathbf{x},t_k)$ as $\mathbf{C}^{l,r}(s,t_k) := \mathbf{C}^{l,r}(x(s),y(s),t_k)$ at time t_k, where $s \in [0,1]$ is the parametric domain, l is the left fold and r is the right fold. Both vocal folds edges (l and r) start from $\mathbf{C}^{l,r}(s=0,t_k) = \mathbf{p}(t_k)$ and end at $\mathbf{C}^{l,r}(s=1,t_k) = \mathbf{a}(t_k)$. They also can be expressed by vector notation as:

$$\mathbf{C}^{l}(t_k) = [\mathbf{p}(t_k)\ \mathbf{c}_2^{l}(t_k)\ \mathbf{c}_3^{l}(t_k)\ \cdots\ \mathbf{a}(t_k)]^T \subset I(\mathbf{x},t_k) \tag{2.2}$$

$$\mathbf{C}^{r}(t_k) = [\mathbf{p}(t_k)\ \mathbf{c}_2^{r}(t_k)\ \mathbf{c}_3^{r}(t_k)\ \cdots\ \mathbf{a}(t_k)]^T \subset I(\mathbf{x},t_k) \tag{2.3}$$

where eq. 2.2 and eq. 2.3 are the left and right folds respectively. Then, the set of the N vocal folds edges extracted from the image sequence $I(\mathbf{x},t)$, $\{\mathbf{C}^{l,r}(t_k) \in \mathbb{R}^M,\ k = 1,\cdots,N\}$, is denoted as $C^{l,r}(t)$ (see eq. 2.4).

$$C^{l,r}(t) = \left\{ \begin{bmatrix} \mathbf{p}^{l,r}(t_1) \\ \mathbf{c}_2^{l,r}(t_1) \\ \mathbf{c}_3^{l,r}(t_1) \\ \vdots \\ \mathbf{a}^{l,r}(t_1) \end{bmatrix}, \begin{bmatrix} \mathbf{p}^{l,r}(t_2) \\ \mathbf{c}_2^{l,r}(t_2) \\ \mathbf{c}_3^{l,r}(t_2) \\ \vdots \\ \mathbf{a}^{l,r}(t_2) \end{bmatrix} \begin{matrix} \cdots \\ \cdots \\ \\ \vdots \\ \cdots \end{matrix}, \cdots, \begin{bmatrix} \mathbf{p}^{l,r}(t_N) \\ \mathbf{c}_2^{l,r}(t_N) \\ \mathbf{c}_3^{l,r}(t_N) \\ \vdots \\ \mathbf{a}^{l,r}(t_N) \end{bmatrix} \right\} \tag{2.4}$$

Mucosal Wave (MW): propagation of the epithelium and superficial layer of lamina propia from the inferior to the superior surface of the vocal folds during phonation. The magnitude and symmetry of the mucosal wave are indicators of tension and pliability of the underlying vocal fold tissue and are essential to the production of good voice quality (Shaw and Deliyski, 2008; Voigt et al., 2010b; Krausert et al., 2011). On the upper surface of the vocal folds, the propagation of the mucosal wave is observed as highlighted reflection changes caused by mucosa upheaval moving laterally across the vocal folds surface during phonation (Voigt et al., 2010b). Additionally, the visualization of the MW is sensitive to the frame rate, therefore for achieving full viewing of the mucosal wave features, the frame rate has to be at least 16 times higher than the frequency of vibration (Shaw and Deliyski, 2008).

Voice onset: beginning of the phonation process when the vocal folds start to oscillate.

Voice offset: ending of the phonation process when the vocal folds cease to oscillate.

Open Quotient (OQ): duration of the opening phase divided by the duration of the glottal cycle.

Closed Quotient (CQ): duration of the closing phase divided by the duration of the glottal cycle.

Speed Quotient (SQ): duration of the opening phase divided by the duration of the closing phase.

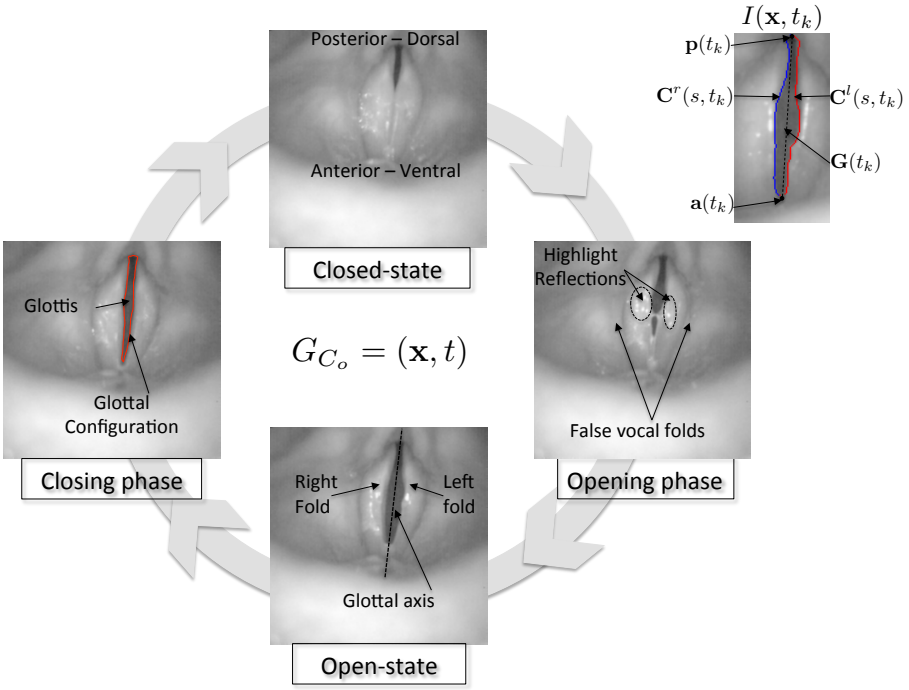


Figure 2.2: Representation of one glottal cycle $G_{C_o}(\mathbf{x}, t)$ with its respective phases: closed-state; opening phase; open-state; closing phase.

2.2 Methods for Direct Observation of Vocal Folds

The need to understand the normal or abnormal anatomy of the laryngeal function, which is the basis of designing treatment strategies, has driven the development of endoscopic laryngeal imaging techniques (Deliyski et al., 2008). Above all, there has been a fascination to understand the vibratory behavior of the vocal folds which has been a subject of great interest in the past. This interest continues today (Woo, 2014). The study of the vibratory behavior of the vocal folds reveals pathological[2] evidences and explains abnormal acoustic manifestations (Yan et al., 2007; Yumoto, 2004). At the same time, it offers to the clinicians one of the best ways to explore the laryngeal functions.

[2]The words pathologic and disorder are used indistinctly around this work to mention all of the behaviors that are not considered as "normal".

The origin of the laryngeal exploration dates back to 1807 when Philipp Bozzini described the first instrument to observe into accessible orifices. This instrument was named Lichtleiter, and it enables the visualization of the lower pharynx and larynx by using artificial light and various mirrors and specula. The Lichtleiter can be considered as the first device to realize an indirect laryngoscopy procedure[3]. Despite the fact that it was not recognized at that time by the scientific community.

It was not until the seminal work of García (García, 1847) during which the use of the indirect laryngoscopy was widespread adopted. Further advances in the clinical use of indirect laryngoscopy came about after the development of instruments and methods for performing transoral laryngeal and airway surgical procedures on awaked patients. Nowadays, the use of indirect laryngoscopy has been almost completely replaced by two techniques that allow to observe and document the complex vibration of the vocal folds with high-resolution and great precision. These two techniques are Laryngeal Videostroboscopy (LVS) and Laryngeal High-Speed Videoendoscopy (LHSV)[4].

2.2.1 Laryngeal Videostroboscopy

During phonation, the vocal folds vibrate at a rate faster than can be perceived by the human eye. Therefore, the use of techniques to create an apparent slow-motion view of the periodic vibratory cycles has been necessary. The process to record an LVS involves the use of a video camera attached to a rigid (transoral) or flexible (transnasal) endoscope where the illumination is provided by a strobe light that flashes at a rate that is synchronized with the patient's fundamental frequency during sustained vowel production. Therefore, the LVS is nothing more than an estimated version of the vibration of the vocal folds that is acquired by sampling its motion.

Figure 2.3 illustrates the complete procedure to record an LVS. (1) and (3) represent the endoscope used to capture the vocal folds motion: rigid (90°) and flexible, respectively; (2) represents the vocal folds; (4) is the real vibratory pattern of the vocal folds; (5) is the stroboscopic light; and (6) is the estimated slow motion version of the vocal folds vibration. Some of the advantages and limitations of Laryngeal Videostroboscopy are mentioned (Mehta and Hillman, 2012a) below:

Advantages

- It can be used with flexible nasofibroscope, providing very good images in color during articulated speech and singing.

[3]Indirect laryngoscopy refers generically to the use of reflected light and images to observe the larynx; this is in contrast with direct laryngoscopy, which is performed under general anesthesia in the operating room (Mehta and Hillman, 2012b).

[4]The terminology used in this work follows the recommendation proposed in (Deliyski et al., 2015a).

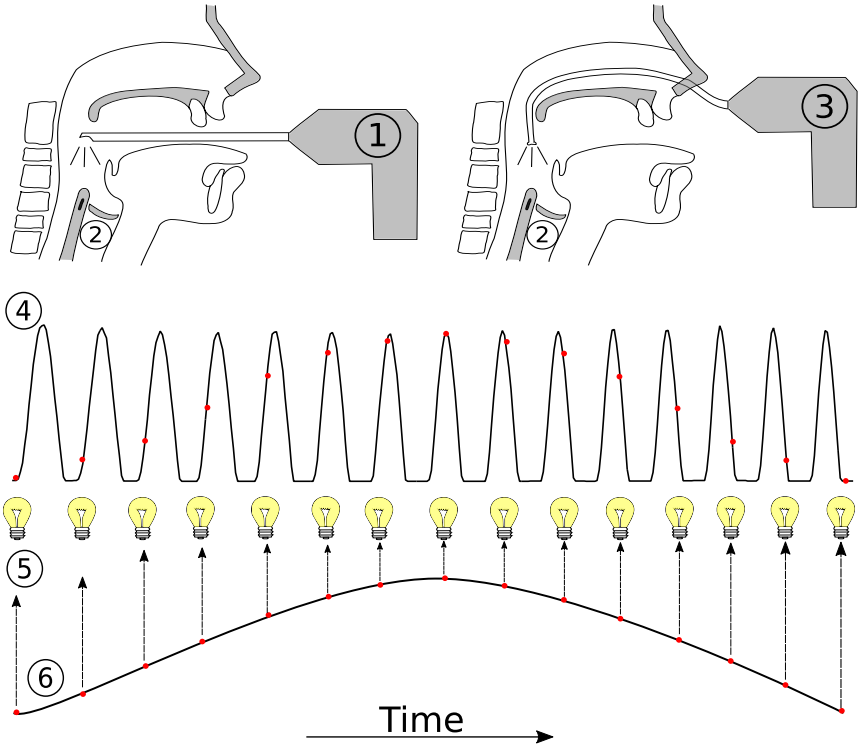


Figure 2.3: Illustration of the stroboscopic sampling. Adapted from (Kendall and Leonard, 2010). (1) and (3) are the rigid and flexible endoscope, respectively; (2) are the vocal folds; (4) represents the real vibratory pattern; (5) illustrates the strobe light; (6) is the estimated version of the vibratory pattern.

- It is possible coupling it with high-definition cameras which provide a higher spatial resolution of vocal folds vibration.

Limitations

- It does not provide a real view of the vocal folds vibratory pattern, so it is restricted to stable and periodic vocal folds vibrations.
- It limits scientific and diagnostic knowledge of the vocal function during voice onset and voice offset.
- It is more sensible to camera rotation, side movement of the laryngoscope and patient movements which produce the delocation of the vocal folds.
- No major technical advancements have been made in recent years regarding LVS.

2.2.2 Laryngeal High-Speed Videoendoscopy

LHSV has revolutionized laryngeal imaging, increasing the understanding of glottal dynamics during the phonation process (Mehta and Hillman, 2012b). LHSV is the only technique capable of acquiring the true intra-cycle vibratory behavior, allowing the study of cycle-to-cycle glottal variations. In LHSV, images are sampled constantly due to the use of a continuous light source (no information loss between frames). Lastly, the image sequence obtained is slowed down to frame rates that can be perceived by human eye.

Nowadays, due to the fast-growth of high-speed technology, it is possible to find cameras that can reach frame rates over "twenty thousand" Frames per Second (fps), recording in color with high spatial resolution and excellent image quality for long durations. With respect to the minimum frame rate requirements of the LHSV for clinical voice assessment, frame rates of 8000 Frames per Second (fps) are recommended with a minimum requirement of 4000 fps (Deliyski et al., 2015b). For LHSV recordings at rates below 4000 fps for women and 3200 fps for men, the videos have to be interpreted with caution.

Figure 2.4 illustrates the principle of sampling in LHSV for two different frame rates. It can be observed that every single cycle is sampled, in contrast to LVS where the samples are taken from different cycles (see Figure 2.3). Some of the advantages and limitations of LHSV are detailed below (Hertegård, 2005; Kendall and Leonard, 2010; Deliyski et al., 2008):

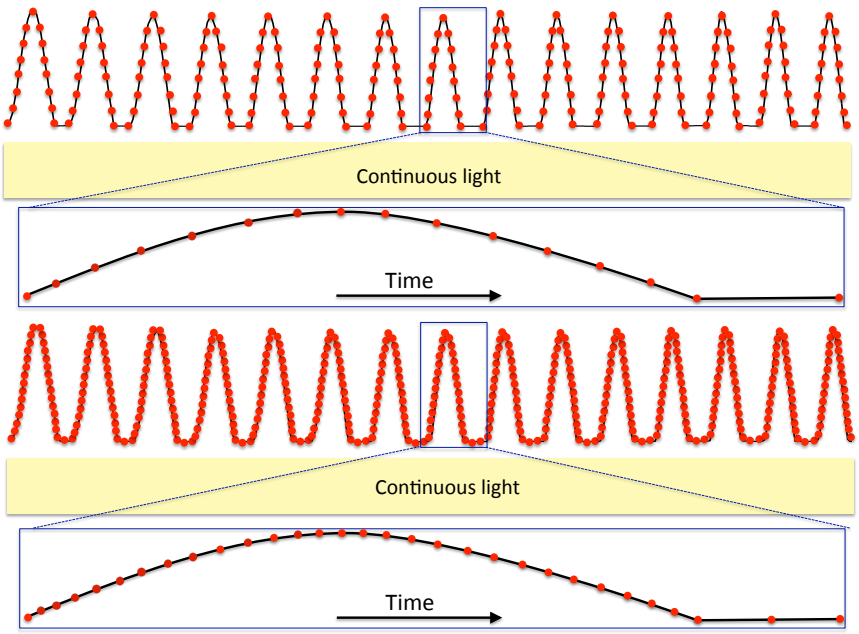


Figure 2.4: Illustrations of the LHSV sampling effect for two different frame rates. Adapted from (Kendall and Leonard, 2010).

Advantages

- It captures the true intra-cycle vibratory behavior of the vocal folds, so aperiodic movements can be visualized.
- It provides a more reliable and accurate objective quantification of the vocal folds vibrations.
- It permits recording and efficiently visualizing transient phonatory events. For instance: phonatory breaks[5], laryngeal spams[6], onset and offset of phonation.
- The combination of LHSV with acoustics and other voice signals may provide complementary, high-precision measures that can improve the clinical practice.
- It is used to examine the basic physiology of different singing styles (Hertegård, 2005).
- It is useful to get insights into tissue vibratory characteristics, the influence of aerodynamical forces and muscular tension, vocal length and evaluation of normal laryngeal functioning in situation of rapid pitch change such as onset and offset of voicing or glides.

Limitations

- It is a high-cost technology which limits its practicality as a clinical tool.
- Due to the huge amount of data acquired, storing and visualization are great problems. For instance, 10 seconds recording data at speed of 10000 fps would require 2 hours and 46 minutes to view the whole recording at a speed of 10 fps.
- It is not possible to provide real-time audiovisual feedback.

2.3 Voice Disorders

A voice disorder occurs when voice quality, pitch, or loudness differ or are inappropriate for an individual's age, gender, cultural background, or geographic location (Titze, 1993; Aronson and Bless, 2009). These disorders can be associated with a deficit in any of the three voice production systems; they can be found at the level of the air pressure system, which is the power source of the phonation, or

[5]Phonatory breaks are short interruptions of the phonatory process.

[6]Laryngeal spams are an uncontrolled/involuntary muscular contraction of the vocal folds which are associated with several voice disorders.

at the phonatory system, which is where the voice is produced, or at the articulatory level, where the vocal projection occurs. This study focusses on the vibration system, specifically at the vocal-folds level.

An abnormal phonation at the vocal folds level is observed as a disruption of one or more of the following vibratory features: periodicity of vibration, vocal folds symmetry, glottal configuration, and mucosal wave (Dejonckere et al., 2001; Kendall and Leonard, 2010). Therefore, it is necessary to establish the criterion of "normality", and categorize as vocal folds disorders all the deviations from those criteria.

2.3.1 Normal Behavior of the Vocal Folds

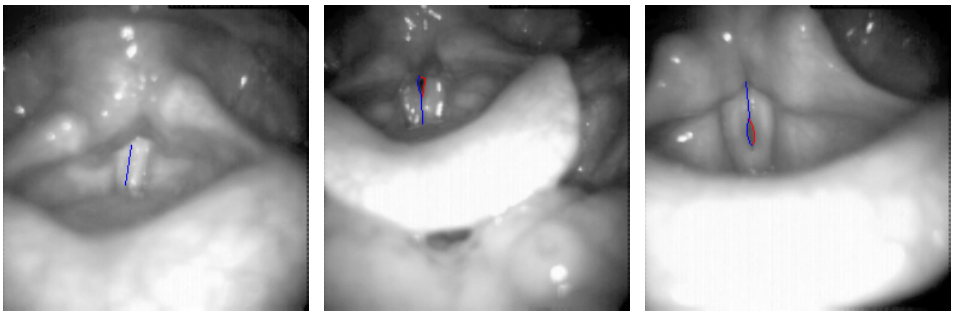
The theoretical idea of a "normal behavior" of the vocal folds can neither be described with a single condition nor with a set of fixed boundaries. Therefore, the following criteria of normality have to be taken carefully, having in mind that some deviations from this "normal behavior" are feasible. It is worth mentioning that the set of criteria have been extracted from different proposals found in the literature along years of research (Kendall and Leonard, 2010; Ahmad et al., 2012; Lohscheller et al., 2013; Manfredi et al., 2006).

A normal periodicity of the vocal folds vibration can be defined as the exact repetition of a spatial-temporal pattern. Thus, irregularity and aperiodicity refer to any changes in this pattern along time. Although irregularity is a feature of many voice disorders, there are minor irregularities in normal voice production that contribute to the human natural sound of voice (Bonilha and Deliyski, 2008). The irregularity of vocal folds vibration is typically the result of an imbalance of the mass or tension between the right and left vocal folds.

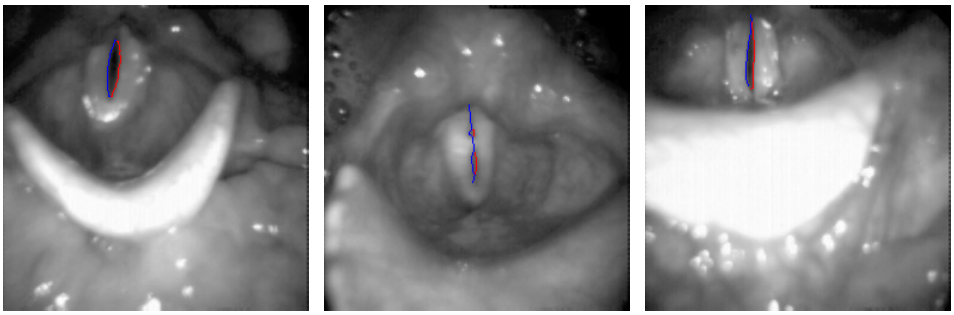
A normal symmetry of vibration of the vocal folds can be defined as the periodic vibration of left and right vocal folds that mirror each other as they oscillate (Mehta et al., 2013). Similarly, significant deviations from such mirrored behavior have been associated with abnormal vibration. The asymmetric movements of the vocal folds imply a lack of equivalent shape, mass, elasticity, and/or viscoelastic properties of the vocal folds. However, there are some minor asymmetric vibrations that are considered as normal voices (Döllinger et al., 2009).

A normal glottal configuration is marked by a complete closure of the vocal folds in the closed-state. Thus, a posterior glottal chink, spindle shape, hourglass configuration, irregular closure, incomplete closure and anterior glottal chink (see Figure 2.5) are considered as abnormal. Glottal configuration is the feature with more variability and is affected by demographic factors as gender and age (Gelfer and Bultemeyer, 1990; Ahmad et al., 2012).

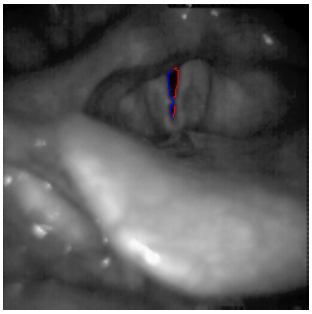
A typical mucosal wave should travel one-half of the width of the superior surface of the vocal folds during normal phonation. A reduced mucosal wave during normal phonation means stiffness. Conversely, a larger than normal mucosal wave signifies flaccidity (Shaw and Deliyski, 2008). The mucosal wave requires



(a) Complete closure (b) Posterior chink (c) Anterior chink



(d) Spindle (e) Irregular closure (f) Incomplete closure



(g) Hourglass closure

Figure 2.5: Common glottal configuration: (a) complete closure; (b) posterior chink; (c) anterior chink; (d) spindle closure; (e) irregular closure; (f) incomplete closure; (g) hourglass.

the relationship of the vocal folds histologic layers to be in balance, hence providing information regarding the underlying layers of the lamina propia and the thyroarytenoid muscle.

Since there is not a general agreement in the classification of the vocal folds disorders, they have been classified indistinctly depending on the author. Nevertheless, when a voice disorder occurs there are two criteria that are commonly used to know the cause of the problem: the disorder has an organic or functional origin.

For that reason, this work will follow such criteria to classify the voice disorders into organic and functional disorders (Godino-Llorente, 2002; Jackson-Menaldi, 2002).

2.3.2 Organic Disorders

Organic disorders are caused by some lesions (physical abnormality), often involving tissues or fluids on the vocal folds. The most common organic disorders with their respective description and illustration (see Figure 2.6) are detailed below:

Nodules (a): vocal folds nodules are symmetric small lesions of the mucosal thickening located halfway between anterior and posterior commissure. The nodules affect the contact of the vocal folds producing an hourglass closure configuration. Additionally, if the nodules are not symmetric, there will be an asymmetric vibration of the vocal folds.

Polyps and cysts (b): vocal folds polyps and cysts are typically single-sided conditions and are mostly located in the anterior part of the vocal folds. From a clinical perspective, polyps and cysts are sometimes very similar: the main difference is that polyps are solid, whereas cysts are fluid-filled structures (Bohr et al., 2014). Concerning the impact on the oscillation, they are an additional mass on the vocal fold. Therefore, they cause changes in the left-right spatial and temporal dynamic symmetries. In addition, cysts and polyps are expected to influence the periodicity of the vibrations and therefore the temporal parameters.

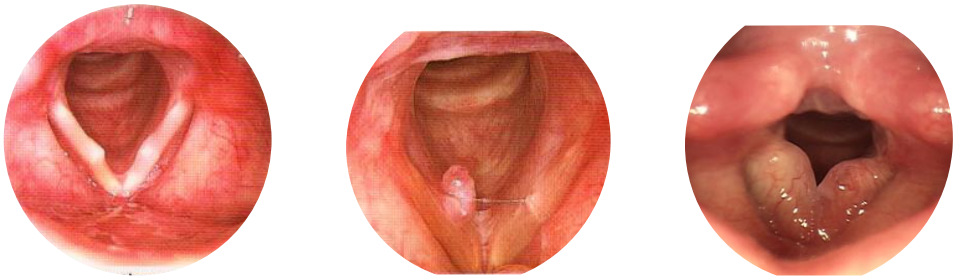
Reinke's edema (c): the Reinke's edema is characterized by the accumulation of fluids directly under the vocal folds epithelium. Concerning the impact on the oscillation, the shape of the glottal configuration during the closed-state is irregular due to the not uniform shape of the folds and the closing state is longer to the opening state affecting the normal vibration of the vocal folds.

Laryngitis (d): the laryngitis is the inflammation of the larynx and is one of the most common conditions. It affects the normal elasticity, viscosity, volume, and tension of the vocal folds. Therefore, it causes irregular and/or reduced vibration decreases the mucosal wave propagation and produces an irregular glottal configuration.

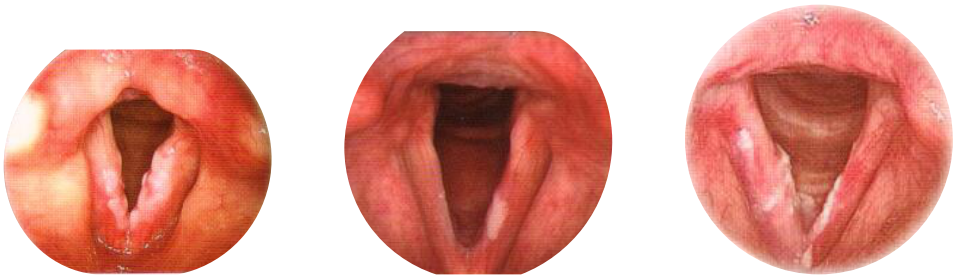
Leukoplakia and erythroplakia (e): they are chronic irritations of the vocal folds observed as a white (leukoplakia) and red (erythroplakia) plaques on the epithelium. The leukoplakia and erythroplakia reduce the mobility of the vocal folds edges, affecting its ability to vibrate.

Carcinoma or laryngeal cancer (f): the carcinoma is produced by the emergence of malignant cells in the tissues of the larynx, originating in most of the cases in the vocal folds (Schultz, 2011). The carcinoma can be observed as a white or red plaques, hence their distinction from premalignant lesions is challenging. Both

have the same aspect, showing irregular or thickened mucosa due to structural changes, affecting the normal vibratory behavior of the vocal folds.



(a) Nodules (b) Polyps (c) Reinke's edema



(d) Laryngitis (e) Leukoplakia (f) Carcinoma

Figure 2.6: Illustration of the Organic disorders. Adapted from (BTP, 2014).

2.3.3 Functionals Disorders

Functional disorders are caused by poor muscle functioning of the vocal folds, resulting in an improper or inefficient use of the vocal mechanism. The functional disorders are characterized by the absence of physical structure lesions and their origin can be either bacterial or by some problem in the nervous system that interacts with the larynx (Cobeta et al., 2013). The most common functional disorders with their respective description are detailed below:

Paralysis and paresis: vocal folds paresis and paralysis result from abnormal nerve input to the laryngeal muscles. Paralysis is the total interruption of nerve impulse, resulting in no movement meanwhile paresis is the partial interruption of nerve impulse, resulting in weak or abnormal motion of laryngeal muscles. Paresis/paralysis are observed as an irregularity on the amplitude of vibration of the vocal folds affecting the symmetry between the left and right folds motion.

Spasmodic dysphonia: in spasmodic dysphonia, the muscles inside the vocal folds experience sudden involuntary movements, known also as spasms, which interfere with the ability of the folds to vibrate and produce voice. The spasmodic dysphonia causes voice breaks and can give the voice a tight, strained quality.

Paradoxical vocal folds movement: the vocal folds behave in a normal fashion almost all of the time but, when an episode occurs, the vocal folds close when they should open, affecting the periodicity of the vocal folds vibration.

Tremors: they are rhythmic, involuntary oscillating movements of the muscles of phonation. It has disabling effect because of fluctuations in the amplitude and fundamental frequency of the voice. Vocal tremor involves not only tremor of the intrinsic muscles of the larynx but also, on occasions, the extrinsic laryngeal, pharyngeal, and palatal muscles, as well as the muscles of the diaphragm, chest wall, and abdomen.

Parkinson disease: it is a neurological disease associated with degenerative lesions of the basal ganglia. The Parkinson disease is observed at the vocal folds level by the reduction of the amplitude of vibration, incomplete vocal closure, vocal tremor, and voice irregularity.

Unlike organic disorders where an appropriate diagnosis can be made based almost solely on a single image of a patient's vocal folds, in the case of functional voice disorders, the diagnostic process is much more complex. This is because the corresponding vocal folds movement can only be diagnosed in the context of overall vibratory behavior, which, to date, is only captured in an adequate manner by LHSV examination. Accordingly, there is significant demand for an objective method to differentiate between functional voice disorders and healthy movement patterns.

2.4 Clinical Applications of the LHSV

The combined use of image-based analysis with acoustic analysis has been studied in the last few years to investigate the correlation between the vocal folds vibratory pattern and the voice quality of the speaker (Yan et al., 2007; Ahmad et al., 2012). In this context, the laryngeal analysis by examination of the physiological vibrational patterns of the vocal folds is an essential approach to understanding the mechanisms of phonation and diagnose voice disorders.

The pioneering works using LHSV date back to 1958 from the seminal works of Timcke, Lenden and Moore (Moore and Von Leden, 1958; Timcke et al., 1958, 1959; Von Leden et al., 1960). They measured the amplitude of vibration of each vocal fold separately and the results were plotted as the variations of the amplitude with respect to time. They used the OQ and SQ parameters to explore the details of laryngeal vibrations during the opening and closing phases. They reported observations of the normal and abnormal vocal folds vibrations during phonation: they found that the changes in air pressure exert a considerable influence on the individual components of the vibratory cycle; they also found that the anatomical configuration of the vocal folds plays a significant role in the production of the vibratory pattern; and they highlighted the importance of the descriptive terms to

measure objectively the vocal folds vibration pattern. It is worth mentioning that the complete procedure was made manually for each frame of the glottal cycle, hence the labor was intensive and is relevant today.

After these seminal works, LHSV has not been mentioned for a long time. It happened again in (Childers et al., 1976), where the development of procedures to extract the glottal waveform and other glottal measurements was reported. The glottal waveform represents the variation of the glottal area along time. This information was used to explicate the exact nature of the vibrational patterns produced by the normal and pathological larynx.

In (Švec and Schutte, 1996) was proposed the use of a high-speed line scanning camera to allow clinicians to observe the vocal folds vibration at a single line, which provided an in-depth and real-time understanding of the vocal folds vibratory function. Subsequent works (Wurzbacher et al., 2006; Lohscheller and Eysholdt, 2008a; Mehta et al., 2013; Lohscheller et al., 2013; Herbst et al., 2014) pointed out the advantages of the high-speed motion of the vocal folds vibration in detecting asymmetries, transients, breaks, opening phase events, closing phase events and irregularities.

Posterior works combine the use of LHSV with biomechanical models to quantify the spatio-temporal vibrations of the vocal folds. In (Döllinger et al., 2002), the two mass model proposed in (Ishizaka and Flanagan, 1972) was used with the LHSV to allow the determination of physiological parameters such as vocal folds tensions and vocal folds masses. The authors in (Tao et al., 2007) use a genetic algorithm to optimize the parameters of the two mass model until the model and the realistic vocal folds had similar dynamic behavior. Then, they extract different parameters including masses, spring constants, and damper constants. In (Schwarz et al., 2008), an automatic optimization procedure was developed for fitting the multi-mass model (Wong et al., 1991) to the observed VF oscillations, with the aim of inferring an approximation of the stiffness and mass distribution along the entire vocal folds. One of the latest works (Pinheiro et al., 2012) uses an optimization method which combines genetic algorithms and a quasi-Newton method to extract some physiological parameters of vocal folds and reproduces some complex behaviors as the ones that occur in different types of pathologies.

The LHSV also have been used to highlight the importance of visualizing the mucosal wave propagation for an accurate diagnosis and optimal treatment of voice disorders. In (Shaw and Deliyski, 2008) was showed the presence of atypical magnitude and symmetry of the mucosal waves in the vocal folds vibration of normal speakers. In (Voigt et al., 2010b) the propagation of the mucosal wave was detected and quantified by combining image processing techniques with physiological knowledge of its lateral movements. The aim of the authors was to replace the subjective assessment of the MW in the clinical environment. The authors in (Krausert et al., 2011) discussed the benefits, the disadvantages, and the clinical applicability of the different mucosal wave measurement techniques. They found the necessity of additional research to broaden the use of the LHSV for an accurate

and objective diagnostic of voice disorders.

Different singing styles have been analyzed with LHSV. For instance: in (Lindestad et al., 2001) the authors studied the mechanism of the bass type of Mongolian throat singing (called Kargyraa). They found that both true vocal folds and false vocal folds vibrate during singing; they also observed that the vibration of the false folds adds subharmonics[7] to the acoustic content. In (Borch et al., 2004) the characteristics of rock singing, also known as distorted singing, were investigated. The authors found some modulations of the vocal folds vibrations by means of periodic or aperiodic motion in the supraglottic mucosa[8] which presumably adds the special expressivity to loud and high tones in rock singing.

The LHSV have been used for clinical voice research purposes. For instance, the applicability of LHSV to diagnose functional voice disorders was demonstrated in (Braunschweig et al., 2008) where the non-stationary activities of the vocal folds during onset were investigated and described by two variables. The first one describes the growth of the vocal folds amplitude during the phonation onset process and the second draws conclusions on voice efficiency with respect to the necessary subglottal pressure and the myoelastic forces. Due to the significant differences in those parameters belonging to the pathological and normal voices, they concluded that it is an objective and stable tool for medical diagnosis. Voigt et al. presented a computer-aided method for automatically and objectively classifying individuals with normal and abnormal vocal folds vibration patterns. First, a set of image processing techniques were employed to visualize the vocal folds dynamics. Later, numerical features were derived, capturing the dynamic behavior and the symmetry of the oscillation of the vocal folds. Lastly, a support vector machine was applied to classify between normal and pathological vibrations. The results indicate that an objective analysis of abnormal vocal folds vibration can be achieved with considerably high accuracy. In (Bohr et al., 2014) a set of parameters were proposed to differentiate between healthy and organic voice disorders in males. The parameters were chosen based on spatio-temporal information of the vocal folds vibration patterns. The spatial parameters provide facts about the opening and closing phase. Meanwhile, the temporal parameters reflect the influence of organic pathologies on the periodicity of vocal folds vibrations. The results obtained suggest that for males, the differences between healthy voices and organic voice disorders may be more pronounced within temporal characteristics that can not be visually detected without LHSV. The authors in (Unger et al., 2015) report a procedure to discriminate between malignant and precancerous lesions by measuring the characteristics of the vocal fold dynamics by means of a computerized analysis of laryngeal high-speed videos. They found that the vocal folds dynamics are significantly affected by the presence of precancerous lesions.

[7] Subharmonics are components of a periodic wave having a frequency that is a submultiple of the fundamental frequency.

[8] Supraglottic mucosa extends along the free edge of the epiglottis and aryepiglottic folds down to the arytenoid cartilages.

LHSV is an active research field with many works published every year. This technique has provided valuable insight into the mechanisms of phonation both in normal and in voice disorders. Additionally, it has proved to be useful in quantifying normal and abnormal glottal vibration patterns.

2.5 Discussion

The laryngeal imaging literature has been reviewed making a special emphasis on the standardization of the concepts and terminology. They have been proved to be useful to understand the mechanism of phonation and to differentiate between healthy and pathological vibratory patterns. The most prominent and accurate technique is the LHSV since it is the sole capable of capturing the true vibratory cycle of the vocal folds motion. For this reason, many innovative works using LHSV have been proposed to understand the fascinating behavior of the vocal folds during phonation and onset. Most of them make use of image processing procedures to compute objective parameters or/and spatial-temporal representations (facilitative playbacks) of the vocal folds vibration. The facilitative playbacks allow the clinicians and the researchers follow the vocal folds dynamics in a more intuitive way by condensing the time-varying information of the LHSV into a few static images, or in an unidimensional temporal sequence.

Despite the progress achieved to describe the vocal folds dynamics using LHSV, it uses in the clinical routine owing to several restrictive factors: LHSV is rather expensive; there are no official guidelines for LHSV footage analysis; the literature is limited by the relatively low or nonexistent correlations among measures of irregularity in vocal folds vibration and acoustic parameters; and the lack of normative parameter values and intervals, which are needed to determine the severity of pathological voice production.

Chapter 3

Facilitative Playback Techniques

"The soul never thinks without a picture"

Aristotle

SUMMARY: Since the pattern of the vocal folds vibration is difficult to evaluate by simply observing the successive frames of video recording, the researchers have introduced the concept of facilitative playbacks to better visualize the features of the vocal folds dynamics. In this chapter, the importance of synthesizing the LHSV information is highlighted. Later on, the most widespread playbacks are presented and divided into two groups based on how the vocal folds motion is assessed.

3.1 Importance of Synthesizing LHSV Information

Using LHSV makes possible to visualize male and female phonation characteristics in most of the clinical scenarios. LHSV characterizes laryngeal tissue dynamics and vocal-folds vibratory features, which are not possible to assess (visualize) using common videoendoscopic and LVS techniques.

LHSV records the motion of the vocal folds at thousands of fps, so their dynamics are difficult to evaluate by simply observing the successive frames recorded. However, with the appropriate image processing techniques, the time-varying data can be synthesized in a few static images, or in an unidimensional temporal sequence. In this way, clinicians or researchers can follow the dynamics of the anatomical features of interest without substituting the rich visual content with scalar numbers. The literature reports some proposals to represent the LHSV information in a more simple way. They are able to objectively identify the presence of organic voice disorders (Bohr et al., 2014), classify functional voice disorders

(Voigt et al., 2010a), vibratory patterns (Lohscheller and Eysholdt, 2008a), discriminate early stage of malignant and precancerous vocal folds lesions (Unger et al., 2015), among others (Lohscheller et al., 2013; Herbst et al., 2014).

These representations improve the quantification accuracy, facilitate the visual perception, and increase the reliability of visual rating while preserving the most relevant characteristics of glottal vibratory patterns. Such representations are named as facilitative playbacks (Deliyski et al., 2008) and, depending on the way they assess the glottal dynamics they can be grouped in local- or global-dynamics playbacks. Table 3.1 presents the main studies carried out to synthesize the vocal folds vibratory patterns.

Author	**Year**	**Playback**	**Dynamics**
Timcke et al.	1958	Glottal Area Waveform (GAW)	Global
Westphal and Childers	1983	Discrete Fourier Transform Analysis (DFTA)	Global
Švec and Schutte	1996	Videokymography (VKG)	Local
Palm et al.	2001	Vibration Profiles (VP)	Global
Neubauer et al.	2001	Empirical Orthogonal Eigenfunctions Analysis (EOF)	Global
Li et al.	2002	Eigenfolds Analysis (EFA)	Global
Zhang et al.	2007	Nonlinear Dynamic Analysis (NDA)	Global
Lohscheller et al.	2007	Vocal Folds Trajectories (VFT)	Local
Yan et al.	2007	Hilbert Transform Analysis (HTA)	Global
Deliyski et al.	2008	Mucosal Wave Kymography (MKG)	Local
Lohscheller and Eysholdt	2008b	Phonovibrogram (PVG)	Global
Sakakibara et al.	2010	Laryngotopography (LGT)	Global
Karakozoglou et al.	2012	Glottovibrogram (GVG)	Global
Unger et al.	2013	Phonovibrographic Wavegram (PVG-wavegram)	Global
Ikuma et al.	2013	Waveform Decomposition Analysis (WDA)	Global
Rahman et al.	2014	Dynamic Time Warping Analysis (DTW)	Global
Chen et al.	2014	Glottaltopogram (GTG)	Global
Herbst et al.	2016	Phasegram Analysis (PGAW)	Global

Table 3.1: Summary of the main studies carried out to synthesize the vocal folds vibratory pattern.

3.2 Local-Dynamics Playbacks

Local-dynamics playbacks analyze the vocal folds behavior along one single line that is computed on a line perpendicular to the main glottal axis. In this category, the most extended playbacks are: Videokymography (VKG), High-Speed Digital Kymograghy (DKG), Vocal Folds Trajectories (VFT), and Mucosal Wave Kymography (MKG). They have been successfully applied to demonstrate the change of glottal dynamics in case of damaged tissues, such as lesions, scars, discoloration of the vocal folds and voice disorders (Deliyski et al., 2008; Švec and Schutte, 2012).

3.2.1 VKG and DKG Playbacks

VKG and DKG synthesize the LHSV in a single image using the same principle, but they differ in how the system delivers the results. Whereas the VKG system delivers the kymographic images directly in real time on a video screen, the DKG exploits software to construct kymographic images from digital high-speed recordings. The VKG and DKG provide a clear visualization of the glottal cycle opening and closing phases, of the mucosal wave traveling across the vocal folds upper surface, and of the displacement of the upper and lower margins of the vocal folds (Švec and Schutte, 1996; Schutte et al., 1998; Švec et al., 2007).

Given a video sequence $I(\mathbf{x},t)$, let us denote a horizontal line at time t_k and position y_j as $\mathbf{KG}(t_k)$, where $\mathbf{KG}(t_k)$ is the row vector $[I(\mathbf{x}_{nj},t_k)\, I(\mathbf{x}_{n-1j},t_k)\, I(\mathbf{x}_{n-2j},t_k) \cdots I(\mathbf{x}_{1j},t_k)]$. Then, the kymographic matrix $I_{DKG}(x,y)$ (or $I^{R,G,B}_{DKG}(x,y)$ for color) is constructed with a set of N vectors $\{\mathbf{KG}(t_k) \in \mathbb{R}^n,\ k = 1,\cdots,N\}$, where each $\mathbf{KG}(t_k)$ is a column vector of $I_{DKG}(x,y)$ (see eq 3.1).

$$I_{DKG}(x,y) = \begin{pmatrix} \overbrace{I(\mathbf{x}_{nj},t_1)}^{\mathbf{KG}(t_1)} & \overbrace{I(\mathbf{x}_{nj},t_2)}^{\mathbf{KG}(t_2)} & \cdots & \overbrace{I(\mathbf{x}_{nj},t_N)}^{\mathbf{KG}(t_N)} \\ I(\mathbf{x}_{n-1j},t_1) & I(\mathbf{x}_{n-1j},t_2) & \cdots & I(\mathbf{x}_{n-1j},t_N) \\ \vdots & \vdots & \ddots & \vdots \\ I(\mathbf{x}_{1j},t_1) & I(\mathbf{x}_{1j},t_2) & \cdots & I(\mathbf{x}_{1j},t_N) \end{pmatrix} \tag{3.1}$$

In order to interpret the kymographic vibratory pattern, Figure 3.1 depicts the schematic view of DKG compared with the traditional displays of the vocal folds oscillations. The first row shows eight phases of a glottal cycle in the frontal section, starting with vocal folds opening and ending with a complete vocal folds closure. The second row presents the same eight phases as viewed from above of the vocal folds using LHSV. The third shows the DKG playback at a position y_j. The kymographic image depicts two cycles of the vocal folds oscillation. The important features observed from the eight phases are: (1) lower margin of the VF starts to open; (2) upper margin of the VF starts to open; (3) lower and upper margins of the VF open; (4) lower margin of VF is maximally open, upper margin of the VF still opens; (5) lower margin of the VF closes and is visible; upper margin

of glottis is maximally open; (6) lower and upper margins of the VF close, mucosal wave propagates on the surface; (7) lower margin of the VF is closed; and (8) upper margin of the VF is closed. Besides the VF, it is possible to observe the motion or none of the ventricular folds (Švec and Šram, 2002).

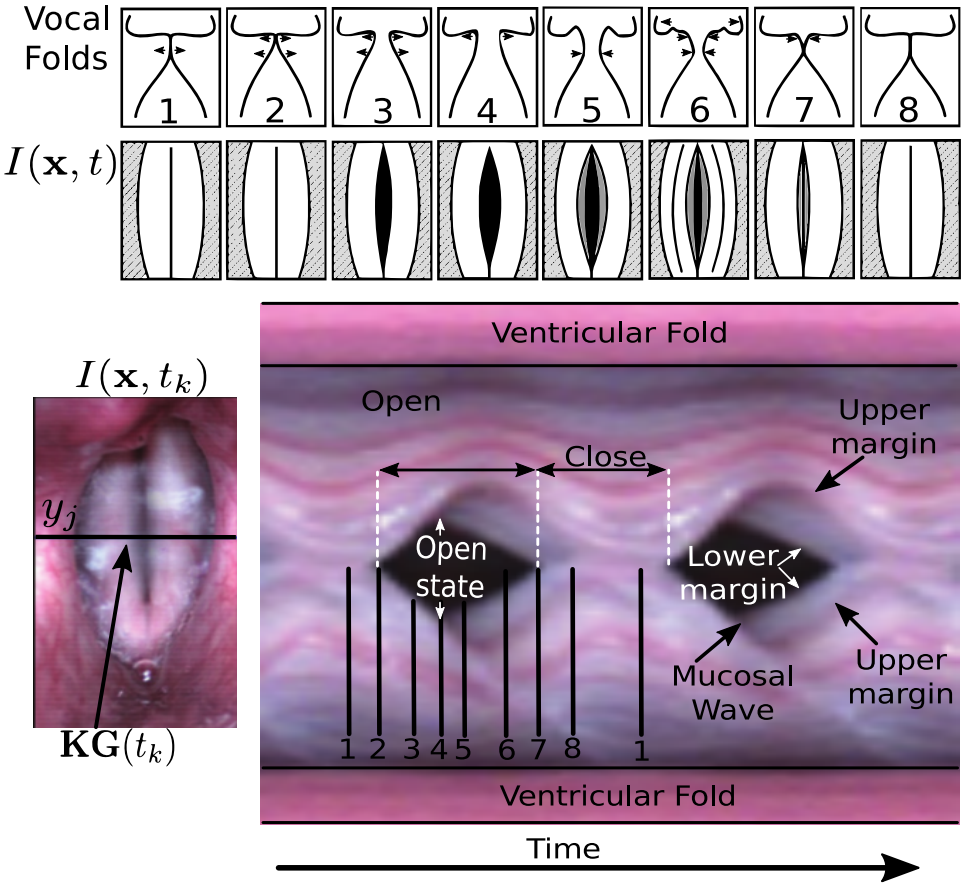


Figure 3.1: Schematical drawing of the successive phases of a glottal cycle in three views. First row: Frontal section of the vocal folds. Second row: Laryngoscopy (superior view of the vocal folds). Third row: High-Speed Digital Kymograghy at the line y_j. Adapted from (Švec and Šram, 2002).

3.2.2 VFT Playback

The VFT $\delta^{l,r}_{seg}(pc,t)$ synthesizes the LHSV in a single image that describes the deflections of the vocal folds edges perpendicular to the glottal main axis (Lohscheller et al., 2007). Hence, the vocal folds edges $C^{l,r}(t)$ have to be computed on advance. Later on, a trajectory line $\mathbf{L}(t_k)$ at time t_k that intersects perpendicularly with $\mathbf{G}(t_k)$ in a predefined point $\mathbf{g}_{pc}(t_k)$ is defined. The current position of $\mathbf{g}_{pc}(t_k)$ is updated every frame to compensate the relative movement of the endoscope, of the larynx,

or of the vocal folds length changes via eq 3.2.

$$\mathbf{g}_{pc}(t_k) = \mathbf{p}(t_k) + (\mathbf{p}(t_k) - \mathbf{a}(t_k)) \left(\frac{pc(\%)}{100\%} \right) \in \mathbf{G}(t_k) \tag{3.2}$$

Later, the intersection between the vocal folds edges $\mathbf{C}^{l,r}(t_k)$ and the trajectory line $\mathbf{L}(t_k)$ is computed by eq 3.3:

$$\mathbf{c}^{l,r}_{pc}(t_k) : \mathbf{c}^{l,r}_{pc}(t_k) \in \mathbf{L}(t_k) \ \wedge \ \mathbf{c}^{l,r}_{pc}(t_k) \in \mathbf{C}^{l,r}(t_k) \tag{3.3}$$

The vocal folds trajectory is obtained by eq 3.4 as:

$$\delta^{l,r}_{seg}(pc,t_k) = \|\mathbf{g}_{pc}(t_k) - \mathbf{c}^{l,r}_{pc}(t_k)\|_2 \ ; \quad k = \{1,2,\ldots,N\} \tag{3.4}$$

where $\delta^{l,r}_{seg}(pc,t_k)$ are the deflections of the vocal folds edges at the point $\mathbf{c}^{l,r}_{pc}(t_k)$ and pc indicates the position of $\mathbf{g}_{pc}(t_k)$ in the glottal main axis. The VFT playback is illustrated in Figure 3.2 and expressed in vector notation at eq 3.5.

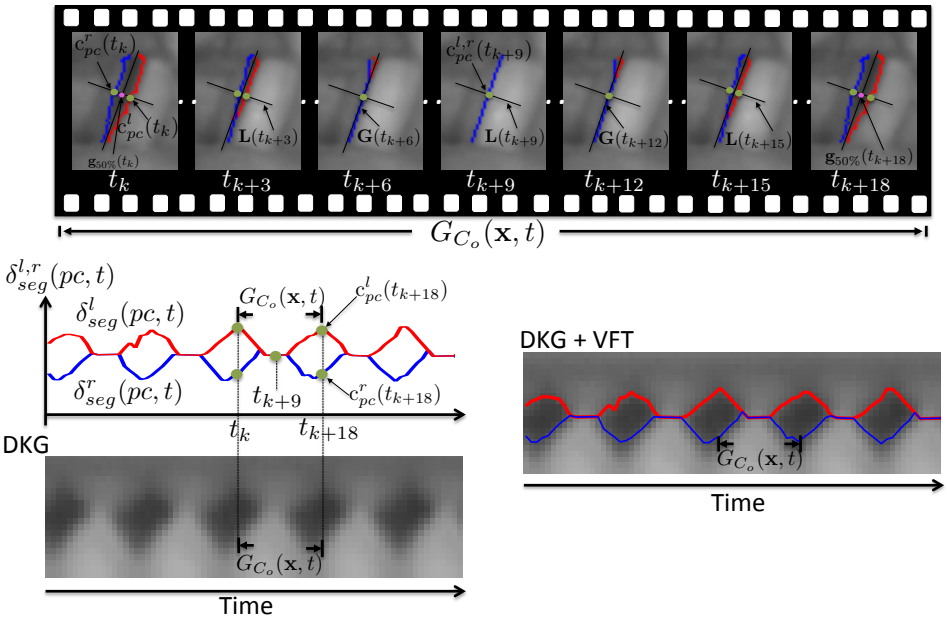


Figure 3.2: Illustration of the VFT playback. First row: the image sequence of one glottal cycle $G_{C_o}(\mathbf{x},t)$. Second row: vocal folds trajectories $\delta^{l,r}_{seg}(pc,t)$, DKG and DKG+VFT playbacks of five glottal cycles.

$$\delta^{l,r}_{seg}(pc,t) = [\delta^{l,r}_{seg}(pc,t_1)\ \delta^{l,r}_{seg}(pc,t_2)\ \cdots\ \delta^{l,r}_{seg}(pc,t_k)\ \cdots\ \delta^{l,r}_{seg}(pc,t_N)] \tag{3.5}$$

The first row in Figure 3.2 represents a particular glottal cycle $G_{C_o}(\mathbf{x},t)$, where $\mathbf{L}(t_k)$ intersects $\mathbf{G}(t_k)$ and $\mathbf{C}^{l,r}(t_k)$ in the points $\mathbf{g}_{50\%}(t_k)$ and $\mathbf{c}^{l,r}_{pc}(t_k)$, respectively.

The vocal folds trajectories describe unambiguously the oscillation pattern of vocal folds vibrations at a specific line, in this example a line located at 50% of the total length of the glottal main axis (pc=50%). It is worth to mention that in those cases where the vocal folds contours cross the glottal main axis, which occurs frequently in asymmetric vocal fold vibrations, the trajectories $\delta_{seg}^{l,r}(pc,t)$ are defined to be negative.

3.2.3 MKG Playback

The MKG is a kymographic image of the mucosal wave along the posterior-anterior axis. The MKG allows a temporal representation of the propagation of the mucosal edges in consecutive glottal cycles during a sustained phonation. The MKG brightness relates to the speed of motion of the mucosal edges, and the color shows the phase of motion (i.e. opening is green and closing is red) (Deliyski et al., 2008; Kendall and Leonard, 2010; Shaw and Deliyski, 2008). Figure 3.3 depicts the MKG playback for an image sequence of four glottal cycles..

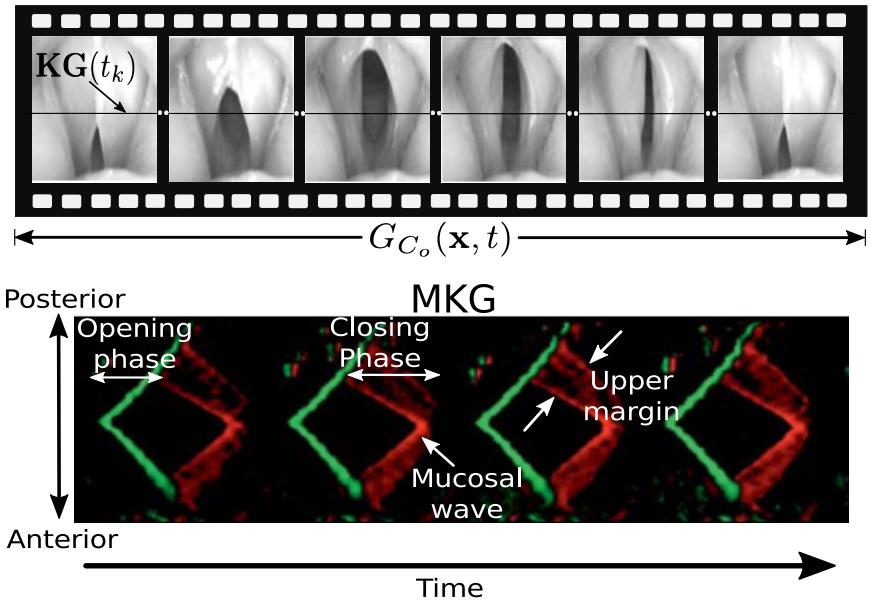


Figure 3.3: Illustration of the MKG playback. First row: six images of a particular glottal cycle G_{C_o}. $\mathbf{KG}(t_k)$ is a horizontal line at time t_k and position y_j. Second row: MKG playback, the green tonalities represent the opening phase and the red tonalities represents the closing phase. Adapted from (Deliyski et al., 2008).

The opening phase is represented with green tonalities and the closing phase with red tonalities. Additionally, the MKG visualizes the upper margin of the vocal folds which appears as a double-edged or thicker curve during the closing phase. The MKG has the potential to assess fine details of the mucosal wave, including the propagation of the mucosal edges during opening and closing phases. However,

there is not detailed explanation in the literature about the techniques used for its implementation.

3.3 Global-Dynamics Playbacks

The global-dynamics playbacks analyse the vocal folds behaviour along the whole glottal length. Most of them are focused on vocal folds edge motion by means of glottal segmentation algorithms. The most widespread and successful playbacks used either by clinicians or researchers are: Glottal Area Waveform (GAW), Phonovibrogram (PVG) and Glottovibrogram (GVG).

