

UNIVERSITY OF PARIS-SUD XI, COMPUTER SCIENCE DEPARTMENT  
UNIVERSITY OF CRETE, COMPUTER SCIENCE DEPARTMENT  
LIMSI-CNRS, AUDIO & ACOUSTICS GROUP

**Glottal source analysis: a combinatory  
study using high-speed videoendoscopy and  
electroglottography**

Sevasti-Zoi Karakozoglou

Master of Science Thesis

Paris, June 2010



UNIVERSITY OF PARIS-SUD XI, COMPUTER SCIENCE DEPARTMENT  
UNIVERSITY OF CRETE, COMPUTER SCIENCE DEPARTMENT  
LIMSI-CNRS, AUDIO & ACOUSTICS GROUP

**Glottal source analysis: a combinatorial study using high-speed  
videoendoscopy and electroglottography**

A thesis submitted by

**Sevasti-Zoi Karakozoglou**

in partial fulfillment for the  
Master of Science Degree

Author:

---

Sevasti-Zoi Karakozoglou

Supervisor:

---

Christophe d'Alessandro  
Directeur de Recherche, LIMSI-CNRS

Committee:

---

---

---

June 2010



# Glottal source analysis: a combinatory study using high-speed videoendoscopy and electroglottography

Sevasti-Zoi Karakozoglou

Master of Science Thesis

University of Paris-Sud XI, Computer Science Department

University of Crete, Computer Science Department

LIMSI-CNRS

Audio & Acoustics Group

## Abstract

Voice is our primary communication tool. This human instrument is important for many professionals such as singers and teachers. With the appearance of various voice disorders, clinical studies dictate the necessity of voice function assessment. Investigation techniques seek to provide for comprehensive information on the voice production mechanism and more specifically, the examination of vocal-folds vibratory behavior and the corresponding acoustic features.

The vocal folds are composed of twin infoldings of mucous membrane stretched horizontally across the larynx. Their vibration modulates the air flow expelled from the lungs during phonation, thus functioning as an acoustic excitation, i.e. the source of voiced speech. Electroglottography is a non-invasive investigation technique which measures the impedance related to the glottal area, that is the area between the vocal cords. Direct observation of the vocal-fold vibrations can only be achieved with invasive imaging techniques, by recording the vocal-fold vibrations from a top-view position in the larynx.

High-speed videoendoscopy is the most promising approach to directly assess vocal-fold vibrations. Automatic segmentation of the glottal area from the acquired data is a major challenge. In this work, we address this problem by proposing a local-based active contours framework for glottis segmentation. The method provides localizing of the glottal area with high accuracy and allows the contour to split and merge, according to the vocal-fold vibratory pattern. The method is fully automatic and parameter selection is performed automatically, based on the sequence statistics and empirical observations performed on a wide range of high-speed sequences.

Manual verification resulted on average in  $\pm 10$  pixel modifications, which correspond to less than 1% of the average glottal area. The vast amount of data that has to be evaluated both quantitatively and qualitatively dictate the need for dimensionality reduction of the spatio-temporal information and efficient representation of the high-speed data in a compact, handy and lossless way. One and two dimensional representations have been used for that reason. A comparison has been conducted on glottal parameters (fundamental frequency, open quotient) estimated on the electroglottographic signal and the extracted glottal area signal, using the electroglottographic signal as reference.

Supervisor

**Christophe d'Alessandro**

**Research Director, LIMSI-CNRS**



# