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
Máster en Ingeniería de Sistemas y Servicios
para la Sociedad de la Información

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

Máster en Ingeniería de Sistemas y Servicios
para la Sociedad de la Información
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

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

Vocal: D

Vocal: D

Vocal: D

Secretario: D

Revisores Remotos: D

Revisores Remotos: D

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 randonné. 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 Videendoscopy (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 Videendoscopy (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 procesamiento de imágenes.

Debido a la gran cantidad de información que debe ser evaluada tanto cualitativa como cuantitativamente es necesario sintetizar esta información espaciotemporal 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 quedó 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 Glottovibrograma (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. Las 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 accéder, 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 sensible 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 minimales 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 inclut 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 Detec- tion 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 <i>Myoelastic-Aerodynamic</i> theory, with a coronal and superior visualization of the vocal folds.	10
1.7	Schematic representation of the vocal tract with its different 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	Schematic 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 Kymography 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 applying 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 re-computing 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 variation 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 function.	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 specularly; 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 <i>good metrics</i> . 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 <i>good metrics</i> 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 $\hat{\delta}_{OF}^{l,r}(pc, t_k)$ during the opening phase.	145
7.7	First row: correlation between OF trajectories and segmentation trajectory for each sequence. Second and third row: $\hat{\delta}_{seg}^{l,r}$ and $\hat{\delta}_{TVL1}^{l,r}$ 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_xGVG $, OFGVG-LK, OFGVG-MT and OFGVG-TV L1 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 $ dGAW $; second row: GAW and GOFW-LK; third row: GAW and GOFW-MT; fourth row: GAW and GOFW-TV L1.	150
7.10	Correlation and SSIM obtained by comparing $ d_xGVG $ 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_xGVG $ 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_xGVG $ 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_xGVG $. The vibratory shape pattern for three consecutive cycles is marked with a continuous red line.	153
7.13	$ d_xGVG $ 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 TV L1-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 <i>good metrics</i> 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 <i>good metrics</i> . 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 Kymography
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	Glottal topogram
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 Glottovibrograph (OFGVG) and Glottal Optical Flow Waveform (GOFW), analyze the global dynamics; and the remaining one, called Optical Flow Kymograph (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 adduction¹ and abduction² 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

¹Adduction is the action to close the vocal folds.

²Abduction 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 propria. The second layer is the vocal ligament and is composed of the intermediate lamina propria and deep lamina propria. 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* (~ 80 Hz) and *mi5* (~ 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 source³.

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.

³It 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 different 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 Manuel 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 based 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-

⁴Translation 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 glissando⁵.

Mechanism M0	Mechanism M1	Mechanism M2	Mechanism M3
Fry	Modal	Falsetto	Whistle
Pulse	Normal	Head (W)	Flageolet
Stroh bass	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).

⁵An 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 , $k = \{1, 2, \dots, N\}$, is denoted as $I(\mathbf{x}, t_k)$. Therefore, the intensity of a given pixel $\mathbf{x}_{ij} = (x_i, y_j)$, $i = \{1, 2, \dots, n\}$ and $j = \{1, 2, \dots, m\}$, 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 space¹ 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

¹Color 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 thyroid 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

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

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:

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

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

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 $I(\mathbf{x}, t)$, $\{\mathbf{C}^{l,r}(t_k) \in \mathbb{R}^M, k = 1, \dots, N\}$, is denoted as $\mathbf{C}^{l,r}(t)$ (see eq. 2.4).

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

Mucosal Wave (MW): propagation of the epithelium and superficial layer of lamina propria 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 pathological² 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.

²The 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 procedure³. 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.

³Indirect 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).

⁴The 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 breaks⁵, laryngeal spasms⁶, 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

⁵Phonatory breaks are short interruptions of the phonatory process.

⁶Laryngeal spasms 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

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 propria 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.

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 (Lindstad 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 subharmonics⁷ 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 mucosa⁸ 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.

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

⁸Supraglottic 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 Kymography (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 $I_{DKG}^{R,G,B}(x, y)$ for color) is constructed with a set of N vectors $\{\mathbf{KG}(t_k) \in \mathbb{R}^n, k = 1, \dots, N\}$, where each $\mathbf{KG}(t_k)$ is a column vector of $I_{DKG}(x, y)$ (see eq 3.1).

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

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 Kymography at the line y_j . Adapted from (Švec and Šram, 2002).

3.2.2 VFT Playback

The VFT $\delta_{seg}^{l,r}(pc, t)$ 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.

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

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:

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

The vocal folds trajectory is obtained by eq 3.4 as:

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

where $\delta_{seg}^{l,r}(pc, t_k)$ are the deflections of the vocal folds edges at the point $\mathbf{c}_{pc}^{l,r}(t_k)$ 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 $G_{Co}(\mathbf{x}, t)$. Second row: vocal folds trajectories $\delta_{seg}^{l,r}(pc, t)$, DKG and DKG+VFT playbacks of five glottal cycles.

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

The first row in Figure 3.2 represents a particular glottal cycle $G_{Co}(\mathbf{x}, t)$, 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 $\mathbf{c}_{pc}^{l,r}(t_k)$, 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_{Co} . $\mathbf{KG}(t_k)$ 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 .

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

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

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

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.

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

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).

$$I_{key}(\mathbf{x}, t) = \arg \max_{k=1 \dots N} I(\mathbf{x}, t_k) \quad (3.9)$$

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.

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

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

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

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 $\delta_{seg}^{l,r}(pc, t_k)$; (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).

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 synthesizes 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 $\mathbf{c}_{pc}^l(t_k)$ and right $\mathbf{c}_{pc}^r(t_k)$ fold at position pc and time t_k .

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

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.

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

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).

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

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

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.

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, support 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 energy¹ 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.

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

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-

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-

giant 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.

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 (Biernond 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).

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

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)

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

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

eq 4.4

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

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

The BBC provides only one equation to recover the two unknown components of $\mathcal{W}(\mathbf{x}, t_k)$. Therefore, it is necessary to introduce an additional constraint encoding a priori information of $\mathcal{W}(\mathbf{x}, t_k)$. 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 Nedeveschi (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

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).

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

where $\mathcal{W}^2(\mathbf{x}_{ij})$ 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-

for 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:

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

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.

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, exemplar-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 professional painting restorators. 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 exemplar-based methods use image statistics and self similarity priors. The statistics of image textures are assumed to be stationary or homogeneous. 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.

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 applying 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 initialization 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:

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).

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	Gaussian 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 (Cerroloza 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 (Pineiro 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 neighbors.
- 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 Marañón 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.