For instance, GAW uses the glottal segmentation to compute a glottal gap area function along time from which several parameters can be estimated (Herbst et al., 2014). Contrariwise PVG and GVG playbacks are 2-D representations of vocal folds vibratory pattern as a function of time, for which the movements of the vocal folds edges along the anterior-posterior axis are summarized into a time-varying image line. In comparison to GVG, PVG allows to distinguish left- and right-fold movements, and is thus more sensitive to the accuracy of glottal main-axis (Karakozoglou et al., 2012).

3.3.1 GAW Playback

The GAW was first introduced in a series of articles (Timcke et al., 1958, 1959; Von Leden et al., 1960) where the variation of the glottal gap area at a function of time was explored to understand the normal and pathologic vibratory behaviour of the vocal folds. Let us consider $I_{seg}(\mathbf{x},t)$ as a segmented LHSV, having the same size of the original video $I(\mathbf{x},t)$. The segmented LHSV is a set of binary images, where 1 is assigned to pixels belonging to the glottal gap area (foreground) and 0 is assigned to pixels belonging to the other laryngeal structures (background). Eq 3.6 computes $I(\mathbf{x},t_k)$ in time t_k.

$$I_{seg}(\mathbf{x},t_k) = \begin{cases} 1 & \text{pixels} \in \text{ glottal gap} \\ 0 & \text{background} \end{cases} \tag{3.6}$$

Therefore, the glottal gap area at t_k is computed via eq 3.7 as follows:

$$A_o(t_k) = \sum_{i=1}^{n}\sum_{j=1}^{m} I_{seg}(\mathbf{x}_{ij},t_k)\ ; \quad k = \{1,2,\cdots,N\} \tag{3.7}$$

then GAW can be expressed in vector notation by eq 3.8, where each of its elements represents the area of the glottal gap in a particular instant of time.

$$\mathbf{GAW}(t) = [A_o(t_1)\ A_o(t_2)\ \cdots\ A_o(t_k)\ \cdots\ A_o(t_N)] \tag{3.8}$$

the GAW playback measures the glottal area function throughout the glottal cycle, being possible to compute some features as: opening and closing phase of

the vocal folds oscillations, maximum and minimum glottal areas, open quotient, closed quotient and speed quotient, among others. Figure 3.4 illustrates a GAW normalized within the interval [0,1]. The peaks represent the open-states of the vibratory cycles meanwhile the valleys represent the closed-states of the vibratory cycles. The maximum and minimum amplitudes of the whole vibratory cycles can be computed by finding the maximum and minimum glottal area respectively. The period of the GAW playback is equivalent to the duration of the glottal cycles, and also to the sum of the opening and closing phase duration.

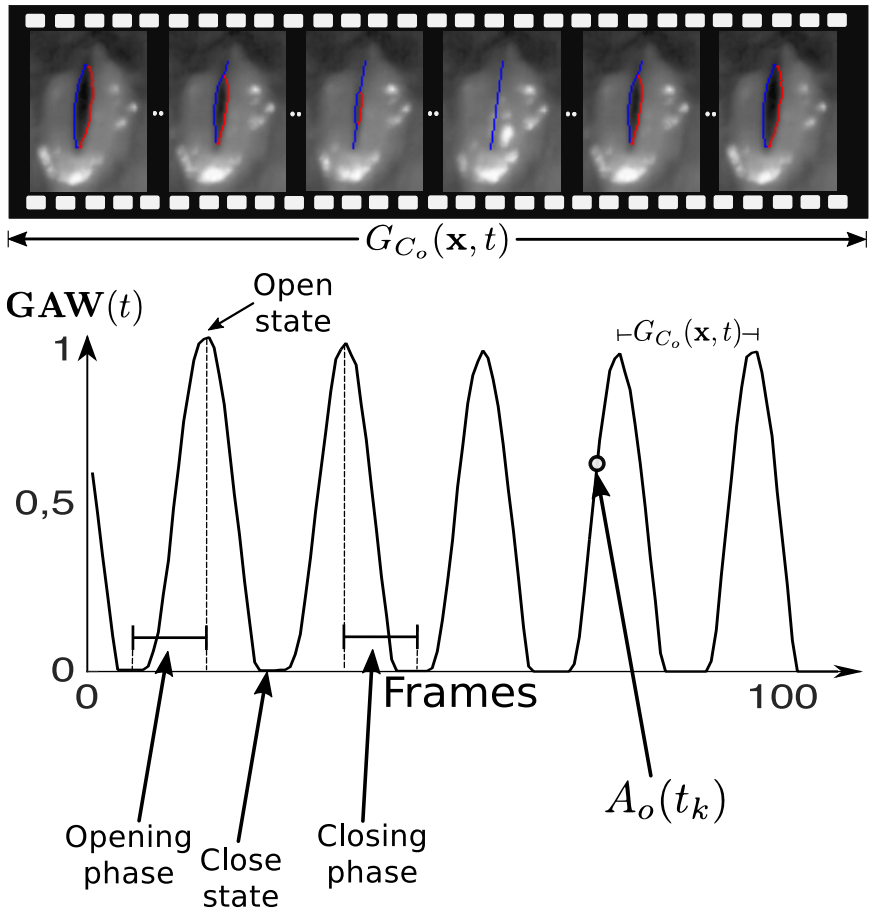


Figure 3.4: Illustration of the GAW playback. First row: six images of a particular glottal cycle G_{C_o}. Second row: GAW playback normalised within the interval [0,1] where 0 represents the minimum area and 1 the maximum area.

3.3.2 PVG Playbacks

The PVG $I_{pvg}(x,y)$ is a further development of spatiotemporal plots of vocal folds vibrations presented in (Neubauer et al., 2001) and of the glottal shape representation proposed by (Westphal and Childers, 1983). PVG is a 2-D diagram introduced

in (Lohscheller and Eysholdt, 2008b) where a set of segmented contours of the moving vocal folds are unambiguously transformed into a set of geometric objects that represents the entire LHSV sequence.

Let us consider that the video sequence $I_{seg}(\mathbf{x},t)$ was computed on advance by any segmentation algorithm. Then, the set of frames with the maximal glottal gap area are identified and named as keyframes $I_{key}(\mathbf{x},t)$ (eq 3.9).

$$I_{key}(\mathbf{x},t) = \underset{k=1\ldots N}{\arg\max} I(\mathbf{x},t_k) \tag{3.9}$$

For each keyframe, $I_{key}(\mathbf{x},t)$, a linear regression line is computed to identify the main orientation of the glottal gap area. The regression line intersects with $\mathbf{C}^{l,r}(t_k)$ at the points $\mathbf{p}(t_k)$ and $\mathbf{a}(t_k)$. Such points are used to split the vocal folds edges into the left $\mathbf{C}^{l}(t_k)$ and right folds $\mathbf{C}^{r}(t_k)$.

Since the vocal folds contours were computed independently, it is necessary to derive a continuous representation of the vocal folds vibrations that links the posterior and anterior point of all images within the LHSV sequence. For doing this, it is assumed that in a single oscillation cycle the positions of the posterior and anterior points for all the intermediate images between the occurrences of two consecutive keyframes do not change dramatically. Therefore, such points are computed approximately by linear interpolation via eq 3.10 where t_O and t_{O+1} indicate two consecutive open-states.

$$\begin{aligned} \mathbf{p}(t_k) &= \mathbf{p}(t_O) + \tfrac{\mathbf{p}(t_{O+1})-\mathbf{p}(t_O)}{t_{O+1}-t_O} \cdot (t_k - t_O)\,; \quad t_O < t_k < t_{O+1} \\ \mathbf{a}(t_k) &= \mathbf{a}(t_O) + \tfrac{\mathbf{a}(t_{O+1})-\mathbf{a}(t_O)}{t_{O+1}-t_O} \cdot (t_k - t_O)\,; \quad t_O < t_k < t_{O+1} \end{aligned} \tag{3.10}$$

By connecting the vocal folds edges $\mathbf{C}^{l,r}(t_k)$ to the approximated position of $\mathbf{p}(t_k)$ and $\mathbf{a}(t_k)$, a continuous representation of the vocal folds edges is obtained also in the parts that are undetected from the segmentation methods (Figure 3.5).

Later, the glottal main axis $\mathbf{G}(t_k)$ and the vocal folds edges $\mathbf{C}^{l,r}(t_k)$ are equidistantly sampled with $pc \in [0,M]$ (Figure 3.6(2)). Then for each image the deflections of the vocal folds edges $\delta_{seg}^{l,r}(pc,t_k)$ are obtained via eq 3.2, $\forall pc \in [0,M]$ and $\forall t \in [1,N]$. $\delta_{seg}^{l,r}(pc,t_k)$ is positive, if the left/right fold contour is correctly located on the ipsilateral side of the glottal main axis. Contrariwise, if the vocal fold edges cross laterally the glottal main axis, $\delta_{seg}^{l,r}(pc,t_k)$ becomes negative. Furthermore, the vocal folds are splitted longitudinally (Figure 3.6(4)) and the left vocal fold is turned 180° around the posterior commissure $\mathbf{p}(t_k)$ (Figure 3.6(5)). Lastly, all the

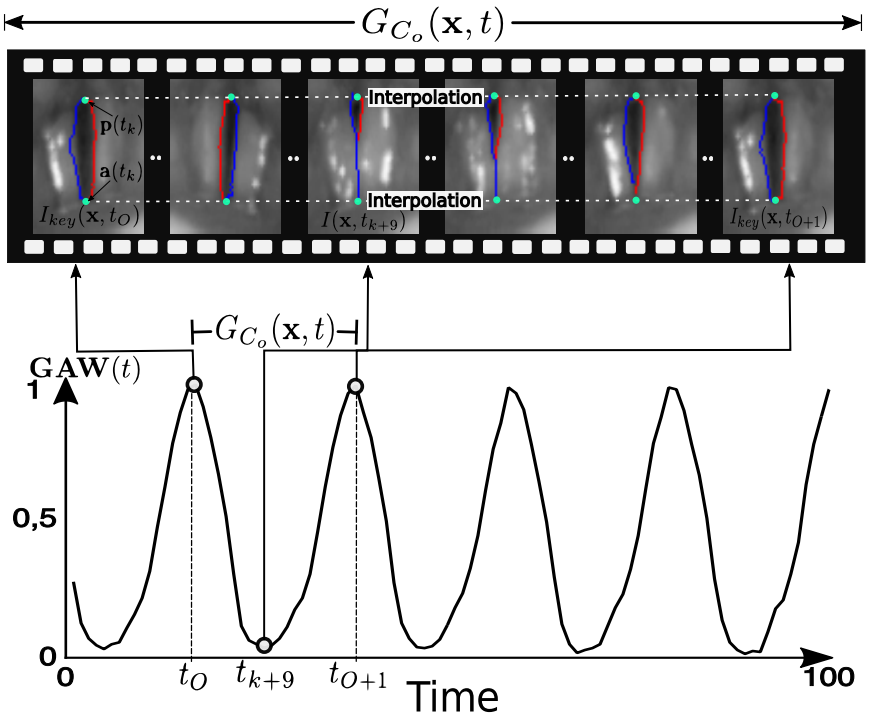


Figure 3.5: Interpolation procedures of the intermediate images within the interval $[t_O, t_{O+1}]$. First row: six images of a particular glottal cycle G_{C_o} explaining the interpolation procedure. Second row: GAW playback normalised within the interval [0,1].

computed $\delta^{l,r}_{seg}(pc,t_k)$ are stored in a matrix $I_{PVG}(x,y) \in \mathbb{R}^{(2M+1)\times N}$ (eq 3.11).

$$I_{PVG}(x,y) = \begin{pmatrix} \delta^l_{seg}(M,t_1) & \delta^l_{seg}(M,t_2) & \cdots & \delta^l_{seg}(M,t_N) \\ \delta^l_{seg}(M-1,t_1) & \delta^l_{seg}(M-1,t_2) & \cdots & \delta^l_{seg}(M-1,t_N) \\ \vdots & \vdots & \ddots & \vdots \\ \delta^l_{seg}(1,t_1) & \delta^l_{seg}(1,t_2) & \cdots & \delta^l_{seg}(1,t_N) \\ \delta^{l,r}_{seg}(0,t_1) & \delta^{l,r}_{seg}(0,t_2) & \cdots & \delta^{l,r}_{seg}(0,t_N) \\ \delta^r_{seg}(1,t_1) & \delta^r_{seg}(1,t_2) & \cdots & \delta^r_{seg}(1,t_N) \\ \vdots & \vdots & \ddots & \vdots \\ \delta^r_{seg}(M,t_1) & \delta^r_{seg}(M,t_2) & \cdots & \delta^r_{seg}(M,t_N) \end{pmatrix} \quad (3.11)$$

In order to visualize $I_{PVG}(x,y)$, each element is represented by color tonalities as shown in Figure 3.6(6) (red represents positive deflections and blue represents negative deflections). The complete procedure to compute PVG is illustrated in Figure 3.6 where 5 glottal cycles are depicted.

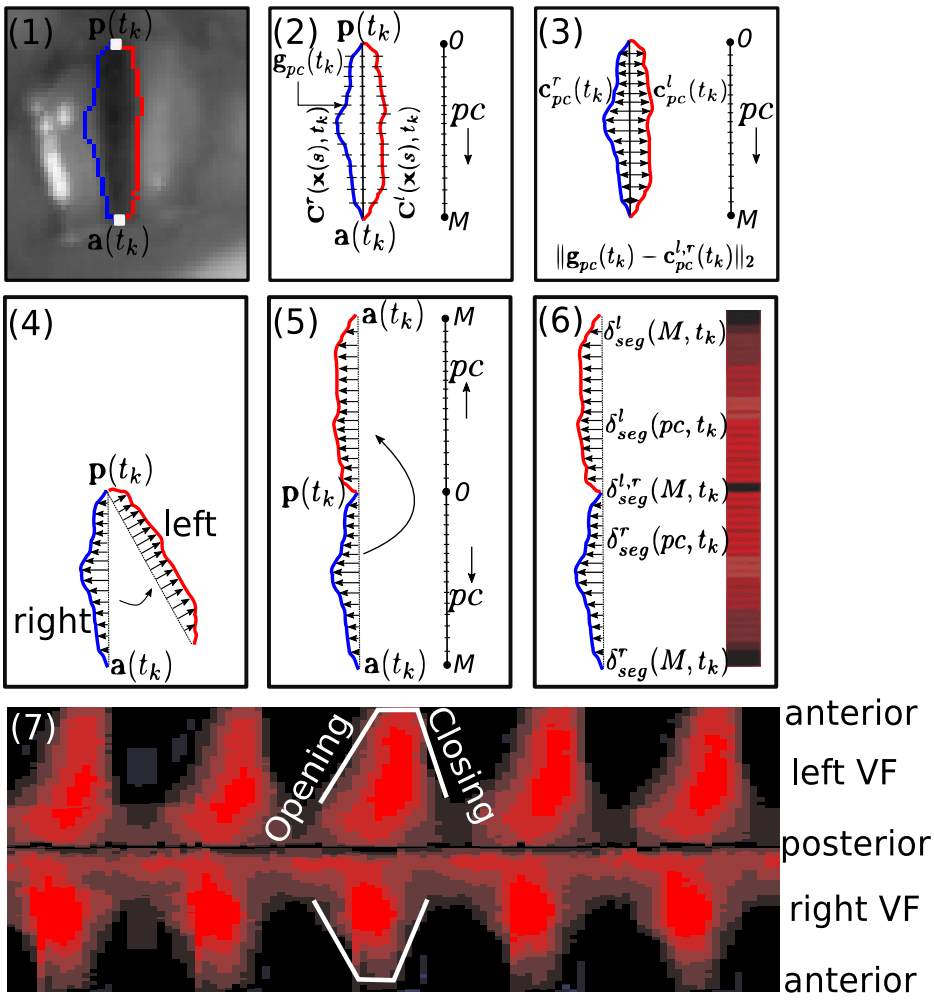


Figure 3.6: Schematic representation of the PVG playback. (1) Segmentation; (2) Resampling of the extracted vocal folds edges; (3) Computation of vocal folds deflections $\delta^{l,r}_{seg}(pc,t_k)$; (4) Splitting of the glottal axis; (5) Virtual turning of the left fold; (6) Color coding of the vocal fold deflections; (7) $I_{PVG}(x,y)$ representing a LHSV with five glottal cycles. Adapted from (Lohscheller and Eysholdt, 2008b).

3.3.3 GVG Playbacks

The GVG playback was proposed in order to solve the difficulties to interpret the PVG and its strongly dependence on the detection of the glottal main axis (Karakozoglou et al., 2012; Döllinger et al., 2011).

The GVG synthesize the LHSV in one single image and its computation uses a similar approach than the PVG formulation. But unlike it, GVG computes the distance between the vocal folds edges themselves. Firstly, the vocal folds edges $C^{l,r}(t)$ are equidistantly sampled with $pc \in [0,M]$. Then, the deflections among

the vocal folds edges are computed by eq 3.12, where $\delta_{GVG}(pc,t_k)$ represents the distance between the left $\mathbf{c}^l_{pc}(tk)$ and right $\mathbf{c}^r_{pc}(t_k)$ fold at position pc and time t_k.

$$\delta_{GVG}(pc,t_k) = \|\mathbf{c}^l_{pc}(t_k) - \mathbf{c}^r_{pc}(t_k)\|_2 \; ; \quad \forall k \text{ and } \forall pc \tag{3.12}$$

Lastly, all the distances are stored in a matrix I_{GVG} (eq 3.13) and normalized within the interval $[0,1]$, with 0 corresponding to zero distance and 1 corresponding to maximal distance. For visualization purposes the matrix is coded with a grayscale map (see Figure 3.7(4)).

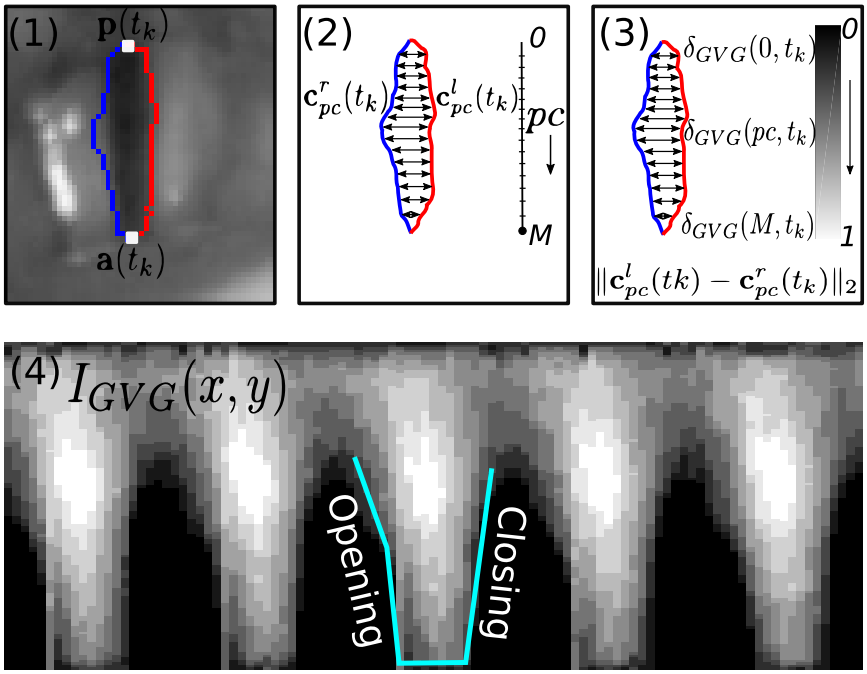


Figure 3.7: Schematic representation of the GVG playback. (1) Segmentation; (2) Resampling of the extracted vocal folds edges; (3) Computation of vocal fold deflections $\delta_{GVG}(pc,t_k)$; (4) $I_{GVG}(x,y)$ representing a LHSV with five glottal cycles.

$$I_{GVG}(x,y) = \begin{pmatrix} \delta_{GVG}(0,t_1) & \delta_{GVG}(0,t_2) & \cdots & \delta_{GVG}(0,t_N) \\ \delta_{GVG}(1,t_1) & \delta_{GVG}(1,t_2) & \cdots & \delta_{GVG}(1,t_N) \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{GVG}(M,t_1) & \delta_{GVG}(M,t_2) & \cdots & \delta_{GVG}(M,t_N) \end{pmatrix} \tag{3.13}$$

The GVG depicts a well-shaped form of the vocal folds vibration even when detection errors occur during the segmentation or glottal main axis detection, providing a more intuitive representation of vocal folds dynamics.

3.4 Discussion

The use of Laryngeal High-Speed Videoendoscopy (LHSV) in combination with image processing techniques is the most promising approach to investigate vocal folds vibration and laryngeal dynamics in speech and singing. The playbacks synthesize the time-varying data which permits to visualize hidden features that are not easily observed from the LHSV.

Despite the great advances that have been reached condensing the data coming from LHSV, many of the aforementioned playbacks present some drawbacks that restrict their applicability since they rely on glottal-area segmentation (GAW, GVG, PVG, PVG-wavegram, VFT, VP, EFA and HTA) which is not a trivial task. The motion analysis is focused only on those points belonging to the glottal contours. Additionally, some of them (GVG, PVG, PVG-wavegram, VFT and VP) rely on the computation of the glottal main axis, which strongly depends on the geometry of the detected glottal area and can be difficult to identify accurately in the presence of a posterior glottal chink. Other Playbacks as: GAW, HTA, NDA and PGAW are based on glottal area waveform computation, so they do not preserve spatial information about vocal folds vibration, limiting their applicability for interpreting spatial vibratory features such as asymmetry. On the other hand playbacks as: VKG, DKG and MKG restrict the information about the dynamics of the vocal folds along one single line. Lastly, GTG and LGT representations are less intuitive to interpret and have not been widely used.

The current challenge is to provide new methods for data visualization to overcome the drawbacks of existing ones, providing simultaneously features that would integrate the time dynamics, such as: velocity, acceleration, instants of maximum and minimum velocity, vocal folds displacements during phonation and motion analysis. These methods should include not only those points belonging to the glottal edges but also those regions that originated such movements. Therefore, a global motion technique capable to synthesize the different patterns that are relevant during the voice production would be desirable.

Part II

State-of-Art in Image Processing and Glottal Segmentation

Chapter 4

Image and Video Processing Techniques

"A picture is worth a thousand words"
Chinese proverb

SUMMARY: In this chapter, a brief review of the basic concepts and definitions related to the most relevant techniques for image and video processing are presented. The examples of the different segmentation and motion estimation algorithms are focused on solving the problem of the glottal gap delimitation, being applied directly to the images in its most basic formulation.

4.1 Review of General Image Segmentation Methods

Image segmentation subdivides an image into its constituent regions or objects and the level of detail of each subdivision depends on the problem being solved (Gonzalez and Woods, 2006).

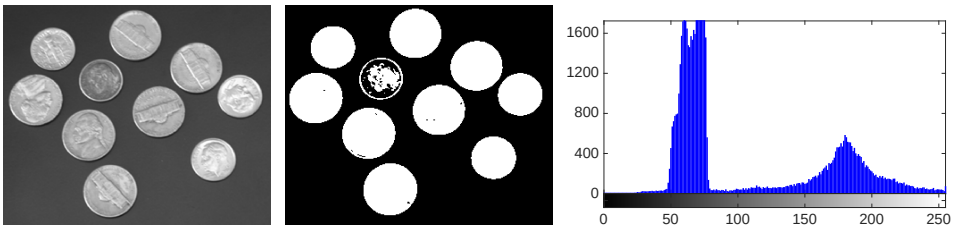
The image segmentation methods can be classified, roughly speaking, into the following categories: *Thresholding, Edge-Based, Region-Based, Classification-Based, Graph-Based and Deformable Models*.

4.1.1 Thresholding

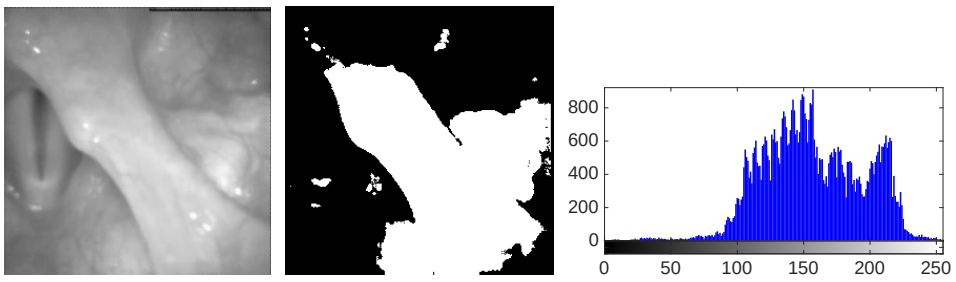
Thresholding is one of the most basic segmentation techniques (Sezgin and Sankur, 2004). The task is to classify the pixels into groups by using a threshold. The pixels with intensity values greater than or equal to the threshold are classified into the first group and the rest of pixels into the second group. The output of the thresholding operation is normally a binary image in which one of the groups

indicates the foreground objects (objects of interest) and the other one corresponds to the background (rest of the image).

The major problem with thresholding is that it considers only the intensity. There is not any relationship between the pixels and there is no guarantee that the pixels identified are contiguous. In addition, it requires the intensity of the image to have a bimodal distribution, which is not common in most images, especially in medical ones. For instance, Figure 4.1 shows two different images with their respective segmentation and intensity distribution. In Figure 4.1a, the intensity has a bimodal distribution. Therefore, the foreground and background can be separated automatically by the Otsu's method (Otsu, 1979). On the other hand, when the intensity distribution is not bimodal, satisfactory segmentations are not possible to obtain (Figure 4.1d).



(a) Original image (b) Segmented Image (c) Histogram



(d) Original image (e) Segmented image (f) Histogram

Figure 4.1: Illustration of the Otsu's thresholding method. First row: coin image thresholded automatically with a value of 126. Second row: laryngeal image thresholded automatically with a value of 163.

4.1.2 Edge-Based

Edges are a set of connected pixels in which there are abrupt changes of intensity. In order to obtain such edges, there are local image processing methods known as edge detectors. In general, edge detectors may be grouped into two categories, gradient and Laplacian (Gonzalez and Woods, 2006).

The gradient is computed by a digital approximation of the derivatives over a neighborhood of a point. There are different discrete differentiation operators pro-

posed to compute such as approximation but the widespread operators in the literature are Roberts, Prewitt, and Sobel. The operators use a pair of 3-by-3 convolution kernels to compute the gradients along the x- and y-directions of the image.

On the other hand, the Laplacian methods search for zero-crossings in the second derivative of the image to find edges. The Laplacian operator is not applied directly to the image since it is sensitive to noise. It is often combined with a Gaussian smoothing kernel, so it is referred to as the Laplacian of a Gaussian (LoG). Then, the edges are obtained by convolving the LoG with the input image.

There are also more advanced techniques proposed in the literature such as Canny (Canny, 1986) and Hough transform (Ballard, 1981). Canny applies a double-thresholding technique to detect strong and weak edges. By using two thresholds, the Canny method performs better in noisy images, and detecting true weak edges is more likely. Meanwhile, Hough transform is a technique which is used to isolate features of a particular shape within an image. It requires the desired features to be specified in some parametric form. However, it is mostly used for the detection of regular curves such as lines, circles, and ellipses.

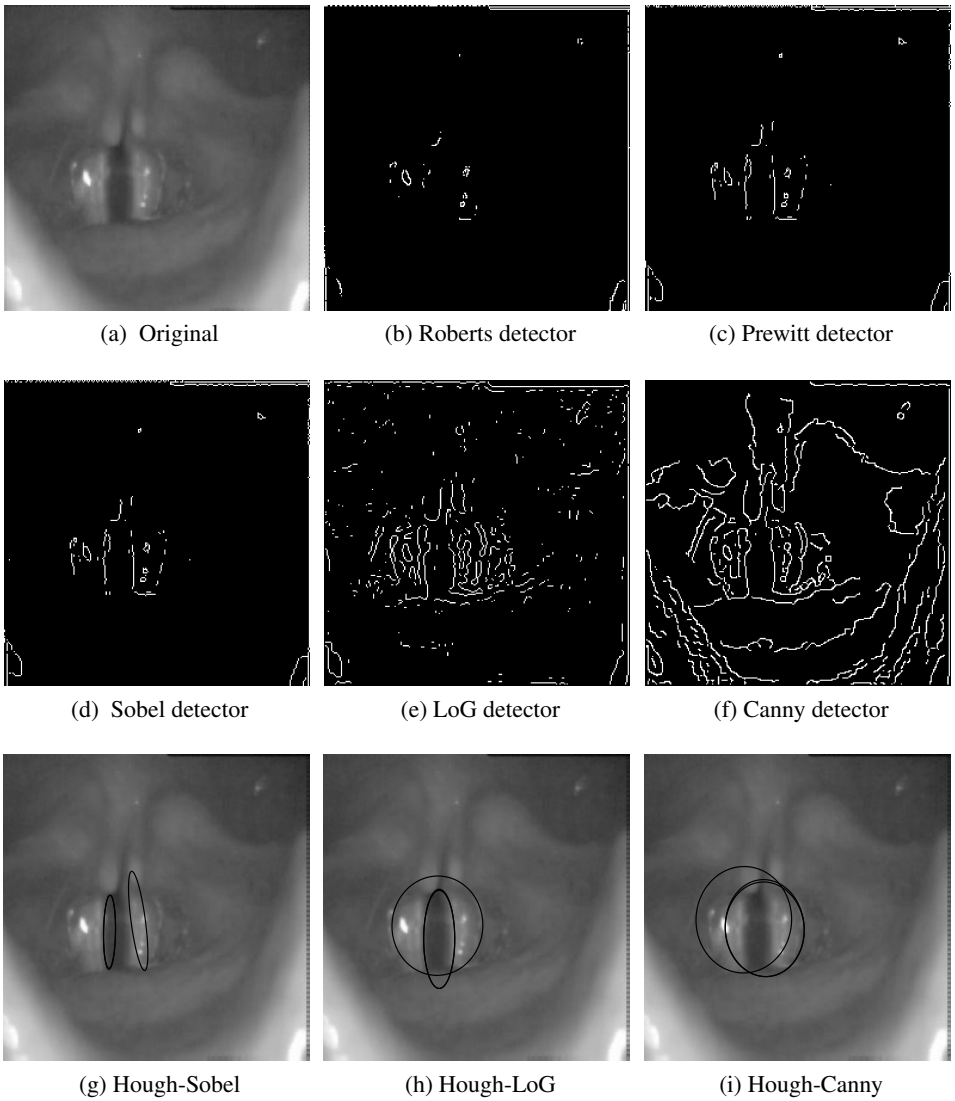
Figure 4.2b, 4.2c, 4.2d, 4.2e, and 4.2f show the results obtained when different edge-based techniques are used to segment a laryngeal image. Meanwhile, Figure 4.2g, 4.2h, 4.2i use edge detector and Hough transform to detect the ellipses with a vertical orientation and major axes between 60 and 90 pixels (Xie and Ji, 2002).

4.1.3 Region-Based

Typically, the Region-Based segmentation algorithm can be divided into two broad categories; Region Growing and Watershed.

- **Region Growing:** The region growing method starts by selecting a set of seed pixels. The selection of the seed pixels can be either manually or automatically and will depend on the nature of the problem. Later, each seed pixel checks its neighbor pixels and adds to its region the neighboring pixels that are satisfying a certain homogeneity criteria, thereby growing the regions (Adams and Bischof, 1994). There are different homogeneity criteria used such as: the difference between the pixel intensity and the mean of the regions (Adams and Bischof, 1994); weighted sum of gradient information and the contrast between the region and the pixel (Xiaohan et al., 1992); and adaptive region growing algorithm that incorporates a homogeneity learning process (Pohle and Toennies, 2001). The region growth should stop when no more pixels satisfy the criterion for inclusion in that region, and the number of regions must be at most the same as the number of seeds planted.

  The issues with this method is that it requires a solid criterion and edges relatively well delimited in order to converge to the region of interest. Furthermore, the algorithm segments objects with inhomogeneous regions into multiple sub-regions, resulting in over-segmentation. Another problem is the



(a) Original (b) Roberts detector (c) Prewitt detector

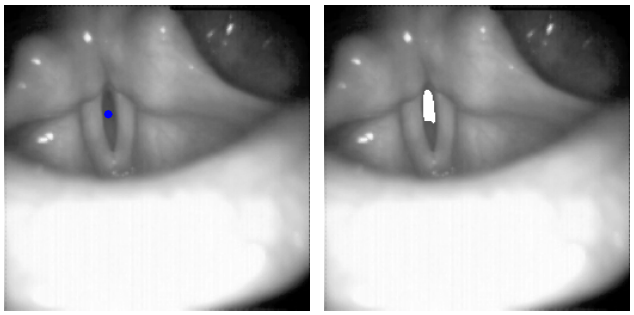
(d) Sobel detector (e) LoG detector (f) Canny detector

(g) Hough-Sobel (h) Hough-LoG (i) Hough-Canny

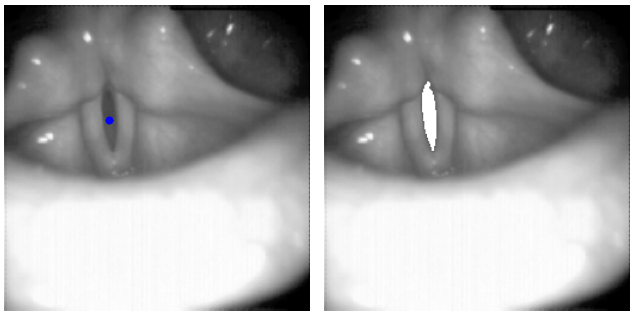
Figure 4.2: A laryngeal image segmented by different edge-based techniques. (a) Original image; (b) Roberts detector; (c) Prewitt detector; (d) Sobel detector; (e) LoG detector; (f) Canny detector; (g) Hough Transform with Sobel detector; (h) Hough Transform with LoG detector; (i) Hough Transform with Canny detector.

critical dependency of the initialization, which in many cases makes difficult a complete automatic procedure. Figure 4.3a and 4.3c show the same laryngeal image, but with a small difference in the initialization of the seed pixels which produces a large difference between the segmentation results.

- **Watershed:** The concept of watersheds comes from the field of topography, referring to the division of a landscape in several basins or water catchment



(a) Original (b) Region Growing



(c) Original (d) Region Growing

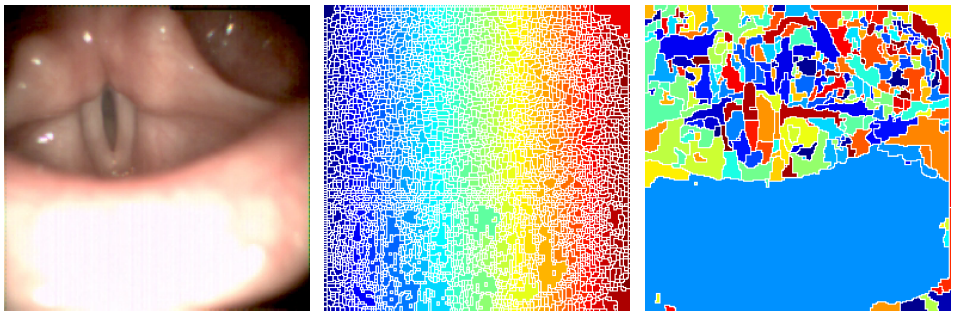
Figure 4.3: Laryngeal image segmented by region growing: (a) and (c) are the same image but with a different seed pixel (circles in blue); (b) and (d) are the respectively region growing segmented images (white regions).

areas. From this point of view, we can consider the image as a topographic surface and the watershed simulates a rain over the image where each pixel represents an altitude as a function of its grey level. The drops that fall over a point flow along the path of steepest descent until reaching a minimum. Such a point is labeled as belonging to the reception basin associated with this minimum. This process is repeated for all the points on the surface, so as a result, the landscape is partitioned into regions or basins separated by dams, called watershed lines or simply watersheds (Roerdink and Meijster, 2000). Gray-scale images can directly be used as the watershed's transformation input, but this usually is not the case because in most cases high values do not indicate edges.

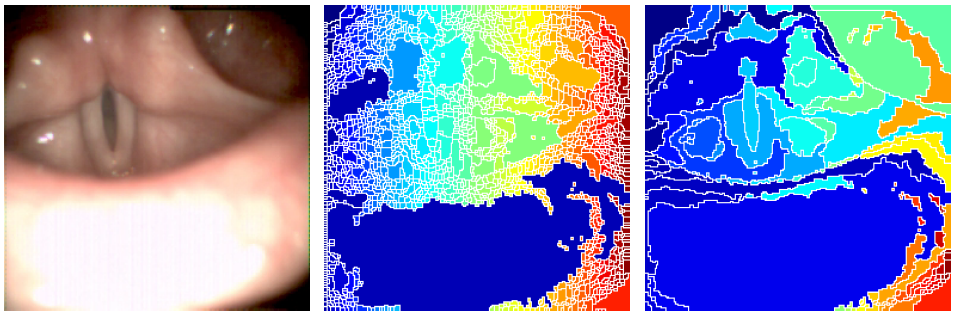
The main advantages of the watershed relies on the fact that the result is a set of well delimited areas, so if we consider that these areas represent the searched objects, we will obtain an accurate edge detection defined by a set of connected pixels. Nevertheless, the watershed transform is usually disappointing, due to the fact that thousands of catchment basins arise where only a few were expected due mainly to noise in the image (Day-Fann and Ming-

Tsong, 2003). A good solution to solve such a problem is to pre-process the initial image to reduce the noise. A widespread technique consists in computing the watershed transform over a thresholding of the gradient image. As a result, the gradient image has its maximum just over the edges of the objects present in the image, so the insignificant edges that appear due to noise are removed. However, this pre-processing does not solve completely the problem, so a post-processing would be required to reach a better solution.

Figure 4.4 shows the results obtained when watersheds are applied to the glottal segmentation problem, 4.4b and 4.4e depict the results after applying the watershed transform to the gray and color image gradient, respectively. Lastly, 4.4c and 4.4f correspond to the post-processing step using (Meyer and Beucher, 1990) and (Osma-Ruiz et al., 2008), respectively.



(a) Original (b) Gray Gradient (c) Marked watershed



(d) Original (e) Color Gradient (f) JND watershed

Figure 4.4: Watershed transformation applied to laryngeal images: (a) and (d) original frames; (b) and (e) watershed transform computed on the gradient of the images; (c) and (f) post-processing step using morphological controlled marked and Just Noticeable Difference (JND), respectively.

4.1.4 Classification-Based

The classification-based techniques segment each of the objects that compose the image based on a set of features that better describe each of them. Some of the

common features used are: texture, shape, brightness, contour energy, curvilinear continuity, among others. The classification-based techniques are typically divided into two broad categories, depending on the nature of the learning: supervised and unsupervised (Duda et al., 2000).

- **Supervised:** The supervised methods use a collection of training examples where each training example is formed by a feature vector and its respective label or also called output. Then the goal is to learn a rule that maps the feature vectors to the labels. Supervised algorithms include linear regression, logistic regression, decision trees, super vector machines, neural networks (Skourikhine et al., 2000), among others.

- **Unsupervised:** The unsupervised methods, in contrast to supervised methods, do not need to train data to segment an image. The objects are segmented based on natural groupings of pixel features (e.g. color, texture, spatial, etc) or possibly even some automatic learn sense. Common unsupervised algorithms include K-means Clustering, Fuzzy C-means Clustering (Macqueen, 1967), Principal Component Analysis (PCA) (Cootes et al., 1995), Mean Shift (Comaniciu and Meer, 2002), among others.

Figure 4.5 shows the results obtained after K-means Clustering in a laryngeal image. Figure 4.5b and Figure 4.5c use only color features with two and ten classes respectively. Meanwhile, Figure 4.5d depicts the result obtained when spatial and color features are combined.

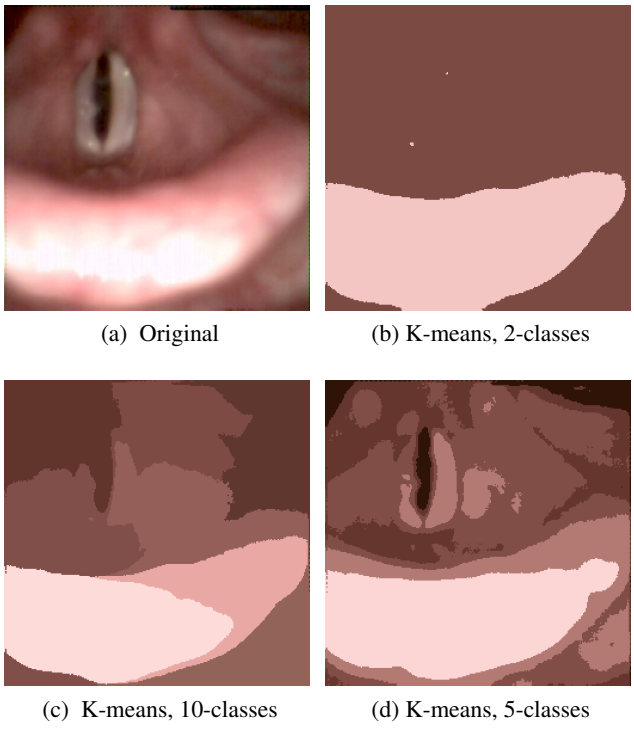
4.1.5 Graph-Based

The graph-based techniques borrow tools from graph theory to separate foreground and background. Graph-based methods represent the images as a graph where each node corresponds to an image pixel or region and their edges connections are weighted with respect to a similarity criterion between the pixels or regions.

The basic principle of most of the graph based segmentation methods is to find a set of disjoint sub-graphs that share a common feature. The graph partition is commonly formulated as an energy[1] minimization problem and can be solved via graph matching, random walker, min-cut/max-flow algorithm, Dijkstra's algorithm and Kruskal's or Prim's algorithm, among others (Boykov et al., 2001; Boykov and Kolmogorov, 2004).

Figure 4.6 depicts examples of interactive graph-cut segmentation applied to a laryngeal image. Figure 4.6b and Figure 4.6e shows the user-drawn markers where red represents the background (laryngeal structures) and blue represents the foreground (glottal gap). The final delimitations obtained are deployed in Figure 4.6c and Figure 4.6f.

[1]Energy is a relative term in image processing. The aim behind using the term 'Energy' is a minimization problem or maximization one.



(a) Original (b) K-means, 2-classes

(c) K-means, 10-classes (d) K-means, 5-classes

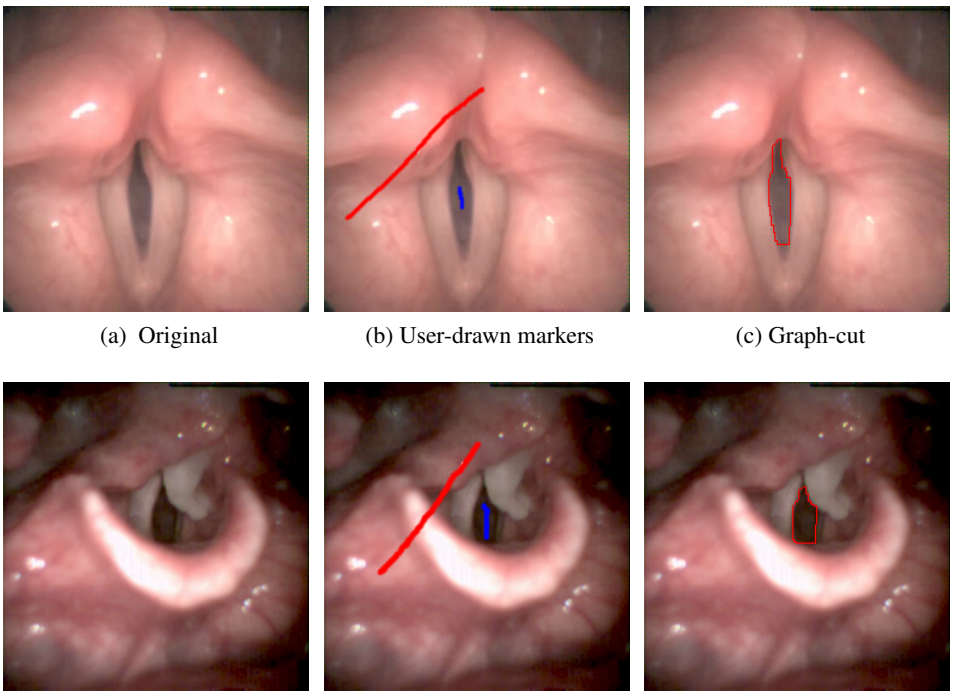
Figure 4.5: Classification-based segmentation applied to laryngeal images: (a) original frame; (b) K-means with two classes and color features; (c) K-means with ten classes and color features; (d) K-means with five classes and color-spatial features.

4.1.6 Deformable Models

Deformable models are curves or surfaces defined within an image domain that can move under the influence of internal energies (or regularization term), which are defined within the curve or surface itself, and external energies (or data term), which are computed from the image data. The internal energies are designed to keep the model smooth during deformation. Meanwhile, the external energies are defined to move the model towards a desired features within an image. Various names, such as snakes, active contours or surfaces, balloons, and deformable contours or surfaces, have been used to refer to the deformable models (Xu et al., 2000).

The deformable models can be classified into two approaches, the parametric and the geometric models, depending on how the model is defined in the shape domain. Additionally, there are deformable models that incorporate previous knowledge of the object features by using deformable shape templates.

- **Parametric Models:** also known as active contours, are curves whose defor-



(a) Original (b) User-drawn markers (c) Graph-cut

(d) Original (e) User-drawn markers (f) Graph-cut

Figure 4.6: Interactive Graph-based segmentation applied to laryngeal images: (a) and (d) original image; (b) and (e) interactive user-drawn markers; red background and blue foreground; (c) and (f) final segmentation using graph-cut method.

mations are determined by the displacement of a discrete number of control points along the curve. The main advantage of parametric models is that they are usually very fast in their convergence, depending on the predetermined number of control points. However, they are topology dependent, which avoid them to split or merge during deformation

In this group the one that has attracted more attention to date is the proposed in (Kass et al., 1988), also known as snake. The snake model is controlled by an external energy to drive the snake towards the pixel with high gradient, which means the edges of the object of interest. In the other hand, the internal energy serves to impose a piecewise smoothness constraint, making the snake contract like an elastic band by introducing tension, and rigidity to keep the snakes points (snaxels) together avoiding a break down.

- **Geometric Models:** also known as implicit models, use the theory of curve evolution and the level set method (Tsai and Osher, 2003) to transform the curves into higher dimensional scalar functions which permits handling topological changes naturally (splitting and merging). Geometric deformable models, proposed originally in (Mumford and Shah, 1989), provide an ele-

gant solution to address the primary limitations of parametric deformable models.

Different types of data terms are used in the geometric models. They can be based on edges (Caselles et al., 1997), region (Chan and Vese, 2001) or both combined (Mumford and Shah, 1989; Paragios and Deriche, 2002). The edges-based uses image gradients to identify object boundaries meanwhile the region-based tries to model the foreground and background regions statistically and find an energy optimum where the model best fits the image.

- **Deformable Shape Models:** use global shape parameters to embody a priori knowledge of expected shape and shape variation of the structures. The deformable shape models are widely used when a set of training samples are available.

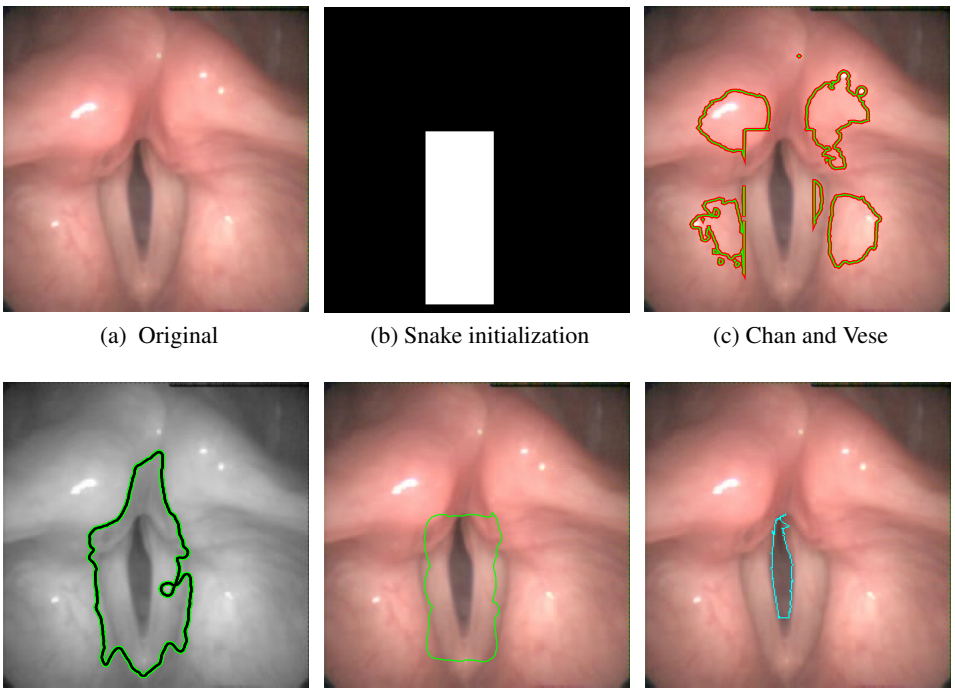
  Most of the deformable shape models learn global modes of variation using PCA. There are others that use pairwise geometric relations between landmarks, or representing shapes as configurations of independently deforming triangles (Felzenszwalb, 2005). The shape in most of the cases is learned by manual extraction of the shape points and only few methods automatically extract the necessary landmark points and their correspondences from the training shapes.

Figure 4.7a depicts a laryngeal image that has been segmented using different deformable models. Figure 4.7c and 4.7d use geometric models based on the formulation proposed in (Xu and Prince, 1998) and (Lankton and Tannenbaum, 2008), respectively. Meanwhile, Figure 4.7e and 4.7f use deformable parametric model. Figure 4.7f is based on (Andrade-Miranda et al., 2013) which has been conceived for the particular problem of glottal segmentation and does not need initialization.

4.2 Review of Motion Estimation Techniques

The motion estimation techniques are the core of numerous applications in computer vision, video processing, robotics and animation. For instance, it is used for object tracking (Yilmaz et al., 2006), human computer interaction (Martínez et al., 2012), temporal interpolation (Lim et al., 2005), spatio-temporal filtering and image compression (Ji et al., 2010).

The main objective of motion estimation algorithms is to precisely and faithfully model the motion in the scene which is typically represented using motion vectors, also known as vector displacements. The motion estimation techniques can be grouped into pixel based methods (direct) and feature based methods (indirect). The direct methods derive motion vectors for each pixel in the scene and can be categorized in Phase Correlation, Block Matching, Pel-Recursive, and Optical Flow (OF). On the other hand, the indirect methods use features matching between frames to compute the motion vectors.



(a) Original (b) Snake initialization (c) Chan and Vese

(d) Lankton and Tannenbaum (e) Xu and Prince (f) Andrade-Miranda et al.

Figure 4.7: Active contours applied to laryngeal images: (a) original image; (b) snake initialization; (c) deformable model based on Chan and Vese (2001) with 300 iterations; (d) deformable model based on Lankton and Tannenbaum (2008) with 300 iterations; (e) Gradient vector flow method proposed by Xu and Prince (1998); (f) active contours based on Andrade-Miranda et al. (2013) with automatic initialization.

4.2.1 Phase Correlation

The Phase correlation exploits the property that translation in the spatial domain has its counterpart in a transform domain, using for instance: the Fourier transform, the Discrete Cosine Transform (DCT) or a Discrete Wavelet Transform (DWT). The result of the phase correlation between two images is a new image which has peak intensities at locations where the two images match the best (Reddy and Chatterji, 1996; Zitová and Flusser, 2003). Since the Phase Correlation uses only phase information, it is relatively insensitive to illumination changes and it achieves excellent robustness against correlated and frequency-dependent noise. However, complex motions are difficult to characterize in the transformed domain.

4.2.2 Block Matching

Block matching divides a frame at time t_k (current frame) into blocks and compares each of the blocks with a corresponding block and its adjacent neighbors in a nearby frame (usually the next frame, t_{k+1}) (Zhu et al., 2002; Changsoo and Hyung-Min, 2013). The similarity between the blocks from the current and next frame are commonly computed via Sum of Square Error (SSE) or Sum of Absolute Difference (SAD). SSE provides a more accurate block matching, but it requires more computation. Meanwhile, SAD provides a fairly good match with a lower computational requirement.

The matching between blocks is computed only inside a region known as the search area. The search area defines the boundary for the motion vectors and limits the number of blocks to evaluate. There are different search strategies, being the most used: full search block matching, three-step search, 2D logarithmic search, one at time search algorithm, sub-pixel motion estimation and hierarchical block matching. The motion vectors are obtained by computing the displacement of the blocks of the current frame with respect to the next frame and all pixels belonging to the same block share the same motion vector. Figure 4.8 depicts graphically the procedure followed to compute the block matching algorithm.

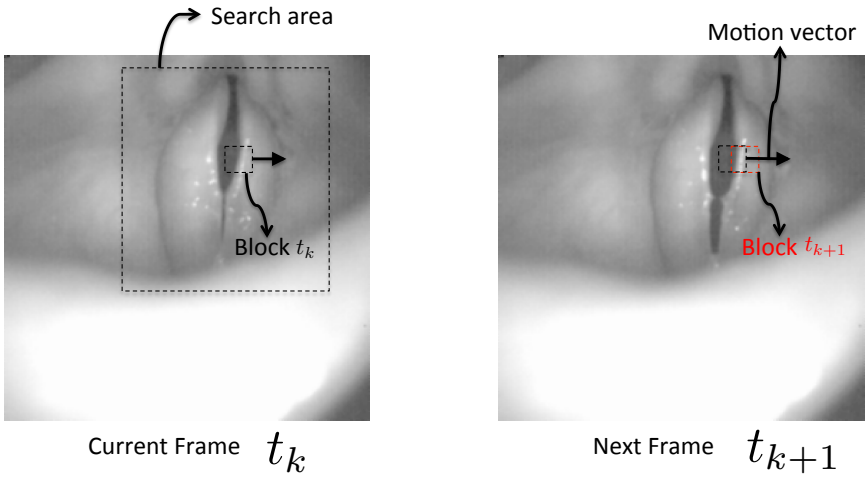


Figure 4.8: Illustration of the block matching algorithm.

4.2.3 Pel-Recursive

The Pel-recursive technique estimates the displacement vectors by recursively minimizing a nonlinear function of the dissimilarity (or also called the displaced frame difference function DFD) between two certain regions located in two consecutive frames where the regions can be a group of pixels or a single pixel. Then, the estimated displacement vector of a pixel is used as an initial estimate for the next pixels and so on. This recursion can be carried out horizontally, vertically, or tem-

porally which means that the estimated displacement vector can be passed to the pixel of the same spatial position within image planes in a temporary neighboring frame (Biemond et al., 1987; Efstratiadis and Katsaggelos, 1990). With respect to the optimization procedure followed to minimize the displaced frame difference function, most of the works are based on traditional approaches as the steepest descent method, and the Newton-Raphson method.

4.2.4 Optical Flow

OF estimation has been used for the last 35 years since the seminal works of Horn-Schunck and Lucas-Kanade (Horn and Schunck, 1981; Lucas and Kanade, 1981), and many innovative methods have been proposed to solve its computation (Beauchemin and Barron, 1995). However, to date, there is no unique method to characterize at minimal computational cost all the possible motion scenarios, including those with disturbing phenomena such as lighting changes, reflection effects, modifications of objects properties, motion discontinuities, or large displacements.

The definition of the OF is originated from a physiological description of the images formed on the retina, which determine that the image is formed due to the change of structured light caused by a relative motion between the eyeball and the scene. In the field of computer vision, Horn-Schunck defined OF in (Horn and Schunck, 1981) as *"the apparent motion of brightness patterns observed when a camera is moving relative to the objects being imaged"*.

Given an image sequence $I(\mathbf{x},t)$, the basic OF assumption is that at any pixel $\mathbf{x}_{ij}$, at time t_k, the intensity $I(\mathbf{x}_{ij},t_k)$ would remain constant during a short interval of time Δt_k, the so-called Brightness Constancy Constraint (BBC) or data term (see eq 4.1).

$$I(\mathbf{x}_{ij},t_k) = I(\mathbf{x}_{ij} + \vec{\mathbf{w}}(\mathbf{x}_{ij},t_k), t_k + \Delta t_k) \quad \forall\, \mathbf{x}_{i,j} \tag{4.1}$$

where $\vec{\mathbf{w}}(\mathbf{x}_{ij},t_k) = (u(\mathbf{x}_{ij},t_k), v(\mathbf{x}_{ij},t_k))$ is the vector displacement of $\mathbf{x}_{ij}$ in a time Δt_k. The vector displacement $\vec{\mathbf{w}}(\mathbf{x}_{ij},t_k)$ has two components: one in the x-axis direction ($u(\mathbf{x}_{ij},t_k)$) and other in the y-axis direction ($v(\mathbf{x}_{ij},t_k)$). Therefore, the total motion field at time t_k is defined as (eq 4.2)

$$\mathcal{W}(\mathbf{x},t_k) = \begin{pmatrix} \vec{\mathbf{w}}(\mathbf{x}_{11},t_k) & \vec{\mathbf{w}}(\mathbf{x}_{12},t_k) & \cdots & \vec{\mathbf{w}}(\mathbf{x}_{1n},t_k) \\ \vec{\mathbf{w}}(\mathbf{x}_{21},t_k) & \vec{\mathbf{w}}(\mathbf{x}_{22},t_k) & \cdots & \vec{\mathbf{w}}(\mathbf{x}_{2n},t_k) \\ \vdots & \vdots & \ddots & \vdots \\ \vec{\mathbf{w}}(\mathbf{x}_{m1},t_k) & \vec{\mathbf{w}}(\mathbf{x}_{m2},t_k) & \cdots & \vec{\mathbf{w}}(\mathbf{x}_{mn},t_k) \end{pmatrix} \tag{4.2}$$

and its components can be defined at the same way respectively by eq 4.3 and

eq 4.4

$$U(\mathbf{x},t_k) = \begin{pmatrix} u(\mathbf{x}_{11},t_k) & u(\mathbf{x}_{12},t_k) & \cdots & u(\mathbf{x}_{1n},t_k) \\ u(\mathbf{x}_{21},t_k) & u(\mathbf{x}_{22},t_k) & \cdots & u(\mathbf{x}_{2n},t_k) \\ \vdots & \vdots & \ddots & \vdots \\ u(\mathbf{x}_{m1},t_k) & u(\mathbf{x}_{m2},t_k) & \cdots & u(\mathbf{x}_{mn},t_k) \end{pmatrix} \tag{4.3}$$

$$V(\mathbf{x},t_k) = \begin{pmatrix} v(\mathbf{x}_{11},t_k) & v(\mathbf{x}_{12},t_k) & \cdots & v(\mathbf{x}_{1n},t_k) \\ v(\mathbf{x}_{21},t_k) & v(\mathbf{x}_{22},t_k) & \cdots & v(\mathbf{x}_{2n},t_k) \\ \vdots & \vdots & \ddots & \vdots \\ v(\mathbf{x}_{m1},t_k) & v(\mathbf{x}_{m2},t_k) & \cdots & v(\mathbf{x}_{mn},t_k) \end{pmatrix} \tag{4.4}$$

The BCC provides only one equation to recover the two unknown components of $\boldsymbol{\mathcal{W}}(\mathbf{x},t_k)$. Therefore, it is necessary to introduce an additional constraint encoding a priori information of $\boldsymbol{\mathcal{W}}(\mathbf{x},t_k)$. Such information comes from the spatial coherency imposed by either local or global constraints (regularization term (Fortun et al., 2015)).

In practice, the BBC assumption is an imperfect photometric expression of the real physical motion in the scene that can not be applied in case of changes in the illumination sources of the scene, shadows, noise in the acquisition process, specular reflections or large and complex deformation. Therefore, several matching costs (also called penalty functions) have been explored to overcome the drawback of the BBC, in particular its sensitivity to noise and illumination changes.

Over the last years, the number of optical flow algorithms with increasingly good performance has grown dramatically and it becomes difficult to summarize all contributions and categorizes. For instance, studies back to the nineties (Beauchemin and Barron, 1995; Barron et al., 1994) classify the OF in six groups: intensity-based differential methods, frequency methods, correlation-based method, multiple motion methods and temporal refinement methods. On the other hand, some of the last studies focus their attention on variational approaches (Mitiche and Aggarwal, 2014; Weickert et al., 2006; Werlberger et al., 2010) since they are versatile, allowing one to model different forms of flow fields by combining different data and regularization terms. But more important, they have shown the most accurate results to the OF problem in the literature (Wedel and Cremers, 2011). Other OF algorithms with outstanding performance are based on discrete optimization (Menze et al., 2015), the main advantage over the continuous approaches is that it does not require differentiation of the energy and can thus handle a wider variety of data and regularization terms. On the counterpart, a trade-off has to be found between the accuracy of the motion labeling and the size of the search space.

4.2.4.1 Evaluation of the Optical Flow Methods

As in computer vision domain, much attention has been paid to the design of appropriate evaluation procedures for OF. The visualization of motion fields provides a

qualitative insight to the accuracy of the estimation. There are two main visualization techniques to assess the OF: via arrow visualization or via color code. The first one represents the motion vector and offers a good intuitive perception of physical motion. On the counterpart, a clean display requires to under-sample the motion field to prevent overlapping of arrows. Meanwhile, the color code visualization associates a color hue to a direction and a saturation to the magnitude of the vector. It allows a dense visualization of the flow field and a better visual perception of subtle differences between neighbor motion vectors. Figure 4.9 shows the visualization of flow fields, following the proposal in (Baker et al., 2011b).

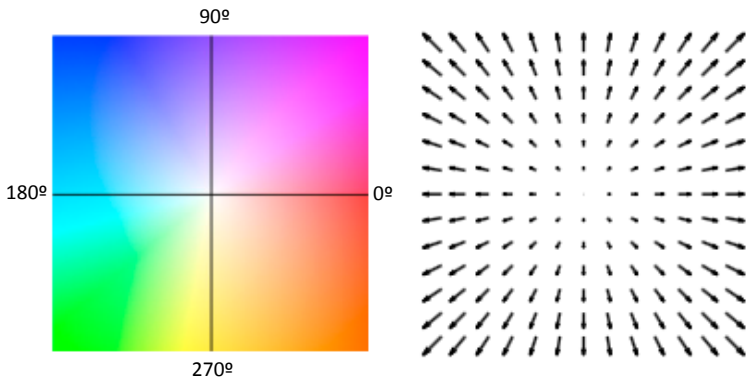


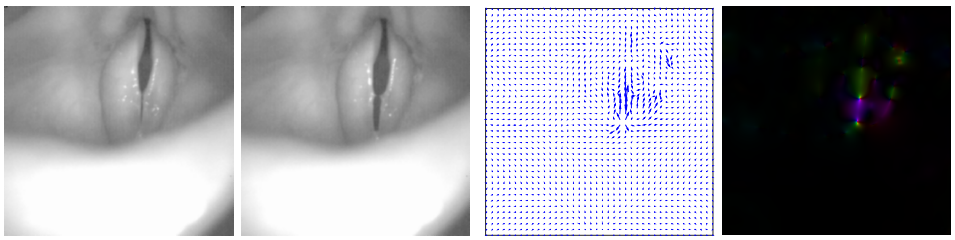
Figure 4.9: The visualization of flow fields. Left side: color code visualization, and right side: arrow visualization. Adapted from (Liu et al., 2011).

Additionally, there are two objective quantitative evaluation methods based on error metrics that are used to compare the performance of the OF methods when a ground truth is available, namely the Angular Error (AE) and the Endpoint Error (EPE) (Baker et al., 2011b). Figure 4.10 depicts the arrow and color code visualization using three different OF formulations: Horn and Schunck (1981), traditional framework; Bruhn et al. (2006), improved version of Horn and Schunck (1981) based on bidirectional multigrid methods; and Drulea and Nedevschi (2013) which is based on correlation transform.

In the next sections, three OF algorithms are presented: Lukas Kanade Optical-Flow (LK-OF) which is a classical framework, Motion Tensor Optical-Flow (MT-OF) that is based on tensor motion and the last one refers to the class of Total Variation L1 methods (TVL1-OF).

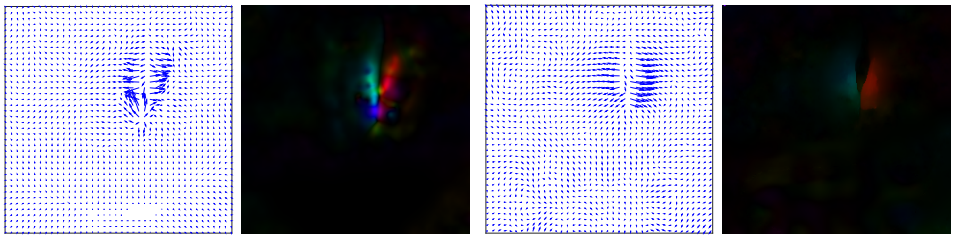
4.2.4.2 Lucas Kanade Optical Flow

LK-OF models the OF locally assuming that each pixel in a local neighborhood Ω has the same motion pattern. To solve the OF at $\mathbf{x}_{ij}$, a weighted least squares method is implemented and the energy functional E_{LK} is minimized (Lucas and



(a) Laryngeal images at time t_k and t_{k+3}

(b) Arrow and color code visualization based on (Horn and Schunck, 1981) OF



(c) Arrow and color code visualization based on (Bruhn et al., 2006) OF

(d) Arrow and color code visualization based on (Drulea and Nedevschi, 2013) OF

Figure 4.10: Arrows and color code visualization. (a) Two laryngeal images taken during the opening phase of the vocal folds at time t_k and t_{k+3}; (b) OF based on Horn and Schunck (1981); (c) improved mathematical formulation of Horn and Schunck (1981) using (Bruhn et al., 2006); (d) OF based on Drulea and Nedevschi (2013).

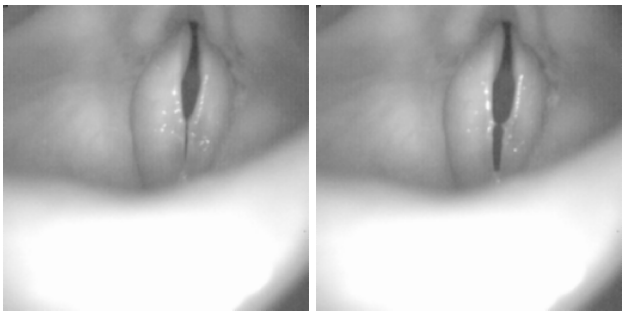
Kanade, 1981).

$$E_{LK} = \sum_{\mathbf{x}_{ij} \in \Omega} \mathscr{W}^2(\mathbf{x}_{ij}) \left(\nabla I(\mathbf{x}_{ij}, t_k) \cdot \vec{\mathbf{w}}(\mathbf{x}_{ij}, t_k) + \frac{\partial I(\mathbf{x}_{ij}, t_k)}{\partial t} \right)^2 \tag{4.5}$$

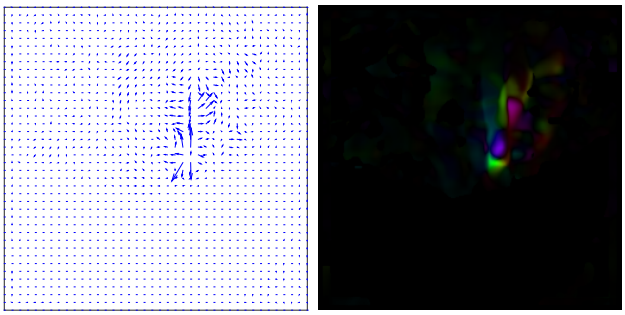
where $\mathscr{W}^2(\mathbf{x}_{ij})$ is the weight function associated with each neighboring pixel that diminishes the importance of distant neighbors. The eq 4.5 simply sums the error of applying the flow vector to the spatial and temporal gradients of all surrounding neighbors. The more inconsistent with spatial and temporal gradients of some neighbors, the higher the error. LK-OF has sub-pixel accuracy and also low computational cost. However, the neighborhood size should be carefully decided to avoid a blurred motion field. Figure 4.13 shows the arrow and color code visualization using the traditional LK-OF and its improved version based on (Bruhn et al., 2006).

4.2.4.3 Motion Tensors

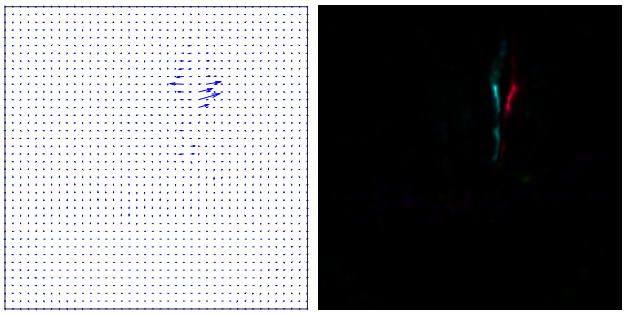
The underlying idea of this technique is to estimate the motion field by combining the 3D Orientation Tensors computed from the image sequence under the con-



(a) Laryngeal images at time t_k and t_{k+3}



(b) Arrow and color visualization based on (Lucas and Kanade, 1981)

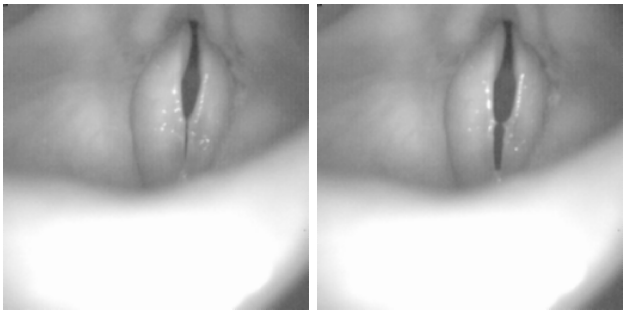


(c) Arrow and color visualization based on (Lucas and Kanade, 1981) with improved mathematical framework (Bruhn et al., 2006)

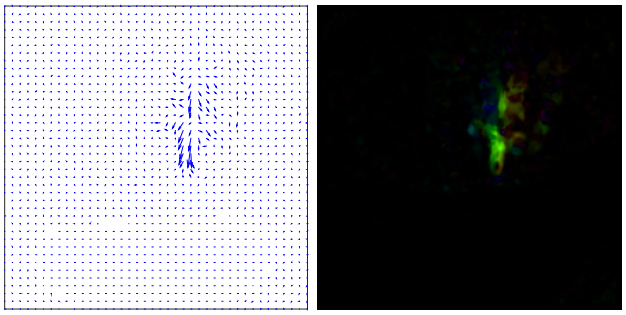
Figure 4.11: Arrows and color code visualization using LK-OF computation

straints of a parametric motion model (Farneback, 2000). The 3D Orientation Tensors are a powerful representation of the local orientations and one way to construct them is by stacking the frames of an image sequence onto each other to obtain a spatiotemporal image volume with two spatial dimensions and a third temporal dimension. From here, it is easy to see that a movement in the image sequence induces structures with certain orientations in the volume. For instance, a translating point is transformed into a line whose direction in the space directly corresponds to its velocity. Normally to avoid the effects of noise and inaccuracies in the ten-

sor estimation and also to solve the aperture problem in some pixels, the use of a motion model to parameterize the coherency motion in small regions is a common approach. However, the disadvantage of such approach is the assumption that the true velocity is at least reasonably consistent with the chosen motion model.



(a) Frame #19 and Frame #22 from LHSV



(b) Arrow and color visualization of the MT-OF

Figure 4.12: Arrows and color code visualization using MT-OF computation

4.2.4.4 TV-L1 Optical Flow

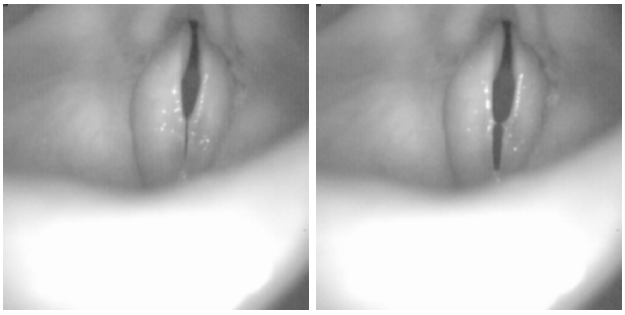
This method minimizes an energy functional, E_{TVL1}, which contains two terms. The first is an image similarity score L_1[2] based on the brightness constancy constraint. The second is a regularization term that adds a smoothness condition by forcing the motion field to be regular using the total variation. The TVL1-OF formulation is expressed by eq 4.6 as:

$$E_{TVL1} = \int \lambda \left| I(\mathbf{x}_{ij} + \vec{\mathbf{w}}(\mathbf{x}_{ij}, t_k), t_k + \Delta t_k) - I(\mathbf{x}_{ij}, t_k) \right| + \left| \nabla \vec{\mathbf{w}}(\mathbf{x}_{ij}, t_k) \right| \qquad (4.6)$$

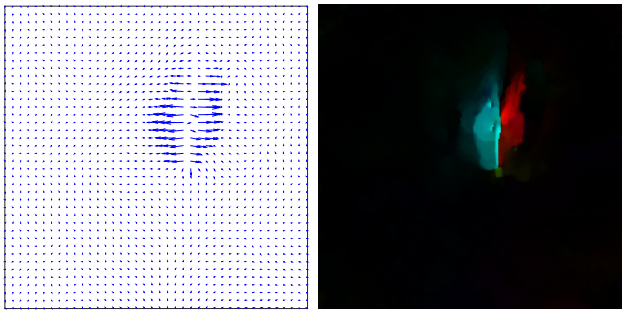
where λ is a free parameter used to balance both terms. The displacement vector $\vec{\mathbf{w}}$ is the minimizer and the strategy followed to solve the energy minimization

[2] L_1 is a metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates.

problem is a convex relaxation approach (Zach et al., 2007; Javier et al., 2013). TVL1-OF formulation provides three main features. First, it allows discontinuities in the flow field, which is desirable when a complex motion is modeled. Secondly, it is sub-pixel accurate. Finally, it is robust to noise thanks to the denoising feature included in the convex relaxation formulation. However, it is more sensitive to large displacements, making necessary the use of multiscale approaches (Javier et al., 2013). The main undesirable effect produced by a multiscale approach is the loss of small and rapidly moving objects in the final estimation of the flow field (Fortun et al., 2015). Figure 4.13 shows the arrow and color code visualization of the TVL1-OF applied to laryngeal images.



(a) Laryngeal images at time t_k and t_{k+3}



(b) Arrow and color visualization based on TVL1-OF

Figure 4.13: Arrows and color code visualization using TVL1-OF computation

4.2.5 Feature-Based Methods

The feature-based methods compute a sparse motion field which means that only some pixels from the whole image have a displacement vector. The pixels used to compute the sparse motion field are salient and distinctive features. In the literature, a large variety of feature extraction methods have been proposed to compute reliable descriptors. Among these descriptors, the scale invariant feature transform (SIFT) descriptor (Lowe, 2004) utilizing local extrema in a series of differences of Gaussian (DoG) functions for extracting robust features, and the speeded-up robust features (SURF) descriptor (Bay et al., 2008) partly inspired by the SIFT descriptor

for a fast computing of distinctive invariant local features, are the most popular and widely used in several applications.

After the feature extraction step, the correspondence between the features in the current image (t_{k+1}) and those detected in the reference image (t_k) are matching. There are different strategies to match the features, for instance: Brute-Force matcher which takes the feature in the reference image and is matched with all other features in the current image using some distance calculation, and the closest one is returned as a good matching. One more efficient method is FLANN which stands for Fast Library for Approximate Nearest Neighbors. FLANN contains a collection of algorithms optimized for fast nearest neighbor search in large datasets and for high dimensional features.

4.3 Review of Inpainting Techniques

Inpainting is the process of restoring missing or damaged areas in an image by assuming that pixels in the known and unknown parts of the image share the same statistical properties or geometrical structures. The inpainting techniques have been used for restoration of photographs, films and paintings, to remove occlusions, such as text, subtitles, stamps and publicity from images. In addition, inpainting can also be used to produce special effects (Guillemot and Meur, 2014). There are 4 categories of inpainting techniques known as diffusion-based methods, examplar-based methods, sparse-based methods and hybrid-based methods.

The diffusion-based methods use parametric models or partial differential equation (PDEs) to smoothly propagate local structures from the exterior to the interior of the damaged region, imitating the gesture of profesional painting restaroators. These methods are well suited for completing straight lines, curves and for inpainting small regions but fail to recover the texture of large areas since they tend to blur the regions.

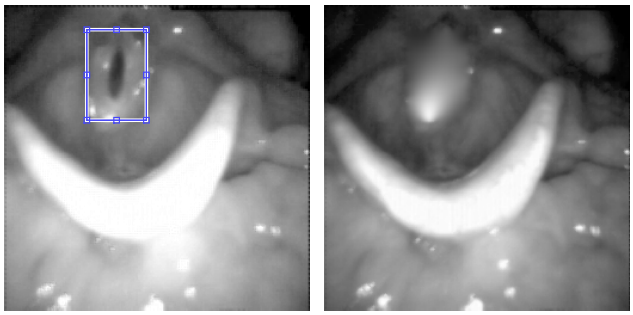
The examplar-based methods use image statistics and self similarity priors. The statistics of image textures are assumed to be stationary or homogeneus. The texture to be synthesized is learned from similar regions in a texture sample by sampling and copying or stitching together patches taken from known parts of the images. These methods have been inspired by local region growing methods and rely on Markov Random Fields modeling of textures.

The sparse-based methods assume that the image or patches are sparsed in a given basis. The basis can be formed by predefined elementary waveforms which are stored in an dictionary matrix. The known and unknown part of the image are assumed to share the same sparse representation.

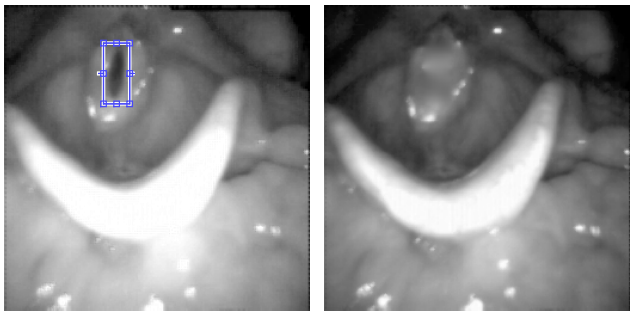
Lastly, the hybrid methods combine the structural diffusion propagation with textural components, having as a result more robust algorithms.

Figure 4.14 depicts two examples of image inpainting using the diffusion-based approach. The original images are showed in Figure 4.14a and 4.14c. The blue rectangles represent the region to be restored. In these particular examples the

glottal gap is replaced by the information that surrounds it (laryngeal structures) in order to simulate a vocal folds in close-state.



(a) Original (b) Diffused inpainting



(c) Original (d) Diffused inpainting

Figure 4.14: Diffusion-based inpainting applied to laryngeal images: (a) and (c) are the same image but with a different region to be inpainted (rectangles in blue); (b) and (d) are the results of appling a Diffusion-based inpainting technique.

4.4 Discussion

The literature intended to solve the segmentation and motion estimation problem is vast, so this chapter only claim to introduce a general classification, emphasizing their use in laryngeal images.

Algorithms such as thresholding ones use the information based on a single pixel and do not take spatial information into account. Additionally, they highly depend on the intensity distribution of the images, which mean that heterogenous objects can not be segmented correctly.

Edge-Based algorithms tend to produce disjoint edges, they are sensitive to noise and have over-segmentation tendency. They alone are not able to solve a complex task such as the glottal segmentation, so the use of additional techniques is necessary (Aghlmandi and Faez, 2012).

Region-Based algorithms require that the target objects to segment have homo-

geneous features. In some cases, an user interaction is needed (region growing). Furthermore, they have problems of over-segmentation (watershed), so it is necessary the use of post-processing steps.

In Classification-Based algorithms, the supervised methods depend on training parameters which are usually setting in a trial-and-error manner. The accuracy of this algorithm largely depends on the selected training samples. Also, they are more tedious to use. On the other hand, the unsupervised methods often produce many objects, particularly for heterogeneous images. The number of objects to be segmented, or also known as classes, is an important parameter that affects their accuracy.

Graph-Based algorithms are computationally expensive and the criterion for a good partition is a challenging task that presents problems of over-segmentation since it uses low-level features such as intensity and edges, which are often corrupted by noise.

The Deformable Models include constraints that make them less sensitive to noise. However, they are sensitive to the initialization, hence a wrong initilization make them converge to a erroneous object. Deformable Shape Models require training samples which make them difficult to implement when different deformations have to be modelled, which is the case of laryngeal images.

In general, thresholding, Region/Edge-Based, graph-based and classification-based algorithms can solve simple medical image segmentation problems where the images are noise free, high contrasted, and have quite homogeneous regions. For complex medical image segmentation problems, deformable models have more potential but they rely on the initialization. The main advantage of thresholding, region-based and edge-based algorithms with respect to classification-based, graph-based and deformable models is that the computational complexities are roughly linear.

On the other hand, the motion estimation techniques have the goal to compute a motion field in which the motion is represented by vector displacements. In Optical Flow, Phase Correlation, Block Matching and Pel-Recursive methods a vector motion is computed for each pixel, which is interesting for motion analysis applications. Contrariwise, the indirect methods use feature descriptors to compute the motion vectors only in pixels with salient features, which make them more accurate in cases of large displacements.

Chapter 5

Glottal Segmentation Techniques

"Truth is ever to be found in the simplicity, and not in the multiplicity and confusion of things".

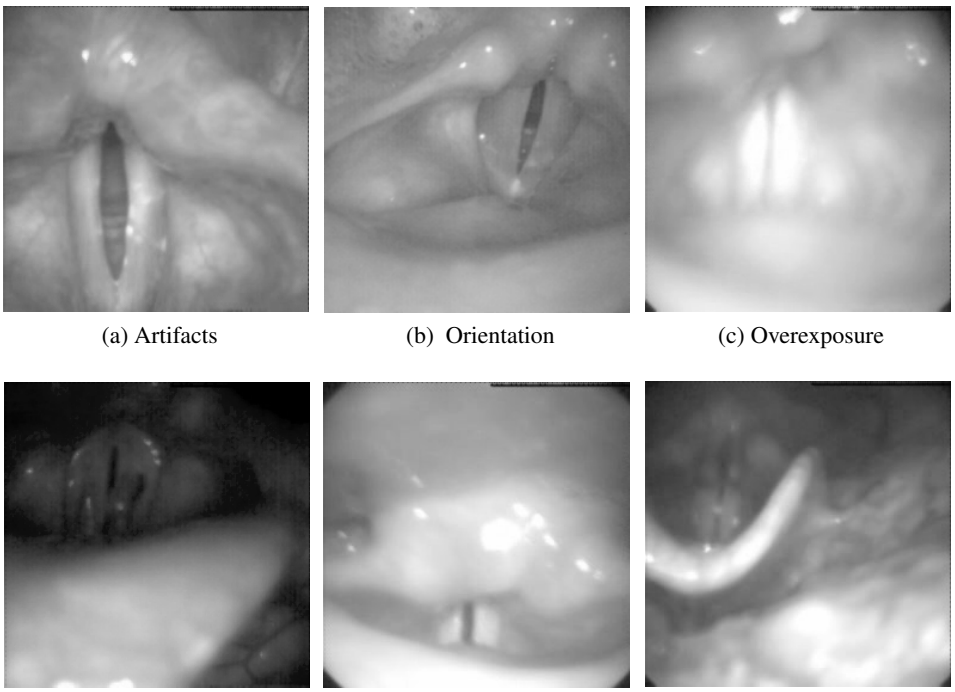
Sir Isaac Newton

SUMMARY: The glottal gap segmentation is the most extended method to synthesize the dynamic behaviour of the vocal folds. However, it is a challenging task, and a number of methods have been proposed. This chapter reviews the literature devoted to solve the problem of the glottal gap segmentation dividing the different approaches into three main stages: Image Enhancement, identification of the Region of Interest (ROI), and Glottal Gap Delimitation.

5.1 Overview

The glottal segmentation is an essential operation for the correct characterization of vocal-folds vibrations which let identify in an objective way different phonation features, 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.

From a pure image processing perspective, the task of tracking the edges of the vocal folds during an entire video sequence appears to be a standard tracking task. Thus, naively, it may seem that once the glottal area in some frame is delineated, it can easily be identified in successive frames, or that by subtracting successive frames the glottal gap will be obtained intuitively. However, there are many reasons why glottal segmentation is not a trivial task, some of them are listed below and depicted in Figure 5.1:



(a) Artifacts (b) Orientation (c) Overexposure

(d) Vocal hemorrhage (e) Aryepiglottic Oclussion (f) Nodule

Figure 5.1: Different laryngeal images during phonation showing different illumination conditions, orientation, depth, occlusion, among others features are showed.

- Inter-video variabilities such as position, orientation, different illumination levels, and depth differences between videos (see Figure 5.1).
- The difficulties presented by the camera rotation during recording, side movement of the laryngoscope, and movements of the patient, which produce the delocation of the vocal folds (Figure 5.1b, 5.1e, 5.1f).
- Reliability against external artifacts introduced due to recording problems. For instance, black elongated artifacts which appear on both the corners and borders of the frames (Figure 5.1a, 5.1c, 5.1e).
- The presence of occlusion effects, meaning that part of the vibrating glottis is hidden under the aryepiglottic fold or by other laryngeal structures (Figure 5.1d, 5.1e, 5.1f).
- Demanding cases as hourglass closure, irregular closure, presences of nodules, polyps and cysts, lesions, scars, presence of mucus, specular reflection, discoloration of the vocal folds, among others (Figure 5.1c, 5.1d, 5.1f)

The literature reports different techniques for the glottis segmentation task. Roughly speaking, they can be grouped depending on the user intervention in semi-

automatic and automatic methods. The semi-automatic techniques let the user interact as many times as needed in order to solve any inconvenience that might appear during the segmentation process. Contrariwise, the automatic techniques process all the data without any previous setting or any user intervention. From a clinical point of view, both methods present advantages and disadvantages but it is worth mentioning that semi-automatic methods are more time consuming for the clinicians, although their accuracy is expected to be better.

With respect to the semi-automatic segmentation, the literature reports different techniques and approaches. For instance, in (Lohscheller et al., 2007; Pinheiro et al., 2014) the user selects an arbitrary set of images within the video sequence and defines one or multiple seed-points belonging to the glottal area; in (Chen et al., 2013; Booth and Childers, 1979) the posterior and anterior commissures are given by the user; in (Mehta et al., 2013) the user defines the glottal midline by indicating the anterior and posterior commissure in the frame with the maximal opening and the user also defines a threshold based on a reference image; in (Blanco et al., 2013) the user searches around the video sequence for the frame with the minimal glottal opening; among others (Larsson et al., 2000; Marendic et al., 2001; Moukalled et al., 2009).

On the other hand, only a few of the existing approaches (Demeyer et al., 2009; Osma-Ruiz et al., 2008; Cerrolaza et al., 2011; Karakozoglou et al., 2012) are designed to be fully automatic. However, in the last few years, the fully automatic glottal segmentation algorithms have become an active research field with growing interest (Ko and Ciloglu, 2014; Schenk et al., 2014, 2015; Andrade-Miranda et al., 2015b; Gloger et al., 2015). Up to now, there is no standardized procedure to automatically segment glottal gap from endoscopic high-speed sequences, in spite of the extensive literature devoted to solve such as problem. The common approach to solve the glottal segmentation, roughly speaking, divides the problem into three main stages: image enhancement, identification of the Region of Interest (ROI), and glottal gap delimitation (see Figure 5.2).

5.2 Image Enhancement

Image enhancement refers to the manipulation or transformation of an image, with the aim of increasing its usefulness or visual appearance. For instance, the modification of intensity values, so as to increase contrast. There are not general criteria behind the enhancement, and often the techniques used for enhancement depend on the application (Gonzalez and Woods, 2006). The most common methods to objectively evaluate the image enhancement are Mean Square Error (MSE) and Peak-Signal-to-Noise-Ratio (PSNR) (Lehmann and Casella, 1998). However, they are not suitable for many applications and they fail to accurately reflect the subtleties of human perception (Wang and Bovik, 2009).

In laryngeal images, the glottis has darker intensity levels than its surrounding tissues. However, they often have low contrast and heterogeneous profiles due to

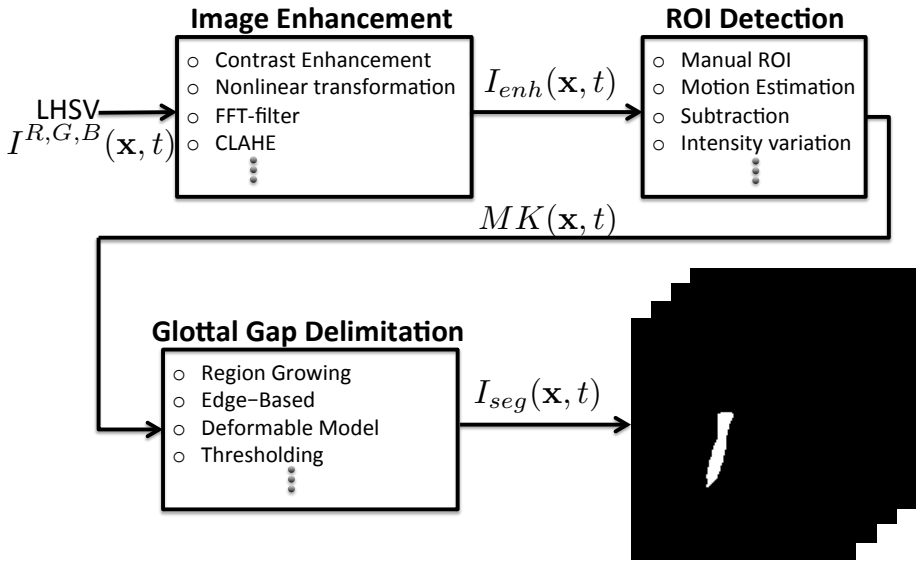


Figure 5.2: Graphic Representation of the three common steps followed to segment the glottal gap.

the illumination conditions. Modeling the histogram of the LHSV with a statistic distribution, such as Rayleigh as in (Yan et al., 2006), or finding the darkest region, produces errors due to the non-uniform contrast of the image, lighting conditions and artifacts due to the recording equipment. For this reason, it is required to simultaneously reduce the effect of the low contrast and to highlight the object of interest (i.e. the glottis). Thus, the use of image enhancing techniques is expected to improve the characteristics of the image for a further processing.

The literature reports the use of different enhancing techniques as a previous step to the glottis segmentation. In (Mendez et al., 2009) the authors combine an anisotropic diffusion with an FFT-based band pass filter in order to obtain a smoother image without losing edge information (second row of Figure 5.3). In (Zhang et al., 2010) a Lagrange interpolation is combined with a Gaussian filter in order to smooth the images, reduce noise and eliminate unwanted details. In (Yan et al., 2012), the authors use a global thresholding to obtain a binary image to eliminate the worthless information. However, this strategy can not be generalized for noisy and poor quality LHSV recordings.

Another alternative is to manipulate the histogram of the image. The most common histogram based processing techniques are the Histogram Equalization (HE), Adaptive Histogram Equalization (AHE), Contrast Limited Histogram Equalization (CLHE), and the Contrast Limited Adaptive Histogram Equalization (CLAHE). CLAHE is used in (Karakozoglou et al., 2012) providing more details in the glottal area while avoiding significant noise introduction (third row of Figure 5.3). CLAHE highlights the details over a small neighborhood preventing the over amplification of noise that can arise from adaptive histogram equalization AHE.

One of the most widespread methods is based on point-wise intensity transformations. The point-wise transformation operates directly over the intensity values of an image, processing each pixel separately and independently. This transformation can be linear, piecewise linear, or nonlinear. Aghlmandi and Faez (2012) establish a methodology for pre-processing LVS as a previous step for edge detection. The authors mention the drawbacks that exist in the acquisition due to the flashing effect at the recording instants, reducing the accuracy of the segmentation algorithm. The same procedure is used in (Skalski et al., 2008) to highlight the glottal area and to reduce the influence of the flashes in LHSV. Figure 5.3 depicts some of the enhancement methods used in the state of art for the glottal gap segmentation.

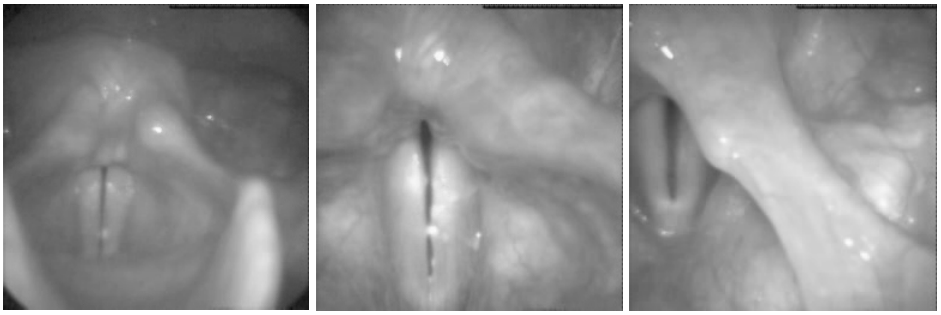
5.3 Region of Interest

A Region of Interest (ROI) is a part of the image that encapsulates important features that can be used for further analysis. ROI detection has been studied for many years. Most algorithms use either Feature-Based or Object-Based approaches. Feature-Based methods find pixels that share similar features to form the ROI. Meanwhile, object-based methods detect the ROI at a higher level than the pixel-by-pixel approach of Feature-Based systems using information such as target shape or structure. Figure 5.4 depicts some examples of ROI detection in six different LHSV sequences using the algorithm presented originally in (Andrade-Miranda and Godino-Llorente, 2014) and extended in (Andrade-Miranda et al., 2015b).

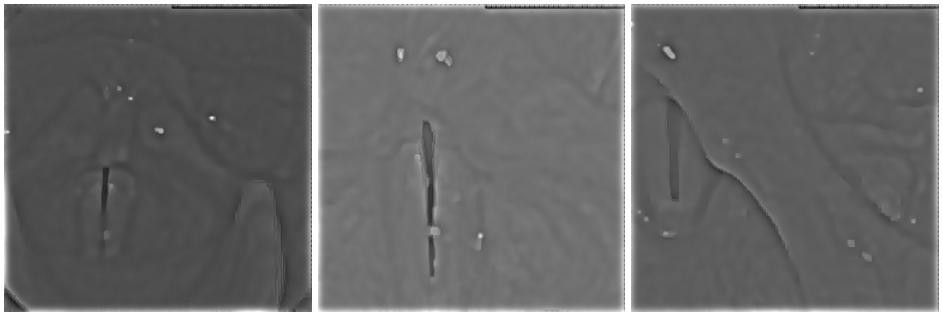
In laryngeal images, the vocal folds, and so the ROI, usually covers less than 25% of the entire image size. Therefore, the ROI detection permits to eliminate the non-relevant information and reduces the number of false detections, so it is an important step to be considered prior to the segmentation process. The literature reports some attempts to detect a ROI. However, most of these studies require user intervention (Palm et al., 2001; Marendic et al., 2001; Yan et al., 2006; Moukalled et al., 2009; Zhang et al., 2010; Yan et al., 2012; Chen et al., 2013) and, even more important, they do not consider the temporal information of the sequence.

In (Skalski et al., 2008), the authors assume that the segmentation of the glottal area from previous frames is available. Then, the values of the pixels where the difference between the current frame and the previous is larger than 20% of the maximum value of the image are set to 1. The authors in (Blanco et al., 2013) also propose an algorithm based on differences between consecutive frames. Other authors as (Larsson et al., 2000; Mendez et al., 2009; Alaoui et al., 2009) use motion estimation techniques to compute the ROI based on the fact that the region with the most salient motion features is the vocal folds.

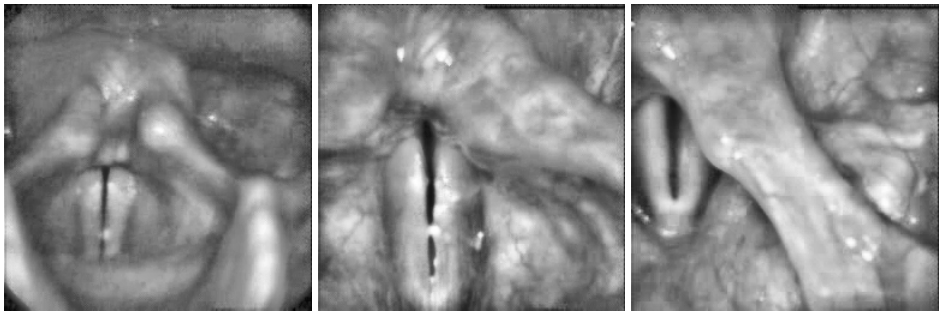
In (Lohscheller et al., 2007; Pinheiro et al., 2014), the user chooses a set of initial seed points in a frame with the vocal folds open. This can be understood as a ROI, since the Region Growing starts its journey from such a manual initialization.



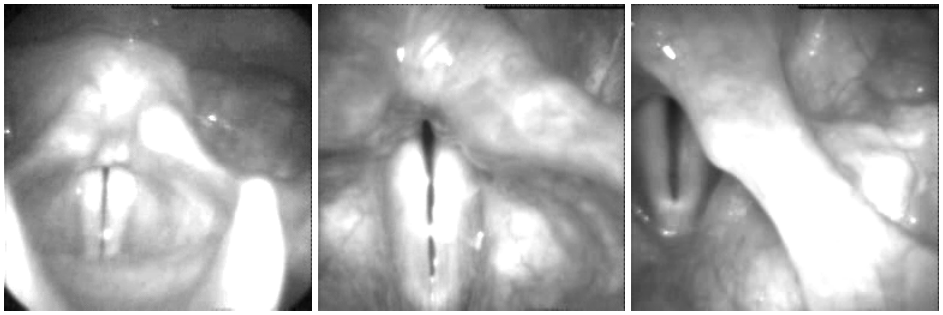
(a) Original



(b) Anisotropic and FFT-filter

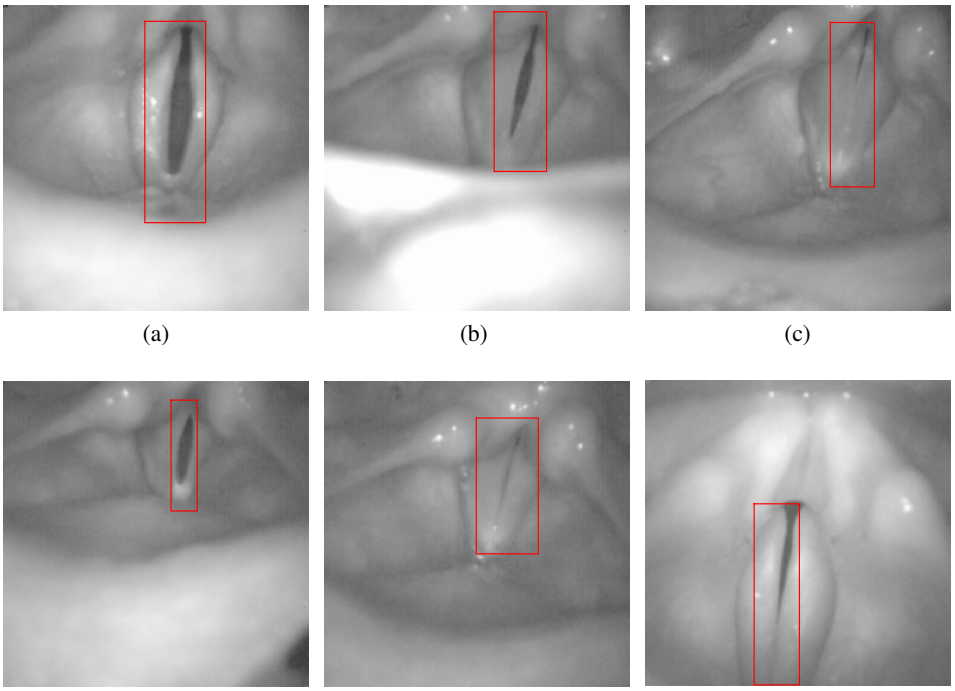


(c) CLAHE



(d) Nonlinear Transformation

Figure 5.3: Visual representation of the different enhancement methods for three different LHSV. First row: original image; second row: anisotropic with FFT-filter (Mendez et al., 2009); third row: CLAHE (Karakozoglou et al., 2012); fourth row: nonlinear transformation with $\beta = 200$ (Skalski et al., 2008).



(a) (b) (c)

(d) (e) (f)

Figure 5.4: Automatic ROI detection of six different LHSV computed based on (Andrade-Miranda and Godino-Llorente, 2014; Andrade-Miranda et al., 2015b). The image in (f) illustrates a minor problem in the detection of the ROI in the posterior commissure.

The method reported in (Karakozoglou et al., 2012) is an Edge-Based morphological processing of some frames extracted from the LHSV, called keyframes. The idea of the morphological operator is to find a large, nearly vertically oriented area and to apply a Sobel filter to detect the strong edges in the vertical direction. Then, a morphological closing operation is carried out over the gradient map to connect small related regions. The regions to be connected are identified by means of connected component analysis. Lastly, the object with the largest area and vertical orientation is chosen. Around the selected area, a rectangle is delineated, representing the ROI.

5.4 Glottal Gap delimitation

The most common techniques reported in the literature are based on Thresholding, Region Growing, Watershed and Deformable models (Parametric and Geometric). However, there are other approaches that include Deformable Shape Model (Cerrolaza et al., 2011), combination of different techniques (Andrade-Miranda et al.,

2015b), feature extraction and training (Gloger et al., 2015), among others (Mendez et al., 2009; Chen et al., 2013; Ko and Ciloglu, 2014).

Some of the main contributions with respect to glottal segmentation are depicted in Table 5.1 and briefly described below:

Author	User	Enhancement	ROI	Glottal Gap Delimitation	Video
Booth and Childers (1979)	✓	—	—	Subtraction and adaptive window	LHSV
Wittenberg et al. (1995)	✓	—	—	Region Growing	LHSV
Larsson et al. (2000)	✓	Contrast enhancement	Motion Estimation and manual ROI	Edge-Based	LHSV
Palm et al. (2001)	✓	—	Manual ROI	Parametric and Deformable Shape	LSV
Marendic et al. (2001)	✓	—	Manual ROI	Parametric Models	LHSV
Yan et al. (2006)	✓	—	Manual ROI	Thresholding Region Growing	LHSV
Lohscheller et al. (2007)	✓	—	Seed points	Region Growing	LHSV
Osma-Ruiz et al. (2008)	X	—	—	Watershed	LSV
Skalski et al. (2008)	✓	Nonlinear transformation	Subtraction	Geometric Models	LHSV
Mendez et al. (2009)	X	Anisotropic FFT-filter	Motion Estimation	Motion Estimation	LSV
Alaoui et al. (2009)	X	—	Motion Estimation	Motion Estimation	LSV
Moukalled et al. (2009)	✓	Histogram Thresholding	Manual $p(t)$, $a(t)$ commissure	Parametric Models	LHSV
Zhang et al. (2010)	✓	Lagrange interpolation	Manual ROI	Differentiation Edge-Based	LHSV
Cerrolaza et al. (2011)	X	—	—	Deformable Shape Models	LSV
Aghlmandi and Faez (2012)	X	Nonlinear transformation	—	Morphological Operators	LHSV
Elidan and Elidan (2012)	✓	—	—	Parametric Models	LSV
Yan et al. (2012)	✓	—	Manual ROI	Parametric Models	LHSV
Karakozoglou et al. (2012)	X	CLAHE	Morphological ROI	Geometric Models	LHSV
Mehta et al. (2013)	✓	—	Manual $p(t)$, $a(t)$ commissure	Thresholding	LHSV
Blanco et al. (2013)	✓	—	Subtraction	Thresholding	LHSV
Chen et al. (2013)	✓	Reflection removal	Manual $p(t)$, $a(t)$ commissure	Simplified Dynamic Programming	LHSV
Andrade-Miranda et al. (2013)	X	Anisotropic Thresholding	—	Parametric Models	LSV
Pinheiro et al. (2014)	✓	—	Seed points	Region Growing	LHSV
Ko and Ciloglu (2014)	X	Reflectance modeling	Intensity variation	Gaussin Mixture Models	LHSV
Gloger et al. (2015)	X	—	—	Training clasification	LSV LHSV
Schenk et al. (2015)	X	Color contrast stretching	Salient region	Geometric Models	LHSV

Table 5.1: Summary of the main studies carried out from glottal segmentation

- The authors in (Booth and Childers, 1979) use a model image that is subtracted from each frame of the video sequence. The objective is to remove as much artifact, or background as possible. They found that the best initial model corresponds to a frame with the glottis completely closed. After that, the posterior and anterior commissures are defined manually and two adaptive windows are used to trace the left and right folds boundaries separately.

- Wittenberg et al. (1995) use the dark pixels of the image as seed points for a Region Growing algorithm. However, this criterion is not appropriate when the images have shades and low contrast.

- In (Larsson et al., 2000), the motion of the mucosa wave is used to compute the ROI. Later, the gray scale levels of the image are adjusted manually and the vocal folds edges are tracked using the maximal derivative of the image.

- In (Palm et al., 2001), a variation of the snakes based method called balloon model is used to improve the behaviour against noise, as well as to obtain some degree of independence with the initialization procedure. The balloon model is defined by vertices and edges. Curve evolution is steered by shrinking and expanding forces, smoothing forces and external forces resulting from image edges. Deformable Shape Models are interconnected with the balloon model to integrate shape constraints for the vocal folds in the process of curve evolution. The main drawback is that all of their parameters have to be tuned carefully to achieve good results.

- The authors in (Marendic et al., 2001) extend the traditional active contour model to solve the glottis segmentation problem using two internal stretching forces to guide the active contour into narrowing posterior and anterior glottal commissures. They adapt a linear filter to a Canny edge detector to reduce the noise and to compute the external energy. They initialize the snake in the current frame using the results from the previous one. The parameters of the internal and external energies are selected empirically.

- In (Yan et al., 2006), the authors combine the use of Thresholding with Region Growing. First, a group of seed points are automatically computed by assuming that the frames follow a Raleigh distribution for both glottis and background regions. Then, the seed points are used for the Region Growing algorithm. This approach segments the glottal regions assuming that the glottal regions are significantly darker than the background tissue.

- In the semi-automatic method (Lohscheller et al., 2007), the user selects within the image sequence an arbitrary number of frames where one or multiple seed points can be placed inside the glottis. Then, a homogeneity criterion is defined using thresholding. The user checks the results and adapts the thresholds until a satisfactory segmentation is obtained. If the results are not according to the expectations, the whole procedure must be repeated again.

- The authors in (Osma-Ruiz et al., 2008) use the Watershed transform followed by a JND based region merging and a linear discriminant analysis which is based on seven binary invariant moments. However, the authors need to use several values in their approach to adapt the parameters and threshold values (i.e. visibility thresholds) either for the region merging as binary invariant moments.

- In (Skalski et al., 2008), the glottis is segmented using active contours and level set methods. The level set method is appropriate for modeling changing topologies since merging and breaking are made automatically, which is observed often in laryngeal images.

- In (Demeyer et al., 2009), the authors propose a framewise glottis segmentation strategy using Region Growing. They find the frames with the most opened glottal gap (keyframes) to start the segmentation. Then, the seed points are determined using the maximal response of a Laplacian of Gaussian filter and the threshold is computed iteratively based on the mean value of the glottis. However, the presence of the artifacts with very low gray values (i.e. outer image regions near the image border areas) can mislead the proposed method and produce inappropriate starting frames.

- The authors in (Mendez et al., 2009) estimate the motion of the vocal folds using the Wiener estimator. The Wiener estimator produces a smooth vector field which is used as a reference for a neighborhood algorithm which eliminates the pixels with less motion. Lastly, a threshold segmentation algorithm is applied to segment the glottal gap. This algorithm has problems to segment the glottis when the vocal folds are closed, and when there is not enough motion between consecutive frames.

- The authors in (Moukalled et al., 2009) use a pair of open curves to segment the glottis, one for the left and the other for the right fold. The inconvenience with this procedure is that it requires adjusting some parameters twice per video. The first one is used to initiate the snake and the second one to verify the segmentation before propagating to the remaining stages.

- In (Zhang et al., 2010) the authors integrate the features of the Lagrange interpolation, differentiation, and Canny detector to segment the glottal gap. First, the frames are smoothed using a Lagrange interpolation. Then, a differentiation is computed along the frames' rows to determine the extreme values of the image intensity as well as the vocal folds edges (the left fold corresponds to a minimal and the right fold a maximal). Lastly, the Canny detector is used to compute a continuous representation of the vocal folds edges.

- In (Cerrolaza et al., 2011), the authors present an automatic glottis segmentation approach using Deformable Shape Models. The approach starts with an

initial coarse segmentation by means of the Region Growing technique. The seed points are determined based on a simple linear relationship between the average gray level of the image and the optimal seed points obtained from the training examples. Lastly, the non-glottal regions are eliminated using the reliability score factor from the trained shape models.

- The authors in (Aghlmandi and Faez, 2012) propose an Edge-Based method which combines morphological operations and Hough transform. The Hough transform connects the lines that are not detected correctly by the morphological operations. Lastly, all the edges that do not belong to the glottis are not included in the final segmentation.

- The authors of (Elidan and Elidan, 2012) present a method that uses a global energy which allows to jointly consider the individual contour evolution in each frame. The global energy promotes the temporal consistency between the segmentations in each frame.

- In (Yan et al., 2012) the authors perform three steps to segment the glottal gap. First, a rough segmentation is performed by global thresholding and followed by the detection of an ellipse-shaped region that approximates the glottal geometry. Secondly, the parameters of the ellipse are estimated using PCA. Lastly, the snake method is applied using the estimated ellipse as an initial contour.

- In (Karakozoglou et al., 2012) a local region-based framework is used to guide an active contour algorithm. The foreground and background are modeled in terms of small regions with constant intensities that depend on their means. The active contour model uses a level set based procedure which allows to split and merge the vocal folds edges.

- In (Blanco et al., 2013), the frame with the minimal glottal opening (reference frame) is chosen manually as a starting point. Then, a binary difference is computed between the reference frame and each frame of the video sequence where the pixels with values greater than a threshold obtained empirically are defined as ROI. Lastly, the minimum gray value within the ROI is used to segment the glottal gap.

- In (Chen et al., 2013), the vocal folds edges are segmented separately by following the paths with the largest absolute gradient along the posterior-anterior commissures. The cost function used is based on the mean and standard deviation of the gray level, and in the gradient distribution of the images.

- The authors in (Andrade-Miranda et al., 2013) propose a fully automatic procedure to segment the glottal gap based on a gradient vector flow. The gradient vector flow creates a vector field over the image which is used for the initialization process and evolution of an active contour algorithm.

- In (Pinheiro et al., 2014), the user selects an arbitrary number of points inside the glottal gap to estimate its mean color intensity. Then, the glottis is separated from the remaining regions using an adaptive thresholding method, which is based on the statistical relationship between each pixel and its neighbords.

- The authors in (Ko and Ciloglu, 2014) present a novel illumination model based on the mean intensity distribution along the longitudinal cross section of the center of the glottis. Then, the new histogram distribution is modeled by a Gaussian Mixture Model (GMM) and the estimated GMM is used to isolate the glottis from the background. However, this method neglects the global drift and therefore does not provide any type of motion compensation.

- The authors in (Gloger et al., 2015) propose a fully automatic method to segment the glottis using local color and shape information. They divide the approach in three modules: training, recognition and segmentation. In the training, 60 different glottis shapes are manually segmented, and a set of descriptors are computed. The recognition module is designed to recognize, delineate and determine the optimal starting glottis regions. The last module segments the glottis based on properties of the previous frame. Hence, the glottis is continuously tracked within vibration cycles of the video by a frame-by-frame-wise segmentation technique.

- The authors in (Schenk et al., 2014) propose a framework that consists of three steps: pre-processing, ROI and seed region detection, and glottis segmentation. The preprocessing deals with problems like non-homogeneous background, illumination artifacts and global drift are dealt with. Then, a ROI and seed regions are automatically computed. Lastly, the generated seed regions are used as initialization for a 3D Geodesic active contour segmentation.

5.5 Discussion

Currently, the task of identifying the glottal gap is carried out by semi-automatic methods. In this context, and with the exponential growth of computer power and the constant improvement of the algorithms used for image processing, the hard task of automatically segmenting the glottal gap has achieved a dramatic advancement. However, many of the techniques found in the literature still have weaknesses that make them impractical in a clinical environment, in which the automatization and reliability are fundamental.

The most common techniques reported in the literature to segment the glottal gap are based on Thresholding, Region Growing and Deformable Models methods. The studies based on Thresholding assume that the glottis has darker intensity levels than the vocal fold tissues (Yan et al., 2006; Mehta et al., 2013). However,

the laryngeal images often have low contrast and heterogeneous profiles. Hence, selecting a global threshold results in an erroneous delimitation of the glottal gap, since the intensity distribution is not bimodal. On the other hand, the studies based on Region Growing requires a solid criterion for the seed selection and relatively well-delimited edges in order to converge towards the glottal gap. Furthermore, the algorithms segment objects with inhomogeneous regions into multiple sub-regions, resulting in over-segmentation (Lohscheller et al., 2007; Pinheiro et al., 2014). With respect to the Deformable Models, they have the advantage to couple appropriately to non-rigid and amorphous contours by an iterative minimization of an energy function. However, the initialization process is not a trivial task. Therefore, many authors use manual procedures to initialize the active contours (Palm et al., 2001; Marendic et al., 2001; Moukalled et al., 2009). Lastly, most of the studies do not take into account the temporal dimension of the problem and they do not consider that the glottis corresponds to less than 25% of the total image, so each frame is treated individually leaving aside the information obtained from the previous frames.

An accurate detection of the glottal gap along time is a very important task to objectively characterize the vibratory patterns of the VF. This is usually carried out synthesizing different representations such as Vibration Profiles (VP), Glottal Area Waveform (GAW), Glottovibrogram (GVG), Phonovibrogram (PVG), among others, and extracting some important measurements as: the symmetry of vibration, the amplitude of vibration, mucosal wave, periodicity, etc. It is known that these parameters are correlated with voice quality and health condition, and help the specialist to evaluate the phonation in an objective way.

Part III

Contribution to the laryngeal High-Speed Video Processing

Chapter 6

Contribution to the Glottal Gap Segmentation

"Imagination is more important than knowledge"

Albert Einstein

SUMMARY: In this chapter two algorithms are proposed to tackle the problem of the glottal gap segmentation. The first one, named as Glottal Segmentation Based on Watershed Transform and Active Contours (**SnW**), uses traditional image segmentation methods such as Region-Based and Deformable Models but adding the temporal information of the videos. The second one, receive the name of Glottal Segmentation Based on Background Subtraction and Inpainting (**InP**), and presents a quasi-automatic framework to accurately segment the glottal area, introducing several techniques never explored before in the state of the art. The method takes advantage of the possibility of a minimal user intervention for those cases where the automatic computation fails. Lastly, a set of validation guidelines are proposed in order to standardize the criteria of accuracy and efficiency of the segmentation algorithms.

6.1 Database Description

In order to demonstrate the strengths and limitations of the proposed methods, several experiments were carried out using two databases which are described below:

Database1 (DB1): This database was provided by Dr. Erkki Bianco and Gilles Degottex (Gilles, 2010) and consists of 22 videos. The LHSV system used to record the videos is a Richard Wolf-ENDOCAM 5562. The vocal folds were

filmed through a rigid endoscope which passes through the mouth, connected to a high-speed camera providing 4000 colored fps with a resolution of 256×256 pixels. The distances between the camera's head in the oropharynx and the vocal folds were variable. The database includes usual phonatory mechanisms such as M1 (the main mechanism used in speech), M2 (a laryngeal mechanism which can be found for high-pitched voice) and particular phonatory situations such as: breathy voice, tense voice, pressed voice, exhaled and inhaled fry. The database comprises only non-pathological phonation. The videos present different illumination levels, contrast, presence of nodules, partial occlusion of the glottis and lateral displacements of the camera.

Database2 (DB2): The database consists of 54 high-speed sequences: 36 females (67%) and 18 males (33%). Each sequence has 400 frames, thus the number of images analyzed is 21600. The recording took place at the ENT service of the Gregorio Marañon Hospital in Madrid. The videos were recorded during a sustained vowel phonation, including in some cases the vocal onset. The high-speed sequences were acquired using the camera system WOLF HRES ENDOCAM 5562 and a rigid endoscope with angle of view of 70°. The light source was the AUTO LP 5132 and all the videos were recorded in color. The sampling rate was 4000 fps and the spatial resolution of 256×256 pixels. The distances between the camera's head in the oropharynx and the vocal folds were variable. The database includes voiced sounds in laryngeal mechanism M1, lesions in the vocal folds (polyps and nodules), paralysis, paresis, postoperative papillary thyroid cancer, patients with multinodular goiter and postoperative diplophonia. The recordings present different illumination levels, contrast, partial occlusion of the glottis and lateral displacements of the camera.

6.2 Glottal Segmentation Based on Watershed and Active Contours

SnW method performs an automatic segmentation of the glottis. The algorithm identifies the ROI which is iteratively updated to be tolerant for camera displacements. In this way, a robust initialization for each frame is obtained. Finally, a procedure that combines watershed and active contours is used to delineated the glottal gap.

SnW method is divided into five main modules: 1) image enhancement, $I_{NLT}(\mathbf{x},t)$; 2) ROI detection; 3) first region merging, $I_{JND}(\mathbf{x},t)$; 4) correlation regions merging, $I_{cor}(\mathbf{x},t)$; and, 5) post-processing, $SnW(\mathbf{x},t)$. Each of these modules generates an intermediate result that is used for the subsequent step. Figure 6.1 summarizes graphically the different steps of the process, and the following subsections detail the procedures followed.

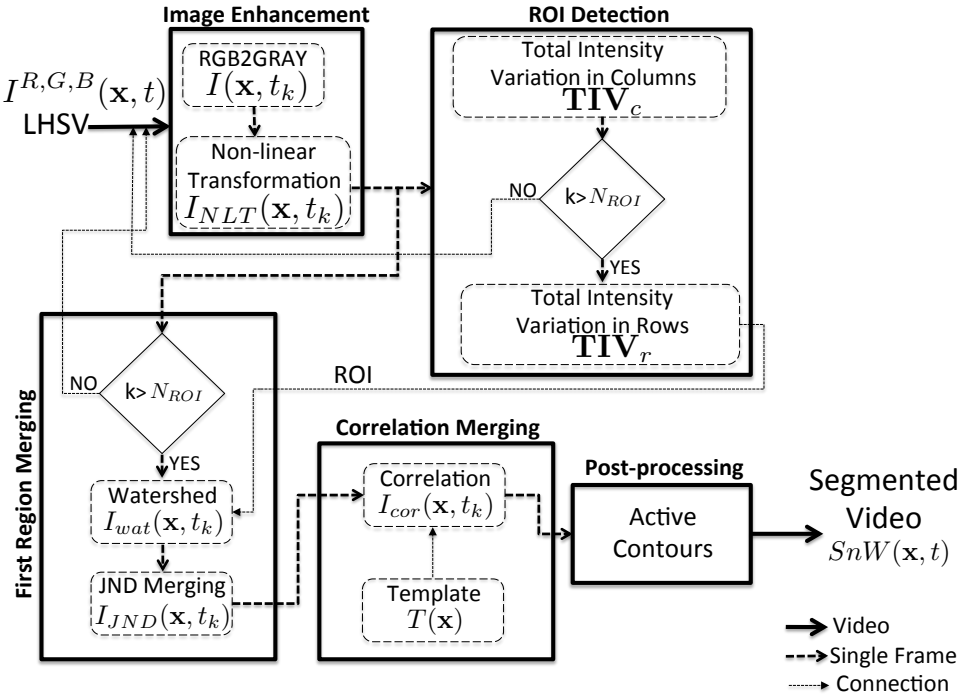


Figure 6.1: Graphic Representation of the different steps followed to segment the glottal gap: image enhancement, ROI detection, first region merging, correlation merging and post-processing. In this case, to differentiate from an arbitrary segmentation $I_{seg}(\mathbf{x},t)$, the final glottal delineation is denoted as $SnW(\mathbf{x},t)$.

6.2.1 Image Enhancement

Firstly, it is necessary to convert the original RGB sequence $I^{R,B,G}(\mathbf{x},t)$ to a grey scale through a transformation according to the model YIQ (Russ, 2002). After such conversion, the luminance Y, is used to generate the new video sequence in the grey scale $I(\mathbf{x},t)$. Then, a similar procedure as the one proposed in (Aghlmandi and Faez, 2012; Skalski et al., 2008) based on non linear transformation is followed (eq 6.1).

$$I_{NLT}(\mathbf{x},t_k) = \begin{cases} 255 & \forall x_i, y_j \mid I(x_i,y_j,t_k) > \overline{L}_j \\ 255 \times \left(\dfrac{I(x_i,y_j,t_k)}{\overline{L}_j} \right)^{\zeta} & \forall x_i, y_j \mid I(x_i,y_j,t_k) \leq \overline{L}_j \end{cases} \qquad (6.1)$$

$$\overline{L}_j = \frac{1}{m\beta} \sum_{i=1}^{m} I(x_i,y_j,t_k)$$

where $I(x_i,y_j,t_k)$ denotes the gray intensity in the pixel $\mathbf{x}_{ij}$; $\overline{L}_j$ is the mean of lighting levels in row j of the image; m is the number of columns in the image; β is an adjustable factor for increasing or reducing the contrast; and ζ is a coefficient

that is commonly set to 1.8. The β parameter is crucial to improve the contrast since wrong values produce results in which it is hard to distinguish between the glottis and the surrounding tissues or in other cases loss of glottis information. The decision of which β is the best option to enhance laryngeal images depends on a trade-off between contrast and information loss.

In order to validate the reliability of the non-linear transformation, several parameters have to be adjusted and some justifications need to be done. Firstly, it is necessary to justify the selection of the enhancement method considering subjective and objective criteria. The quality of the image enhancement techniques is difficult to assess since evaluating enhancement techniques is still an open problem. In (Tian and Kamata, 2008) an interesting framework was proposed combining three measures: Peak-Signal-to-Noise-Ratio (PSNR), Edge Overlapping Ratio (EOR) and Mean Segment Overlapping Ratio (MSOR), corresponding to three image features including intensity, edge, and segment. The PSNR is used to describe the intensity changes before and after enhancement and can be computed as (eq 6.2):

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) \tag{6.2}$$

where MSE is the mean square error between the intensity of the original image $I(x_i, y_j)$ and the intensity of the enhanced image $I_{NLT}(x_i, y_j)$ (eq 6.3).

$$MSE = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\|I(x_i, y_j) - I_{NLT}(x_i, y_j)\|^2 \tag{6.3}$$

EOR measures how close are the edges maps of $I(x_i, y_j)$ and $I_{NLT}(x_i, y_j)$. The edges maps are computed using any of the edge detectors found in the literature with the same setting in both images. The EOR is computed by the following equation (eq 6.4):

$$EOR = \frac{|E_{NLT}(x,y) \cap E(x,y)|}{|E(x,y)| + (|E_{NLT}(x,y)| - |E_{NLT}(x,y) \cap E(x,y)|)} \tag{6.4}$$

where $E(x,y)$ and $E_{NLT}(x,y)$ are the edges maps for the original and enhanced image, respectively, and $|\,.\,|$ denotes in this case the cardinality, in other words, the number of edge pixels. Lastly, MSOR describes how similar are the segmentation of $I(\mathbf{x}, t_k)$ and $I_{NLT}(\mathbf{x}, t_k)$ by comparing the overlapping of the different segments found. Suppose that after segmenting $I(\mathbf{x}, t_k)$, by using any segmentation algorithm, it has m_a segments $A = \{a_1, a_2, \cdots, a_{m_a}\}$ and $I_{NLT}(\mathbf{x}, t_k)$ has n_b segments $B = \{b_1, b_2, \ldots, b_{n_b}\}$. Then, MSOR is computed by equation eq 6.5 as:

$$MSOR(A,B) = \frac{1}{n_b}\sum_{a\in A}\max_{b\in B}\left(\frac{|(b \cap a)|}{|a| + (|b| - |(b \cap a)|)}\right) \tag{6.5}$$

The objective measure proposed in (Tian and Kamata, 2008) is applied to 110 images, extracted from the 22 videos (DB1). Then, considering the literature, three

enhancement methods are compared: anisotropic with FFT (Mendez et al., 2009), CLAHE (Karakozoglou et al., 2012) and non-linear transformation (Aghlmandi and Faez, 2012; Skalski et al., 2008). The non-linear transformation is tested with different values of β with an incremental step of 30 from 100 up to 300. The results obtained are presented in Figure 6.2 and summarized in Table 6.1. The first graphic describes the intensity changes before and after enhancement (PSNR); the second describes the similarity between edges (EOR); and, lastly, MSOR describes the similarity between regions. For LHSV, well defined edges and well delimited regions (EOR, MSOR) should be prioritized to facilitate the latter segmentation step. After analyzing the objective results and considering also a subjective evaluation based on visual inspection of the contrast over 110 images (see some examples in Figures 5.3 and Figure 6.3), the non-linear transformation with parameter $\beta = 200$ is chosen because it keeps a good balance between objective and subjective visual inspection.

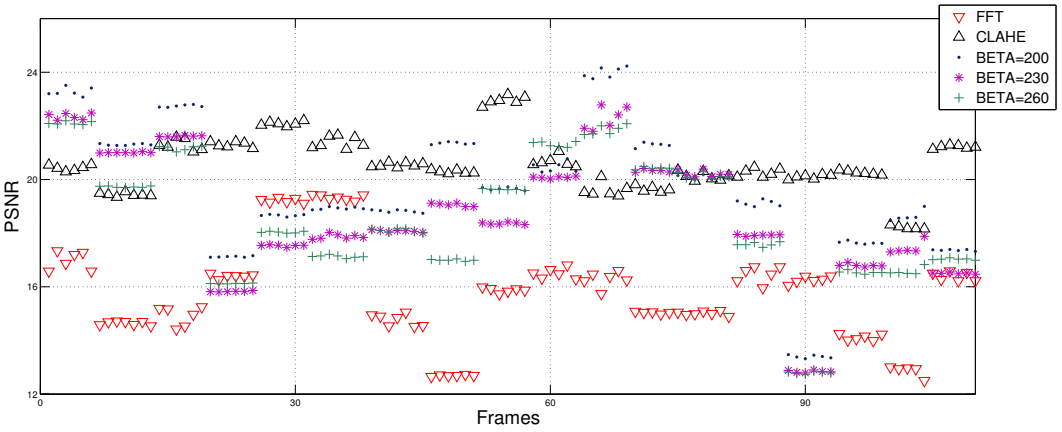
	FFT	CLAHE	$\beta = 140$	$\beta = 170$	$\beta = 200$	$\beta = 230$	$\beta = 260$
PSNR	15,80	20,57	30.06	24,55	19,66	18,69	18,52
EOR	0,11	0,34	0,59	0,51	0,46	0,32	0,20
MSOR	0,18	0,11	0,11	0,13	0,17	0,14	0,12

Table 6.1: Summary of the results reported in Figure. 6.2 for the different image enhancement techniques used.

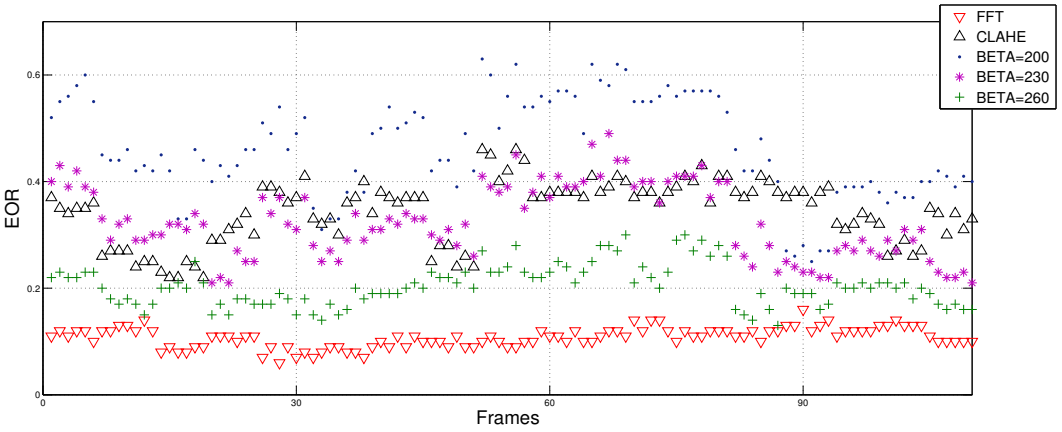
6.2.2 ROI Localization

Since the displacements of the glottis are small between consecutive frames, images taken at consecutive time instants are strongly correlated among them. Thus translation movements in a short period of time are almost null. However, due to the involuntary movements of the camera or the patient, the recordings present small displacements of the focus that are more significant as the number of evaluated frames increases. Considering the aforementioned, establishing a criterion based on the change of the spatial intensity profile to detect the ROI each N_{ROI} frames is a good choice. The squared area to be tracked is selected adaptively based on the variations of the image intensity and the inter-frame disparity for an appropriated set of frames, reducing the effect of the transversal shifts. By taking advantage of the continuous light source used to record the LHSV, the area with the largest variability within the image can be identified. This is done by analyzing the cumulative intensity variation of each frame in the x and y coordinates and, at the end, the area with the highest variability in time is identified as the glottis.

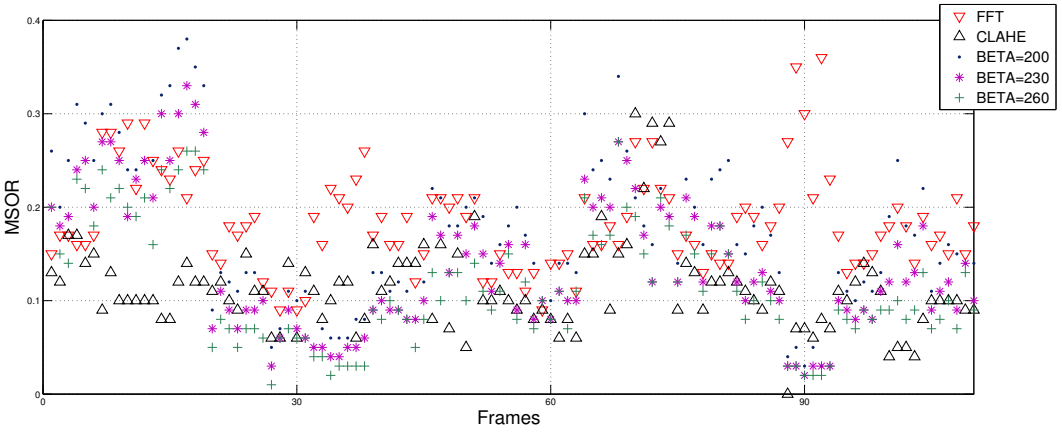
When delineating the ROI, it is also important to consider the periodic reflections (highlights) that could appear in the image and that would increase the size of the ROI. However, the non-linear transformation done in the image enhance-



(a) PSNR comparison



(b) EOR comparison



(c) MSOR comparison

Figure 6.2: Comparison of the pre-processing algorithms. The objective evaluations applied to 110 HSDI images extracted from the 22 videos (DB1): (a) PSNR graph; (b) EOR graph; (c) MSOR graph.



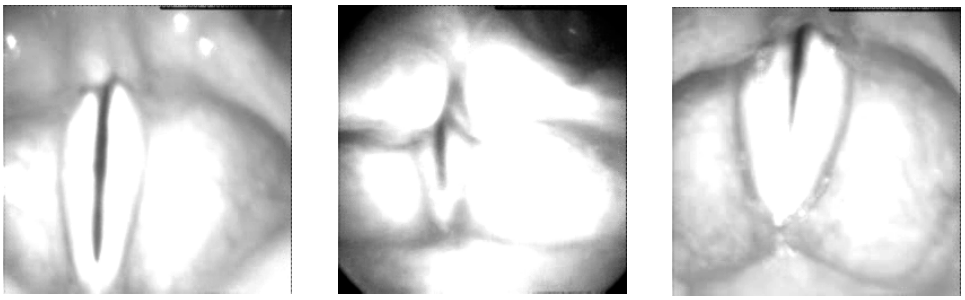
(a) $\beta = 140$



(b) $\beta = 170$



(c) $\beta = 200$



(d) $\beta = 230$

Figure 6.3: Visual representation of the non linear transformation using different values of β.

ment step mitigates the influence of flashes as it has already been demonstrated in previous studies (Aghlmandi and Faez, 2012; Skalski et al., 2008). Figure 6.4 summarizes the complete procedure followed to obtain the ROI and the concepts of total intensity variation in columns (**TIV**$_c$), and the total intensity variation in rows (**TIV**$_r$) are introduced bellow.

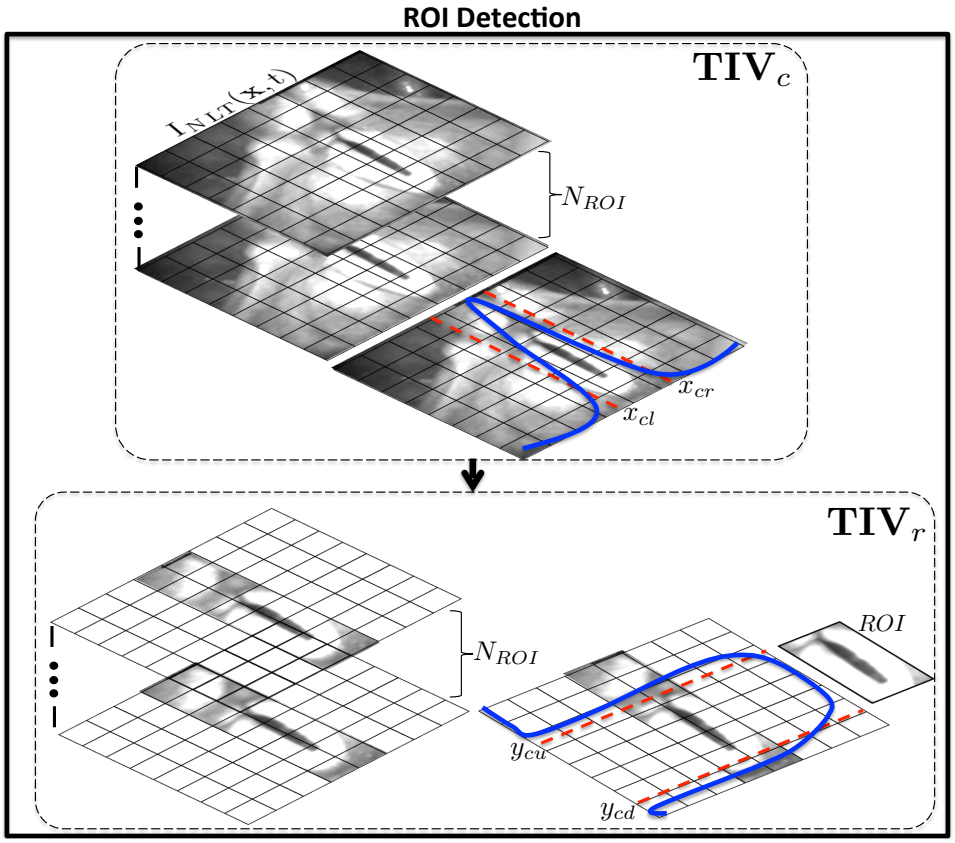


Figure 6.4: ROI localization. Upper panel: procedure to obtain **TIV**$_c$; bottom panel: procedure to compute **TIV**$_r$ and the final ROI.

Total Intensity Variation in Columns (TIV$_c$)

The first intensity variations to be analyzed are those related to the columns of the laryngeal images. The reason to start the analysis in the columns stands on the fact that the main axis of the glottal gap is usually located in a quasi-vertical position (with a slope of more than 30 degrees with respect to the horizontal axis). Hence, the information arising from the cumulative intensity variation in the horizontal axis is more significant than in the vertical one. In order to obtain the Total Intensity Variation in Columns (**TIV**$_c$), it is necessary to define two additional terms: the intensity variation matrix $S_c(x,t)$ and the average intensity variation vector

($\mathbf{AIV}_c$). In $S_c(x,t)$, each row represents the intensity variation of the columns for each frame. The eq 6.6 describes the mathematical procedure to compute $S_c(x,t)$.

$$S_c(x,t) = \frac{\sum_{j=1}^{n} I(x_i, y_j, t_k)}{n} \qquad \begin{matrix} \forall x_i, 1 \leq x_i \leq m; \\ \forall t_k, 1 \leq t_k \leq N_{ROI}; \end{matrix} \tag{6.6}$$

where $I(x,y,t)$ is the LHSV sequence with its respective x, y and t coordinates, n and m are the number of rows and columns of each frame respectively. Lastly, N_{ROI} is the number of frames that are used to find the ROI, this value is adjustable with a value not exceeding the maximum number of frames in the video N.

The average intensity variation $\mathbf{AIV}_c(x)$ (eq 6.7) is a vector in which each of its elements represents the horizontal intensity variation for the N_{ROI} frames evaluated. Finally, the total intensity variation in columns $\mathbf{TIV}_c$ (eq 6.8) is computed through analysis of the intensity variation of each frame with respect to the average intensity variation of the N_{ROI} frames by means of the Mean Absolute Error (MAE). For the ROI problem the most interesting points are those reporting the highest error.

$$\mathbf{AIV}_c(x) = \frac{\sum_{k=1}^{N_{ROI}} S_c(x_i, t_k)}{N_{ROI}} \qquad \forall x_i, 1 \leq x_i \leq m; \tag{6.7}$$

$$\mathbf{TIV}_c(x) = \frac{\sum_{k=1}^{N_{ROI}} |S_c(x_i, t_k) - \mathbf{AIV}_c(x_i)|}{N_{ROI}} \quad \forall x_i, 1 \leq x_i \leq m; \tag{6.8}$$

The eq 6.8 represents the region with the largest variability in the N_{ROI} frames under consideration, and its behavior resembles to a Gaussian-like function whose center coincides with the main axis of the glottal gap (see Figure 6.4). In order to obtain the cut-off points on the x-axis, $\mathbf{TIV}_c$ is fitted to a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ using the Non-linear Least Squares method (Björck, 1996). The mean of the gaussian will be the column with the largest intensity change and the standard deviation will determine the size of the ROI. The cut-off points in the x-axis are obtained using eq 6.9.

$$x_{cl} = \mu_x - \kappa_x \sigma_x; \qquad x_{cr} = \mu_x + \kappa_x \sigma_x; \qquad TI_x = [x_{cl}, x_{cr}] \tag{6.9}$$

where x_{cl} and x_{cr} are the left and right cut-off borders respectively, $\kappa_x \sigma_x$ is the standard deviation and TI_x is the tolerance interval that indicates the width of the ROI in the x-axis.

Total Intensity Variation in Rows ($\mathbf{TIV}_r$)

The Total Intensity Variation in Rows ($\mathbf{TIV}_r$) is computed following the same criteria used previously for $\mathbf{TIV}_c$ but with slight differences, since $\mathbf{TIV}_r$ uses the

reduced area obtained in the previous step as a starting point, and further evaluates the variation in rows. The method used to find out the up and down cut-off points, y_{cu} and y_{cd} in the y-axis is analogue to its counterpart in the x-axis (eq 6.10). Then, the ROI is defined as the region enclosed by the pairwise points: (x_{cl}, y_{cu}) and (x_{cr}, y_{cd}).

$$y_{cu} = \mu_y - \kappa_y \sigma_y; \qquad y_{cd} = \mu_y + \kappa_y \sigma_y; \qquad TI_y = [y_{cu}, y_{cd}] \tag{6.10}$$

The $\mathbf{TIV}_r$ computation deals with two complex scenarios; the first is when the glottis is divided in two or more regions. This problem does not affect the normal performance of the ROI detection despite of the presence of extra valleys in the $\mathbf{TIV}_r$ since an average movement is computed for N_{ROI} frames reducing the effect of the valleys. The second scenario is even more demanding and corresponds to the presence of a glottal chink. Here, depending on the top and down cut-off points in the y axis (y_{cu} and y_{cd}) , some information in the posterior part could be lost. Nonetheless, this scenario does not commit the general performance of the algorithm since there is an optimal range for no loss of information, as will be shown in the next subsection. Figure 6.5 shows a LHSV in which both scenarios are presented.

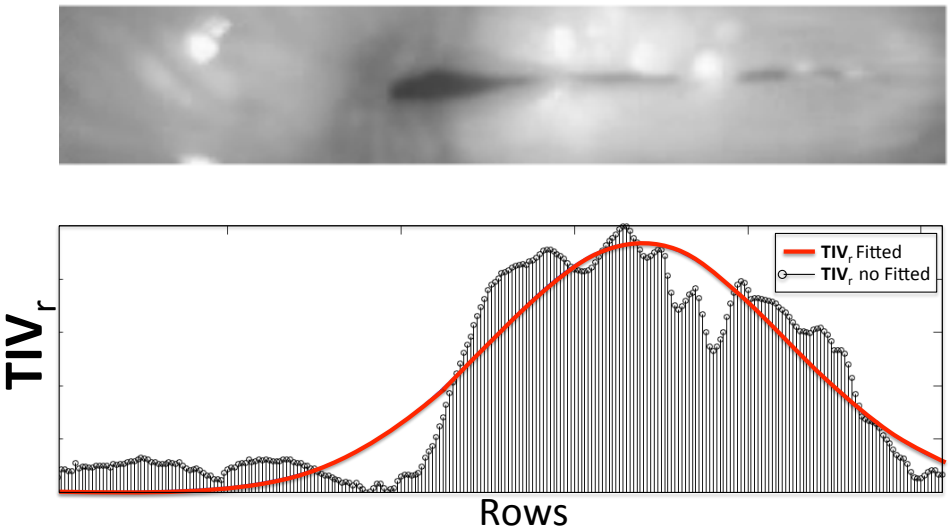


Figure 6.5: $\mathbf{TIV}_r$ for N_{ROI}=100 frames. The LHSV sequence has a glottal chink and the glottis is splitted in two parts, illustrating one of the most demanding cases. The frame is rotated in horizontal position for a better visualization.

Adaptive ROI for Motion Compensation

The displacements of the endoscope affect the alignment of the HSDI image pixels along time, leading to difficulties to track the dynamic characteristics of the la-

ryngeal structures. Thus, methods to compensate these displacements are needed. Deliyski (2005) argues about the importance of a procedure to compensate the distortion originated by the endoscope displacements in the synthesis of a VKG. Since it causes intermittent changes of the voicing pattern, it makes difficult the interpretation of the results. The movement of the vocal folds (70 - 400 Hz) is much faster than the one originated by the endoscope ($\sim$15 Hz), so the motion caused by the endoscope is indistinguishable in one glottal cycle. In order to clarify this, let us consider the case of the vocal folds vibrating with a fundamental frequency $f_o = 100\,Hz$ (period T_o, equal to 10 ms) and an endoscope displacement with a frequency of $f_e = 15\,Hz$ (period T_e, equal to 66.67 ms). In this scenario the movement of the endoscope would be noticeable after at least 6 glottal cycles. With this in mind, the ROI have to be recomputed every N_{ROI} frames to compensate the camera motion and to reduce the false detections.

The value of N_{ROI} is undoubtedly one of the most important parameters of the proposed methodology to accurately detect the ROI and for the motion compensation. This parameter could take any value between 1 and the total number of frames N. However a value close to 1 limits the possibility to characterize the motion that is present. Contrariwise, if the value of N is close to the total number of frames, non valuable information is added to the ROI increasing the false detections. Additionally, N_{ROI} provides reliability against the camera and/or patient displacement. With a small value of N_{ROI} the algorithm becomes more robust against movements, avoiding the effects of those displacements that are related to the endoscope.

In order to demonstrate this fact, Figure 6.6 shows 4 frames in different instants of the LHSV with their corresponding $\mathbf{TIV}_c$ plots. The $\mathbf{TIV}_c$ was computed from $I(\mathbf{x},t)$ and without using the gaussian fitting in order to emphasize the effect of the displacements. The instants of time under consideration are: t = 30, 1000, 2000, 2975. It is possible to check by simple inspection, that increasing N_{ROI} deviates $\mathbf{TIV}_c$ from the gaussian pattern. The explanation to this phenomenon is related with the horizontal motion of the camera during the recording. This causes additional peaks that do not belong to the ones produced by the vocal folds motion, so an erroneous gaussian fitting and wrong cut-off points are generated. An important conclusion obtained from these examples is referred to the average position of the glottis: the lobe with the maximum peak in $\mathbf{TIV}_c$ will be the average position of the glottal gap.

Through experimentation, it is observed that the minimum N_{ROI} to achieve a robust ROI is that containing at least one complete glottal cycle. For instance, with a high-speed data rate ($LHSV_{rate}$) of 6665 frames per second (0.15 ms per frame), and a fundamental frequency of phonation of $\sim$236 Hz (period equal to 4.23 ms) the minimum value of N_{ROI} to be chosen is approximately 28. Figure 6.7 shows three different images belonging to three different LHSV recordings, each of them with their respective $\mathbf{TIV}_c$ plots. The plots present a high complexity due to the presence of occlusions caused by the laryngeal structures. However, this fact does not affect the proposed procedure since it analyses the average intensity

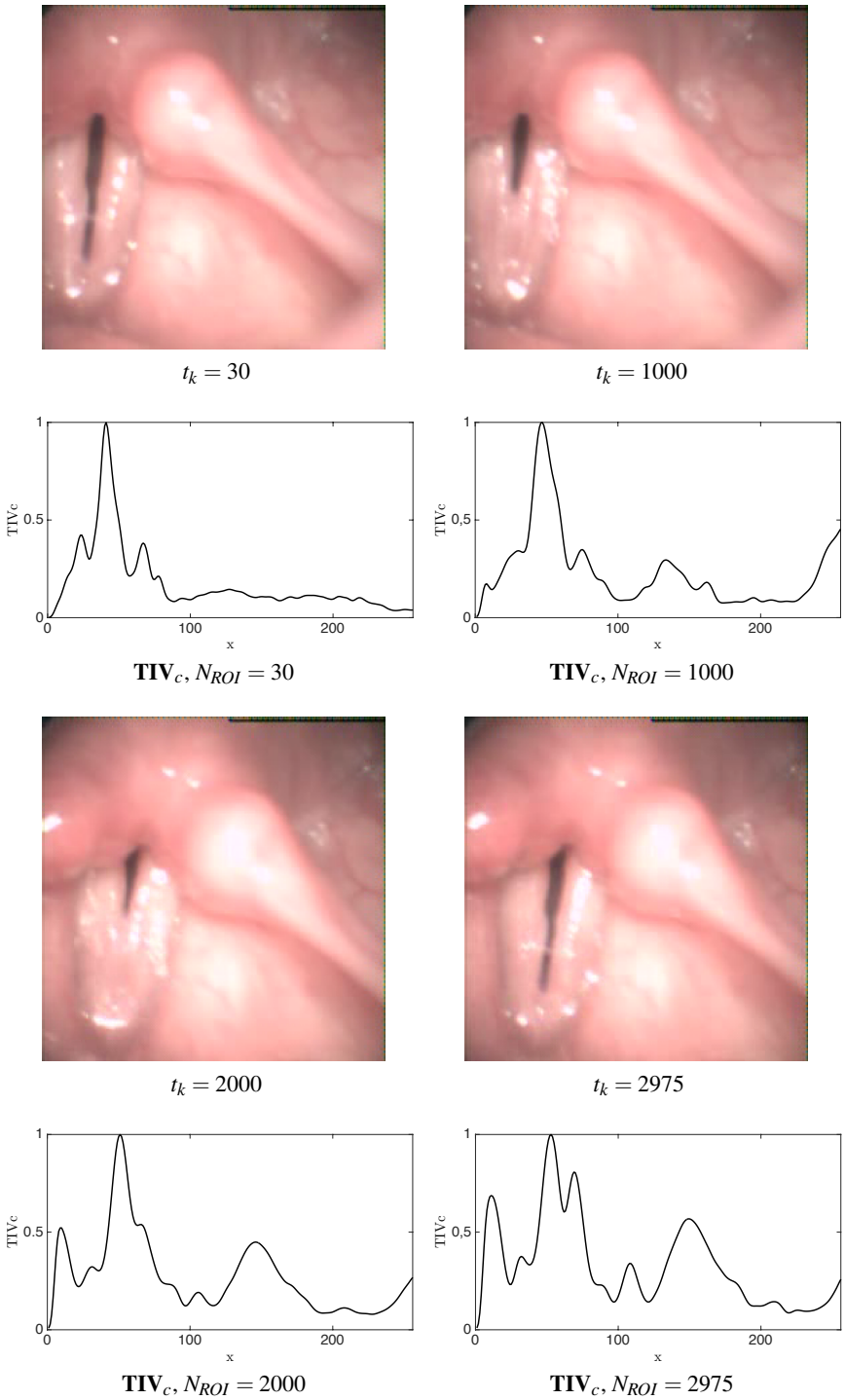


Figure 6.6: Effect of the transversal motion in $\mathbf{TIV}_c$. The importance of recomputing the ROI is illustrated plotting different $\mathbf{TIV}_c$ without gaussian fitting for different values of N_{ROI}: N_{ROI}=30, N_{ROI}=1000, N_{ROI}=2000, and N_{ROI}=2975.

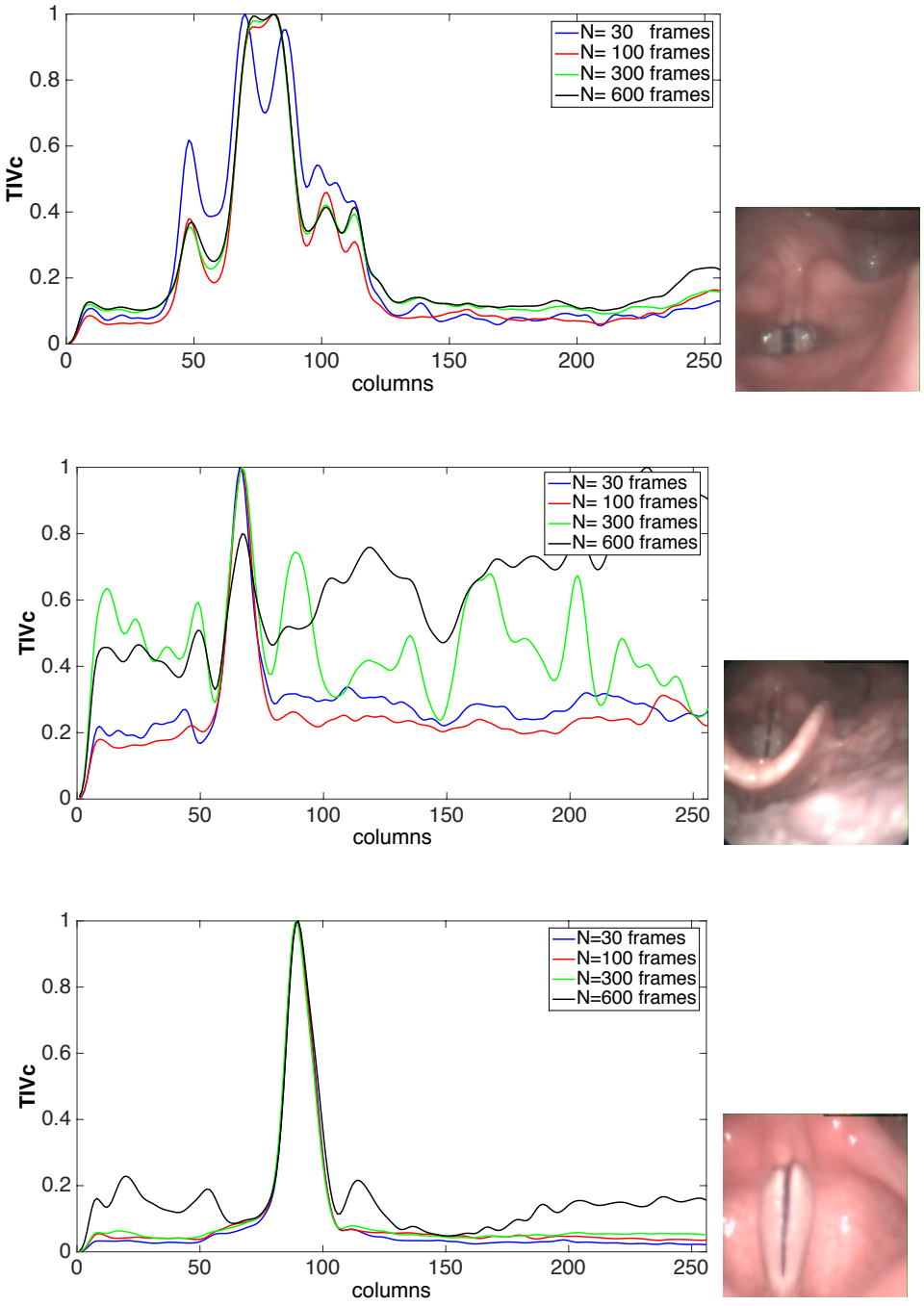


Figure 6.7: Evaluation of the effect of different N_{ROI} settings. Graphical representation of the varaition of N_{ROI} for three different LHSV sequences $\mathbf{TIV}_c$ for $N_{ROI} = 30$(blue line), $N_{ROI} = 100$(red line), $N_{ROI} = 300$(green line), $N_{ROI} = 600$(black line).

position of the glottal gap for a set of N_{ROI} consecutive frames. In the first row of Figure 6.7, with $N = 30$, there are two peaks that represent the two vocal folds, and the valley in the middle is the glottal space during the opening phase of the glottal cycle, meaning that a complete cycle has not been reached ($\sim$41 frames per cycle). Meanwhile, with $N_{ROI} = 100$ (red plot) a complete glottal cycle is included, and the width of the Gaussian completely covers the glottis. Increasing N_{ROI} leads to many fluctuations close to the maximum peak. The second example in Figure 6.7 ($\sim$38 frames per cycle) is even more demanding because there is a significant occlusion and the glottal gap is not clearly visible. Nonetheless, due the minimal presence of lateral movements, the maximum peak is conserved for all the different values of N_{ROI}, but in concordance with the increase of the number of frames, $\mathbf{TIV}_c$ starts losing its Gaussian-like shape. The third row in Figure 6.7 ($\sim$17 frames per cycle) shows an ideal case without camera movements, in which the performance of the algorithm is not affected by the choice of N_{ROI}. The optimal N_{ROI} is in the range between one glottal cycle (G_{C_o} eq 6.11) and one motion cycle of the endoscope (MC_e eq 6.12).

$$G_{C_o} = T_o / LHSV_{rate} \tag{6.11}$$

$$MC_e = T_e / LHSV_{rate} \tag{6.12}$$

$$N_{ROI} \in [G_{C_o}, MC_e] \tag{6.13}$$

Lastly, it is necessary to determine the cut-off points, analyzing the overlapping between $\mathbf{TIV}_c$ and $\mathbf{TIV}_r$ curves from the LHSV sequence. The coefficients that regulate the cut-off points are κ_x and κ_y. Both coefficients can be set indifferently, depending on whether the ROI is to be more bounded in the x or y axis. For instance, a TI_x of 99.7% ($\mu_x \pm 3\sigma_x$) would include non-relevant information to the ROI. Conversely, if TI_x is set to 68% ($\mu_x \pm \sigma_x$) the ROI would over-adjust to the glottis area, causing loss of information. Based on the experimentation around the 22 LHSV sequences and considering no loss of information, a good tradeoff to fix κ_x and κ_y is in the range $[2,3]$. In this work both coefficients were fixed with a value of 2.5.

6.2.3 First Region Merging

After reducing the size of the area to be analyzed, the next step is the identification of the glottis boundaries. The algorithm used for this purpose is the one described in (Osma-Ruiz et al., 2008), which is based on a watershed transform computed over the gradient of the images combined with a Just Noticeable Difference (JND) based region merging. The watershed transform creates a set of well delimited objects, opening the possibility to identify their individual features and statistics to find out those that belong to the glottis. Nevertheless, the results of the watershed

transform is disappointing, due to the fact that thousands of objects arise when only the glottis is expected. This problem is solved by pre-processing the gradient image with a thresholding, and by merging objects with one or more features in common (Bleau and Leon, 2000; Hernandez et al., 2005). A thresholding with a value of 10 is applied to the magnitude of the image gradient and those pixels with a value below 10 are assigned to 0, so they are converted into minima that can only belong to the internal part of any region. This simple thresholding removes most of the regions that appear due to the intrinsic noise originated by the LHSV acquisition and the ones produced for the tissues texture. The threshold applied to the gradient image has been chosen to avoid removing significant edges of the image. The watershed transform $I_{wat}(\mathbf{x},t_k)$ is computed by eq 6.14, where $\mathbf{T}_{wat}$ represents the watershed operator and $\|\nabla I_{NLT}(\mathbf{x},t_k)\|$ the magnitude of the gradient of the image enhanced.

$$I_{wat}(\mathbf{x},t_k) = \begin{cases} \mathbf{T}_{wat}[\|\nabla I_{NLT}(\mathbf{x},t_k)\|] & \text{if } \|\nabla I_{NLT}(\mathbf{x},t_k)\| > 10 \\ 0 & \text{otherwise} \end{cases} \tag{6.14}$$

On the other hand, the merging criterion is based on a fixed threshold over a cost function that decides if two regions have to be merged or not. The chosen cost function is calculated using the JND of different gray levels of the image and has been theoretically defined in (Day-Fann and Ming-Tsong, 2003). The JND represents the sensibility of the human visual system to perceive the changes of luminance. The human visual system is not able to differentiate certain changes in luminance. For instance, assuming that the luminance is expressed in tonalities of gray, the visual system can not distinguish between a gray level of 80 and a gray level of 85. Additionally, this insensibility do not follow a linear behaviour, being the eye less sensitive to the change of luminance in the dark levels than in the bright ones. The function for evaluating the visibility threshold of the JND is described by eq 6.15 and graphically established in Figure 6.8.

$$JND(I(\mathbf{x}_{ij})) = \begin{cases} D_0 \cdot \left(1 - \sqrt{\frac{I(\mathbf{x}_{ij})}{127}}\right) + 3 & \text{if } I(\mathbf{x}_{ij}) \leq 127 \\ \gamma \cdot (I(\mathbf{x}_{ij}) - 127) + 3 & \text{otherwise} \end{cases} \tag{6.15}$$

$I(\mathbf{x}_{ij}) \in [0,255]$ is the intensity value of a given pixel $\mathbf{x}_{ij}$, and the parameters D_0 and γ depend on the viewing distance between a tester and the monitor. D_0 denotes the visibility threshold when the background is 0 and γ denotes the slope of the line that models the JND visibility threshold function at higher background luminance. The values of D_0 and γ are set to 17 and $3/128$, based on the subjective experiments done in (Chou and Li, 1995). The merging cost function used to fuse the region is computed by eq 6.16:

$$F_c = [|mR_1 - mR_2| - \min(JND(mR1), JND(mR2)) + 255] \tag{6.16}$$

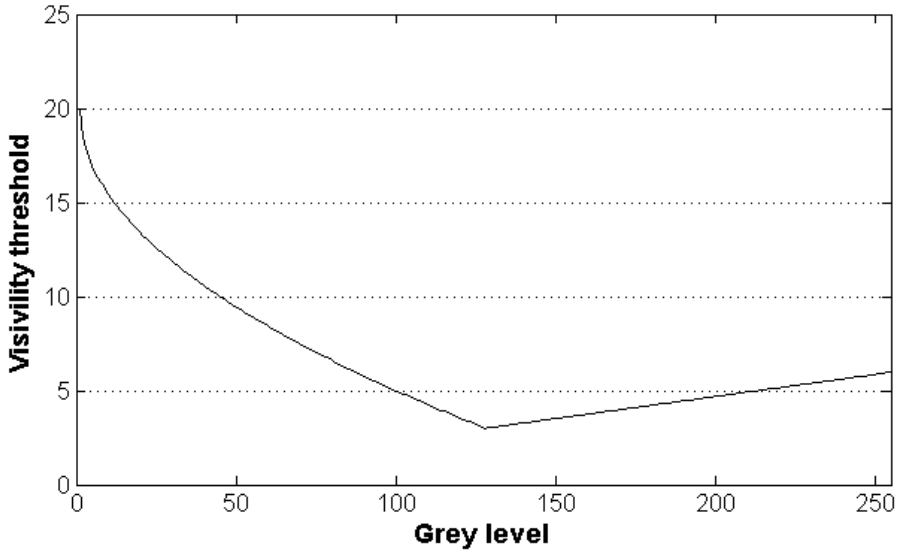
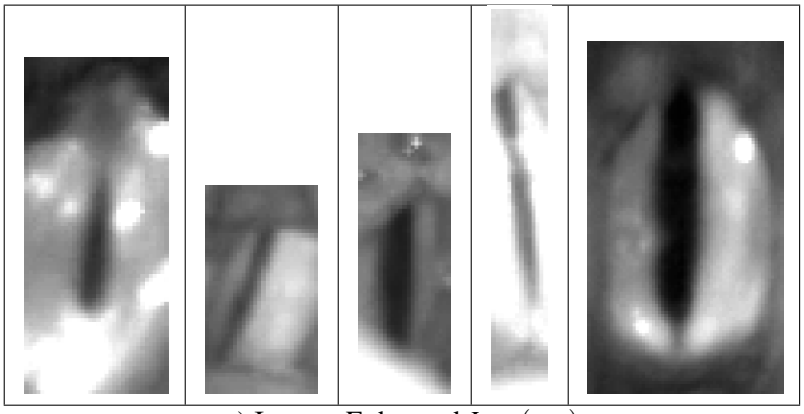


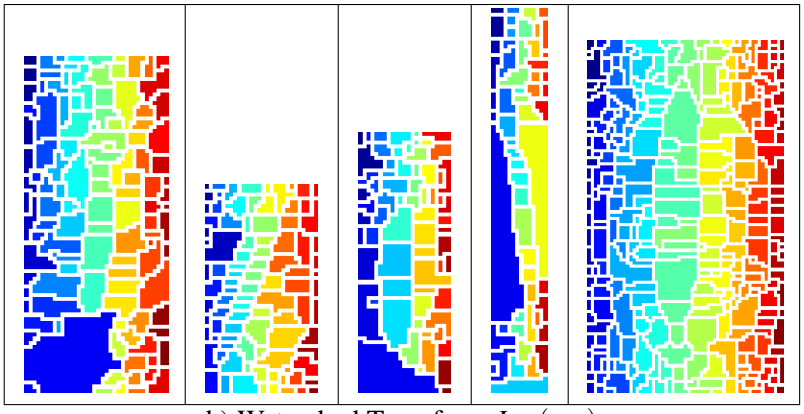
Figure 6.8: Visibility threshold of the human visual system as a function of the grey level in the image.

where $mR1$ and $mR2$ are the average values of the gray level of two neighbour regions, and min is the minimum JND between the average of both regions. The goal of the merging cost function is to combine all the regions in which F_c is below a specific threshold. This is because, under this merging threshold, the human vision system considers that the average grey level of the basins is the same, so it is not able to discriminate between them. Based on the experimentation and considering the thresholding used in (Osma-Ruiz et al., 2008) for LVS videos, the threshold value is set to 265 for all the LHSV.

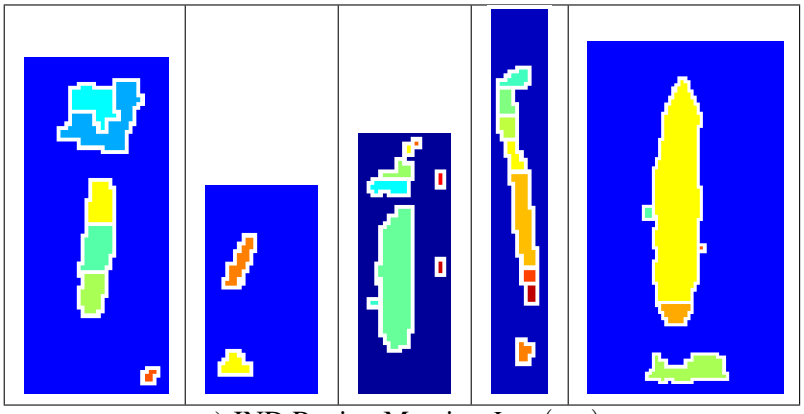
Lastly, the JND function is slightly modified in order to reduce the brighter regions. Firstly, the N/N_{ROI} frames with the maximal glottal opening (also known as keyframes, $I_{key}(\mathbf{x},t)$) are computed based on eq 3.9. Since, the intensity distribution of the glottis and background in the $I_{key}(\mathbf{x},t)$ has a cuasi-bimodal behavior, it is feasible to reduce part of the meaningless information by Otsu's method (Otsu, 1979). This algorithm performs automatically a clustering-image thresholding assuming that the image contains two classes of pixels (glottis and background). For all the values over the Otsu's threshold, a F_c of 265 has been assigned so the bright regions of the image (background) belong to an unique region and the amount of information is drastically reduced. However, there is still over-segmentation caused by the intensity disparity inside the glottis and also by the presence of regions that despite not being part of the glottis, they have some of their intensity features. Figure 6.9 shows some examples of each of the step followed in the first region merging procedure, including its final result $I_{JND}(\mathbf{x},t_k)$.



a) Images Enhanced $I_{NLT}(x,y)$



b) Watershed Transform $I_{wat}(x,y)$



c) JND Region Merging $I_{JND}(x,y)$

Figure 6.9: Illustration of the first region merging. First row: image enhanced; second row: watershed transform after thresholding the magnitude of the gradient; third row: the watershed transform after the JND cost fucntion.

6.2.4 Correlation Regions Merging

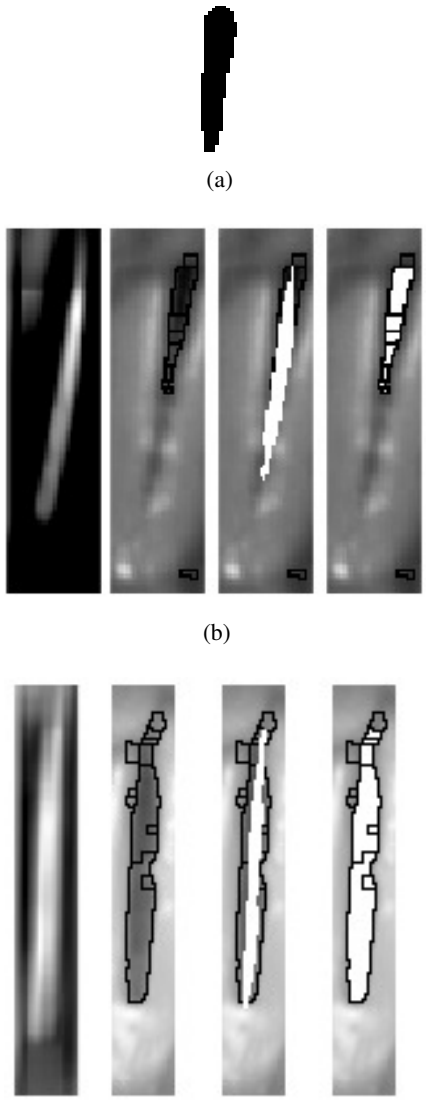
The next step consists of another merging process. The goal now is to join all the regions that correlate with a standard template obtained from the database.

The standard template $T(\mathbf{x})$ was obtained empirically based on manual segmentations carried out by one expert in all the frames of the database with the maximal glottal opening. The potential templates are built with white background and a black foreground. The white background acts like an edge enhancer to highlight the glottis contours. The test involves the use of different glottis shapes, resizing, warping and small rotations. After an intensively evaluation of all of these features, a standard template that better correlates with the available data in DB1 is obtained. The standard template resizes automatically depending of the ROI size, ensuring that is not affected by a different zoom of the glottis. The standard template obtained for a ROI of 40x148 has a size of 12x42 (see Figure 6.10a). This template is used as a baseline for the correlation merging.

The standard template is correlated with each frame using the Normalize Correlation Coefficient (CC) (Edwards, 1976), providing values within the range $[-1,1]$. If both images are absolutely identical the value is 1; if they are completely uncorrelated, 0; and if they are completely anti-correlated, -1 (for example if one image is the negative of the other). The CC has been selected due to its invariance with respect to the intensity and because its similitude matrix provides valuable information about the glottis and vocal folds position. The threshold for a good matching is established in 0.45 in the similitude matrix. Below this value the glottis is considered fully closed. The regions obtained by the cross-correlation are intersected with the results of the first merging process, and the overlapped objects are merged. The second region merging, $I_{cor}(\mathbf{x},t_k)$, is computed by eq 6.17 where $CC(\ .\ )$ represents the correlation coefficient operation and $T(\mathbf{x})$ is the standard template.

$$I_{cor}(\mathbf{x},t_k) = \begin{cases} 1 & \text{if } CC(I_{JND}(\mathbf{x},t_k),T(\mathbf{x})) > 0.45 \\ 0 & \text{otherwise} \end{cases} \tag{6.17}$$

Figure 6.10b and Figure 6.10c show the complete correlation and merging procedure carried out to detect the glottis. The first image represents the similitude matrix; the second represents the first region merging; the third one the overlapping between the cross-correlation and the first region merging; and, finally, the fourth represents the results of the second merging process. However, due to the inter-video variability the merging process could not be enough to obtain a reliable glottis segmentation. In order to clarify this idea, the last image in Figure 6.10b represents an example in which the glottis presents an irregular shape and it is not well delimited in the anterior part. This phenomenon appears due to the difficulties of establishing an unique criterion for the region merging (threshold of the cost function) even when there exists large inter video changes in the illumination conditions. For that reason, a post-processing stage is required to refine and smooth the segmentation obtained from the second merging region.



(a)

(b)

(c)

Figure 6.10: Merging steps. (a) Standard template found empirically based on manual segmentation, $T(\mathbf{x})$; (b) and (c) show from left to right: two different frames of two different sequences, similitude matrix, first region merging, cross-correlation overlapping and correlation region merging.

6.2.5 Post-Processing: Localizing Region-Based Active Contours

The experimentation carried out has shown that the anterior and posterior part of the glottis are not always accurately segmented during the correlation merging step, producing in some cases a wrong delineation of those regions (see Figure 6.10c). Therefore, a post-processing step that uses the result of the correlation regions

merging $I_{cor}(\mathbf{x}, t_k)$ as initialization for an active contours algorithm is proposed.

The active contour method proposed in (Lankton and Tannenbaum, 2008) models the foreground and background in terms of smaller local regions. This framework allows a correct conversion in cases of inhomogeneity, common in laryngeal images. The analysis of local regions leads to the construction of a family of local energies at each point along the initial curve. In order to optimize the local energies, each point of the curve is considered separately and moves to minimize the energy computed in its own local region.

The energies can be modeled in three different ways: the uniform modeling energy, the means separation energy and the histogram separation energy. We choose the Chan Vesel-model (Chan and Vese, 2001), which models the interior and exterior of a region as constant intensities represented by their means. Since the post-processing step is only a refined version of the previous steps, the number of iterations of the active contour and the radius of the local regions can be fixed without an extensive analysis of the database. A radius of 5 pixels is enough for a refined procedure. Meanwhile, 100 interactions ensure full convergence to the glottis. Figure 6.11 depicts 12 examples of segmented frames ($SnW(\mathbf{x},t)$) with their respective intermediate results.

6.2.6 Evaluation of the ROI Performance

It is hard to decide which method is the best for detecting the ROI and even harder to compare the performance between them since all have been evaluated with different databases and some of them are based on manual initialization (Palm et al., 2001; Marendic et al., 2001; Yan et al., 2006; Lohscheller et al., 2007; Zhang et al., 2010; Mehta et al., 2013; Pinheiro et al., 2014). However, in order to provide some subjective notions of the the results obtained, the algorithm described in (Karakozoglou et al., 2012) is implemented and compared with the proposal. The main reason for choosing (Karakozoglou et al., 2012) for comparison is based on the fact that it is also fully automatic, the final segmentation has great accuracy, and the framework followed to solve the segmentation presents some similarities with the **SnW** method.

After an in-depth analysis of the algorithm proposed in (Karakozoglou et al., 2012), two aspects need to be considered to improve its results. The first one is related to the choice of the area in the landmarks, since in pathological cases the glottis might be divided in two parts, and the algorithm might identify only one of them. This is depicted in the first and second row of Figure 6.12a. In both landmarks the glottis is splitted into two parts, which means that one of the regions will be discarded leading to an incomplete segmentation. The second drawback occurs when there are large artifacts with vertical orientation, like those depicted in the third row of Figure 6.12a. These artifacts are common due to reflections of the light inside the tube of the endoscope. Contrariwise, the methodology proposed for ROI detection solves the first problem computing the maximal intensity variation of a region, which means that it is not affected by glottis splitting. Regarding the

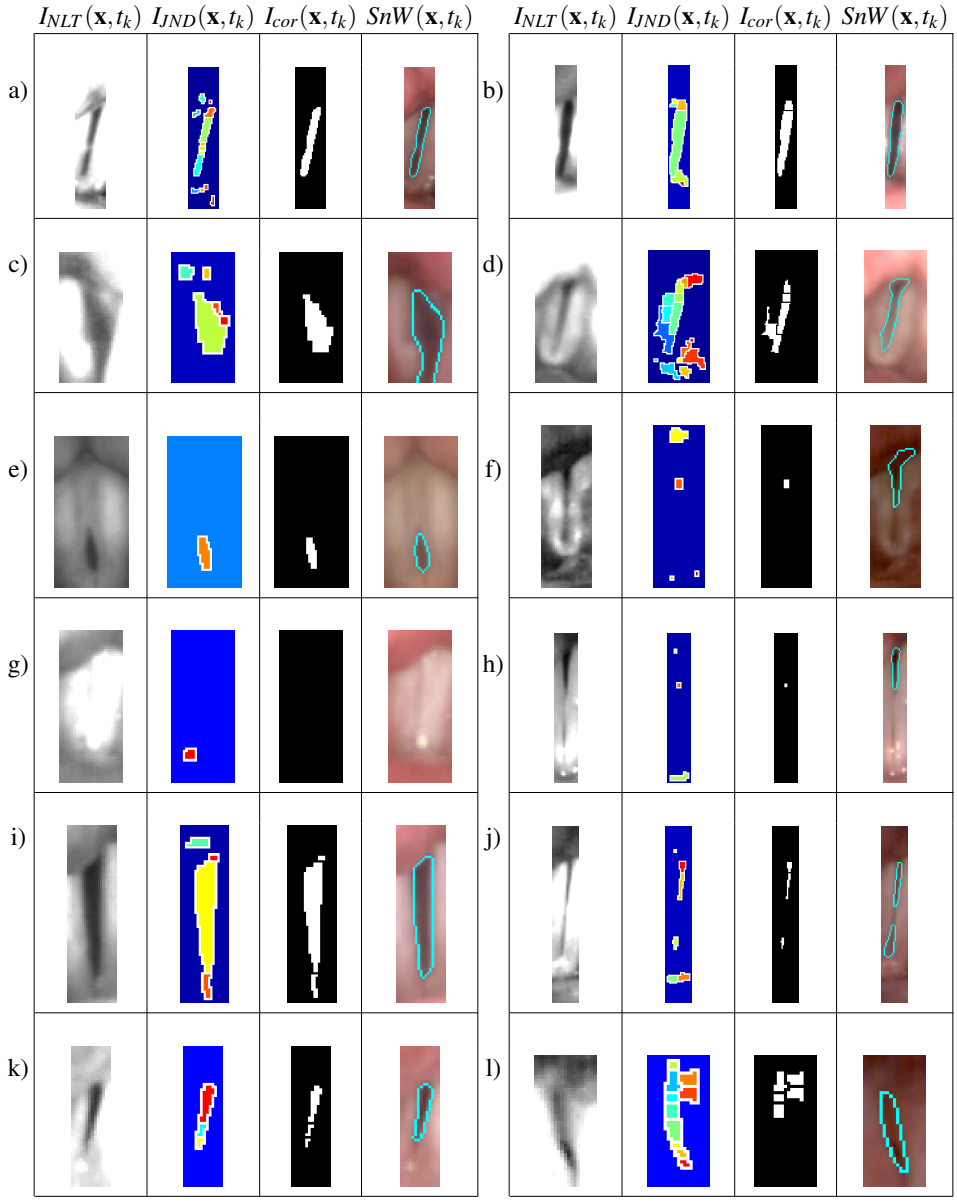
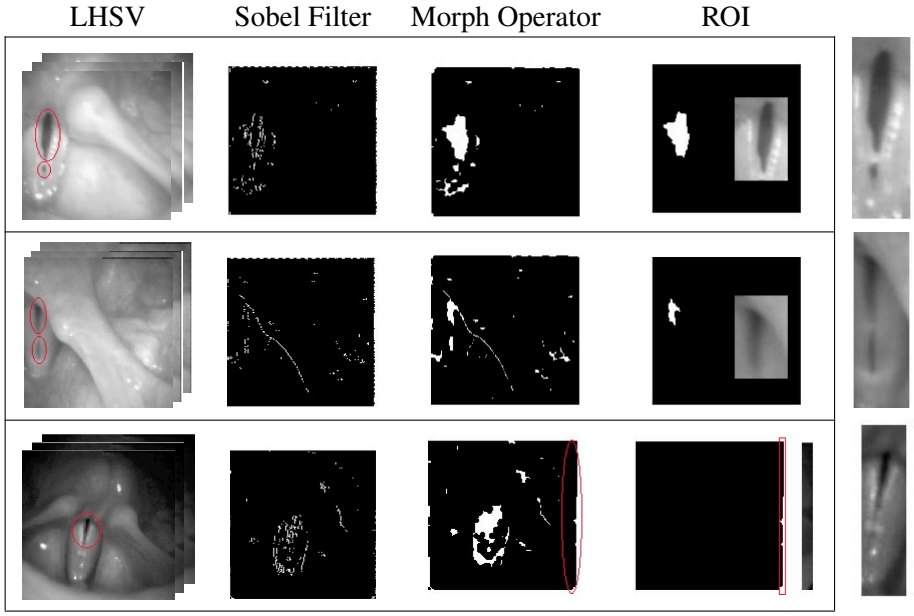


Figure 6.11: Complete methodology representation. From left to right: enhanced image $I_{NLT}(\mathbf{x},t_k)$; segmentation obtained after watershed and first region merging $I_{JND}(\mathbf{x},t_k)$; second region merging $I_{cor}(\mathbf{x},t_k)$; final delimitation of the glottis after 100 iterations $SnW(\mathbf{x},t_k)$.

second drawback: no false detections are produced since there are no changes in the intensity along time in such regions.

However, the approach based on intensity variation is not suitable to analyse all different kinds of phonatory conditions. For example, as the approach analyses the

variance of pixel intensities to identify the ROI it is not suitable to identify the glottis correctly during the voice onset interval, and in cases of vocal folds paralysis and paresis. In those cases the proposal of Karakozoglou et al. has a better performance. One way to tackle the problem of the voice onset is by back-propagating the information obtained during the steady phonation to the onset interval. As a matter of comparison with respect to the approach found in (Karakozoglou et al., 2012), the ROI detection obtained using intensity variation is depicted in Figure 6.12b. Additionally, some results of the ROI based on intensity variation are depicted in Figure 6.13 with its respective $\mathbf{TIV}_c$ and $\mathbf{TIV}_r$ plots.



a) Karakozoglou et al. (2012) b)

Figure 6.12: ROI detection using two approaches. a) ROI detection according to (Karakozoglou et al., 2012); b) final ROI obtained using the approach based on intensity variation.

6.2.7 Drawbacks of SnW Technique

A complete framework is proposed to automatically segment and track the glottal area from laryngeal high-speed video recordings over time. Initially the position of the glottis is identified using a variance criterion regarding the pixel intensity. From this information a region of interest (ROI) is obtained. Within this ROI, the glottis is segmented using a watershed approach. This intermediate segmentation result is subsequently corrected using a glottis template. The final result is obtained using region based active contours.

Despite the good performance of **SnW**, some of the parameters used were obtained empirically (Standard Template selection, Watershed Merging criteria,

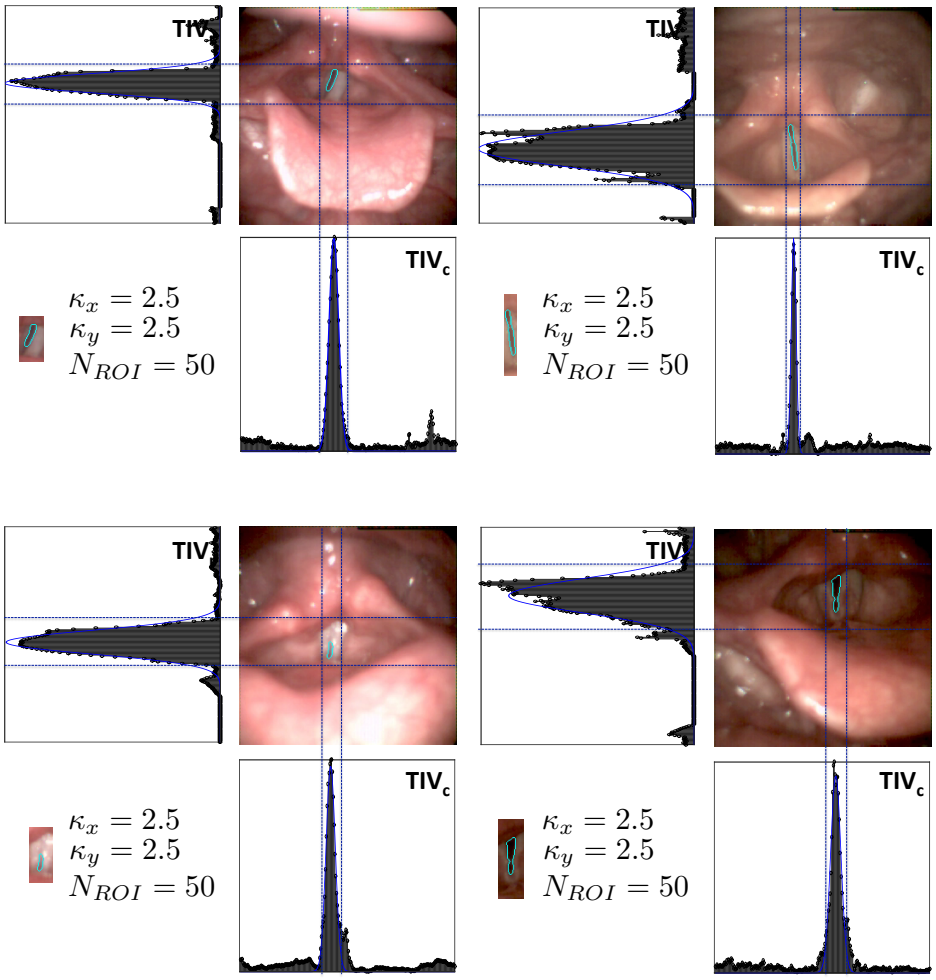


Figure 6.13: Examples of the results obtained using a ROI based on intensity variation with their respective $\mathbf{TIV}_c$ and $\mathbf{TIV}_r$ plots.

Correlation Thresholding) or involving a degree of compromise with the objective measures (wnhancement parameter β). The standard template fails when the glottal opening increases its size abruptly since the correlation among the template and the previous merging step produces an initialization smaller than the one expected. The solution to this problem could stand on increasing the number of iterations in the post-processing step.

Another problem is related to the enhancement method since it is difficult to generalise a value of β when different illumination level and contrast are present on the LHSVs. As a matter of fact, Figure 6.11f depicts a frame in which the flashes have not been completely eliminated after the enhancement step. There-

fore, an erroneous delineation of the vocal folds edges is produced in the posterior commissure of the glottis.

Additionally, other parameters such as the ones used for the watershed merging criteria were set subjectively. Therefore, more evaluation of the presented framework and also some degree of user intervention is needed to guarantee its applicability in a clinical environment.

6.3 Glottal Segmentation by Background Modeling and Inpainting

In view of the problems mentioned in section 6.2.7, a quasi-automatic method to accurately segment the glottal area is presented which introduces several techniques not explored before in the state of the art.

InP proposes a novel approach that smooths the textures of the background (laryngeal structures) and foreground (glottal gap), detects the ROI using the temporal intensity information, and segments the glottis by creating an adaptive background model. Furthermore to the automatic segmentation, the method provides the chance of a minimal user interaction to improve the results in those cases in which the results are not those desired.

InP method is divided into three main modules: 1) image enhancement; 2) ROI detection; and, 3) glottal gap delimitation. Each of these modules generates an intermediate result that is used for the subsequent step. Figure 6.14 summarizes graphically the different steps of the process, and the following subsections detail the procedures followed.

6.3.1 Image Enhancement

The method followed performs a color equalization for each instant t_k, followed by a procedure that isolates the specular reflections and a bilateral filtering. The three sub-steps are showed graphically in Figure 6.15 and described below.

6.3.1.1 Color Equalization

During the recording of a laryngeal video sequence, complex reflectance phenomena appears due to intrinsic surface properties. These reflectances appear because light source and viewing direction are almost identical. Thereby, wet mucosa surfaces perpendicular to the viewing direction are showed as white flashing spots (JungHwan et al., 2007). Therefore, the white flashing spots have to be highlighted via color equalization as a prior step to mitigate them.

Since histogram equalization is a non-linear process, each color channel ($I^R(\mathbf{x},t_k)$, $I^G(\mathbf{x},t_k)$ and $I^B(\mathbf{x},t_k)$) can not be equalized independently. Therefore, the equalization has to be applied in such a way that the intensity values are equalized without disturbing the color balance of the image. Thus, it is necessary to con-

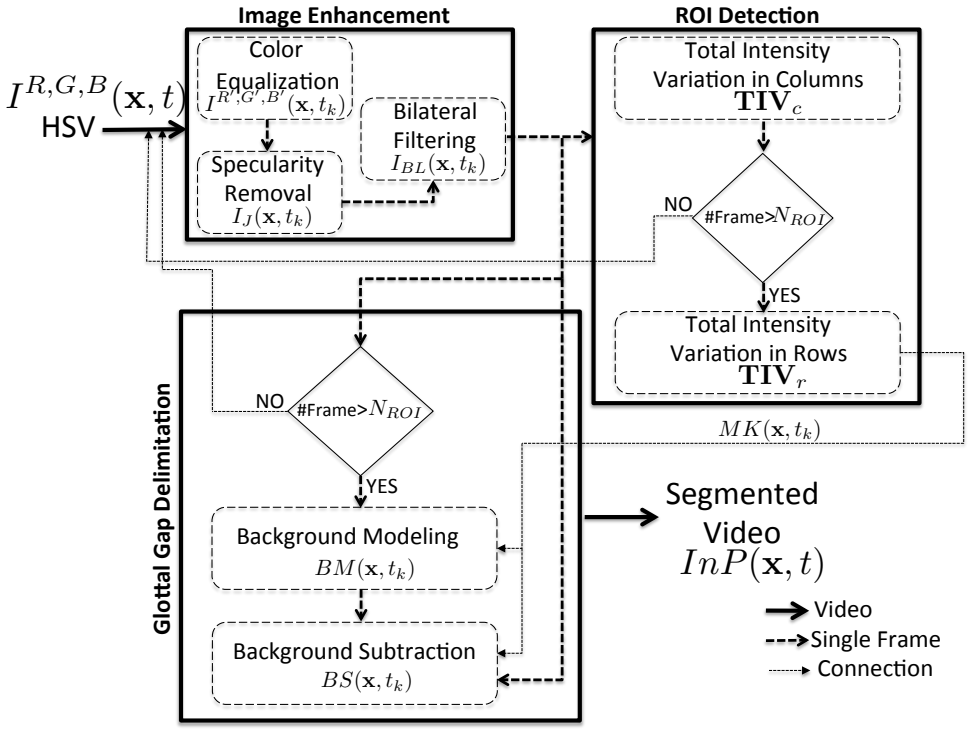


Figure 6.14: Graphic representation of the steps followed to segment the glottal gap.

vert $I^{R,G,B}(\mathbf{x},t_k)$ into another color space that separates intensity values from color components. For our purpose, the color space *YCbCr* was selected since it is defined by a transformation from an associated *RGB* color space that separates the luminance information, Y, from the chrominance Cb and Cr (see eq 6.18).

$$I^{Y,Cb,Cr}(\mathbf{x},t_k) = \mathbf{T}_{YCbCr}[I^{R,G,B}(\mathbf{x},t_k)] \tag{6.18}$$

where $I^{Y,Cb,Cr}(\mathbf{x},t_k)$ is the transformed *YCbCr* image, and $\mathbf{T}_{YCbCr}$ is the operator that maps the color spaces. Then, a transformation $\mathbf{T}_{HEq}$ is performed on the intensity plane Y, producing a new image with a flat histogram. Such transformation is a histogram equalization and is defined by eq 6.19.

$$I^{Y'}(\mathbf{x},t_i) = \mathbf{T}_{HEq}[I^{Y}(\mathbf{x},t_k)] = (L-1)\int_0^{lv} \mathscr{P}_{lv}(z)\,\mathrm{d}z \tag{6.19}$$

where $I^{Y'}(\mathbf{x},t_k)$ represents the equalized image, $I^{Y}(\mathbf{x},t_k)$ is the original intensity image, lv represents the different gray levels, $lv \in [0,L]$, L being the total number of gray levels in $I^{Y}(\mathbf{x},t_k)$, and $\mathscr{P}_{lv}$ the probability of an occurrence of a pixel at level lv. Lastly, the image in *Y'CbCr* space is taken back to the *RGB* color space

using a transformation $\mathbf{T}_{RGB}$ (eq. 6.20).

$$I^{R',G',B'}(\mathbf{x},t_k) = \mathbf{T}_{RGB}[I^{Y',Cb,Cr}(\mathbf{x},t_k)] \tag{6.20}$$

the image $I^{R',G',B'}(\mathbf{x},t_k)$ is the color equalized image in the *RGB* space, and $I^{Y',Cb,Cr}(\mathbf{x},t_k)$ is the color image equalized only on the intensity plane Y.

6.3.1.2 Specularity Removal

When the illuminant color is known and the reflectance of the surface can be represented with a dichromatic model, one simple way to isolate the specular reflection effects is by linearly transforming the *RGB* color space rotating its coordinate axes. This rotation is such that one of the axes becomes aligned with the direction of the effective source color $\mathbf{s} \in \mathbb{Z}^{+3}$. This transformation defines a new color space, which is referred as the *SUV* color space (Mallick et al., 2006). The *SUV* transformation is defined by eq 6.21 using a rotation matrix $\mathbf{R} \in \mathrm{SO}(3)$[1]. The rotation matrix satisfies the condition that $[\mathbf{R}]\mathbf{s} = [1,0,0]^T$.

$$I^{S,U,V}(\mathbf{x},t_k) = \mathbf{R}[I^{R',G',B'}(\mathbf{x},t_k)] \tag{6.21}$$

where $I^{S,U,V}(\mathbf{x},t_k)$ is the transformed image in the *SUV* color space. *SUV* is a source-dependent color space, since it depends on the effective source color vector of the image. It has two important properties: first, it separates the diffuse and specular reflection effects; second, the S channel encodes the entire specular component and an unknown fraction of the diffuse component, while the remaining two channels (U and V) are independent of specular invariants.

The aforementioned procedure was followed to eliminate the specularity. The source vector $\mathbf{s}$ was set up to $[255,255,255]$ that corresponds to white light in the *RGB* color space and the rotation matrix $\mathbf{R}$ was computed by aligning the R-axis with the source light $\mathbf{s}$. The $\mathbf{R}$ matrix is obtained by eq 6.22, where $(\theta_\mathbf{s},\phi_\mathbf{s})$ are the elevation and azimuthal angles of the source vector $\mathbf{s}$ in the RGB coordinate system. $\mathbf{R}_G(\theta_\mathbf{s})$ and $\mathbf{R}_B(\phi_\mathbf{s})$ are a clockwise rotation about the G-axis and B-axis by an angle $\theta_\mathbf{s}$ and $\phi_\mathbf{s}$, respectively.

$$\mathbf{R} = [\mathbf{R}_G(-\theta_\mathbf{s})][\mathbf{R}_B(\phi_\mathbf{s})] \tag{6.22}$$

Lastly, a monochromatic free specularity image, $I_J(\mathbf{x},t_k)$, is computed according to eq 6.23, combining the two pure diffuse channels $I^U(\mathbf{x},t_k)$ and $I^V(\mathbf{x},t_k)$.

$$I_J(\mathbf{x},t_k) = \sqrt{I^U(\mathbf{x},t_k)^2 + I^V(\mathbf{x},t_k)^2} \tag{6.23}$$

[1] SO(3) is the set of all orthogonal matrices of size 3 with determinant +1.

6.3.1.3 Bilateral Filtering

The monochromatic image $I_J(\mathbf{x},t_k)$ contains discontinuities in the surface across the diffuse information that has not been accurately propagated. To solve such problem, a post-processing step based on a bilateral filter (Paris et al., 2009) is performed.

The bilateral filter is defined by eq 6.24, where $I_{BL}(\mathbf{x},t_k)$ stands to the filtered image, $I_J(\mathbf{x},t_k)$ is the image to be filtered, Ω is a window centered in $\mathbf{x}$, $\mathbf{x}'$ represents a pixel in the Ω window, $\mathcal{W}_\mathbf{x}$ is a normalization term, $\mathcal{G}_{\sigma_r}$ is a Gaussian function for smoothing the differences in intensities, and $\mathcal{G}_{\sigma_s}$ is a Gaussian function for smoothing the differences in coordinates.

$$I_{BL}(\mathbf{x},t_k) = \frac{1}{\mathcal{W}_\mathbf{x}} \sum_{\mathbf{x}'\in\Omega} I_J(\mathbf{x}',t_k)\mathcal{G}_{\sigma_r}(\|I_J(\mathbf{x}',t_k) - I_J(\mathbf{x},t_k)\|)\mathcal{G}_{\sigma_s}(\|\mathbf{x}'-\mathbf{x}\|) \qquad (6.24)$$

eq 6.24 has the advantages of a simple formulation (each pixel is replaced by an average of its neighbors), it depends only on two parameters (σ_s and σ_r) and it can be used in a non-iterative manner which makes the parameters easy to set since their effect is not cumulative over several iterations.

Since the bilateral filter was included only to smooth small features, the value of σ_s was set to a constant value of 8 and σ_r to a constant value of 15. Both values performed consistently well for all our experiments, thus their setting do not require manual intervention.

The result of the contrast enhancement step is a gray scale image, $I_{BL}(\mathbf{x},t_k)$, free of specularity and with a marked difference in the contrast of the laryngeal structures and the glottis.

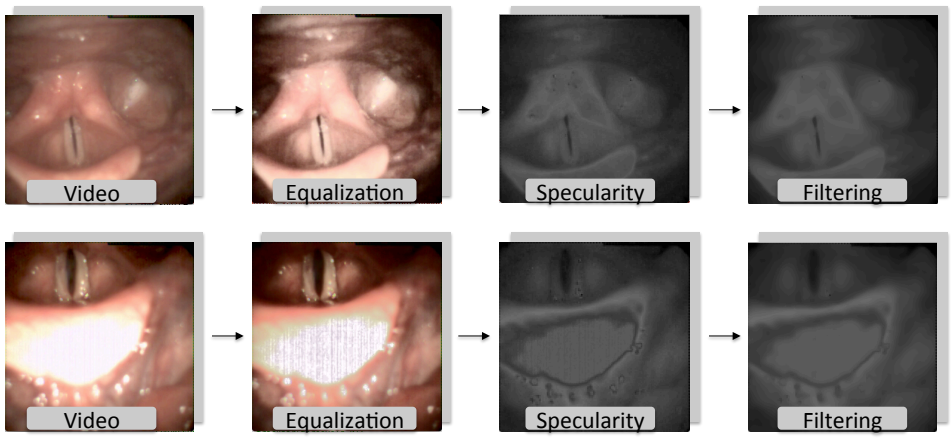


Figure 6.15: Contrast enhancement procedure for two different LHSV. From left to right: original LHSV; color equalization; image free of specularity; and image after bilateral filtering

6.3.2 ROI Localization

The procedure used for the ROI detection is the same as the one described in section 6.2.2 with the difference that it is applied to the $I_{BL}(\mathbf{x},t_k)$ image. The ROI is defined as the region enclosed by the pairwise points: (x_{cl},y_{cu}) and (x_{cr},y_{cd}) and is also used to generate a binary mask $MK(\mathbf{x},t_k)$ (eq 6.25) which will be employed for the glottal gap delimitation step.

$$MK(\mathbf{x},t_k) = \begin{cases} 1 & \text{if } x \in [x_{cl},x_{cr}] \text{ and } y \in [y_{cu},y_{cd}] \\ 0 & \text{contrariwise} \end{cases} \tag{6.25}$$

6.3.3 Glottal Gap Delimitation

This module is based on an inpainting algorithm which is used to create a background model. The background model is extracted with the glottis occluded to later perform a subtraction for each incoming frame. In order to exemplify the procedure, the results of the steps followed to obtain the final segmentation are shown in Figure 6.16.

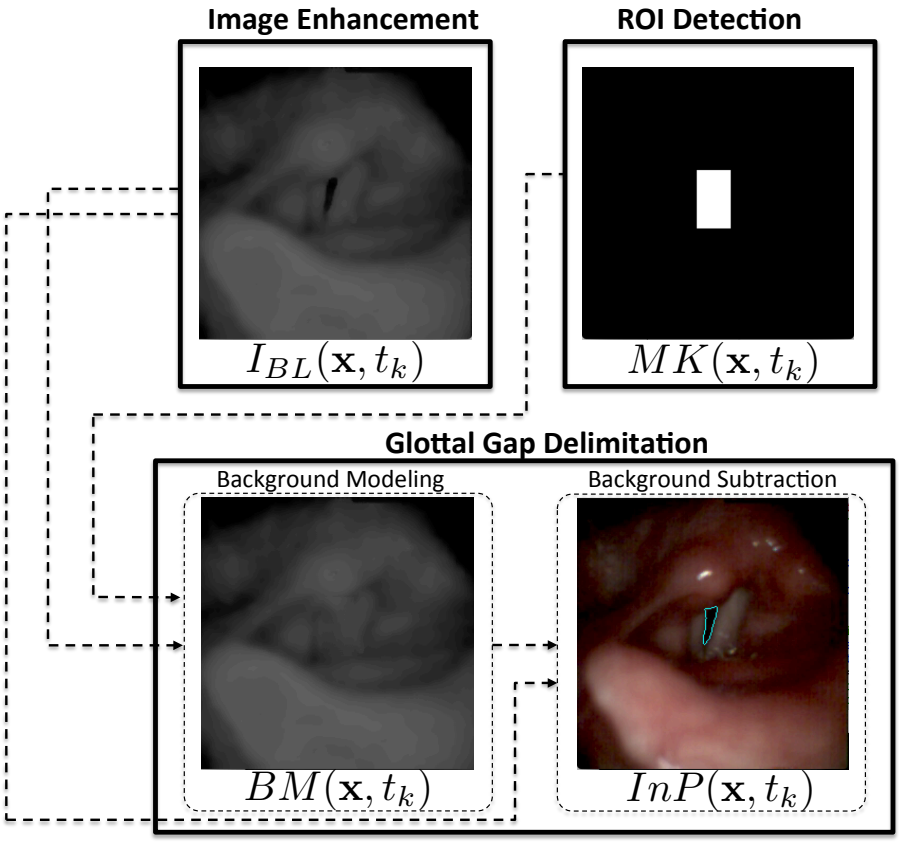


Figure 6.16: Complete framework of the glottal gap delimitation.

6.3.3.1 Background Modeling Using Inpainting

In this method, the object detection is achieved by making a representation of the scene called *background model*, and then finding deviations from it for each incoming frame. Any significant change in the image region from the background model is supposed to represent a moving object (Yilmaz et al., 2006).

Let us define the background model as a flat image, $BM(\mathbf{x},t_k)$, composed only by the information of the tissues that surrounds the glottal gap. $BM(\mathbf{x},t_k)$ is computed for each t_k using an inpainting procedure that combines a binary mask $MK(\mathbf{x},t_k)$ and the enhanced image $I_{BL}(\mathbf{x},t_k)$.

The inpainting technique is guided by the assumption that pixels in the known and unknown parts of the image share the same statistical properties or geometrical structures. The method used is based on (Telea, 2004), which has the immediate advantage of well-developed theoretical and numerical results. This technique propagates the local image structures of $I_{BL}(\mathbf{x},t_k)$ from the external part to the interior of the mask $MK(\mathbf{x},t_k)$, "imitating" the gesture of a professional painting restorer. $MK(\mathbf{x},t_k)$ is updated every N_{ROI} frames to compensate the drift of the camera.

6.3.3.2 Background Subtraction

A subtraction operation is computed between the enhanced image, $I_{BL}(\mathbf{x},t_k)$ and the background model, $BM(\mathbf{x},t_k)$ (eq 6.26) which is carried out only inside the ROI.

$$BS(\mathbf{x},t_k) = I_{BL}(\mathbf{x},t_k) - BM(\mathbf{x},t_k) \tag{6.26}$$

Since $BM(\mathbf{x},t_k)$ has higher intensities than $I_{BL}(\mathbf{x},t_k)$, the negative values obtained from the subtraction will correspond to the glottal gap. Therefore, an initial threshold, Th_{INC}, is introduced to identify the noticeable motion produced by the vocal folds movement. If $BS(\mathbf{x},t_k)$ is lower than Th_{INC}, the motion of the pixel $\mathbf{x}$ is considered significant. In our case, a value of -8 performed consistently well for all the experiments. However, Th_{INC} is not enough to segment accurately the glottal gap so another thresholding procedure is used. This second threshold is denoted as Th_{ADJ} and is obtained by an iterative procedure described in (Ridler and Calvard, 1978).

Lastly, in order to smooth and eliminate spurious information in the glottal gap segmentation, it was necessary to perform basic morphological operations to remove isolated pixels and fill holes. Algorithm 1 explains in detail the procedure followed to compute the background model and the background subtraction, and Figure 6.17 shows some results of the glottal gap segmentation.

6.3.4 User Intervention: Including a Manual ROI

There are two scenarios in which the automatic segmentation fails to correctly delimitate the vocal folds edges. The first occurs when there are partial or no motion

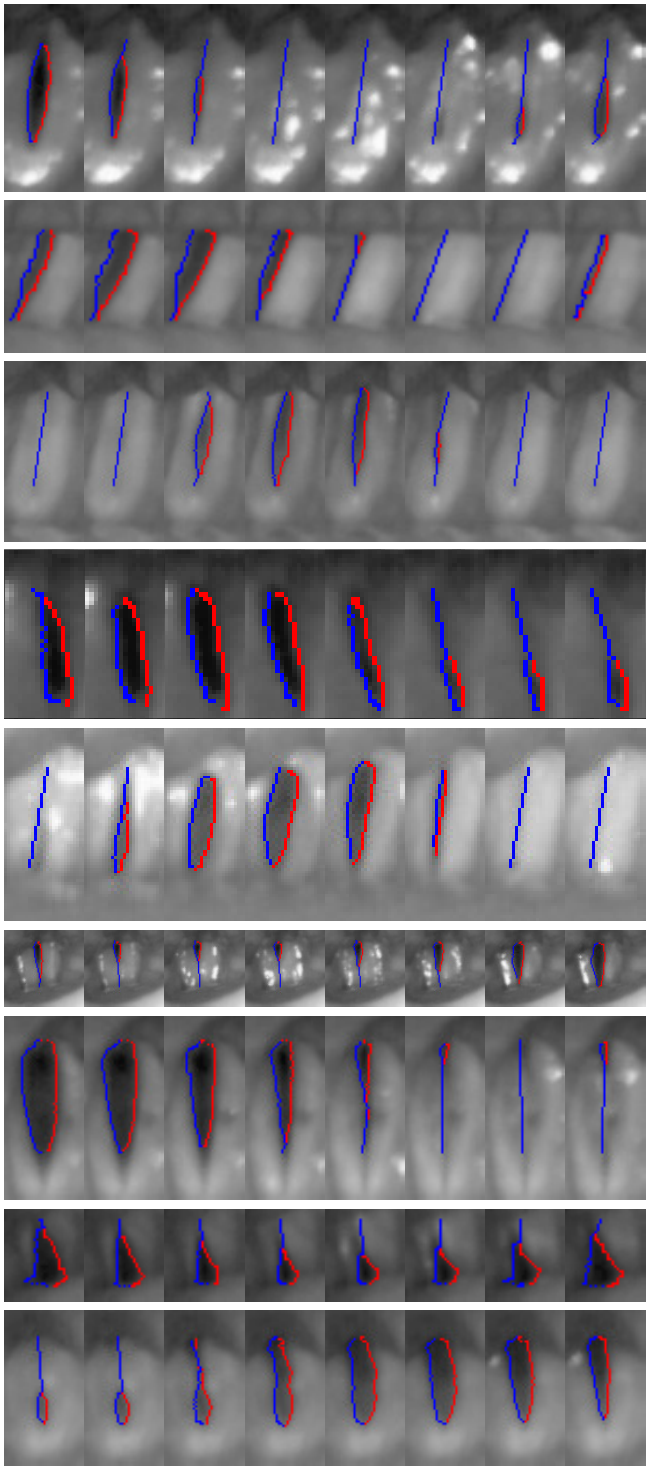


Figure 6.17: Glottal gap segmentation of 9 LHSVs in the instants of time t_k= 1, 3, 5, 7, 9, 11, 13, 15 using **InP** method.

Algorithm 1: Pseudocode for background modeling and subtraction

input : ROI, $I_{BL}(\mathbf{x},t_k)$, Th_{INC}
output: Foreground ($InP(\mathbf{x},t_k)$)
$MK(\mathbf{x},t_k) = createMasK(ROI)$;
$BM(\mathbf{x},t_k) = inpainting(I_{BL}(\mathbf{x},t_k), MK(\mathbf{x},t_k), method : Telea)$;
foreach x in the ROI **do**
  $BS(\mathbf{x},t_k) = I_{BL}(\mathbf{x},t_k) - BM(\mathbf{x},t_k)$;
  if $BS(\mathbf{x},t_k) < Th_{INC}$ **then**
    $InP(\mathbf{x},t_k) = I_{BL}(\mathbf{x},t_k)$;
  else
    $InP(\mathbf{x},t_k) = 255$;
  end
end
$Th_{ADJ} = AdaptiveThreshold(InP(\mathbf{x},t_k))$;
foreach x in the ROI **do**
  if $InP(\mathbf{x},t_k) < Th_{ADJ}$ **then**
    $InP(\mathbf{x},t_k) = 1$;
  else
    $InP(\mathbf{x},t_k) = 0$;
  end
end
$InP(\mathbf{x},t_k) = MorphologicOperation(InP(\mathbf{x},t_k))$;

in the vocal folds. This can be commonly seen during the phonation onset and in presence of total or partial paralysis of the vocal folds. In both cases the ROI detection fails to correctly identify the glottal area, producing false detections or an incomplete segmentation. The second scenario is related to glottal gap orientation. Despite the automatic method to compute the ROI is tolerant to the glottal orientation, its accuracy decreases as the angle formed by the glottal main axis and the principal vertical axis increases, producing an erroneous delineation in the anterior part of the vocal folds.

The manual procedure begins by finding the frame with the maximum aperture of the glottis. Later on, such frame is presented to the user and the ROI is manually selected. This ROI can change its size as many times as necessary or in a time frame considered by the user. Additionally, a slide bar is added to let the user modify the threshold (Th_{USR}) of the subtraction operation. The final segmentation is displayed during the whole procedure, refreshing the results when the ROI or Th_{USR} parameters are being adjusted. Since the vocal folds vibration occurs within the range from posterior to anterior commissures along the vertical direction, a recommended practice is to consider these commissures as reference points to define the ROI. On the other hand, the width of the ROI has to be limited inside the region covered by the vocal folds.

As happened with the automatic ROI detection procedure, the manually chosen ROI has to be able to dynamically compensate the glottal drift. For this reason, a block matching algorithm is applied to each corner of the manual ROI, allowing the ROI to increase or decrease its area automatically according to the glottal gap changes. In order to reduce the computational burden, the block matching algorithm is computed only every N_{USR} frames where N_{USR} is set up by the user. The graphical representation of this procedure is shown in Figure 6.18.

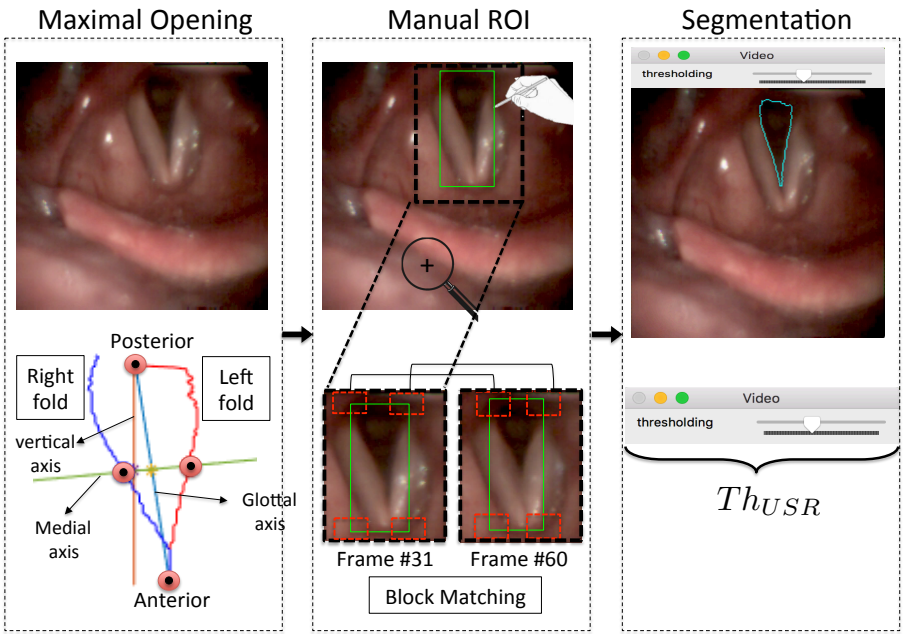


Figure 6.18: Graphical representation of the manual intervention process. Top left panel: frame with the maximal opening; bottom left panel: representation of the vocal folds reflections; medium panel: user interaction and manual ROI; top right panel: segmentation results; bottom right panel: detail of the slide bar to manually modify the threshold.

In 16 out of 54 LHSV (30% of DB2), user interaction was needed for a correct segmentation. Table 6.2 provides a summary of the findings of these 16 LHSV. The manual intervention was required in cases of paralysis, vocal folds occlusions, paresis, or wrong camera orientation.

Figure 6.19 shows four typical erroneous ROI detections with their respective GVG, PVG, VKG and GAW playbacks. In Figure 6.19a the problem is originated by the orientation of the vocal folds, causing an incomplete glottal gap detection. Figure 6.19b shows a paralysis of both vocal folds. In this case, the ROI is not able to capture the motion, selecting a region with a bigger area and causing over-segmentation. Figure 6.19c represents a unilateral paralysis of the left fold. And Figure 6.19d shows a partial paralysis of the right fold. Both effects can only be observed in the PVG and VKG. In this example the automatic segmentation fails

	Paralysis left	Paralysis right	Bilateral paralysis	vocal folds occlusion	camera orientation	Paresis	Total
Female	–	2	2	–	–	2	6
Male	2	1	1	2	1	3	10

Table 6.2: Overview of the findings in the 16 LHSV that required manual intervention.

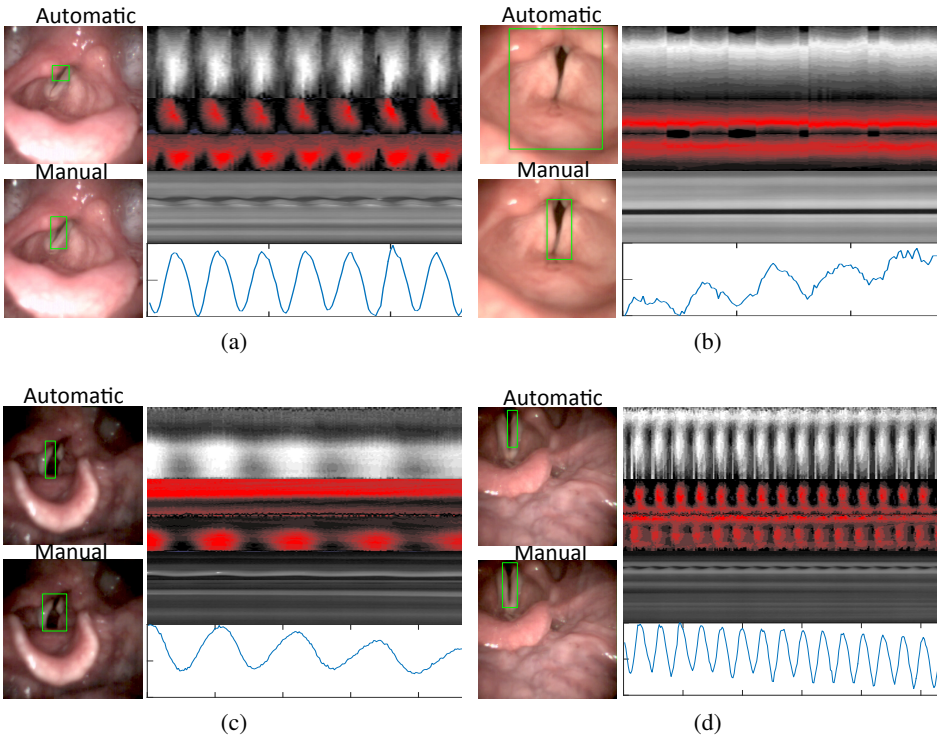


Figure 6.19: Four HSVs with their respective GVG, PVG, VKG and GAW playbacks. Automatic and manual ROIs are represented by a green rectangle. a) Error due to the glottal gap orientation; b) paralysis of both vocal folds; c) unilateral paralysis of the left fold; d) partial paralysis of the right fold.

since the ROI partially captures the region in motion of the vocal folds, producing under-segmentation.

6.4 Accuracy Assessment of the Vocal Folds Segmentation

Assessing the glottal segmentation is not trivial due to the huge amount of frames to evaluate and the need to take into account the spatial-temporal information of

the video sequences. The evaluation is even more complicated having in mind that there are neither standard metrics to evaluate the distinct algorithms nor public databases that could be used for benchmark and comparison purposes. In view of these limitations, a set of guidelines to measure the accuracy and efficiency of the segmentation algorithms are presented. These guidelines are divided in three groups according to their nature: *analytical*, *subjective*, and *objective*. The combined analysis of these guidelines provides more robust criteria to decide which is the most appropriate method to delineate the glottal gap.

The proposed guidelines are used to compare three glottal gap segmentation algorithms: two automatic, and one semiautomatic. The semiautomatic one is the open software presented in (Birkholz, 2016) that implements the algorithm proposed in (Lohscheller et al., 2007). For simplicity, this method is denoted as Seed Region Growing (SrG). Meanwhile, the two automatic methods are the ones described in section 6.2 (**SnW**) and section 6.3 (**InP**).

6.4.1 Analytical Methods: Assessing the Efficiency of the Vocal Folds Segmentations

The *analytical* methods assess the segmentation independently of the final results. In other words, they are only applicable to evaluate the general performance of the algorithms. Some of the properties of the segmentation algorithms are: processing strategy, processing complexity, resource efficiency, segmentation time, and segmentation resolution (Zhang et al., 1996). These properties are generally independent of the segmentation accuracy. For this reason, they should be investigated together with the subjective and objective measures. Since the segmentation time is a critical aspect to translate the results to the clinical setting, it will be used as a reference to compare the efficiency of the glottal segmentation algorithms.

6.4.2 Subjective Evaluation: Accuracy of the Vocal folds Detection by Playbacks Analysis

A simple *subjective* way to evaluate the glottal segmentation is by visual inspection. However, it requires a frame by frame intensive evaluation over a large set of images and with the contribution of several experts to minimize the inter evaluation bias. The subjective approach used in the literature grade the segmentation and the video quality on a 0-5 point ordinal scale (Lohscheller et al., 2007; Karakozoglou et al., 2012). The main drawback of this approach is that this evaluation does not consider the inherent spatio-temporal information, but only the spatial.

A more complete way to subjectively analyze the vocal folds segmentation requires evaluating the information provided by the playbacks. Thus, three subjective trials to assess the accuracy of the glottal segmentation are used: *segmentation quality*, *readability of the playbacks* (GVG, PVG, GAW), and *shape similarity* betweenVFT and VKG. All the trials are ranked in a 0-5 point ordinal scale where 0 is very bad and 5 is very good.

Regarding the *segmentation quality* trial, a video sequence that has already been segmented (first row of Figure 6.20) is shown to the expert. Then, he ranks an average quality of the whole sequence observed.

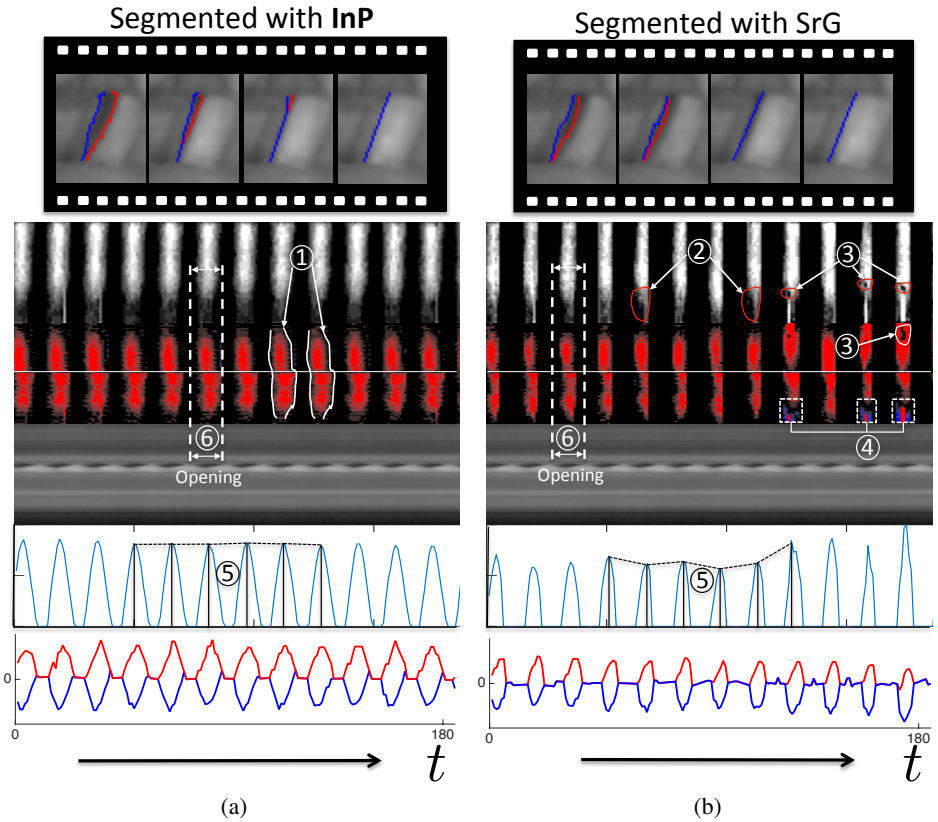


Figure 6.20: LHSV of a patient after a carcinoma surgery with its respective playbacks: GVG, PVG, VKG, GAW and VFT. a) Segmented with **InP**; b) segmented with SrG. (1) Vibratory pattern; (2) errors in the anterior or posterior part of the glottis; (3) playbacks discontinuities; (4) main glottal axis crossing; (5) glottal area waveform; (6) opening state length.

The *readability assessment* trial has the goal to detect errors in the segmentation using the information provided by the GVG, PVG and GAW playbacks. In addition to the aforementioned playbacks, VKG is included since it helps to verify if the cycle lengths of the playbacks are correct. The readability of the playbacks is ranked based on six criteria, which are detailed next:

- Vibratory pattern (1): the shape of the vibratory pattern must keep a quasi-similar behavior along the LHSV for a normal phonation. This characteristic is observed in GVG and PVG.
- Errors in the anterior or posterior part of the glottis (2): most of the seg-

mentation algorithms fail to segment correctly the anterior and posterior part of the glottal gap. These problems are easily detected in GVG and PVG, showing discontinuities in the anterior or posterior part of the playbacks.

- Playbacks discontinuities (3): discontinuities in the playbacks are due to segmentation errors, and are observed as holes in the inner part of the vibratory pattern. These errors are visible either with GVG or PVG.

- Main glottal axis crossing (4): since PVGs compute the motion of the vocal folds with respect to the glottal main axis (red for positive, and blue for negative displacements), they provide clues about unusual vibratory behaviors.

- Glottal area waveform (5): the shape of the GAW during the different glottal cycles has to be uniform along time for normal phonation. Deviations of this uniformity are usually clear examples of over or under-segmentation. GAW in Figure 6.20b shows the effect of the number of pixels detected belonging to the glottis for each glottal cycle on the shape of the GAW.

- Opening state length (6): considering that VKG playback facilitates the visual assessment of the length of the glottal cycle in one line, it is possible to get a general idea of how well the instants of total closing were detected. Hence, the width of the opening-state in GVG has to be comparable in size with the VKG width in the same position *pc*.

The third trial is called *shape similarity* and takes advantage of the information of the VKGs. Here, the shape of VKG and VFT are compared in the same position *pc*. The similarity between both shapes will determine the accuracy of the segmentation. In order to assist the expert with the visual assessment, both playbacks are overlapped as shown in Figure 6.21. The dashed lines in the middle and bottom panel with white color are an approximation of the expected VKG shape. These lines are displayed to provide an idea of the shape differences between VKG and VFT for this particular example.

6.4.3 Objective Supervised Evaluation: Accuracy of the Vocal Folds Detection via Ground-Truth Comparison

The *objective* supervised evaluation compares the segmented image against a reference manually identified that will be considered as a ground-truth (Manual Segmentation (MaN)). The degree of similarity between the human and machine generated images determines the quality of the segmentation. The objective supervised methods provide a finer evaluation of the segmentation accuracy. Contrariwise, manually generating a reference image is a difficult, subjective, and time-consuming task.

The literature refers different kinds of objective metrics that have been used before to assess the glottal segmentation: DICE and an area error (Gloger et al.,

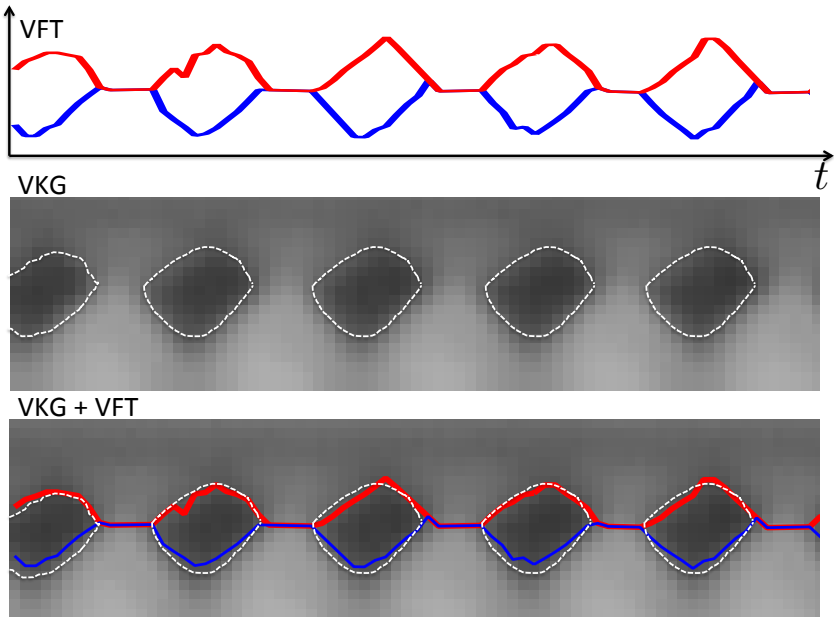


Figure 6.21: VKG and VKT playbacks of a patient after a carcinoma surgery. First row: vocal folds displacement trajectories; second row: videokymogram; third row: VKG and VKT overlapping. The dashed lines in white show the correct delimitation of the VKG.

2015); a multipoint scale comparison (Karakozoglou et al., 2012); mean square error (Ko and Ciloglu, 2014); and tracking and comparison of some points of interest (Lohscheller et al., 2007). However, it is certainly unreasonable to expect all the metrics be valid for the glottal segmentation problem, since each metric have sensitivities to different properties of the segmentation and thus can discover different types of error (Taha and Hanbury, 2015b).

Therefore, there is a need to standardize the procedure to objectively assess the accuracy of glottal gap segmentation by finding the most suitable metrics for such task. These metrics receive the name of *good metrics* and have to fulfill the following properties:

- Contour accuracy: the segmentations have to provide boundary delimitations as exact as possible. The metrics that are more sensitive to point positions, as distance based metrics, are more suitable to evaluate the segmentation.

- Degree of overlapping: the segmentations have to provide a correct location of the segmented object (alignment between segmentation and ground-truth). This aspect is important to rank correctly the instants of total closure. The suitable metrics for this property are the overlap based.

- Complex Boundary: the segmentations lead with non-regular shapes, thus

the metrics that are sensitive to pixels positions are more suitable to evaluate the final results. The most suitable metrics are the distance based ones.

- Background dominates: all the metrics based on a true negative factor (pixels correctly detected as background) have to be avoided. Such metrics are biased with respect to the ratio between the number of foreground pixels (glottis) and the number of background pixels (glottal structures), producing a class imbalance when the background represents the largest part, as occurs in the glottal segmentation.

- Over and under-segmentation penalization: the metrics have to penalize equally the over and under-segmentation.

- High class imbalance: when the segmentation process produces small regions, metrics with chance adjustment are recomended

- Outlier sensitivity: sometimes automatic segmentations have outliers in form of a small set of pixels outside of the right target area. The Outlier sensitivity describes metrics that penalize such outliers.

In order to find out the *good metrics*, an initial set of 18 metrics that have been used previously in the literature to evaluate different segmentation problems are computed and evaluated pairwise: **InP** vs. MaN, SrG vs. MaN, and **SnW** vs. MaN, for 760 images. The 18 initial metrics computed can be categorized depending on their nature and their definition as: overlap based, pair-counting based, information theory based, probabilistic based, and spatial distance based.

The first group computes the degree of overlap between two segmentations. To this group belongs: DICE or overlap index, Jaccard (JAC), true positive rate (TPR) or sensitivity, true negative rate (TNR) or specificity, F1-Score (FMS), false positive rate (FPR), false negative rate (FNR), positive predictive value or precision (PPV), global constancy error (GCE), and object-level consistency error (OCE) (Dice, 1945; Polak et al., 2009).

The second group measures the similarity between clusterings. One of its important properties is that is not based on labels, and thus can be used to evaluate clusterings as well as classifications. The metrics implemented in this work were Rand Index (RI) and Adjusted Rand Index (ARI) (Rand, 1971; Hubert and Arabie, 1985).

The third group computes a measure of information content for each segmentation. The variation of information (VI) figure of merit belongs to this group. It measures the amount of information that one segmentation shares with the other (Meilă, 2003).

The fourth group describes metrics defined as functions of statistics calculated from the pixels in the overlapped regions of the segmentations. The metrics included are: Cohen Kappa coefficient (KAP) and the area under the ROC curve (AUC) (Cohen, 1960).

Lastly, the metrics based on spatial distance are defined as functions of the euclidean distances between the pixels that belong to the ground-truth and the pixels of the automatic segmentation. The metrics used in this category are: Hausdorff distance (HD), average Hausdorff distance (AHD) and Pratt Index (Abdou and Pratt, 1979; Taha and Hanbury, 2015a).

In those cases with no unique metric fulfilling all the properties at the same time, a combination of more than one metric will be necessary. Also, a good practice is to reject metrics that have similar definitions to avoid redundant information. Table 6.3 shows in rows the metrics, and in columns the guidelines that have to be followed. The three last columns summarize the results of the three trials, μ represents the average accuracy of 760 images analyzed, and ε_{close} rates how many times an image was ranked with 0. A zero ε_{close} can be understood as not overlapping or no segmented images, which is related with the error introduced at the closed instants.

Additionally, Table 6.3 shows a check (✓) to denote a metric that is recommended for the corresponding property; a cross (X) denotes that the metric is not recommended; and empty cells denote neutrality. The *good metrics* are the ones that have at least one check without crosses. The metrics that satisfy this statement are highlighted in yellow: DICE, JAC, FMS, ARI, KAP, AHD and Pratt. In order to avoid redundancy, JAC and FMS were excluded because they provide a similar ranking than DICE coefficient (JAC and FMS are derived from the DICE equation). Following the same criteria, AHD and Pratt are metrics based on distance errors, so one of them may be excluded. Since AHD does not rank the similarity between segmentations in a range scale (as Pratt does, between 0 and 1), we consider that is less intuitive, and it has also been excluded. Thus, the metrics that best suit the guidelines are: one based on overlapping (DICE), one based on pair-counting (ARI), one based on probabilistic means (KAP), and one based on distance (Pratt).

Lastly, in order to verify the concordance between the metrics, pairwise Pearson's correlation coefficients were calculated. The 760 ranks obtained for each metric are correlated between them and deployed on Table. 6.4. The results show a great correlation between the four metrics which mean that any of them can be chosen as a *good metrics*. For the purpose of objectively evaluate the accuracy of the glottal gap segmentation only DICE and Pratt are used.

6.5 Results

The subjective evaluation was carried out for the whole database DB2 by an expert used to deal with laryngeal images. Meanwhile, the objective supervised evaluation was carried out on 38 different high-speed recordings from DB2. From each movie a sequence of 20 frames was manually segmented MaN, leading to a total of 760 images analyzed. It is worth mentioning that all experiments were executed on a MacBook Pro with a 2.4 GHz Intel core i5 processor and 8 GB of RAM memory.

	Contour accuracy	Degree of overlapping	Complex boundary	Background dominates	Segmentation penalization	High class imbalance	Outlier sensitivity	MaN vs **InP**		MaN vs SrG		MaN vs **SnW**	
								μ	ε_{close}	μ	ε_{close}	μ	ε_{close}
DICE		✓					✓	0.70	0.07	0.57	0.23	0.52	0.32
JAC		✓					✓	0.60	0.07	0.48	0.23	0.45	0.32
TPR		✓			X			0.75	0.07	0.57	0.23	0.60	0.32
TNR		✓		X				0.92	0.07	0.76	0.23	0.71	0.28
FMS		✓					✓	0.71	0.07	0.57	0.23	0.54	0.29
FPR		✓			X			0.07	0.17	0.23	0.25	0.28	0.10
FNR		✓		X				0.24	0.17	0.42	0.13	0.39	0.20
PPV		✓			X			0.73	0.07	0.61	0.23	0.48	0.32
GCE		✓			X			0.07	0.10	0.23	0.13	0.28	0.10
OCE		✓		X				0.64	0.07	0.54	0.23	0.48	0.29
RI				X				0.91	0.07	0.76	0.23	0.71	0.28
ARI						✓		0.70	0.07	0.57	0.23	0.52	0.28
VI				X			✓	0.09	0.10	0.25	0.13	0.30	0.10
KAP						✓	✓	0.70	0.07	0.57	0.23	0.52	0.28
AUC				X			✓	0.84	0.07	0.67	0.23	0.65	0.28
HD	✓		✓				X	6.69	0.11	6.05	0.17	8.71	0.14
AHD	✓		✓				✓	0.71	0.11	0.59	0.17	1.32	0.14
Pratt	✓		✓				✓	0.72	0.07	0.60	0.23	0.54	0.28

Table 6.3: Summary of the 18 metrics with the selection guidelines. Each row corresponds to one of the metrics. The first seven columns correspond to the properties evaluated to be part of the *good metrics* set. A check (✓) denotes that the metric is recommended for the corresponding property; a cross (X) denotes that the metric is not recommended; and empty cells denote neutrality. The last three columns are the average values of each metric for the three assessments.

		ARI	DICE	KAP	Pratt
MaN vs **InP**	ARI	1	1	1	0.95
	DICE	1	1	1	0.95
	KAP	1	1	1	0.95
	Pratt	0.95	0.95	0.95	1
MaN vs SrG	ARI	1	1	1	0.97
	DICE	1	1	1	0.97
	KAP	1	1	1	0.97
	Pratt	0.97	0.97	0.97	1
MaN vs **SnW**	ARI	1	0.99	1	0.97
	DICE	0.99	1	0.99	0.97
	KAP	1	0.99	1	0.97
	Pratt	0.97	0.97	0.97	1

Table 6.4: Pearson's Correlation coefficients among the *good metrics*. Correlations correspond to MaN vs. **InP**, MaN vs. SrG and MaN vs. **SnW** trials.

6.5.1 Analytical Assessment

The SnW algorithm was completely implemented in Matlab® with a computation cost of 0.58 fps. SrG was written in C++ using the cross-platform GUI library wxWidgets 2.8.12. Since SrG is only available for the Windows® platform, it was necessary to run the software in a virtual machine with Windows 7. The segmentation time of 400 high-speed images took approximately 3.1 s without considering the previous user interaction. The user adjusted values of the thresholds were in average 53.4±16.9 s. Lastly, the **InP** was implemented in C++ using the OpenCV library, and the segmentation results were exported and plotted in Matlab for a better visualization of the results. The segmentation time for the fully automatic procedure of 400 high-speed images took approximately 23 s. while the user intervention took less than 25.4±12.7 s for each high-speed sequence. Table 6.5 summarizes the segmentation times of the three algorithms.

s/image	**InP**	SrG	**SnW**
Automatic	0.057±0.01	0.007±0.0018	0.58±0.09
User interaction	25.4±12.7	53.4±16.9	—

Table 6.5: Segmentation times of the three algorithms (in fps) of 400 high-speed images.

6.5.2 Subjective Assessment

Five playbacks were synthesized to analyze the accuracy of the vocal folds deflections for the 38 LHSV: GVG, PVG, GAW, VFT and VKG. They were rated in a 0-5 point scale.

Concerning to the segmentation quality trial, **InP** was rated 4.1±0.6 (mean value±std deviation), SrG with 3.8±0.8 and **SnW** with 2.2±0.4. However, the experiments reported values up to 4.7 (mean) using SrG complemented with a strong user intervention.

Regarding to the readability of the playbacks, **InP** and SrG have again a similar performance, ranking 3.4±0.5 and 3.3±0.6 respectively. Contrariwise, **SnW** only reached values of 2.7±0.3, making difficult the interpretation of the GVG and PVG playbacks.

Lastly, the shape similarity between VKG and VFT was 3.8±0.7, 3.7±0.8 and 3.5±0.4 for **InP**, SrG and **SnW** respectively. The subjective findings for the whole database are summarized in Table. 6.6 for the three segmentation techniques. The results suggest that the best segmentation accuracy is obtained with **InP** followed closely by the SrG method.

In order to better illustrate the results of the subjective evaluation, a subset of 10 LHSVs is assessed and their medical findings summarized in Table 6.7. (P0) shows a no symmetric pattern between the left and right vocal folds, (P1) has a bocio multinodular in the right vocal fold, (P2) and (P7) represents a normal voice

	Quality	Readability	Shape
InP	4.1±0.6	3.4±0.5	3.8±0.7
SrG	3.8±0.8	3.3±0.6	3.7±0.8
SnW	2.2±0.4	2.7±0.3	3.5±0.4

Table 6.6: Subjective assessments used to evaluate the segmentation performance (in a 0-5 point scale).

production with a glottal chink, (P3) represents a normal phonation with a complete posterior closure, the presence of polyps and nodules can be clearly identified in (P4) and (P8) respectively, (P5) depicts a closure defect in the anterior part of the vocal folds and (P6) displays the vibratory pattern of a patient with vagal paraganglioma.

	sex	age	medical finding
(P0)	female	58	no symmetric
(P1)	female	41	bocio multimodular
(P2)	female	28	normal, glottal chink
(P3)	female	59	normal
(P4)	female	29	polyp
(P5)	female	84	no anterior close
(P6)	male	82	vagal paraganglioma
(P7)	female	54	normal, glottal chink
(P8)	female	45	nodule
(P9)	male	105	normal

Table 6.7: Summary of the clinical information for a subset of 10 HSV taken from the database DB2.

The glottal segmentation of four frames corresponding to the video sequences (P5) and (P7) are depicted in Figure 6.22. MaN is used as a baseline and is presented in the first row for comparison purposes. The glottal contours are shown in blue and red for the right and left vocal folds respectively. For these particular examples, **InP** and SrG present almost the same segmentation results and both of them are highly correlated with the one obtained in MaN. Contrariwise, **SnW** presents problems of over-segmentation and also can not segment correctly the glottal gap in presence of a glottal chink.

With respect to the use of the playbacks, the GVGs and PVGs for **InP**, SrG and **SnW** are depicted in Figure 6.23, respectively. Meanwhile, the overlapping between VKG and the trajectories of the vocal folds for 4 oscillation cycles in the medial axis are showed in Figure 6.24. As it can be observed, the playbacks obtained using **InP** and SrG have well defined vibratory patterns that can be easily readable for all the subset videos. However, they have slight differences in the duration of the glottal cycles, in the presence or not of a glottal chink and in the main

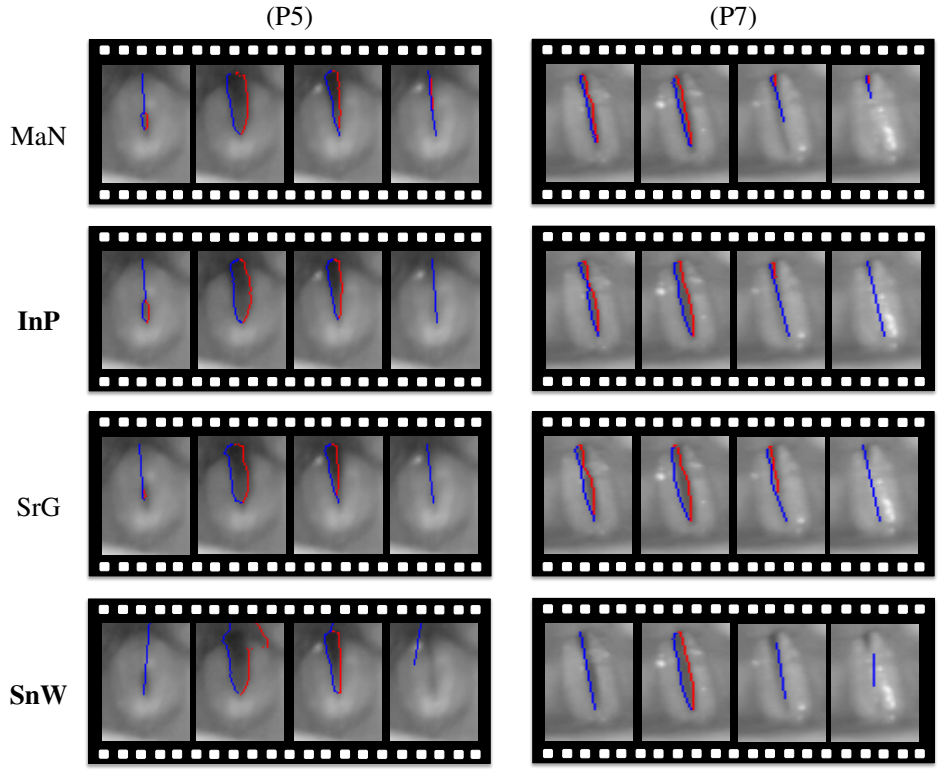


Figure 6.22: Segmented frames corresponding to a dysphonic voice (P5) and to a normal voice production (P7). First row: manual segmentation (MaN); second row: segmentation based on inpainting (**InP**); third row: segmentation based on region growing (SrG); fourth row: segmentation based on snakes and watershed (**SnW**).

glottal axis crossing. For instance, the patient (P2) presents a glottal chink that is observed either with **InP** or SrG but its size changes depending on the method used for the segmentation. Additionally, extra artefacts in blue can be observed in the SrG PVG which mean crossing of the main glottal axis. For this particular patient the correct segmentation is the one obtained using **InP** since there are no crossing and the size of the glottal chink is correct. Contrariwise, (P4) shows an example in which the segmentation based on SrG performs better that the one obtained with **InP** since the correct vibratory pattern is one without glottal chink.

On the other hand, the playbacks obtained with **SnW** are not legible for all the videos ((P2), (P4), (P6), (P8) and (P9)) and also present unexpected vibratory patterns with respect to their medical finding ((P0) and (P7)). Therefore, the segmentation obtained via **SnW** is less accurate in comparison with **InP** and SrG when the GVG and PVG are assessed.

The results of the subjective rating of the 10 LHSVs are summarized in Figure 6.25 on a 5-point ordinal-scale. The three segmentation algorithms are de-

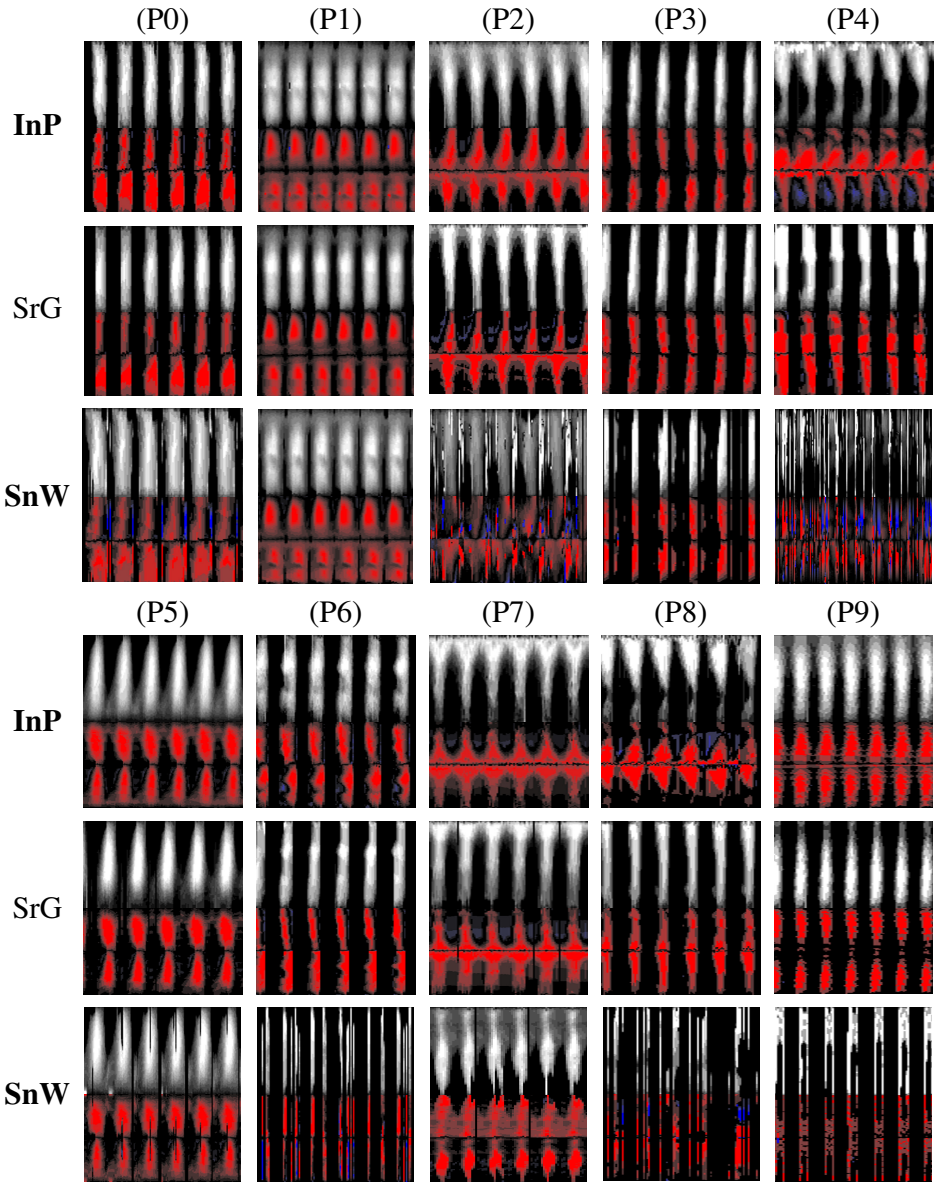


Figure 6.23: GVG and PVG playbacks corresponding to the 10 HSVs presented in Table 6.7.

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**.

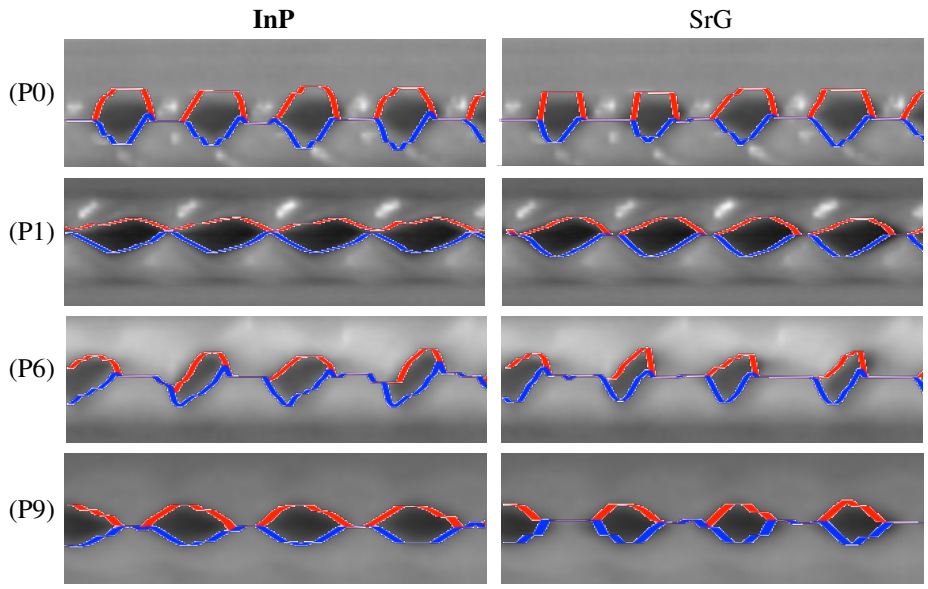


Figure 6.24: Overlapping between vocal fold trajectories and VKG in the medial axis using **InP** and SrG. Four glottal cycles are shown for each video sequence.

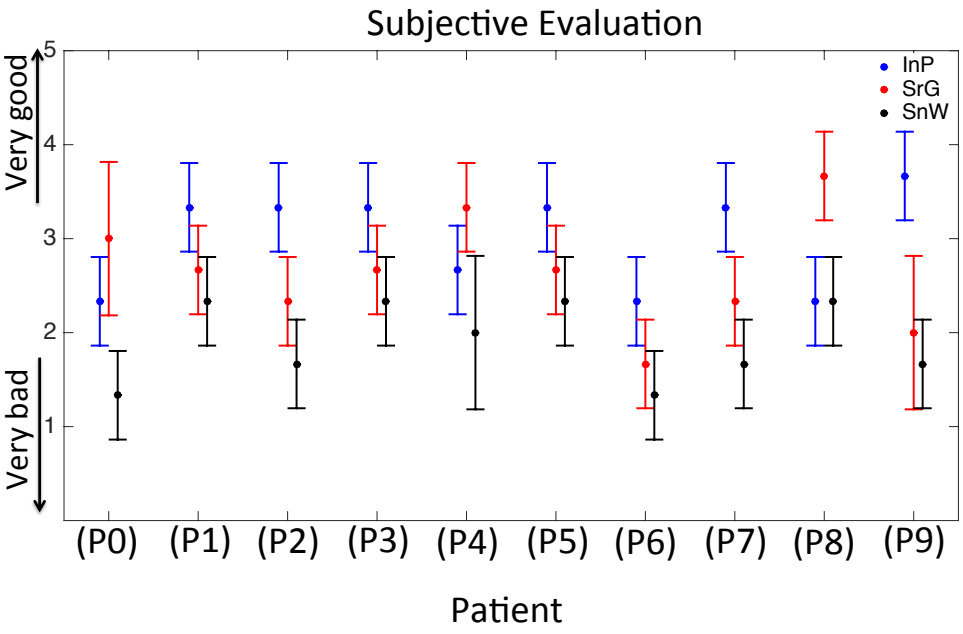


Figure 6.25: Segmentation subjective assessment of 10 patients on a 5-point scale.

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

μ represents the average accuracy obtained from 760 images analyzed, and ε_{close} rates how many times an image is ranked with 0. The accuracy improvement with respect to μ is computed as the percentage difference between **InP** and SrG, and SnW respectively. Meanwhile, The accuracy improvement with respect to ε_{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 μ and 25% in ε_{close}.

InP	**Accuracy Improvement (%)**			
	DICE μ	Pratt μ	DICE ε_{close}	Pratt ε_{close}
SrG	13	12	16	16
SnW	18	18	21	25

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**.

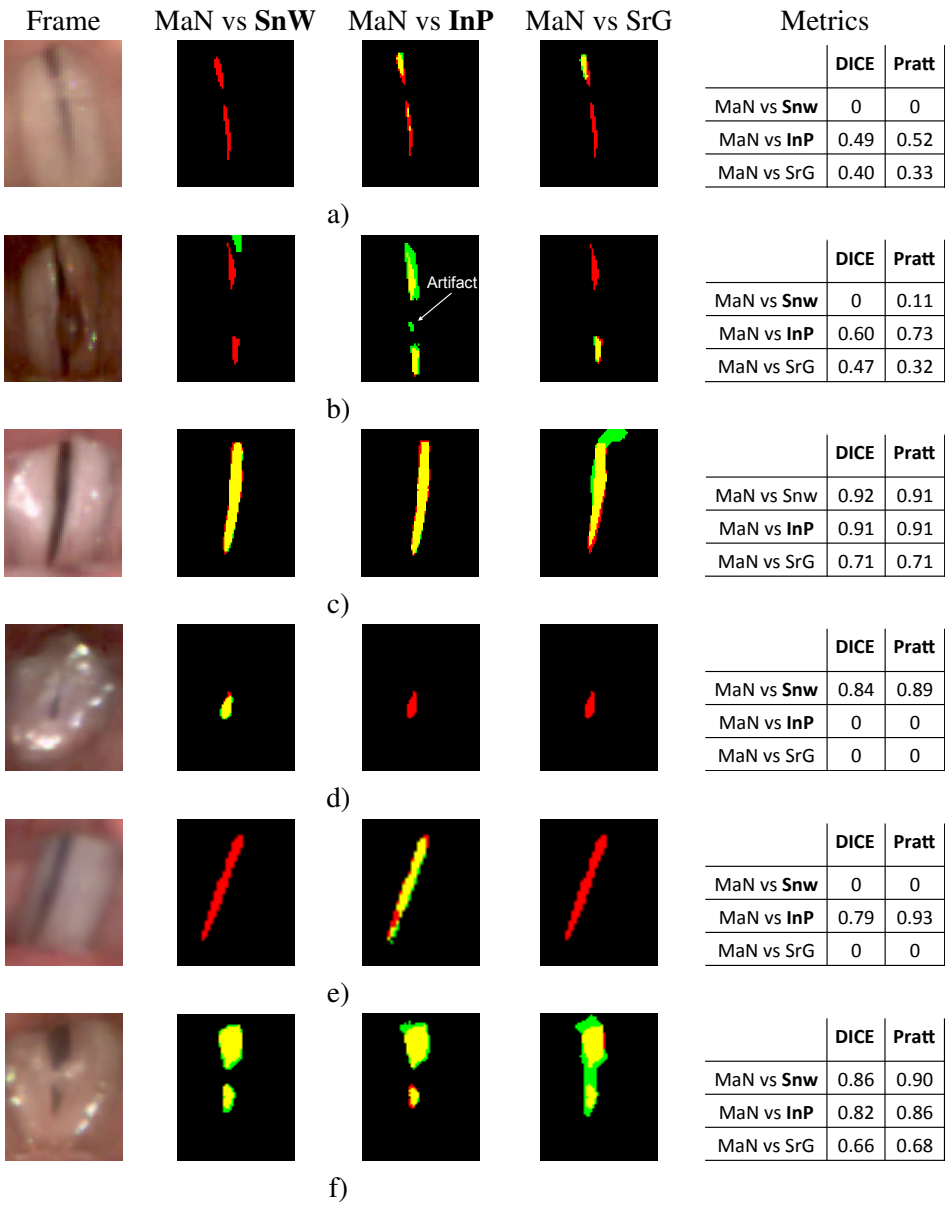
In Figure 6.26c, **SnW** and **InP** have similar results with values over 0.9, demonstrating an accurate segmentation. In the fourth frame, **SnW** is the only method able to segment correctly the anterior part of the glottis having rankings of 0.84 and 0.89 for DICE and Pratt, respectively.

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, respectevely.

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 aproximation 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



	DICE	Pratt
MaN vs **Snw**	0	0
MaN vs **InP**	0.49	0.52
MaN vs SrG	0.40	0.33

	DICE	Pratt
MaN vs **Snw**	0	0.11
MaN vs **InP**	0.60	0.73
MaN vs SrG	0.47	0.32

	DICE	Pratt
MaN vs Snw	0.92	0.91
MaN vs **InP**	0.91	0.91
MaN vs SrG	0.71	0.71

	DICE	Pratt
MaN vs **Snw**	0.84	0.89
MaN vs **InP**	0	0
MaN vs SrG	0	0

	DICE	Pratt
MaN vs **Snw**	0	0
MaN vs **InP**	0.79	0.93
MaN vs SrG	0	0

	DICE	Pratt
MaN vs **Snw**	0.86	0.90
MaN vs **InP**	0.82	0.86
MaN vs SrG	0.66	0.68

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 $\mathbf{p}(t)$ and anterior $\mathbf{a}(t)$ part of the glottal main axis for two consecutive frames during the opening phase.

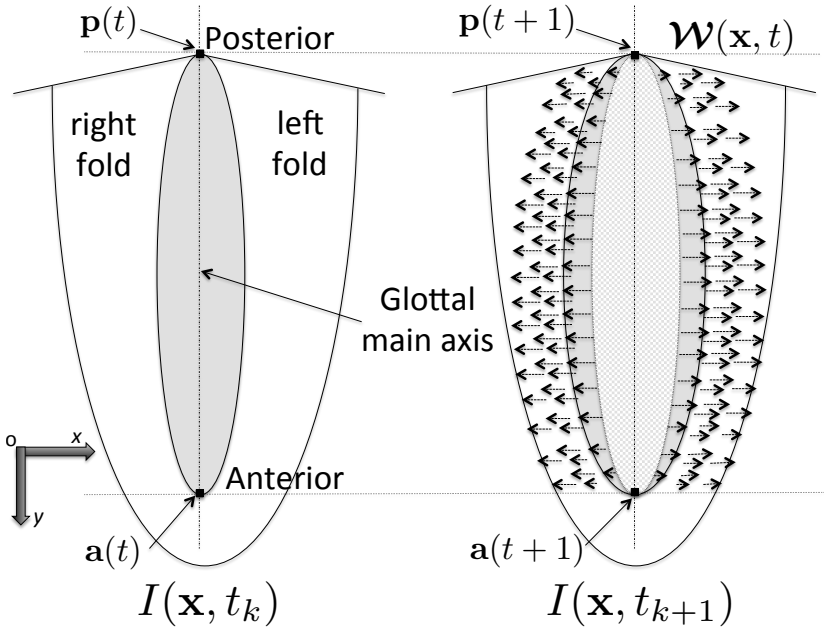


Figure 7.1: Illustration of a synthetic motion field $\mathcal{W}(\mathbf{x},t)$ located among the posterior ($\mathbf{p}(t+1)$) and anterior ($\mathbf{a}(t+1)$) part of the vocal folds during two consecutive instants of time, t_k and t_{k+1}.

Additionally, the fluctuations over time of the motion field $\mathcal{W}(\mathbf{x},t)$ have to reflect the glottal dynamics solely. In order to prove this fact, the magnitude changes of $U(\mathbf{x},t)$ 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.

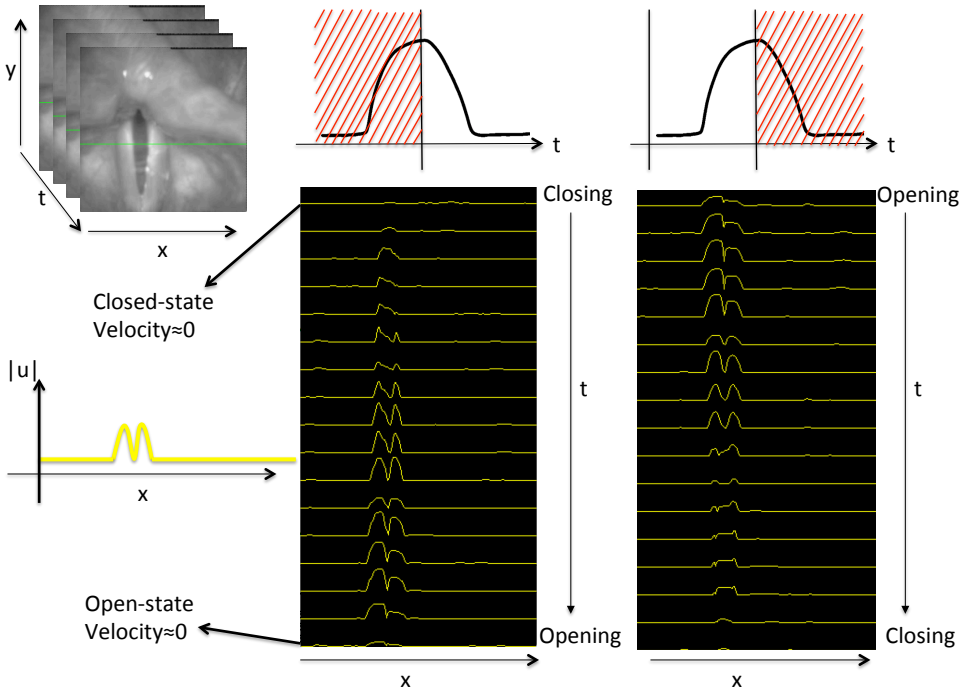


Figure 7.2: Fluctuation of u along one line for a complete glottal cycle

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 $\times$ 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[1].

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).

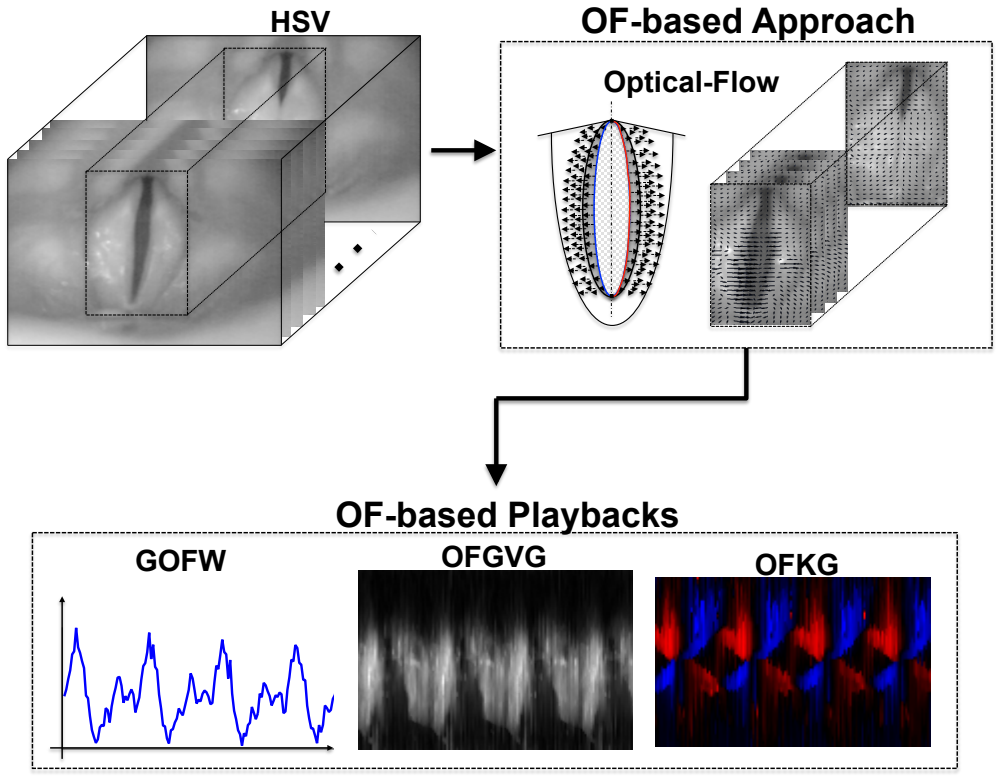


Figure 7.3: Graphical representation of the procedure followed to compute the new playbacks.

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 ($U(\mathbf{x},t_k)$). 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

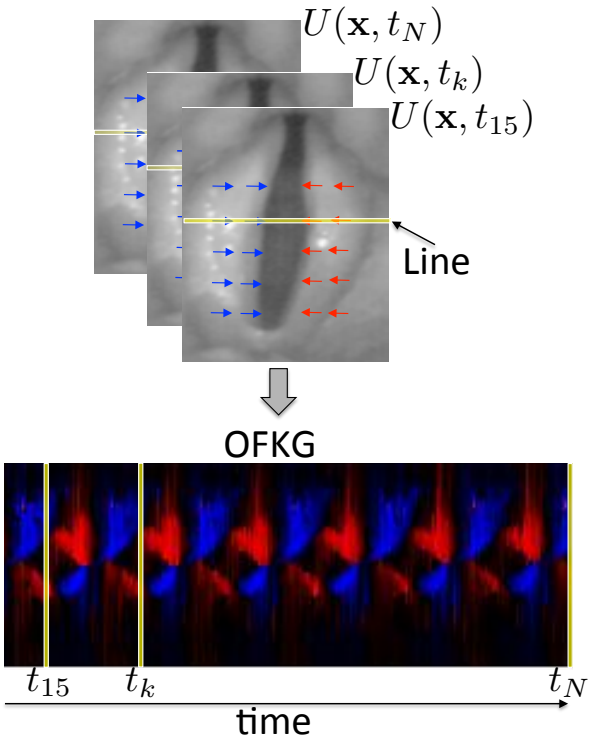


Figure 7.4: Schematic view of OFKG playback for the line represented in yellow, which is located in the median part of the vocal folds; The new local playback distinguishes the direction of motion (rightwise: red; leftwise: blue).

Algorithm 2: Pseudocode for OFKG playback

input : ROI, $I(\mathbf{x},t)$, Line
output: OFKG
foreach k *in* $I(\mathbf{x},t_k)$ **do**
  $I(\mathbf{x},t_k) \leftarrow ROI(I(\mathbf{x},t_k))$
  $I(\mathbf{x},t_{k+1}) \leftarrow ROI(I(\mathbf{x},t_{k+1}))$
  $[U(\mathbf{x},t_k),V(\mathbf{x},t_k)] \leftarrow computeOpticalFlow(I(\mathbf{x},t_k),I(\mathbf{x},t_{k+1}))$
  foreach $u(x_i,Line,t_k)$ *in* $U(\mathbf{x},t_k)$ **do**
    if $\theta(u(x_i,Line,t_k),v(x_i,Line,t_k)) \in [-\pi/2,\pi/2]$ **then**
      $OFKG(t_k,x_i) \leftarrow colorCode(|u(x_i,Line,t_k)|,blue)$
    else
      $OFKG(t_k,x_i) \leftarrow colorCode(|u(x_i,Line,t_k)|,red)$
    end
  end
end

by averaging each row of $U(\mathbf{x},t_k)$ 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

input : ROI, $I(\mathbf{x},t)$
output: OFGVG
foreach k *in* $I(\mathbf{x},t_k)$ **do**
  $I(\mathbf{x},t_k) \leftarrow ROI(I(\mathbf{x},t_k))$
  $I(\mathbf{x},t_{k+1}) \leftarrow ROI(I(\mathbf{x},t_{k+1}))$
  $[U(\mathbf{x},t_k),V(\mathbf{x},t_k)] \leftarrow computeOpticalFlow(I(\mathbf{x},t_k),I(\mathbf{x},t_{k+1}))$
  foreach *Row in* $U(\mathbf{x},t_k)$ **do**
    $$OFGVG(t_k,y_j) \leftarrow \frac{\sum_{i=1}^{n} \left|U(x_i,y_j,t_k)\right|}{n} \quad \forall j \in m$$
  end
end

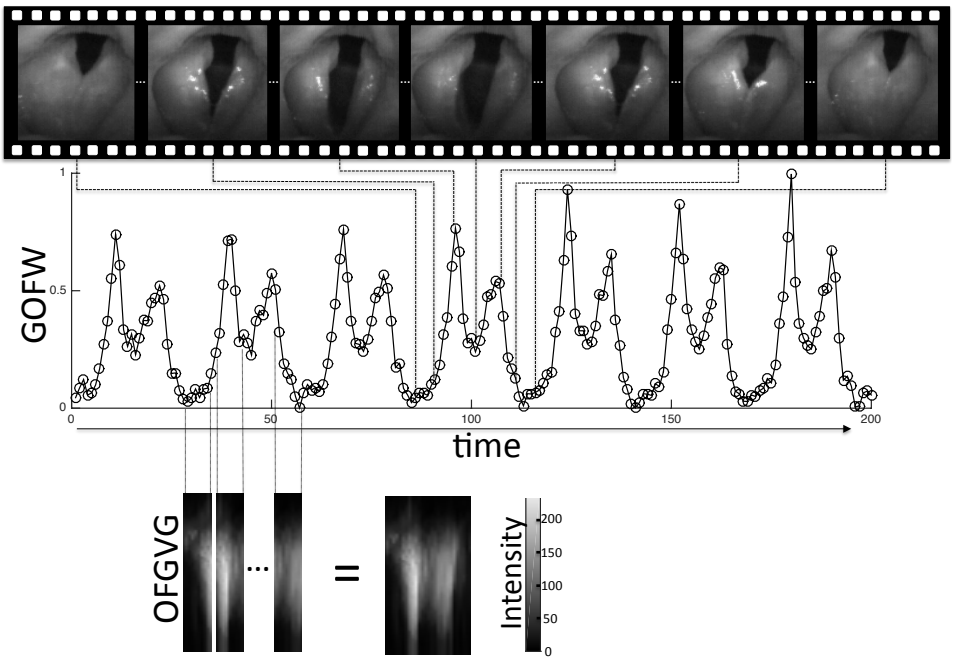


Figure 7.5: First row: frames representation of one glottal cycle. Second row: schematic view of GOFW. Each point in the playback (dark circles) is obtained by averaging the absolute magnitude of $U(\mathbf{x},t_k)$. Third row: schematic view of one OFGVG cycle. Dark regions indicate no velocity ($u(\mathbf{x}_{ij},t_k) = 0$).

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 $U(\mathbf{x},t_k)$. 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, $I(\mathbf{x},t)$
output: GOFW
foreach k *in* $I(\mathbf{x},t_k)$ **do**
  $I(\mathbf{x},t_k) \leftarrow ROI(I(\mathbf{x},t_k))$
  $I(\mathbf{x},t_{k+1}) \leftarrow ROI(I(\mathbf{x},t_{k+1}))$
  $[U(\mathbf{x},t_k),V(\mathbf{x},t_k)] \leftarrow computeOpticalFlow(I(\mathbf{x},t_k),I(\mathbf{x},t_{k+1}))$
  $GOFW(t_k) \leftarrow \dfrac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left|U(x_i,y_j,t_k)\right|}{n\times m}$
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 $\mathbf{L}(t_k)$ at time t_k, which intersects perpendicularly with glottal main axis $\mathbf{G}(t_k)$ in a predefined point $\mathbf{g}_{pc}(t_k)$ is defined and updated every image using eq 3.2. Following, the intersection between the vocal folds edges $\mathbf{C}^{l,r}(t_k)$ and trajectory line $\mathbf{L}(t_k)$ is computed, $\{\mathbf{c}^{l,r}_{pc}(t_k) : \mathbf{c}^{l,r}_{pc}(t_k) \in \mathbf{L}(t_k) \wedge \mathbf{c}^{l,r}_{pc}(t_k) \in \mathbf{C}^{l,r}(t_k)\}$. Then, the displacement trajectories $\hat{\delta}^{l,r}_{OF_{\mathcal{W}}}(pc,t_k)$ at t_k and position pc is defined by eq 7.1 as:

$$\hat{\delta}^{l,r}_{OF_{\mathcal{W}}}(pc,t_k) = \mathcal{W}(\mathbf{c}^{l,r}_{pc}(t_k)) \tag{7.1}$$

In view of the aforementioned, two additional trajectories can be derived from eq 7.1: $\hat{\delta}^{l,r}_{OF_u}(pc,t_k) = U(\mathbf{c}^{l,r}_{pc}(t_k))$ and $\hat{\delta}^{l,r}_{OF_v}(pc,t_k) = V(\mathbf{c}^{l,r}_{pc}(t_k))$. However, as the glottal edges have a motion pattern mainly perpendicular to the glottal axis, $\hat{\delta}^{l,r}_{OF_v}(pc,t_k)$ is negligible. Hence $\hat{\delta}^{l,r}_{OF_{\mathcal{W}}}(pc,t_k)$ reflects primarily the fluctuations along t_k produced by $\hat{\delta}^{l,r}_{OF_u}(pc,t_k)$. From now, both terms are used indistinctly and denoted for

simplicity as $\hat{\delta}^{l,r}_{OF}(pc,t_k)$. The graphical procedure followed to plot $\hat{\delta}^{l,r}_{OF}(pc,t)$ is described in Figure 7.6 and expressed in vector notation in eq 7.2.

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

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

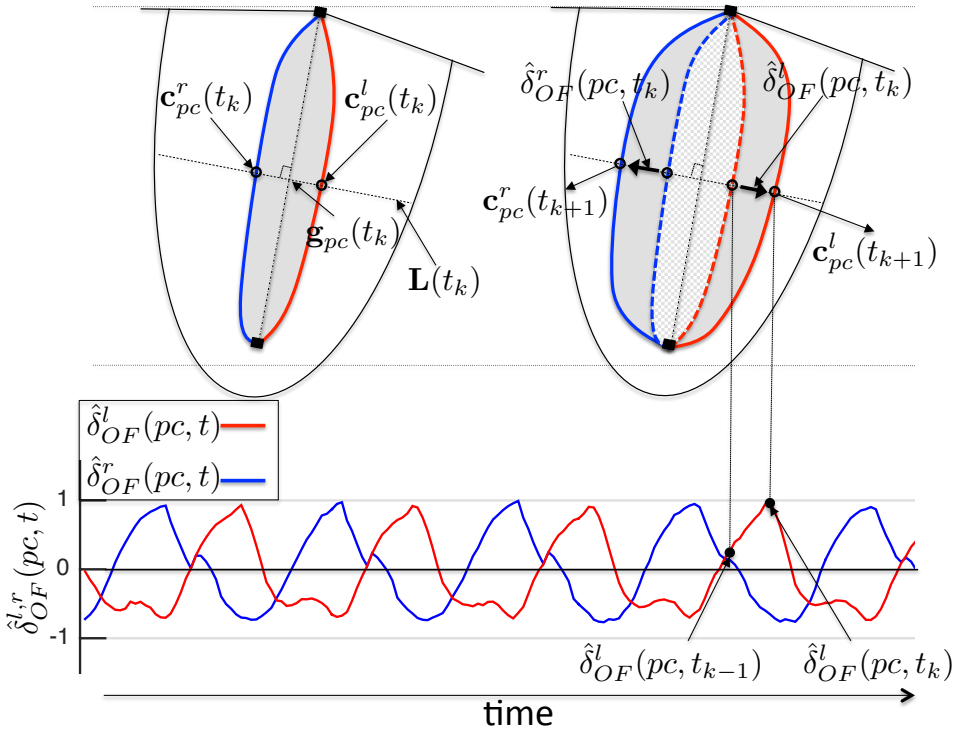


Figure 7.6: Schematic procedure to compute $\hat{\delta}^{l,r}_{OF}(pc,t_k)$ during the opening phase.

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 results: 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.

$$\hat{\delta}^{l,r}_{seg}(pc,t_k) = \mathbf{c}^{l,r}_{pc}(t_{k+1}) - \mathbf{c}^{l,r}_{pc}(t_k) \tag{7.3}$$

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

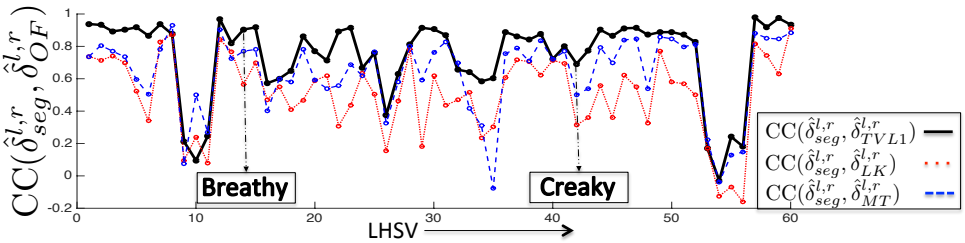
The second assessment tries to find out the similarities of traditional playbacks with OF playbacks by visually analyzing their common features and quantifying their resemblance through two metrics: Structural Similarity Index (SSIM) (Wang et al., 2004) and Normalize Correlation Coefficient (CC).

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 $C^{l,r}(t)$ 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.

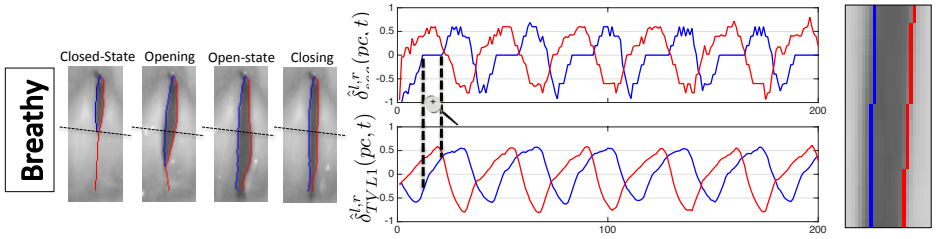
7.6 Results

7.6.1 Comparison Among Segmentation and OF Displacement Trajectories

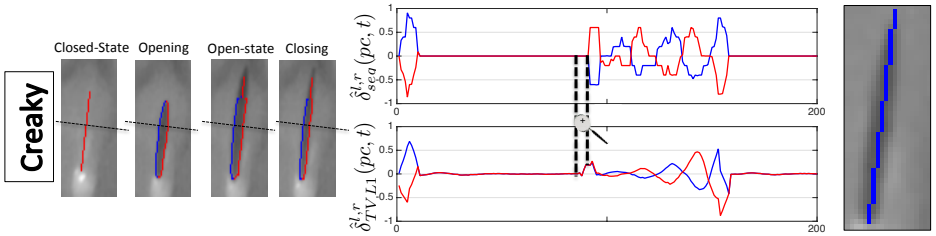
The correlation between the segmentation trajectory and OF-based trajectories is depicted in Figure 7.7a ($\text{CC}(\hat{\delta}^{l,r}_{seg}, \hat{\delta}^{l,r}_{TVL1})$, $\text{CC}(\hat{\delta}^{l,r}_{seg}, \delta^{l,r}_{LK})$ and $\text{CC}(\hat{\delta}^{l,r}_{seg}, \hat{\delta}^{l,r}_{MT})$). Each point of the graphic corresponds to the correlation of one LHSV sequence. Best correlations are obtained when $\hat{\delta}^{l,r}_{TVL1}$ is compared with $\hat{\delta}^{l,r}_{seg}$. The average correlation, $\overline{\text{CC}}(\hat{\delta}^{l,r}_{seg}, \hat{\delta}^{l,r}_{TVL1})$, achieved for that case is 0.74 while the average correlations $\overline{\text{CC}}(\hat{\delta}^{l,r}_{seg}, \hat{\delta}^{l,r}_{LK})$ and $\overline{\text{CC}}(\hat{\delta}^{l,r}_{seg}, \hat{\delta}^{l,r}_{MT})$ only reached values of 0.51 and 0.63, respectively. The greatest correlation is 0.98 which belongs to $\text{CC}(\hat{\delta}^{l,r}_{seg}, \hat{\delta}^{l,r}_{TVL1})$. Meanwhile, $\text{CC}(\hat{\delta}^{l,r}_{seg}, \hat{\delta}^{l,r}_{LK})$ and $\text{CC}(\hat{\delta}^{l,r}_{seg}, \hat{\delta}^{l,r}_{MT})$ 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 $\text{CC}(\hat{\delta}^{l,r}_{seg}, \hat{\delta}^{l,r}_{TVL1})$ with values below to 0.5, representing 13% of the videos in the database.



(a) Correlation between $\hat{\delta}^{l,r}_{seg}(pc,t)$ and $\hat{\delta}^{l,r}_{OF}(pc,t)$ with $pc = 50\%$



(b) Breathy phonation



(c) Creaky phonation

Figure 7.7: First row: correlation between OF trajectories and segmentation trajectory for each sequence. Second and third row: $\hat{\delta}^{l,r}_{seg}$ and $\hat{\delta}^{l,r}_{TVL1}$ are compared for two phonatory tasks: breathy and creaky (sequences selected on first panel). The left panel shows four frames of each LHSV with their respective segmentation and trajectories. The right panel shows the close up of two frames with segmentation errors corresponding to the interval in dashed lines.

In order to understand the differences between $\hat{\delta}^{l,r}_{TVL1}$ and $\hat{\delta}^{l,r}_{seg}$, 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) $\hat{\delta}^{l,r}_{TVL1}$ presents some fluctuations originated from a vibration of the vocal folds, while $\hat{\delta}^{l,r}_{seg}$ 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 $\hat{\delta}^{l,r}_{seg}(pc,t)$.

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 $|d_x\text{GVG}|$ and three OFGVG. Similarities between $|d_x\text{GVG}|$ 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 $|d_x\text{GVG}|$ 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 $|d_x\text{GVG}|$.

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 ($|d\text{GAW}|$). The GOFW provides valuable information about the total velocity of the vocal folds motion for each instant of time. Additionally, if $|d\text{GAW}|$ 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 $|d\text{GAW}|$ 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 $|d\text{GAW}|$ 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-

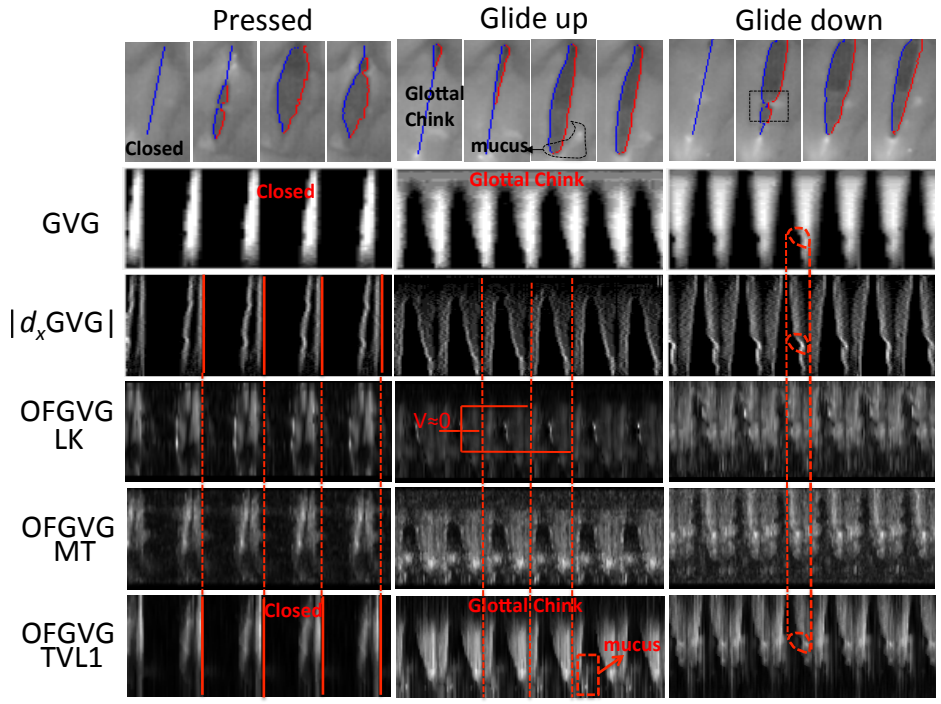


Figure 7.8: Illustration of GVG, $|d_x\text{GVG}|$, OFGVG-LK, OFGVG-MT and OFGVG-TVL1 playbacks for three different phonatory tasks (pressed, glide up and glide down).

tude values are different depending on the OF used. In the glissando task, $|d\text{GAW}|$ and GOFW have a discrepancy with respect to maximal velocity occurrence. In $|d\text{GAW}|$, 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 $|d_x\text{GVG}|$ and OFGVGs. The correlation between $|d_x\text{GVG}|$ 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

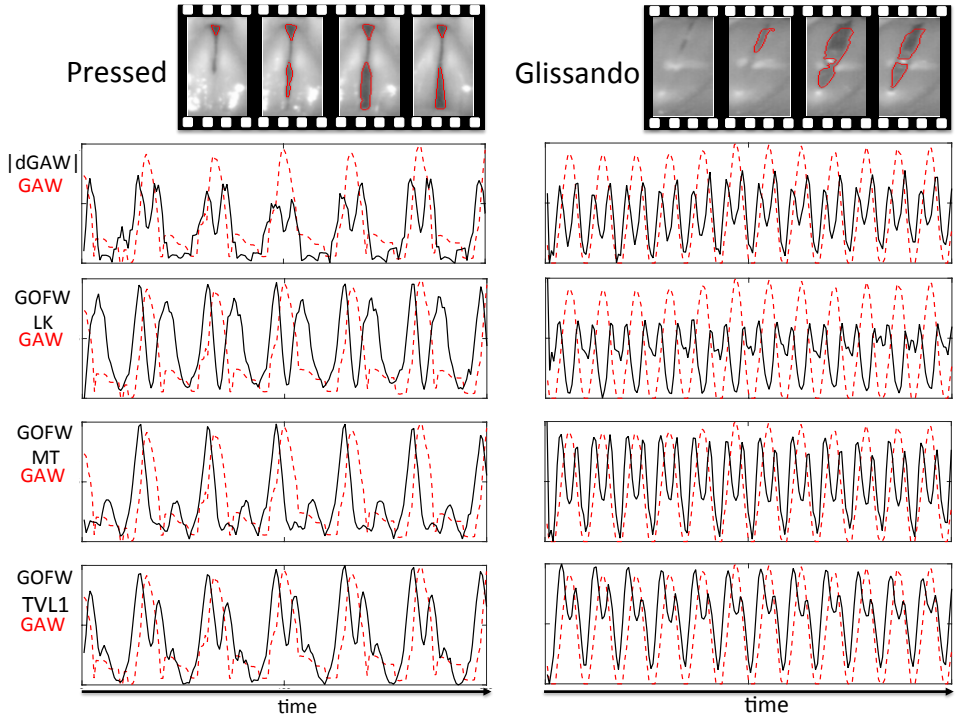


Figure 7.9: GAW vs GOFW representation for a pressed and glissando task. First row: GAW and $|d\text{GAW}|$; second row: GAW and GOFW-LK; third row: GAW and GOFW-MT; fourth row: GAW and GOFW-TVL1.

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 $|d_x\text{GVG}|$ 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, $|d_x\text{GVG}|$ and OFGVG playbacks using TVL1-OF of the entire DB3.

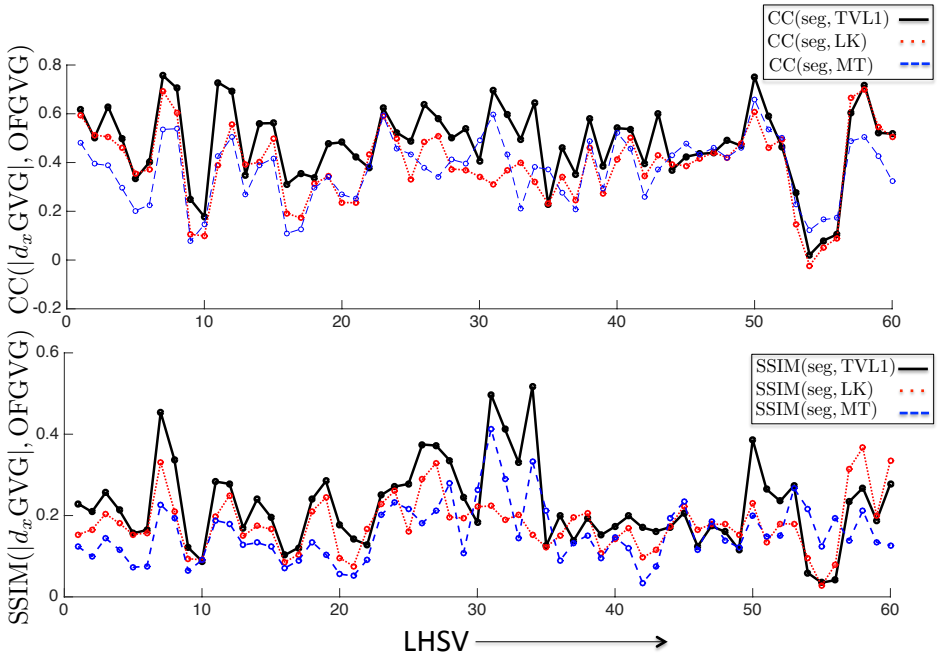


Figure 7.10: Correlation and SSIM obtained by comparing $|d_x\text{GVG}|$ with each OFGVG. The horizontal axis represents the video sequence in the DB3 database, and the vertical axis the value of the metrics.

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

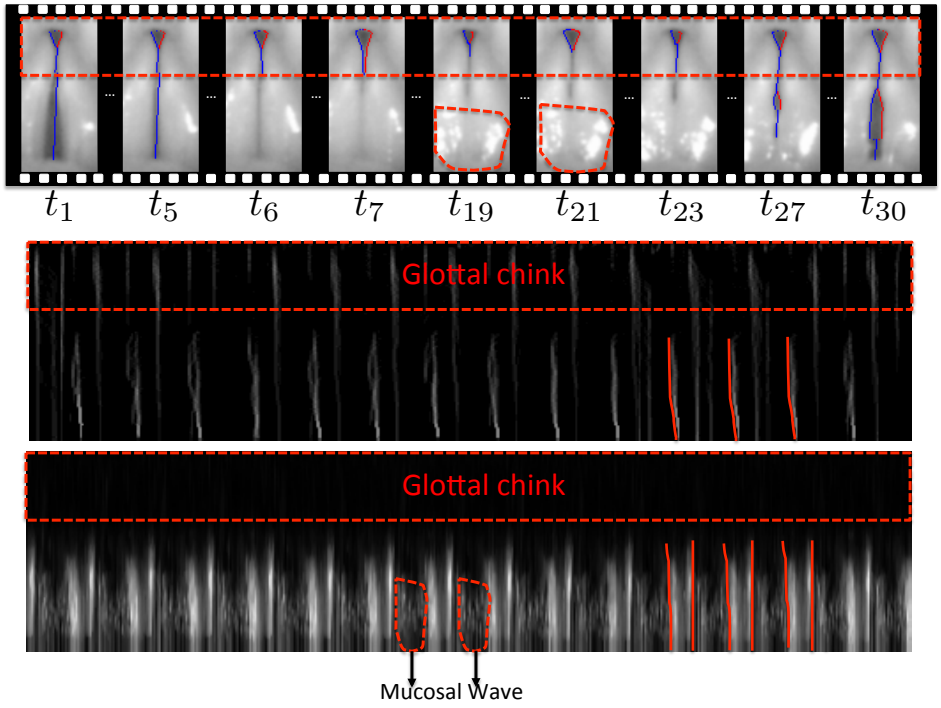


Figure 7.11: Upper panel: nine segmented frames, the rectangle dotted with red correspond to the space between the margin of the ROI and to the area with a glottal chink; middle panel: $|d_x\text{GVG}|$ playback; lower panel: OFGVG with a vertical length that depends on the ROI size. The effect caused by the mucosal wave motion and the vibratory shape pattern for three consecutive cycles are marked with dotted and continuous red lines respectively.

change over time. Meanwhile, OFKG is displayed with low intensity tonalities ($u(pc1,t) \approx 0$). 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-TVL1 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$_{OF}$,

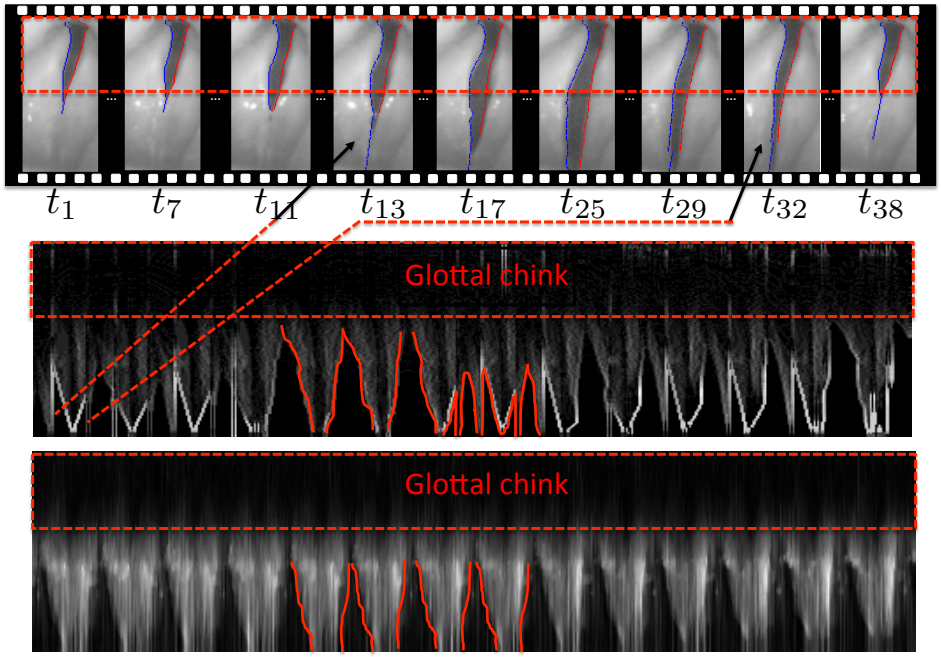


Figure 7.12: Upper panel: nine segmented frames, the areas dotted with red correspond to the posterior glottal chink; middle panel: $|d_x\text{GVG}|$ playback; lower panel: OFGVG with a vertical length that depends on the ROI size. The misleading calculation of the distance between edges is observed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

is computed using the whole motion field $U(\mathbf{x},t)$ inside a ROI. Meanwhile, the second, named OFGVG$_{seg}$, uses only the displacement vectors $U(C^{l,r}(t))$ 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.

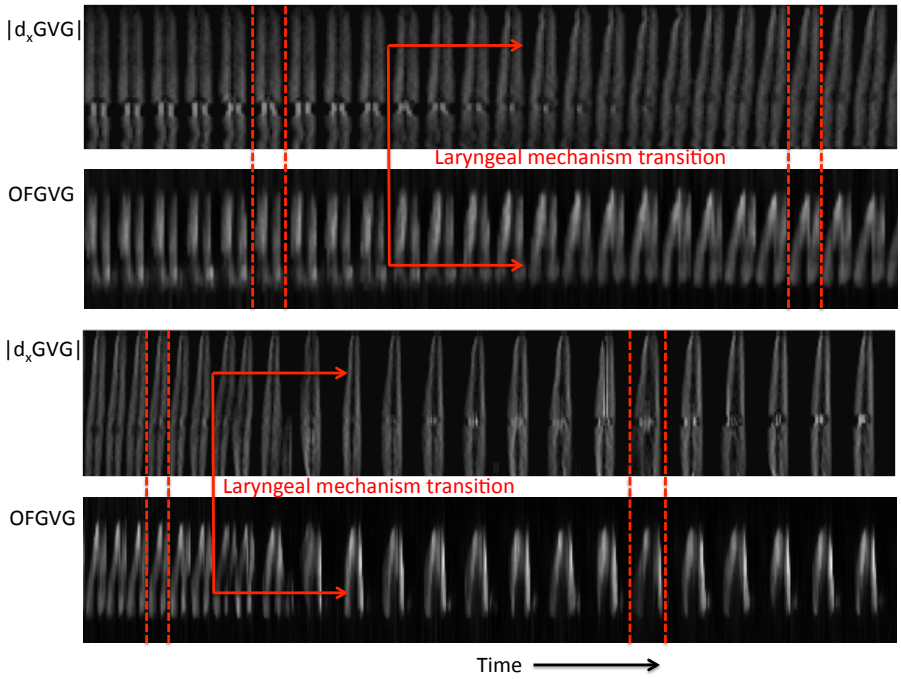


Figure 7.13: $|d_x\text{GVG}|$ and OFGVG visualization of peculiar vocal-folds vibratory movements during glissando with a laryngeal-mechanism transition. Upper panel: 24 glottal cycles. Lower panel: 23 glottal cycles. The laryngeal mechanism transition is pointed out with red arrows and the dashed lines in red indicate different glottal cycles.

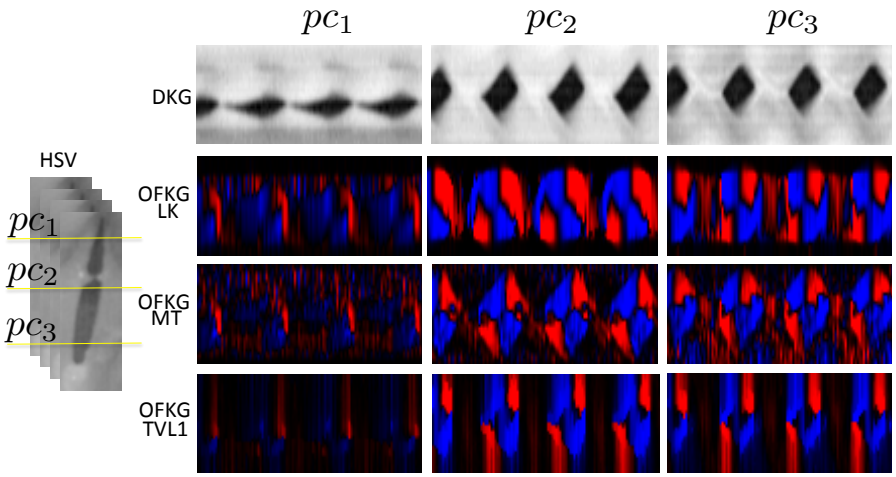


Figure 7.14: Illustration of DKG and OFKG at three different positions of the LHSV sequence. First row: VKG playback; second row: OFKG using LK-OF; third row: OFKG using MT-OF; fourth row: OFKG using TVL1-OF.

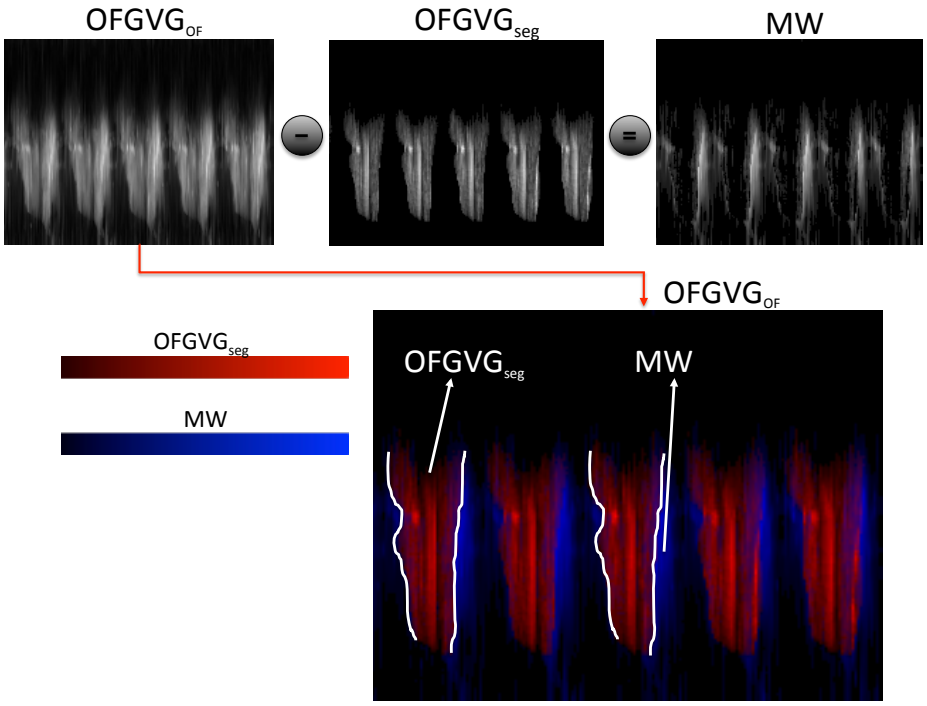


Figure 7.15: Illustration of the mucosal wave contribution in an OFGVG playback. First row: OFGVG of the whole vocal folds surface motion (OFGVG$_{OF}$), OFGVG of the vocal folds edges (OFGVG$_{seg}$) and mucosal wave propagation (MW). Second row: overlapping of both contributions (in red, the edges motion, and in blue, mucosal wave surface motion)

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 of 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 β in the

enhancement was determined to find the best trade-off between contrast and information loss; the standard template was determined from manual segmentations of the glottal gap by finding the one that best correlates with the videos on the database; and the Watershed Merging Threshold, Correlation Threshold and the cut-off points were selected based on experimentation around the database. Thus, the accuracy of **SnW** will highly depend on how these six parameters (two for the ROI) behave to the different illumination level and contrast presented on the LHSVs. Despite **SnW** was conceived to be fully automatic it is possible to let the user interact with these six degrees of freedom to solve any inconvenience that might appear during the segmentation process. However, the task of tuning six variables is time consuming and not the most appropriated for a clinical application.

At this point is important to mention the benefits of using semi-automatic or automatic methods to segment the glottal gap. The semi-automatic techniques let the user interact as many times as needed in order to solve any inconvenience that might appear during the segmentation process. Contrariwise, the automatic techniques process all the data without any previous setting or any user intervention. From a clinical point of view, both methods present advantages and disadvantages but it is worth mentioning that semi-automatic methods are more time consuming for the clinicians, although their accuracy is expected to be better. Therefore, it is necessary to provide novel frameworks making a trade-off between automatic and semi-automatic with minimal interaction to segment the glottal gap.

In this sense, **InP** proposes an approach that smooths the textures of the background (laryngeal structures) and foreground (glottal gap), detects the ROI using the temporal intensity information, and segments the glottis by creating an adaptive background model. Furthermore to the automatic segmentation, the method provides the chance of a minimal user interaction to improve the results in those cases in which they are not those desired. The smoothing process also called contrast enhancement, has the goal to eliminate the specularity effect that appears due to intrinsic surface properties by combining techniques as Color Equalization, Specularity Removal, and Bilateral Filtering. The final result is a grayscale image, free of specularity and with a marked difference in the contrast of the laryngeal structures and the glottis. The ROI detection is the same used in **SnW** but over the images obtained after the smoothing process. The final glottal gap delimitation is achieved by making a representation of the scene called background model. The background model represents an image in which the glottis has been completely occluded by means of an inpainting algorithm. Then, a background subtraction is performed between the background model and each incoming frame. Any significant change in the image region from the background model is supposed to represent a moving object in our case the glottal gap. There are some particular cases in which a correct automatic segmentation can not be achieved, being necessary an user intervention. The manual procedure begins finding the frame with the maximum aperture of the glottis. Later on, such frame is presented to the user and

the ROI is manually selected. This ROI can change its size as many times as necessary or in a time frame considered by the user. Additionally, a slide bar is added to let the user modify the threshold of the subtraction operation. Therefore, **InP** has only one degree of freedom which makes it more suitable for clinical applications. Although specularity removal, background subtraction, and inpainting techniques are common in the image processing field, to the best of our knowledge, this is the first time that they have been used for the segmentation of the vocal folds.

Secondly, the problem of evaluating the glottal gap segmentation has been addressed. Assessing the glottal segmentation is not trivial due to the huge amount of frames to evaluate, and the need to take into account the spatial-temporal information of the video sequences. The evaluation is even more complicated having in mind that there are neither standard metrics to evaluate the distinct algorithms nor public databases that could be used for benchmark and comparison purposes. In view of these limitations, we propose a set of guidelines to evaluate the vocal folds segmentation accuracy based on a subjective and objective analysis. These guidelines are divided into three groups according to their nature: analytical, subjective, and objective. The analytical assesses the segmentation independently of the final results. In other words, are only applicable to evaluate the general performance of the algorithms. The subjective analyzes the information provided from the facilitative playbacks by proposing three subjective trials to assess the accuracy of the glottal segmentation: segmentation quality, readability of the playbacks (GVG, PVG, GAW), and shape similarity between VFT and VKG. Lastly, the objective evaluation compares the segmented image against a reference manually identified, also known as ground-truth. The degree of similarity between the human and machine generated images determines the quality of the segmentation. Using the set of new guidelines, the accuracy and efficiency of **InP** and **SnW** are compared against Seed Region Growing (SrG) segmentation. Based on the results, it can be concluded that the best results in the subjective and objective evaluation 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**. Additionally, **InP** does not require an intensive user interaction, which suggests its appropriateness be transferred to the clinical environment.

Lastly, a new method to synthesize the dynamical information of the vocal folds based on Optical Flow computation is proposed. The main reason to use OF techniques is that they allow the possibility to track unidentified objects solely based on its motion, with no need of additional segmentation techniques. Therefore, not only the points belonging to the glottal edges are included but also those regions that originated such movements. Three new playbacks are proposed: 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. These new ways for data visualization have the goal to overcome the drawbacks of existing playbacks, providing simultaneously features that integrate the time dynamics, such as velocity,

acceleration, instants of maximum and minimum velocity, vocal folds displacements during phonation and motion analysis. The proposed OF-based playbacks have demonstrated a great correlation in shape with the traditional playbacks such as DKG, GAW, and GVG, 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 spatiotemporal representations when segmentation is not available, or when it is not reliable enough due to failures in the glottal-edges detection. The only drawback identified is that, compared to the traditional playbacks, those based on OF are slightly blurred due to the effect introduced by the mucosal wave.

8.2 Future Works

In this work, a new way of analyzing laryngeal images based on OF has been explored but there are still many different tests and experiments that need to be addressed in future works:

- In contrast to glottal area segmentation, the movement of the vocal folds can also be traced in the anterior-posterior direction using OF. Therefore, it is necessary to find out alternative ways to synthesize the information extracted from the OF. For instance, one alternative will be the use of PCA to reduce the dimensionality of the vector motion field similarly as proposal in (Chen et al., 2014) with uses the intensity of LHSV. Furthermore, it will be interesting to include the information of the y-axis and also to consider the deflections of the left and right folds separately.

- Another line of work is related to the classification of functional voice disorders using the OF playbacks. For instance, a set of numerical features, as the ones presented in (Voigt et al., 2010a; Unger et al., 2015; Döllinger et al., 2011) for traditional playbacks, can be derived from the OF playbacks which would capture the dynamic behavior and the symmetry of oscillating vocal folds. Some of the features could be the cycle duration, OFGVG contour, inter-cycle and intra-cycle symmetry. Additionally, the OF techniques can be combined with the ones based on glottal gap segmentation.

- From our experience, we know that is difficult to validate the motion field obtained via OF due to the lack of ground-truth in vivo data. But, we also know that is desirable to have an idea about the accuracy of the OF algorithm compared to reality and not only by comparison with the conventional segmentation techniques. Therefore, we want to explore the use of some kind of mechanical phantom to generate a reproducible motion, for example with moistened silicon vocal folds and recreate the laryngeal conditions (lighting, reflections). Firstly, record the motion of a smooth surface (comparable to in vivo vocal folds) and later apply a prominent pattern to the surface (reliable

to track by OF techniques). It would be very interesting to learn how these two measurements compare.

- When building the OFGVG and OFKG , the average velocity along the vocal folds and perpendicular to the glottal axis is taken. This also includes the mucosal wave activity that cannot be separated from the real motion of vocal fold edges directly. For that reason in section 7.6.3, the individual contribution of the mucosal wave was studied. However, the objective detection and quantification of mucosal wave propagation have to be studied in more detail since its existence and magnitude provides valuable information about the coupling between the mucosa and the subjacent vocal folds muscle. Additionally, it is widely held that mucosal wave activity is a useful indicator of the quality of voice production and the presence of voice disorders (Krausert et al., 2011; Shaw and Deliyski, 2008; Voigt et al., 2010b)

- Currently, we are presenting three different visualizations which are evaluated separately. It might be helpful to show how these methods provide complementary information and combining them put additional information to the context. In addition, it will be very interesting to make a systematic comparison with other glottal-activity signals such as Electroglottography (EGG) in order to provide a more complete assessment of vocal folds vibration.

- We are also interested in investigating the influence of the ventricular folds in some particular phonation, as the one in the bass type of Mongolian throat singing, using OF techniques to understand how much the presence of ventricular fold vibration contributes to a change in voice quality.

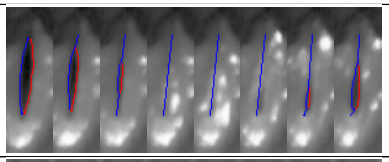
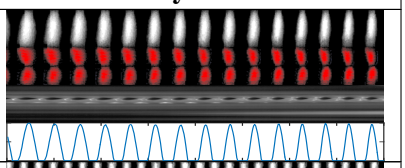
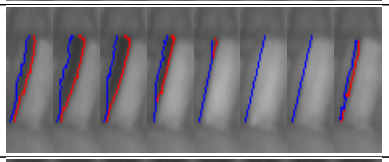
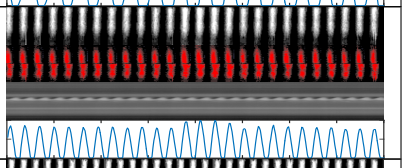
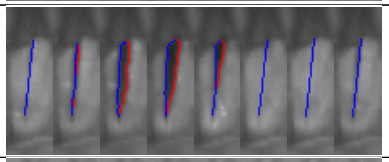
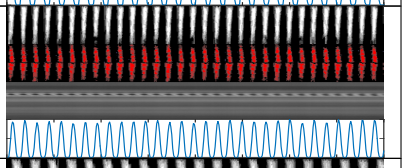
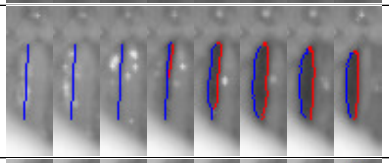
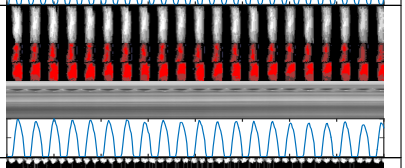
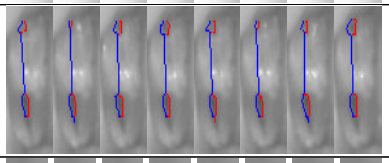
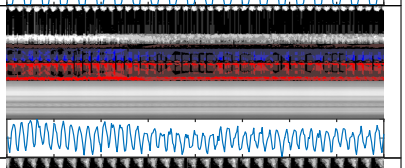
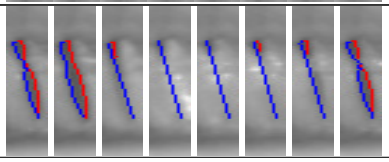
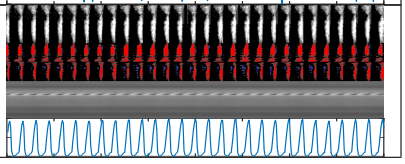
- In this work, we only use the segmentation and OF approaches to estimate the motion of the vocal folds. However, we are interested in exploring other techniques as registration to characterize the kinematics of the vocal folds. The image registration is one of the fundamental tasks within image processing which let to find an optimal geometric transformation between two corresponding image. Therefore, we want to explore if it is possible to model the deformation of the vocal folds for each particular laryngeal mechanism using geometric transformations.

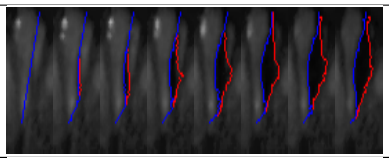
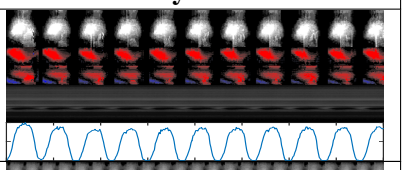
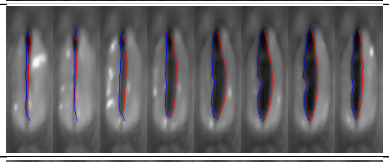
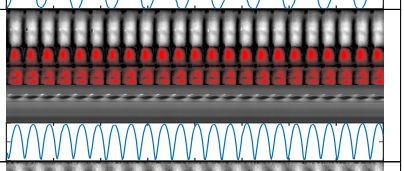
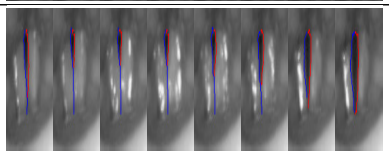
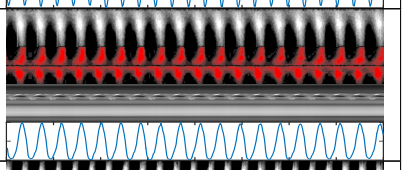
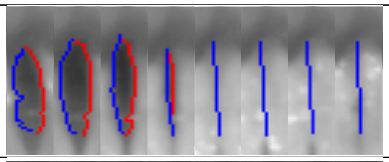
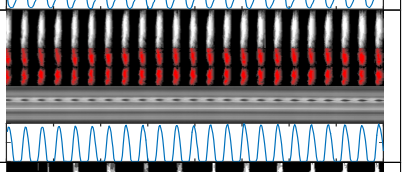
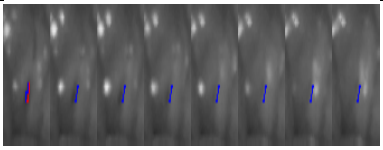
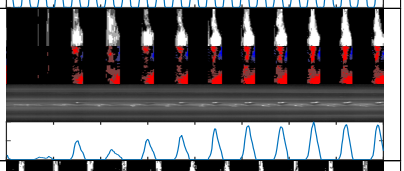
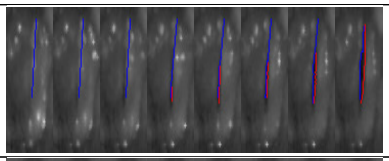
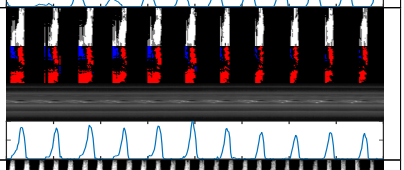
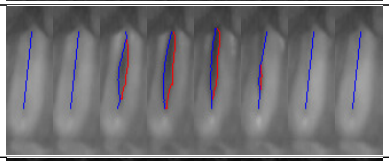
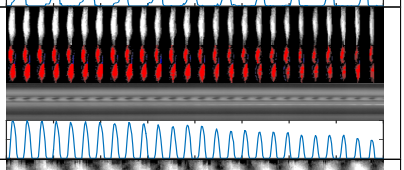
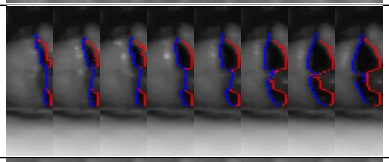
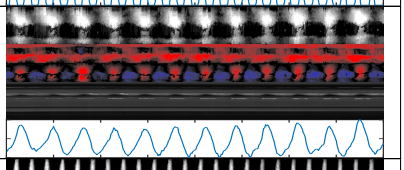
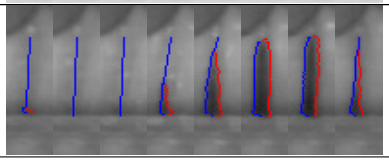
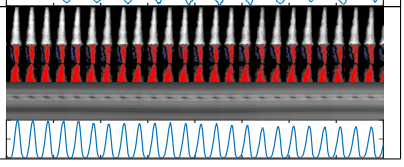
Part V

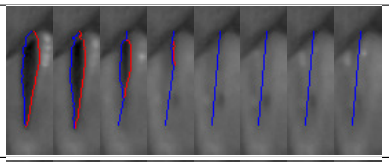
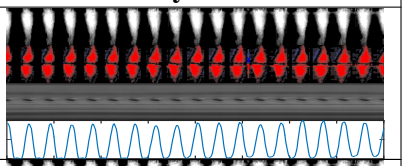
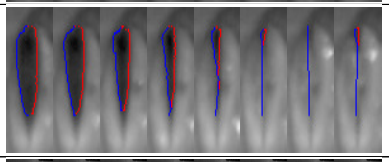
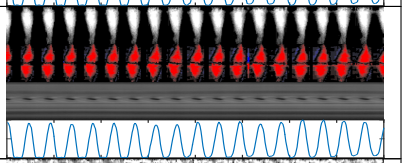
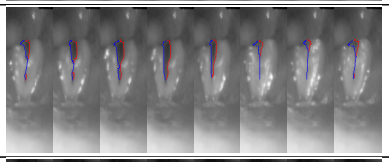
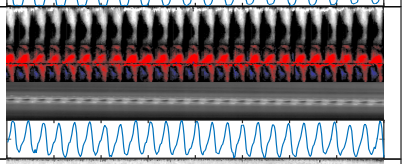
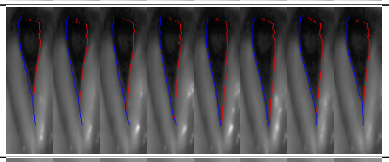
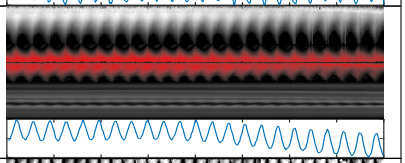
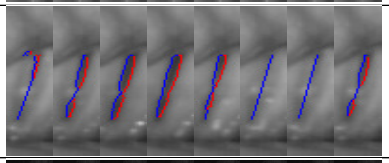
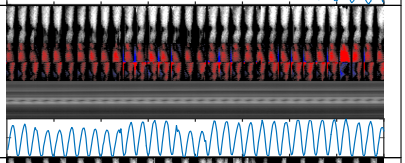
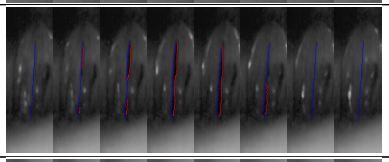
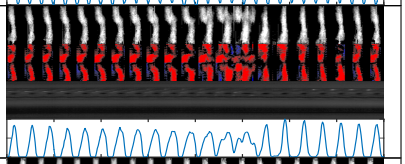
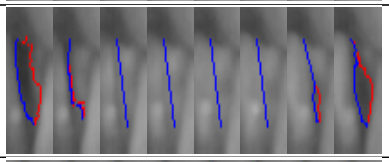
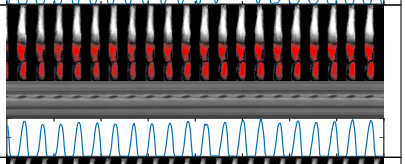
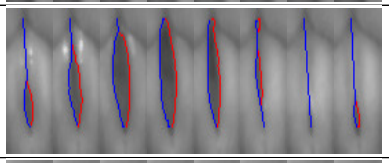
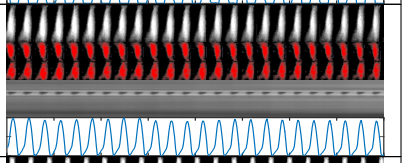
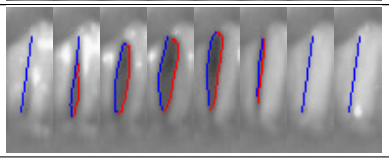
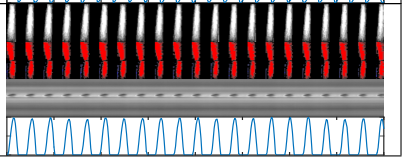
Appendices

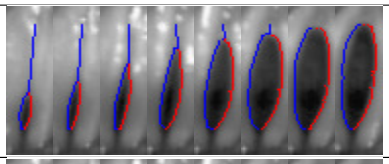
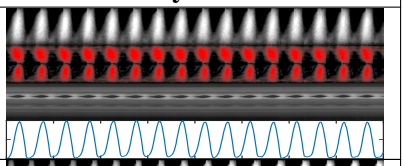
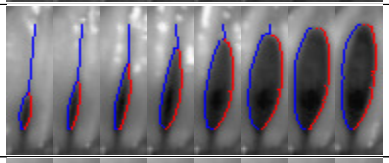
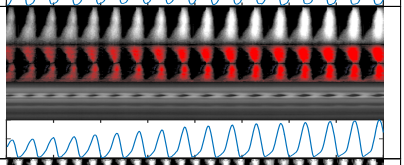
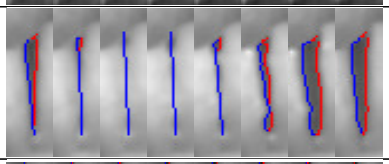
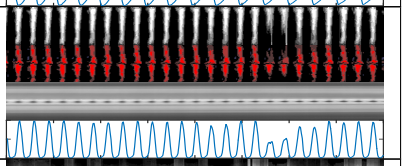
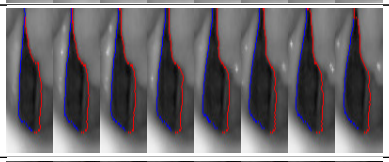
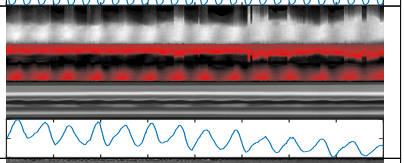
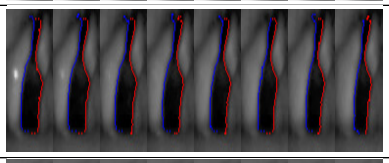
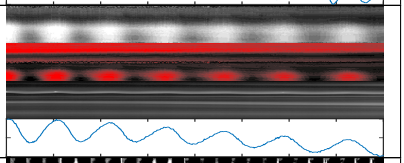
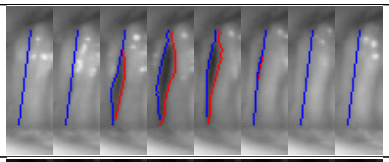
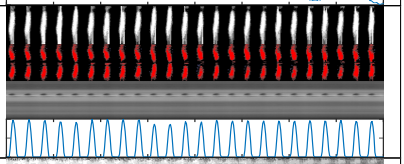
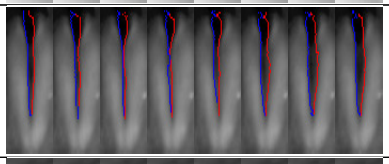
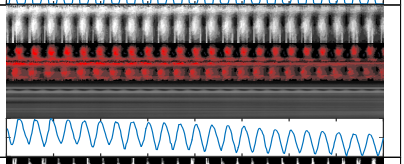
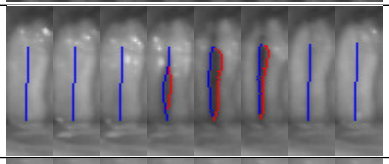
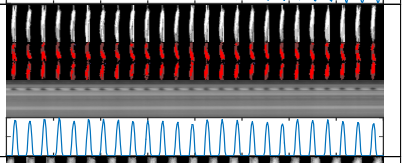
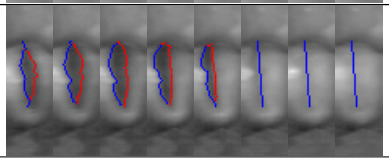
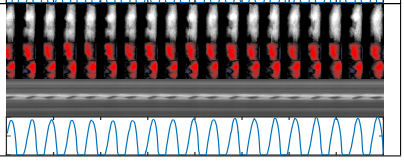
Appendix A

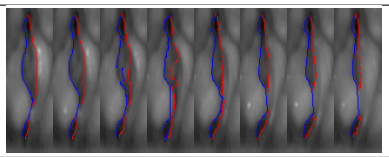
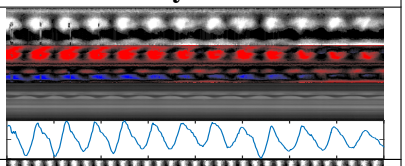
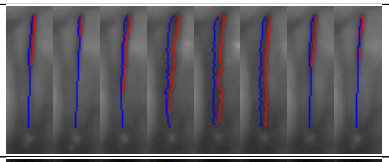
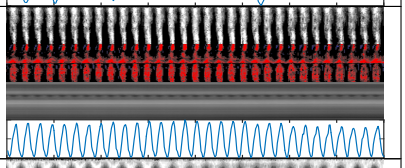
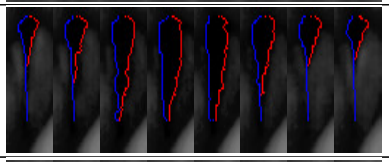
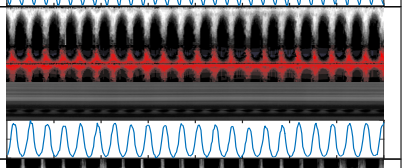
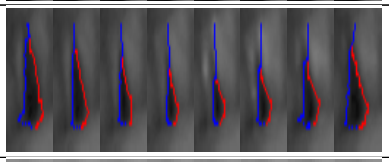
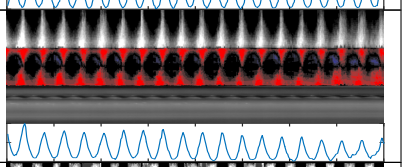
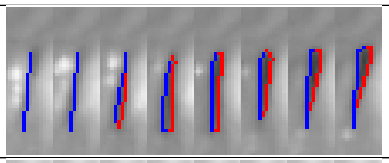
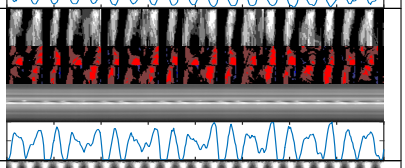
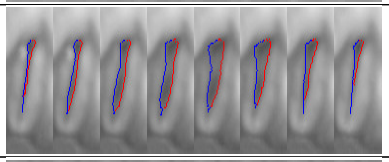
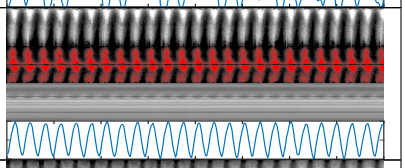
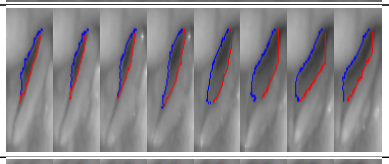
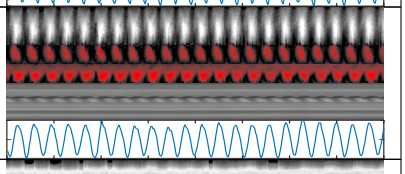
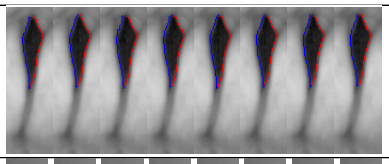
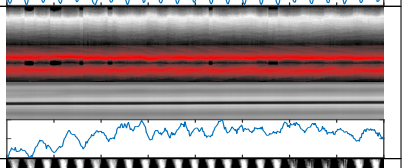
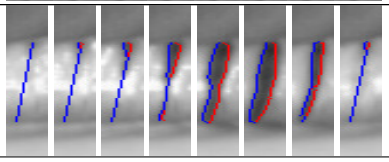
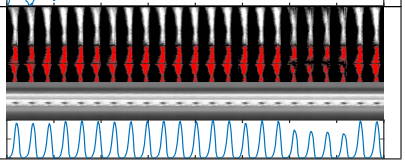
Playbacks Computed Using InP

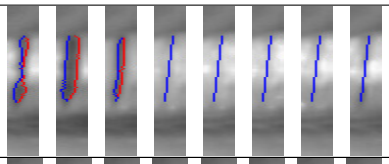
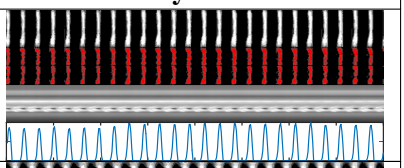
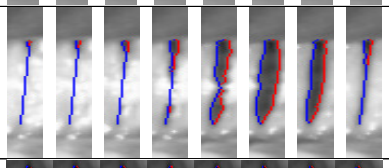
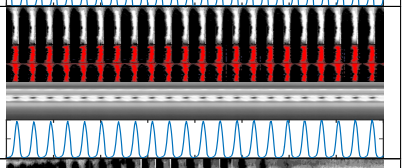
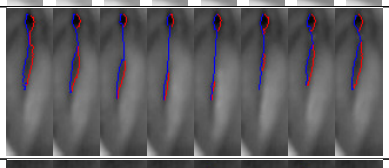
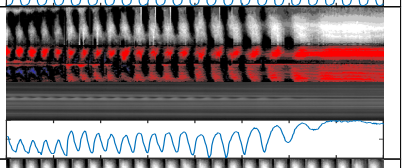
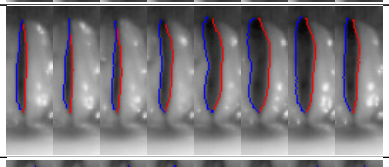
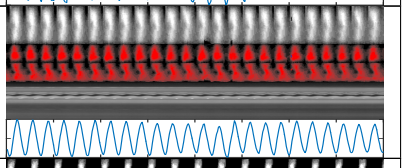
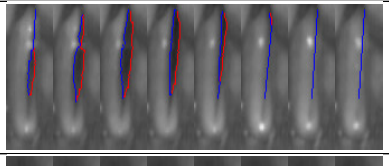
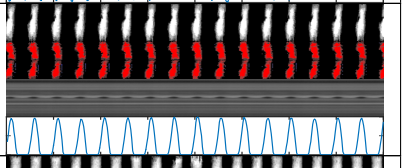
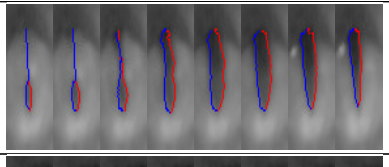
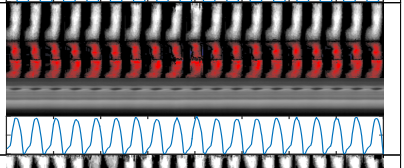
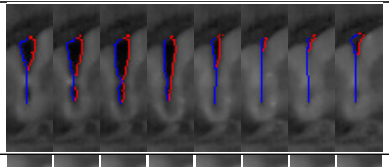
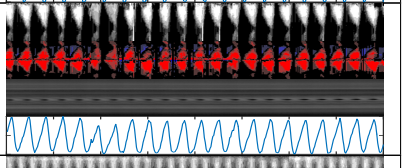
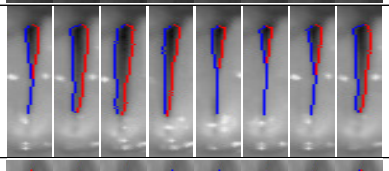
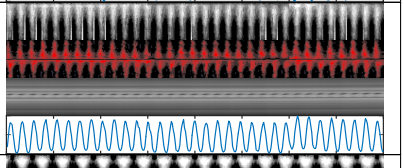
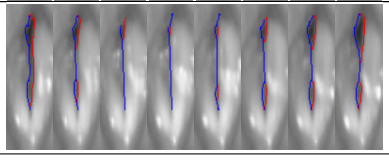
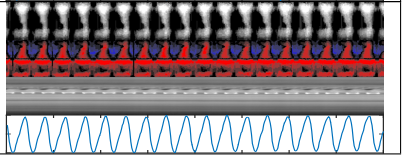
Index	Frames	Playbacks
1		
2		
3		
4		
5		
6		

Index	Frames	Playbacks
7		
8		
9		
10		
11		
12		
13		
14		
15		

Index	Frames	Playbacks
16		
17		
18		
19		
20		
21		
22		
23		
24		

Index	Frames	Playbacks
25		
26		
27		
28		
29		
30		
31		
32		
33		

Index	Frames	Playbacks
34		
35		
36		
37		
38		
39		
40		
41		
42		

Index	Frames	Playbacks
43		
44		
45		
46		
47		
48		
49		
50		
51		

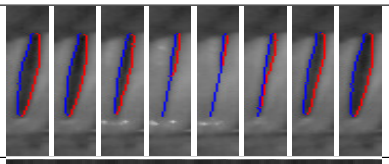
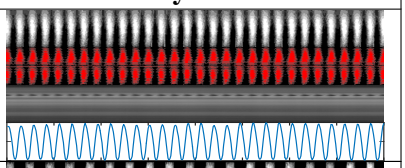
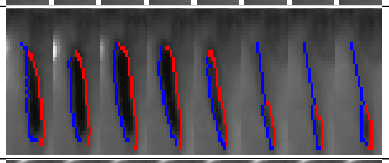
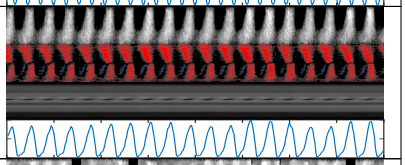
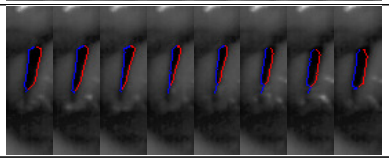
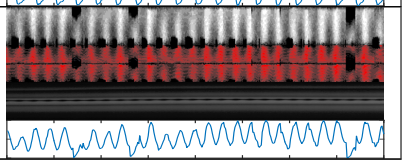
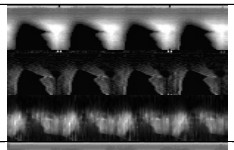
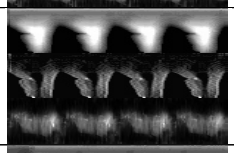
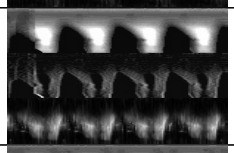
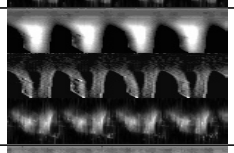
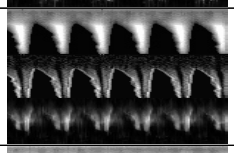
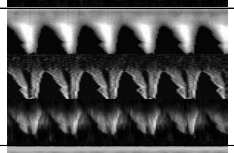
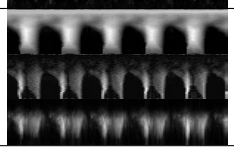
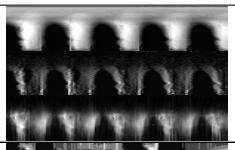
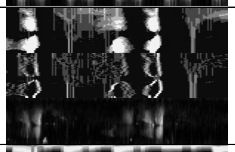
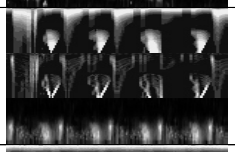
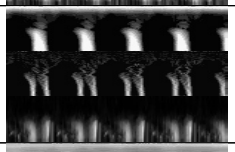
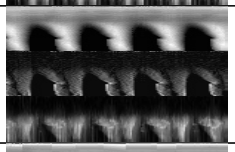
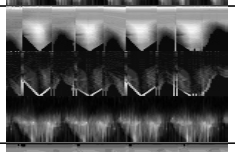
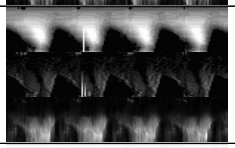
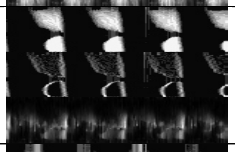
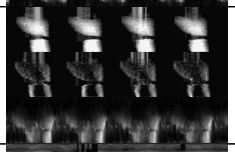
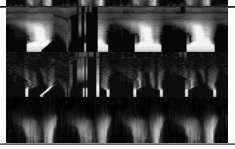
Index	Frames	Playbacks
52		
53		
54		

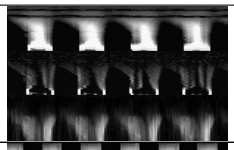
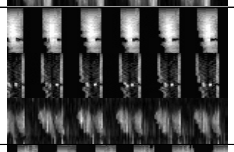
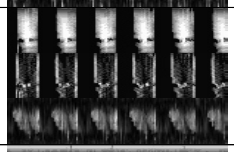
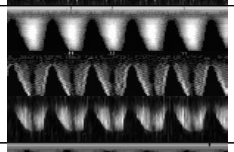
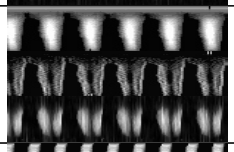
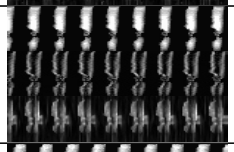
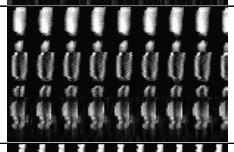
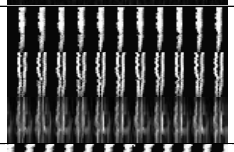
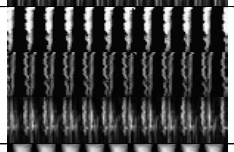
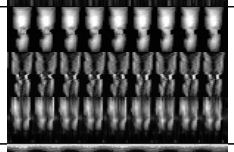
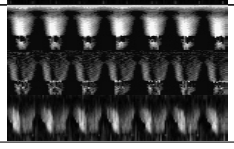
Table A.1: Used recordings from the database DB2. The LHSV sequences indices are shown, as well as the glottal segmentation at time t_k=1, 5, 10, 15, 20. The last column from top to down presents the corresponding GVG, PVG, DKG and GAW playbacks.

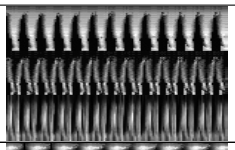
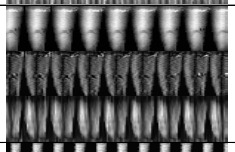
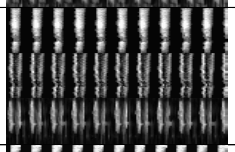
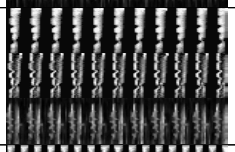
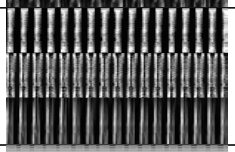
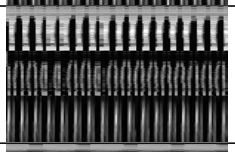
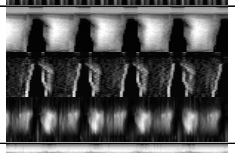
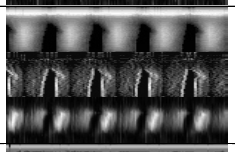
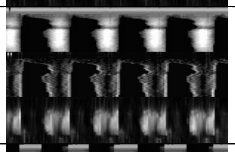
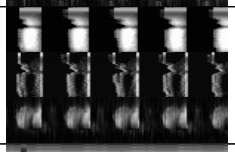
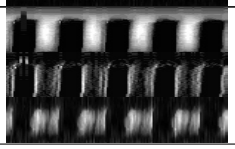
Appendix B

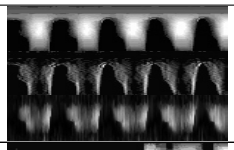
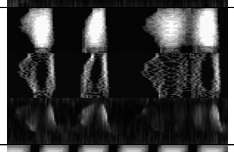
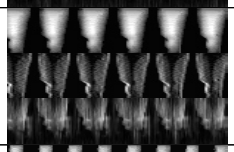
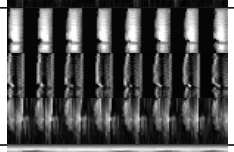
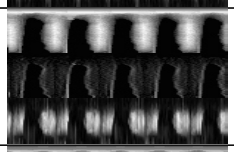
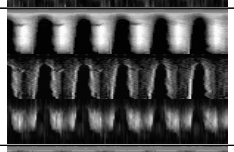
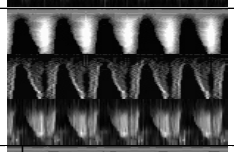
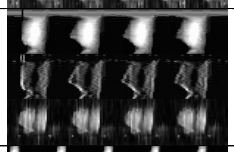
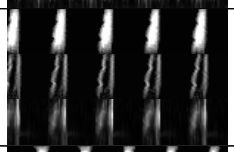
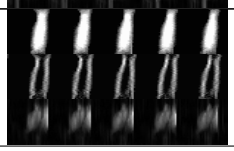
OFGVG Playback Thumbnails

Index	Description	Glottal Movement	Thumbnail
1	normal	anterior-to-posterior	
2	normal	anterior-to-posterior	
3	normal	anterior-to-posterior	
4	normal	anterior-to-posterior	
5	normal	middle-to-edges	
6	normal	middle-to-edges	
7	breathy	middle-to-edges	

Index	Description	Glottal Movement	Thumbnail
8	breathy	anterior-to-posterior	
9	creaky	anterior-to-posterior	
10	pressed	complex vibration	
11	pressed	complex vibration	
12	breathy	middle-to-edges	
13	breathy	middle-to-edges	
14	breathy	anterior-to-posterior	
15	breathy	anterior-to-posterior	
16	creaky	anterior-to-posterior	
17	creaky	anterior-to-posterior	
18	pressed	anterior-to-posterior	

Index	Description	Glottal Movement	Thumbnail
19	pressed	anterior-to-posterior	
20	pitch D3 in M1	middle-to-edges	
21	pitch D3 in M1	middle-to-edges	
22	pitch D3 in M1	anterior-to-posterior	
23	pitch D3 in M1	anterior-to-posterior	
24	pitch A3 in M1	anterior-to-posterior	
25	pitch A3 in M1	anterior-to-posterior	
26	pitch D4 in M1	anterior-to-posterior	
27	pitch D4 in M1	anterior-to-posterior	
28	glissando	anterior-to-posterior	
29	glissando	anterior-to-posterior	

Index	Description	Glottal Movement	Thumbnail
30	glissando	anterior-to-posterior	
31	pitch A3 in M2	anterior-to-posterior	
32	pitch D4 in M2	anterior-to-posterior	
33	pitch D4 in M1	anterior-to-posterior	
34	pitch A4 in M2	anterior-to-posterior	
35	pitch A4 in M2	anterior-to-posterior	
36	breathy	anterior-to-posterior	
37	breathy	anterior-to-posterior	
38	normal	anterior-to-posterior	
39	normal	posterior-to-anterior	
40	breathy	anterior-to-posterior	

Index	Description	Glottal Movement	Thumbnail
41	breathy	anterior-to-posterior	
42	creaky	medial-to-edges	
43	creaky	medial-to-edges	
44	glide down	posterior-to-anterior	
45	glide down	posterior-to-anterior	
46	breathy	posterior-to-anterior	
47	breathy	posterior-to-anterior	
48	glide up	posterior-to-anterior	
49	glide up	posterior-to-anterior	
50	pressed	anterior-to-posterior	
51	pressed	anterior-to-posterior	

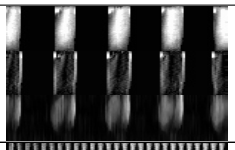
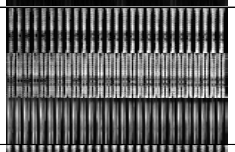
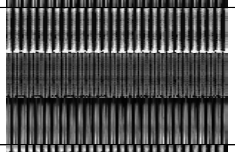
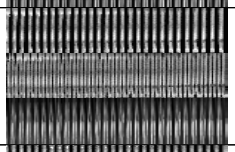
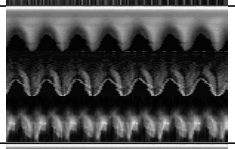
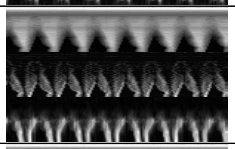
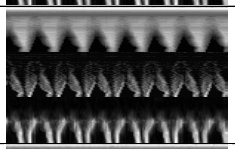
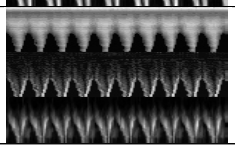
Index	Description	Glottal Movement	Thumbnail
52	pressed	posterior-to-anterior	
53	pitch F4# in M2 soft vibrato	posterior-to-anterior	
54	pitch F4# in M2 soft vibrato	posterior-to-anterior	
55	pitch F4# in M2 soft vibrato	posterior-to-anterior	
56	pitch F4# in M2 soft vibrato	posterior-to-anterior	
57	breathy glide down	posterior-to-anterior	
58	breathy glide down	posterior-to-anterior	
59	breathy glide down	posterior-to-anterior	
60	breathy glide up	posterior-to-anterior	

Table B.1: Used recordings from the database DB3. The LHSV sequences indices are shown, as well as the used laryngeal mechanisms and the glottal vibration pattern. The last column from top to down presents a small portion of the corresponding GVG, $|d_x\text{GVG}|$ and OFGVG-TVL1 thumbnail.

Appendix C

Scientific Production

Journal Articles

[J1] **Andrade-Miranda Gustavo** and Juan I. Godino-Llorente, *Glottal Gap tracking by a continuous background modeling using inpainting*, Medical & Biological Engineering & Computing. **In Press**. (JCR, IF=1.79).

[J2] **Andrade-Miranda Gustavo**, Henrich Bernardoni Nathalie, and Juan I. Godino-Llorente, *Synthesizing the motion of the vocal folds using optical flow based techniques*, Biomedical Signal Processing and Control, Vol 34, April 2017, Pages 25-35. (JCR, IF=1.52).

[J3] Laureano Moro-Velázquez, Jorge Andrés Gómez-García, Juan Ignacio Godino-Llorente, and **Andrade-Miranda Gustavo**, *Modulation Spectra Morphological Parameters: A New Method to Assess Voice Pathologies according to the GRBAS Scale*, BioMed Research International, vol. 2015, Article ID 259239, 13 pages. (JCR, IF=2.17)

[J4] **Andrade-Miranda Gustavo** , Juan I Godino-Llorente, Laureano Moro-Velázquez and Jorge Andrés Gómez-García, *An AutomaticMethod to Detect and Track the Glottal Gap from High Speed Videoendoscopic Images*, BioMedical Engineering OnLine, vol 14, June 2015, 26 pages. (JCR, IF=1.38)

Peer-Reviewed Conference Articles

[C1] **Andrade-Miranda, Gustavo**, Bernardoni, Nathalie Henrich and Godino-Llorente, Juan Ignacio. *A new method to present high-speed data for laryngeal assessment based on Optical Flow computation*, ICVPB - 10th International Conference on Voice Physiology and Biomechanics, Viña del Mar, Chile, March 14 - 17, 2016.

[C2] **Andrade-Miranda, Gustavo**, Bernardoni, Nathalie Henrich and Godino-Llorente, Juan Ignacio. *A New Technique for Assessing Glottal Dynamics in Speech and Singing by Means of Optical-Flow Computation*, INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresdren, Germany.

[C3] **Andrade-Miranda, Gustavo**, Bernardoni, Nathalie Henrich and Godino-Llorente, Juan Ignacio. *Optical-Flow Kymograms and Glottovibrograms: A new way to present high-speed data for laryngeal assessment*, MAVEBA 2013 - 9th International workshop, Models and analysis of vocal emissions for biomedical applications, Firenze, Italy, September 2-4, 2015.

[C4] **Andrade-Miranda, Gustavo**, Juan Ignacio Godino-Llorente. *ROI detection in high speed laryngeal images*, Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium, pages 477-480, Beijing, China, April 29- May 2, 2014.

[C5] **Andrade-Miranda, Gustavo** and Juan Ignacio Godino-Llorente. *Glottal gap tracking using temporal intensity variation and active contours*, MAVEBA 2013 - 8th International workshop, Models and analysis of vocal emissions for biomedical applications, pages 77-80, Firenze, Italy, December 16-18, 2013.

[C6] **Andrade-Miranda Gustavo** and Juan Ignacio Godino-Llorente. *Automatic glottal tracking from high-speed digital images using a continuous normalized cross correlation*, INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, pages 1144-1148, Lyon, France, August 25-29, 2013.

[C7] **Andrade-Miranda Gustavo**, N. Saenz, V. Osma and J. Godino, *A new approach for the glottis segmentation using snakes*, BIOSTEC, 6th International Joint Conference on Biomedical Engineering Systems and Technologies. pages 318-322, Barcelona, Spain, February 11-14, 2013.

Workshops, Symposia, and Seminars

[S1] **Andrade-Miranda Gustavo**, Juan Ignacio Godino-Llorente and Bernardoni, Nathalie Henrich. *Optical Flow Glottovibrogram: Synthesizing the Vocal Fold Vibrations for Visualization and Analyzing the Laryngeal Dynamics*, URSI, XXXI Symposium Nacional de la Unión Científica Internacional de Radio. pág 99, Madrid, España, Septiembre 5-7, 2016.

[S2] **Andrade-Miranda Gustavo**, Juan Ignacio Godino-Llorente. *Detección de la región de interés en imágenes laríngeas de alta velocidad*, JRBP 2013

- VII Jornadas de Reconocimiento Biométrico de Personas, pág. 145-151, Escuela Politécnica Superior de Zamora, España, Septiembre 12-13, 2013.

[S3] Andrade-Miranda Gustavo , Juan Ignacio Godino-Llorente, *Seguimiento automático de la apertura glottal a partir de imágenes digitales de alta velocidad usando correlación cruzada adaptiva*, JVHC 2013 - I jornadas multidisciplinarias de usuarios de la voz, el habla y el canto, pág 143-151, Palmas de Gran Canaria, España, Junio 27-28, 2013.

[S4] Andrade-Miranda Gustavo, N. Saenz, V. Osma and J. Godino, *Glottis segmentation from laryngeal images using snakes*, in 2nd Workshop de Tecnologías Multibiométricas para la identificación de personas.

References

"and what with little sleep and much reading
his brains got so dry that he lost his wits"

Miguel de Cervantes Saavedra

ABDOU, I. E. and PRATT, W. K. Quantitative design and evaluation of enhancement/thresholding edge detectors. *Proceedings of the IEEE*, vol. 67, pages 753–763, 1979.

ADAMS, R. and BISCHOF, L. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16(6), pages 641–647, 1994.

AGHLMANDI, D. and FAEZ, K. Automatic Segmentation of Glottal Space from Video Images Based on Mathematical Morphology and the hough Transform. *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 2(2), pages 223–230, 2012.

AHMAD, K., YAN, Y. and BLESS, D. Vocal fold vibratory characteristics in normal female speakers from high-speed digital imaging. *Journal of Voice*, vol. 26(2), pages 239–253, 2012.

ALAOUI, E. I., MENDEZ, A., IBN-ELHAJ, E. and GARCIA, B. Keyframes detection and analysis in vocal folds recordings using hierarchical motion techniques and texture information. In *16th IEEE International Conference on Image Processing (ICIP)*, pages 653–656. Cairo, Egypt, 2009.

ANDRADE-MIRANDA, G., BERNARDONI, N. H. and GODINO-LLORENTE, J. I. A new technique for assessing glottal dynamics in speech and singing by means of optical-flow computation. In *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2182–2186. Dresden, Germany, 2015a.

ANDRADE-MIRANDA, G. and GODINO-LLORENTE, J. I. Roi detection in high speed laryngeal images. In *11th International Symposium on Biomedical Imaging (ISBI)*, pages 477–480. Beijing, China, 2014.

ANDRADE-MIRANDA, G., GODINO-LLORENTE, J. I., MORO-VELÁZQUEZ, L. and GÓMEZ-GARCÍA, J. A. An automatic method to detect and track the glottal gap from high speed videoendoscopic images. *BioMedical Engineering OnLine*, vol. 14(1), page 100, 2015b.

ANDRADE-MIRANDA, G., HENRICH, N. and GODINO-LLORENTE, J. I. Optical-flow kymograms and glottovibrograms: A new way to present high-speed data for laryngeal assessment. In *9th International workshop, Models and Analysis of Vocal Emissions for Biomedical Applications*, MAVEBA, pages 71–74. Firenze, Italy, 2015c.

ANDRADE-MIRANDA, G., SÁENZ-LECHÓN, N., NICOLÁS, O.-R. and GODINO-LLORENTE, J. I. A new approach for the glottis segmentation using snakes. In *International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS)*. Barcelona, Spain, 2013.

ARONSON, A. E. and BLESS, D. *Clinical voice disorders*. Thieme Medical Publishers, New York, USA, 2009.

BAILLY, L., HENRICH, N., MÜLLER, F., ROHLFS, A.-K. and HESS, M. Ventricular-fold dynamics in human phonation. *Journal of Speech, Language, and Hearing Research*, vol. 57(4), pages 1219–1242, 2014.

BAKER, S., SCHARSTEIN, D., LEWIS, J., ROTH, S., BLACK, M. and SZELISKI, R. Optical flow benchmark. http://vision.middlebury.edu/flow/, 2011a.

BAKER, S., SCHARSTEIN, D., LEWIS, J. P., ROTH, S., BLACK, M. J. and SZELISKI, R. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, vol. 92(1), pages 1–31, 2011b.

BALLARD, D. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, vol. 13(2), pages 111 – 122, 1981.

BARRON, J. L., FLEET, D. J. and BEAUCHEMIN, S. S. Performance of Optical Flow Techniques. *International Journal of Computer Vision*, vol. 1(12), pages 43–77, 1994.

BAY, H., ESS, A., TUYTELAARS, T. and GOOL, L. V. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, vol. 110(3), pages 346 – 359, 2008.

BEAUCHEMIN, S. S. and BARRON, J. L. The computation of optical flow. *ACM Computing Surveys*, vol. 27(3), pages 433–466, 1995.

BIEMOND, J., LOOIJENGA, L., BOEKEE, D. and PLOMPEN, R. A pel-recursive wiener-based displacement estimation algorithm. *Signal Processing*, vol. 13(4), pages 399 – 412, 1987.

BIRKHOLZ, P. Glottalimageexplorer - an open source tool for glottis segmentation in endoscopic high-speed videos of the vocal folds. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2016*. Dresden, Germany, 2016.

BJÖRCK, A. *Numerical Methods for Least Squares Problems*. Society for Industrial and Applied Mathematics, Amsterdam, Netherlands, 1st edition, 1996.

BLANCO, M., CHEN, X. and YAN, Y. A Restricted, Adaptive Threshold Segmentation Approach for Processing High-Speed Image Sequences of the Glottis. *Engineering*, vol. 5(10B), pages 357–362, 2013.

BLEAU, A. and LEON, L. Watershed-based segmentation and region merging. *Computer Vision and Image Understanding*, vol. 77(3), pages 317 – 370, 2000.

BLUE TREE PUBLISHING. Organic voice disorders. `http://www.bluetreepublishing.com`, Last checked: 20/01/2017, 2014.

BOHR, C., KRÄCK, A., DUBROVSKIY, D., EYSHOLDT, U., SVEC, J., PSYCHOGIOS, G., ZIETHE, A. and DÖLLINGER, M. Spatiotemporal Analysis of High-Speed Videolaryngoscopic Imaging of Organic Pathologies in Males. *Journal of Speech, Language, and Hearing Research*, vol. 57(4), pages 1148–1161, 2014.

BONILHA, H. S. and DELIYSKI, D. D. Period and glottal width irregularities in vocally normal speakers. *Journal of Voice*, vol. 22(6), pages 699–708, 2008.

BOOTH, J. R. and CHILDERS, D. G. Automated analysis of ultra high-speed laryngeal films. *IEEE Transactions on Biomedical Engineering*, vol. 26(4), pages 185–192, 1979.

BORCH, D. Z., SUNDBERG, J., LINDESTAD, P. A. and THALÉN, M. Vocal fold vibration and voice source aperiodicity in 'dist' tones: a study of a timbral ornament in rock singing. *Logopedics Phoniatrics Vocology*, vol. 29(4), pages 147–153, 2004.

BOYKOV, Y. and KOLMOGOROV, V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26(9), pages 1124–1137, 2004.

BOYKOV, Y., VEKSLER, O. and ZABIH, R. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(11), pages 1222–1239, 2001.

BRAUNSCHWEIG, T., FLASCHKA, J., SCHELHORN-NEISE, P. and DÖLLINGER, M. High-speed video analysis of the phonation onset, with an application to the diagnosis of functional dysphonias. *Medical Engineering and Physics*, vol. 30(1), pages 59–66, 2008.

BROX, T., BRUHN, A., PAPENBERG, N. and WEICKERT, J. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In *Computer Vision - ECCV 2004* (edit by T. Pajdla and J. Matas), vol. 3024, pages 25–36. Springer Berlin Heidelberg, 2004.

BRUHN, A., WEICKERT, J., KOHLBERGER, T. and SCHNÖRR, C. A multigrid platform for real-time motion computation with discontinuity-preserving variational methods. *International Journal of Computer Vision*, vol. 70(3), pages 257–277, 2006.

CANNY, J. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8(6), pages 679–698, 1986.

CASELLES, V., KIMMEL, R. and SAPIRO, G. Geodesic active contours. *International Journal of Computer Vision*, vol. 22(1), pages 61–79, 1997.

CERROLAZA, J. J., OSMA, V., VILLANUEVA, A., GODINO, J. I. and CABEZA, R. Full-AutoMatic Glottis Segmentation with active shape Models. In *7th international workshop, Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, vol. 9, pages 35–38. Firenze, Italy, 2011.

CHAN, T. F. and VESE, L. A. Active contours without edges. *IEEE Transactions on Image Processing*, vol. 10(2), pages 266–277, 2001.

CHANGSOO, J. and HYUNG-MIN, P. Optimized hierarchical block matching for fast and accurate image registration. *Signal Processing: Image Communication*, vol. 28(7), pages 779 – 791, 2013.

CHEN, G., KREIMAN, J. and ALWAN, A. The glottaltopogram: A method of analyzing high-speed images of the vocal folds. *Computer Speech and Language*, vol. 28(5), pages 1156 – 1169, 2014.

CHEN, J., GUNTURK, B. K. and KUNDUK, M. Glottis segmentation using dynamic programming. In *Medical Imaging 2013: Image Processing*, vol. 8669, pages 86693L–86693L–9. 2013.

CHILDERS, D. G., PAIGE, A. and MOORE, P. Laryngeal vibration patterns machine-aided measurements from high-speed film. *Archives of Otolaryngology*, vol. 102(7), pages 407–410, 1976.

CHOU, C.-H. and LI, Y.-C. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5(6), pages 467–476, 1995.

COBETA, I., NÚÑEZ, F. and FERNÁNDEZ, S. *Patología de la voz*. Marge Books, Barcelona, Spain, 1st edition, 2013.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, vol. 20(1), pages 37–46, 1960.

COMANICIU, D. and MEER, P. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24(5), pages 603–619, 2002.

COOTES, T. F., TAYLOR, C. J., COOPER, D. H. and GRAHAM, J. Active shape models; their training and application. *Computer Vision and Image Understanding*, vol. 61(1), pages 38–59, 1995.

DAY-FANN, S. and MING-TSONG, H. A watershed-based image segmentation using jnd property. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pages 377–380. Hong Kong, China, 2003.

DEJONCKERE, H. P., BRADLEY, P., CLEMENTE, P., CORNUT, G., CREVIER-BUCHMAN, L., FRIEDRICH, G., VAN DE HEYNING, P., REMACLE, M. and WOISARD, V. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *European Archives of Oto-Rhino-Laryngology*, vol. 258(2), pages 77–82, 2001.

DELIYSKI, D. Endoscope motion compensation for laryngeal high-speed videoendoscopy. *Journal of Voice*, vol. 19(3), pages 485–96, 2005.

DELIYSKI, D., HILLMAN, R. E. and MEHTA, D. D. Laryngeal high-speed videoendoscopy: Rationale and recommendation for accurate and consistent terminology. *Journal of Speech, Language, and Hearing Research*, vol. 58(5), pages 1488–1492, 2015a.

DELIYSKI, D., PETRUSHEV, P. P., BONILHA, H., GERLACH, T. T., MARTIN-HARRIS, B. and HILLMAN, R. E. Clinical implementation of laryngeal high-speed videoendoscopy: Challenges and evolution. *Folia Phoniatrica et Logopaedica*, vol. 60(1), pages 33–44, 2008.

DELIYSKI, D., POWELL, M. E., ZACHARIAS, S. R., GERLACH, T. T. and ALESSANDRO DE ALARCON. Experimental investigation on minimum frame rate requirements of high-speed videoendoscopy for clinical voice assessment. *Biomedical Signal Processing and Control*, vol. 17, pages 21 –28, 2015b.

DEMEYER, J., DUBUISSON, T., GOSSELIN, B. and REMACLE, M. Glottis segmentation with a high-speed glottography: a fully automatic method. In *3rd Advanced Voice Function Assessment International Workshop*, AVFA. Madrid, Spain, 2009.

DICE, L. Measures of the amount of ecologic association between species. *Ecology*, vol. 26(3), pages 297–302, 1945.

DÖLLINGER, M., HOPPE, U., HETTLICH, F., LOHSCHELLER, J., SCHUBERTH, S. and EYSHOLDT, U. Vibration parameter extraction from endoscopic image

series of the vocal folds. *IEEE Transactions on Biomedical Engineering*, vol. 49(8), pages 773–81, 2002.

DÖLLINGER, M., LOHSCHELLER, J., MCWHORTER, A. and KUNDUK, M. Variability of normal vocal fold dynamics for different vocal loading in one healthy subject investigated by phonovibrograms. *Journal of Voice*, vol. 23(2), pages 175–181, 2009.

DÖLLINGER, M., LOHSCHELLER, J., SVEC, J. G., MCWHORTER, A. and KUNDUK, M. *Advances in Vibration Analysis Research*, chapter 22. Support Vector Machine Classification of Vocal Fold Vibrations Based on Phonovibrogram Features, pages 435–456. InTech, 2011.

DRULEA, M. and NEDEVSCHI, S. Motion estimation using the correlation transform. *IEEE Transactions on Image Processing*, vol. 22(8), pages 3260–3270, 2013.

DUDA, R. O., HART, P. E. and STORK, D. G. *Pattern Classification*. Wiley-Interscience, New York, USA, 2nd edition, 2000.

EDWARDS, A. *An introduction to linear regression and correlation*. A series of books in psychology. Freeman, San Francisco, USA, 1976.

EFSTRATIADIS, S. N. and KATSAGGELOS, A. K. A model-based pel-recursive motion estimation algorithm. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pages 1973–1976. Albuquerque, USA, 1990.

ELIDAN, G. and ELIDAN, J. Vocal folds analysis using global energy tracking. *Journal of Voice*, vol. 26(6), pages 760–768, 2012.

FARNEBACK, G. Fast and accurate motion estimation using orientation tensors and parametric motion models. In *15th International Conference on Pattern Recognition (ICPR)*, vol. 1. Barcelona, Spain, 2000.

FELZENSZWALB, P. F. Representation and Detection of Deformable Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(2), pages 208 – 220, 2005.

FORTUN, D., BOUTHEMY, P. and KERVRANN, C. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, vol. 134, pages 1 – 21, 2015.

FREEPIK. General scheme of the voice production apparatus. `http://www.freepik.com/free-vector/organs-body_836934.htm#term=lungs&page=1&position=10`, Last checked: 25/01/2017, 2016.

GARCÍA, M. *Mémoire sur la voix humaine*. Duverger, 1847.

GELFER, M. P. and BULTEMEYER, D. Evaluation of vocal fold vibratory patterns in normal voices. *Journal of Voice*, vol. 4(4), pages 335–345, 1990.

GILLES, D. *Glottal source and vocal-tract separation, Estimation of glottal parameters, voice transformation and synthesis using a glottal model*. PhD Thesis, Université Paris IV - Pierre et Marie Curie, 2010.

GLOGER, O., LEHNERT, B., SCHRADE, A. and VOLZKE, H. Fully automated glottis segmentation in endoscopic videos using local color and shape features of glottal regions. *IEEE Transactions on Biomedical Engineering*, vol. 62(3), pages 795–806, 2015.

GODINO-LLORENTE, J. I. *Estrategias para la detección automática de patología laríngea a partir del registro de la voz*. PhD Thesis, Universidad Politécnica de Madrid, 2002.

GONZALEZ, R. C. and WOODS, R. E. Image segmentation. In *Digital Image Processing*, chapter 10. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 3rd edition, 2006.

GUILLEMOT, C. and MEUR, O. L. Image inpainting: Overview and recent advances. *IEEE Signal Processing Magazine*, vol. 31(1), pages 127–144, 2014.

HENRICH, N. *Etude de la source glottique en voix parlée et chantée: modélisation et estimation, mesures acoustiques et électroglottographiques, perception*. PhD Thesis, Université Paris-Orsay, 2001.

HENRICH, N. Mirroring the voice from Garcia to the present day: Some insights into singing voice registers. *Logopedics Phoniatrics Vocology*, vol. 31(1), pages 3–14, 2006.

HENRICH, N. *La voix humaine: vibrations, résonances, interactions pneumo-phono-résonantielles*. Accreditation to supervise research, Université Grenoble Alpes, 2015.

HENRICH, N., ROUBEAU, B. and CASTELLENGO, M. On the use of electroglottography for caracterisation of the laryngeal mechanisms. In *Stockholm Music Acoustics Conference*. Citeseer, Stockholm, Sweden, 2003.

HERBST, C. T., LOHSCHELLER, J., ŠVEC, J. G., HENRICH, N., WEISSENGRUBER, G. and FITCH, W. T. Glottal opening and closing events investigated by electroglottography and super-high-speed video recordings. *The Journal of Experimental Biology*, vol. 217(6), pages 955–963, 2014.

HERBST, C. T., UNGER, J., HERZEL, H., ŠVEC, J. G. and LOHSCHELLER, J. Phasegram analysis of vocal fold vibration documented with laryngeal high-speed video endoscopy. *Journal of Voice*, vol. 30(6), pages 771.e1–771.e15, 2016.

HERNANDEZ, S. E., BARNER, K. E. and YUAN, Y. Region merging using homogeneity and edge integrity for watershed-based image segmentation. *Optical Engineering*, vol. 44(1), pages 017004–017004–14, 2005.

HERTEGÅRD, S. What have we learned about laryngeal physiology from high-speed digital videoendoscopy? *Current Opinion in Otolaryngology and Head and Neck Surgery*, vol. 13(3), pages 152–156, 2005.

HIXON, T. J., WEISMER, G. and HOIT, J. D. *Preclinical Speech Science: Anatomy, Physiology, Acoustics, and Perception*, chapter 3. Laryngeal function and Speech Production, pages 81–170. Plural Pub, 2008.

HORN, B. K. and SCHUNCK, B. Determining optical flow: a retrospective. *Artificial Intelligence*, vol. 17, pages 185–203, 1981.

HOWARD, D. M. and MURPHY, D. T. *Voice Science Acoustics and Recording*. Plural Publishing, 2008.

HUBERT, L. and ARABIE, P. Comparing partitions. *Journal of Classification*, vol. 2(1), pages 193–218, 1985.

IKUMA, T., KUNDUK, M. and MCWHORTER, A. J. Advanced waveform decomposition for high-speed videoendoscopy analysis. *Journal of Voice*, vol. 27(3), pages 369 – 375, 2013.

ISHIZAKA, K. and FLANAGAN, J. L. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Labs Technical Journal*, vol. 51(6), pages 1233–1268, 1972.

JACKSON-MENALDI, M. C. A. *La voz patológica*. Médica Panamericana, Madrid, Spain, 1st edition, 2002.

JAVIER, S., MEINHARDT-LLOPIS, E. and FACCIOLO, G. TV-L1 Optical Flow Estimation. *Image Processing On Line*, vol. 1(1), pages 137–150, 2013.

JI, H., LIU, C., SHEN, Z. and XU, Y. Robust video denoising using low rank matrix completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1791–1798. 2010.

JUNGHWAN, O., SAE, H., JEONGKYU, L., WALLAPAK, T., JOHNNY, W. and PIET C, D. G. Informative frame classification for endoscopy video. *Medical Image Analysis*, vol. 11(2), pages 110 – 127, 2007.

KARAKOZOGLOU, S. Z., HENRICH, N., D'ALESSANDRO, C. and STYLIANOU, Y. Automatic glottal segmentation using local-based active contours and application to glottovibrography. *Speech Communication*, vol. 54(5), pages 641–654, 2012.

KASS, M., WITKIN, A. and TERZOPOULOS, D. Snakes: Active contour models. *International Journal of Computer Vision*, vol. 1(4), pages 321–331, 1988.

KENDALL, K. and LEONARD, R. *Laryngeal Evaluation: Indirect Laryngoscopy to High-speed Digital Imaging*. Thieme, New York, USA, 2010.

KO, T. and CILOGLU, T. Automatic segmentation of high speed video images of vocal folds. *Journal of Applied Mathematics*, vol. 2014, page 16, 2014.

KRAUSERT, C. R., OLSZEWSKI, A. E., TAYLOR, L. N., MCMURRAY, J. S., DAILEY, S. H. and JIANG, J. J. Mucosal wave measurement and visualization techniques. *Journal of Voice*, vol. 25(4), pages 395 – 405, 2011.

LANKTON, S. and TANNENBAUM, A. Localizing region-based active contours. *IEEE Transactions on Image Processing*, vol. 17(11), pages 2029–2039, 2008.

LARSSON, H., STELLAN, H., LINDESTAD, P.-A. and HAMMARBERG, B. Vocal fold vibrations: High-speed imaging, kymography, and acoustic analysis: A preliminary report. *The Laryngoscope*, vol. 110(12), pages 2117–2122, 2000.

LAVER, J. *The Phonetic Description of Voice Quality*, vol. 31 of *Cambridge Studies in Linguistics*. Cambridge University Press, 2009.

LEHMANN, E. and CASELLA, G. *Theory of Point Estimation*. Springer Verlag, New York, USA, 1998.

LI, L., GALATSANOS, N. P. and BLESS, D. Eigenfolds: a new approach for analysis of vibrating vocal folds. In *3rd International Symposium on Biomedical Imaging (ISBI)*. Washington DC, United States, 2002.

LIM, S. H., APOSTOLOPOULOS, J. G. and EL GAMAL, A. Optical flow estimation using temporally oversampled video. *IEEE Transactions on Image Processing*, vol. 14(8), pages 1074–1087, 2005.

LINDESTAD, P. A., SÖDERSTEN, M., MERKER, B. and GRANQVIST, S. Voice source characteristics in mongolian "throat singing" studied with high-speed imaging technique, acoustic spectra, and inverse filtering. *Journal of Voice*, vol. 15(1), pages 78 – 85, 2001.

LIU, C., YUEN, J. and TORRALBA, A. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 33(5), pages 978–994, 2011.

LOHSCHELLER, J. and EYSHOLDT, U. Phonovibrogram visualization of entire vocal fold dynamics. *The Laryngoscope*, vol. 118(4), pages 753–758, 2008a.

LOHSCHELLER, J. and EYSHOLDT, U. Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing

the underlying laryngeal dynamics. *IEEE Transactions on Medical Imaging*, vol. 27(3), pages 300–309, 2008b.

LOHSCHELLER, J., TOY, H., ROSANOWSKI, F., EYSHOLDT, U. and DOLLINGER, M. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Medical Image Analysis*, vol. 11(4), pages 400–413, 2007.

LOHSCHELLER, J., ŠVEC, J. G. and DÖLLINGER, M. Vocal fold vibration amplitude, open quotient, speed quotient and their variability along glottal length: kymographic data from normal subjects. *Logopedics Phoniatrics Vocology*, vol. 38(4), pages 182–192, 2013.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60(2), pages 91–110, 2004.

LUCAS, B. D. and KANADE, T. An iterative image registration technique with an application to stereo vision. In *7th International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 2, pages 674–679. Vancouver BC, Canada, 1981.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. California, USA, 1967.

MALLICK, S., ZICKLER, T., BELHUMEUR, P. and KRIEGMAN, D. Specularity removal in images and videos: A PDE approach. In *Computer Vision – ECCV 2006* (edit by A. Leonardis, H. Bischof and A. Pinz), vol. 3951 of *Lecture Notes in Computer Science*, pages 550–563. Springer Berlin Heidelberg, 2006.

MANFREDI, C., BOCHHI, L., BIANCHI, S., MIGALI, N. and CANTARELLA, G. Objective vocal fold vibration assessment from videokymographic images. *Biomedical Signal Processing and Control*, vol. 1(2), pages 129–136, 2006.

MARENDIC, B., GALATSANOS, N. and BLESS, D. A new active contour algorithm for tracking vibrating vocal fold. In *IEEE International Conference on Image Processing (ICIP)*, pages 397–400. Thessaloniki, Greece, 2001.

MARTÍNEZ, F., MANZANERA, A. and ROMERO, E. A motion descriptor based on statistics of optical flow orientations for action classification in video-surveillance. In *Multimedia and Signal Processing: Second International Conference (CMSP)*, pages 267–274. Springer Berlin Heidelberg, 2012.

MEHTA, D. D., DELIYSKI, D. D., QUATIERI, T. F. and HILLMAN, R. E. Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings. *Journal of Speech, Language, and Hearing Research*, vol. 54(1), pages 47–54, 2013.

MEHTA, D. D. and HILLMAN, R. E. Current role of stroboscopy in laryngeal imaging. *Current Opinion in Otolaryngology and Head and Neck Surgery*, vol. 20(6), pages 429–36, 2012a.

MEHTA, D. D. and HILLMAN, R. E. The evolution of methods for imaging vocal fold phonatory function. *SIG 5 Perspectives on Speech Science and Orofacial Disorders*, vol. 22(1), pages 5–13, 2012b.

MEILĂ, M. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, vol. 2777, pages 173–187. 2003.

MENDEZ, A., ALAOUI, E. I., GARCÍA, B., IBN-ELHAJ, E. and RUIZ, I. Glottal space segmentation from motion estimation and gabor filtering. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009*, pages 1–4. 2009.

MENZE, M., HEIPKE, C. and GEIGER, A. *Discrete Optimization for Optical Flow*, pages 16–28. Springer International Publishing, 2015.

MEYER, F. and BEUCHER, S. Morphological segmentation. *Journal of Visual Communication and Image Representation*, vol. 1(1), pages 21 – 46, 1990.

MITICHE, A. and AGGARWAL, J. *Computer Vision Analysis of Image Motion by Variational Methods*, vol. 10 of *Springer Topics in Signal Processing*. Springer International Publishing, 2014.

MOORE, P. and VON LEDEN, H. Dynamic variations of the vibratory pattern in the normal larynx. *Folia Phoniatrica et Logopaedica*, vol. 10(4), pages 205–238, 1958.

MOUKALLED, H. J., DELIYSKI, D. D., SCHWARZ, R. R. and WANG, S. Segmentation of laryngeal high-speed videondoscopy in temporal domain using paired active contours. In *6th International Workshop, Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, pages 137–140. Firenze, Italy, 2009.

MUMFORD, D. and SHAH, J. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, vol. 42(5), pages 577–685, 1989.

NEUBAUER, J., MERGELL, P., EYSHOLDT, U. and HERZEL, H. Spatio-temporal analysis of irregular vocal fold oscillations: Biphonation due to desynchronization of spatial modes. *The Journal of the Acoustical Society of America*, vol. 110(6), pages 3179–3192, 2001.

OSMA-RUIZ, V., GODINO-LLORENTE, J. I., SÁENZ-LECHÓN, N. and FRAILE, R. Segmentation of the glottal space from laryngeal images using the watershed transform. *Computerized Medical Imaging and Graphics*, vol. 32(3), pages 193–201, 2008.

OTSU, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 9(1), pages 62–66, 1979.

PALM, C., LEHMANN, T., BREDNO, J., NEUSCHAEFER-RUBE, C., KLAJMAN, S. and SPITZER, K. Automated analysis of stroboscopic image sequences by vibration profile. In *5th International Workshop on Advances in Quantitative Laryngology, Voice and Speech Research*. Aachen, Germany, 2001.

PARAGIOS, N. and DERICHE, R. Geodesic active regions and level set methods for supervised texture segmentation. *International Journal of Computer Vision*, vol. 46(3), pages 223–247, 2002.

PARIS, S., KORNPROBST, P., TUMBLIN, J. and DURAND, F. Bilateral filtering: Theory and applications. *Foundations and Trends in Computer Graphics and Vision*, vol. 4(1), pages 1–73, 2009.

PINHEIRO, A. P., DAJER, M. E., HACHIYA, A., MONTAGNOLI, A. N. and TSUJI, D. Graphical evaluation of vocal fold vibratory patterns by high-speed videolaryngoscopy. *Journal of Voice*, vol. 28(1), pages 106 – 111, 2014.

PINHEIRO, A. P., STEWART, D. E., MACIEL, C. D., PEREIRA, J. C. and OLIVEIRA, S. Analysis of nonlinear dynamics of vocal folds using high-speed video observation and biomechanical modeling. *Digital Signal Processing*, vol. 22(2), pages 304–313, 2012.

POHLE, R. and TOENNIES, K. D. Segmentation of medical images using adaptive region growing. In *Proc. SPIE, Medical Imaging 2001: Image Processing*, vol. 4322, pages 1337–1346. San Diego, USA, 2001.

POLAK, M., ZHANG, H. and PI, M. An evaluation metric for image segmentation of multiple objects. *Image and Vision Computing*, vol. 27(8), pages 1223 – 1227, 2009.

RAHMAN, A. I. A., SALLEH, S.-H., AHMAD, K. and ANUAR, K. Analysis of vocal fold vibrations from high-speed digital images based on dynamic time warping. *International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, vol. 8(6), pages 306 – 309, 2014.

RAND, W. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, vol. 66(336), pages 846–850, 1971.

REDDY, B. S. and CHATTERJI, B. N. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, vol. 5(8), pages 1266–1271, 1996.

RIDLER, T. W. and CALVARD, S. Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 8(8), pages 630–632, 1978.

ROERDINK, J. B. and MEIJSTER, A. The watershed transform: Definitions, algorithms and parallelization strategies. *Journal of Fundamental Informaticae*, vol. 41(1,2), pages 187–228, 2000.

ROUBEAU, B. *Mécanismes vibratoires laryngés et contrôle neuro-musculaire de la fréquence fondamentale*. PhD Thesis, Université Paris-Orsay, 1993.

ROUBEAU, B., HENRICH, N. and CASTELLENGO, M. Laryngeal Vibratory Mechanisms: The Notion of Vocal Register Revisited. *Journal of Voice*, vol. 23(4), pages 425–438, 2009.

RUSS, J. C. *Image Processing Handbook*. CRC Press, Inc., Boca Raton, FL, USA, 4th edition, 2002.

SAADAH, A. K., GALATSANOS, N. P., BLESS, D. and RAMOS, A. Deformation analysis of the vibrational patterns of the vocal folds. In *Bildverarbeitung für die Medizin*. 1996.

SAKAKIBARA, K.-I., IMAGAWA, H., KIMURA, M., YOKONISHI, H. and TAYAMA, N. Modal analysis of vocal fold vibrations using laryngotopography. In *11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 917–920. Makuhari, Japan, 2010.

SCHENK, F., AICHINGER, P., ROESNER, I. and URSCHLER, M. Automatic high-speed video glottis segmentation using salient regions and 3d geodesic active contours. *Annals of the BMVA*, vol. 2015(3), pages 1–15, 2015.

SCHENK, F., URSCHLER, M., AIGNER, C., ROESNER, I., AICHINGER, P. and BISCHOF, H. Automatic glottis segmentation from laryngeal high-speed videos using 3d active contours. In *Medical Image Understanding and Analysis (MIUA)*, pages 111–116. 2014.

SCHULTZ, P. Vocal fold cancer. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, vol. 128(6), pages 301 – 308, 2011.

SCHUTTE, H. K., ŠVEC, J. G. and SRAM, F. First results of clinical application of Videokymography. *The Laryngoscope*, vol. 108(8), pages 1206–1210, 1998.

SCHWARZ, R., DÖLLINGER, M., WURZBACHER, T., EYSHOLDT, U. and LOHSCHELLER, J. Spatio-temporal quantification of vocal fold vibrations using high-speed videoendoscopy and a biomechanical model. *The Journal of the Acoustical Society of America*, vol. 123(5), pages 2717–32, 2008.

SEZGIN, M. and SANKUR, B. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, vol. 13(1), pages 146–168, 2004.

SHAW, H. S. and DELIYSKI, D. D. Mucosal wave: A normophonic study across visualization techniques. *Journal of Voice*, vol. 22(1), pages 23 – 33, 2008.

SIMPSON, B. and ROSEN, C. *Operative Techniques in Laryngology*, chapter 1. Anatomy and Physiology of the Larynx, pages 3–8. Springer Berlin Heidelberg, 2008.

SKALSKI, A., ZIELINKI, T. and DELIYSKI, D. Analysis of vocal folds movement in high speed videoendoscopy based on level set segmentation and image registration. In *International Conference on Signals and Electronic Systems (ICSES)*, pages 223–226. Krakow, Poland, 2008.

SKOURIKHINE, A. N., PRASAD, L. and SCHLEI, B. R. Neural network for image segmentation. In *International Symposium on Optical Science and Technology*, pages 28–35. San Diego, United States, 2000.

SUNDBERG, J. *Comprehensive Human Physiology: From Cellular Mechanisms to Integration*, vol. 1, chapter 53. The human voice, pages 1095–1104. Springer Berlin Heidelberg, 1996.

TAHA, A. A. and HANBURY, A. An efficient algorithm for calculating the exact hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37(11), pages 2153–2163, 2015a.

TAHA, A. A. and HANBURY, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, vol. 15(1), pages 1–28, 2015b.

TAO, C., ZHANG, Y. and JIANG, J. J. Extracting Physiologically Relevant Parameters of Vocal Folds From High-Speed Video Image Series. *IEEE Transactions on Biomedical Engineering*, vol. 54(5), pages 794–801, 2007.

TELEA, A. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, vol. 9(1), pages 23–34, 2004.

TIAN, L. and KAMATA, S. An iterative image enhancement algorithm and a new evaluation framework. In *IEEE International Symposium on Industrial Electronics*, pages 992–997. Cambridge, UK, 2008.

TIMCKE, R., VON LEDEN, H. and MOORE, P. Laryngeal vibrations: measurements of the glottic wave. I. The normal vibratory cycle. *Archives of Otolaryngology*, vol. 68(1), pages 1–19, 1958.

TIMCKE, R., VON LEDEN, H. and MOORE, P. Laryngeal vibrations: measurements of the glottic wave. II. Physiologic variations. *Archives of Otolaryngology*, vol. 69(4), pages 438–444, 1959.

TITZE, I. R. Comments on the myoelastic - aerodynamic theory of phonation. *Journal of Speech, Language, and Hearing Research*, vol. 23(3), pages 495–510, 1980.

TITZE, I. R. *Principles of voice production*. Prentice-Hall Inc, 1993.

TSAI, R. and OSHER, S. Level set methods and their applications in image science. *Communications in Mathematical Sciences*, vol. 1(4), pages 1–20, 2003.

UNGER, J., LOHSCHELLER, J., REITER, M., EDER, K., BETZ, C. S. and SCHUSTER, M. A Noninvasive Procedure for Early-Stage Discrimination of Malignant and Precancerous Vocal Fold Lesions Based on Laryngeal Dynamics Analysis. *Cancer Research*, vol. 75(1), pages 31–39, 2015.

UNGER, J., MEYER, T., HERBST, C. T., FITCH, W. T. S., DÖLLINGER, M. and LOHSCHELLER, J. Phonovibrographic wavegrams: Visualizing vocal fold kinematics. *The Journal of the Acoustical Society of America*, vol. 133(2), pages 1055–1064, 2013.

VAN DEN BERG, J. Myoelastic-aerodynamic theory of voice production. *Journal of Speech, Language, and Hearing Research*, vol. 1(3), pages 227–244, 1958.

VAN DEN BERG, J., ZANTEMA, J. T. and DOORNENBAL, P. On the air resistance and the bernoulli effect of the human larynx. *The Journal of the Acoustical Society of America*, vol. 29(5), pages 626–631, 1957.

VOIGT, D., DÖLLINGER, M., BRAUNSCHWEIG, T., YANG, A., EYSHOLDT, U. and LOHSCHELLER, J. Classification of functional voice disorders based on phonovibrograms. *Artificial Intelligence in Medicine*, vol. 49(1), pages 51–59, 2010a.

VOIGT, D., DÖLLINGER, M., EYSHOLDT, U., YANG, A., GÜRLEN, E. and LOHSCHELLER, J. Objective detection and quantification of mucosal wave propagation. *The Journal of the Acoustical Society of America*, vol. 128(5), pages EL347–EL353, 2010b.

VON LEDEN, H., MOORE, P. and TIMCKE, R. Laryngeal Vibrations: Measurements of the Glottic Wave Part III. The Pathologic Larynx. *Archives of Otolaryngology*, vol. 71(4), pages 16–35, 1960.

ŠVEC, J. G. and SCHUTTE, H. K. Videokymography: High-speed line scanning of vocal fold vibration. *Journal of Voice*, vol. 10(2), pages 201 – 205, 1996.

ŠVEC, J. G. and SCHUTTE, H. K. Kymographic imaging of laryngeal vibrations. *Current Opinion in Otolaryngology and Head and Neck Surgery*, vol. 20(6), pages 458–465, 2012.

ŠVEC, J. G. and ŠRAM, F. Kymographic imaging of the vocal folds oscillations. In *7th International Conference on Spoken Language Processing*, vol. 2, pages 957–960. 2002.

ŠVEC, J. G., ŠRAM, F. and SCHUTTE, H. K. Videokymography in Voice Disorders: What to Look For? *Annals of Otology, Rhinology and Laryngology*, vol. 116(3), pages 172–180, 2007.

WANG, Z., BOVIK, A., SHEIKH, H. and SIMONCELLI, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, vol. 13(4), pages 600–612, 2004.

WANG, Z. and BOVIK, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, vol. 26(1), pages 98–117, 2009.

WEDEL, A. and CREMERS, D. Optical flow estimation. In *Stereo Scene Flow for 3D Motion Analysis*, pages 5–34. Springer London, 2011.

WEICKERT, J., BRUHN, A., BROX, T. and PAPENBERG, N. A survey on variational optic flow methods for small displacements. In *Mathematical Models for Registration and Applications to Medical Imaging* (edit by O. Scherzer), vol. 10, pages 103–136. Springer Berlin Heidelberg, 2006.

WERLBERGER, M., POCK, T. and BISCHOF, H. Motion estimation with nonlocal total variation regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2464–2471. San Francisco, United States, 2010.

WESTPHAL, L. and CHILDERS, D. Representation of glottal shape data for signal processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31(3), pages 766–769, 1983.

WITTENBERG, T., MOSER, M., TIGGES, M. and EYSHOLDT, U. Recording, processing, and analysis of digital high-speed sequences in glottography. *Machine Vision and Applications*, vol. 8(6), pages 399–404, 1995.

WONG, D., ITO, M., COX, N. and TITZE, I. R. Observation of perturbations in a lumped-element model of the vocal folds with application to some pathological cases. *The Journal of the Acoustical Society of America*, vol. 89(1), pages 383–394, 1991.

WOO, P. Objective measures of laryngeal imaging: What have we learned since Dr. paul moore. *Journal of Voice*, vol. 28(1), pages 69–81, 2014.

WURZBACHER, T., SCHWARZ, R., DÖLLINGER, M., HOPPE, U., EYSHOLDT, U. and LOHSCHELLER, J. Model-based classification of nonstationary vocal fold vibrations. model-based classification of nonstationary vocal fold vibrations. *The Journal of the Acoustical Society of America*, vol. 120(2), pages 1012–1027, 2006.

XIAOHAN, Y., YLA-JAASKI, J., HUTTUNEN, O., VEHKOMAKI, T., SIPILA, O. and KATILA, T. Image segmentation combining region growing and edge detection. In *11th International Conference on Pattern Recognition (IAPR)*, vol. 3, pages 481–484. The Hague, Netherlands, 1992.

XIE, Y. and JI, Q. A new efficient ellipse detection method. In *16th International Conference on Pattern Recognition (ICPR)*, vol. 2, pages 957–960. IEEE, Quebec, Canada, 2002.

XU, C., PHAM, D. L. and PRINCE, J. L. Image segmentation using deformable models. In *Handbook of Medical Imaging. Vol.2 Medical Image Processing and Analysis*, pages 175–272. 2000.

XU, C. and PRINCE, J. L. Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing*, vol. 7(3), pages 359–369, 1998.

YAN, Y., CHEN, X. and BLESS, D. Automatic tracing of vocal-fold motion from high-speed digital images. *IEEE Transactions on Biomedical Engineering*, vol. 53(7), pages 1394–1400, 2006.

YAN, Y., DAMROSE, E. and BLESS, D. Functional analysis of voice using simultaneous high-speed imaging and acoustic recordings. *Journal of Voice*, vol. 21(5), pages 604–616, 2007.

YAN, Y., DU, G., ZHU, C. and MARRIOTT, G. Snake based automatic tracing of vocal-fold motion from high-speed digital images. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 593–596. Kyoto, Japan, 2012.

YILMAZ, A., JAVED, O. and SHAH, M. Object tracking: A survey. *ACM Computing Surveys*, vol. 38(4), page 13, 2006.

YUMOTO, E. Aerodynamics, voice quality, and laryngeal image analysis of normal and pathologic voices. *Current Opinion in Otolaryngology and Head and Neck Surgery*, vol. 12(3), pages 166–173, 2004.

ZACH, C., POCK, T. and BISCHOF, H. A duality based approach for realtime tv-l1 optical flow. In *Proceedings of the 29th DAGM Conference on Pattern Recognition*, pages 214–223. Springer-Verlag, Berlin, Heidelberg, 2007.

ZHANG, H., FRITTS, J. E. and GOLDMAN, S. A. A survey on evaluation methods for image segmentation. *Pattern Recognition*, vol. 29(8), pages 1335–1346, 1996.

ZHANG, Y., BIEGING, E., TSUI, H. and JIANG, J. J. Efficient and effective extraction of vocal fold vibratory patterns from high-speed digital imaging. *Journal of Voice*, vol. 24(1), pages 21 – 29, 2010.

ZHANG, Y., JIANG, J. J., TAO, C., BIEGING, E. and MACCALLUM, J. K. Quantifying the complexity of excised larynx vibrations from high-speed imaging using spatiotemporal and nonlinear dynamic analyses. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 17(4), pages 1–10, 2007.

ZHU, C., LIN, X. and CHAU, L.-P. Hexagon-based search pattern for fast block motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12(5), pages 349–355, 2002.

ZITOVÁ, B. and FLUSSER, J. Image registration methods: a survey. *Image and vision computing*, vol. 21, pages 977–1000, 2003.

–Qu'est-ce que signifie "apprivoiser"? – dit le Petit Prince–
–C'est une chose trop oubliée – dit le Renard–,
ça signifie "créer des liens..."

Tu n'es encore pour moi, qu'un petit garçon tout semblable à cent mille petits garçons. Et je n'ai pas besoin de toi. Et tu n'as pas besoin de moi non plus. Je ne suis pour toi qu'un renard semblable à cent mille renards. Mais, si tu m'apprivoises, nous aurons besoin l'un de l'autre. Tu seras pour moi unique au monde. Je serai pour toi unique au monde.....

Le Petit Prince... Antoine de Saint Exupéry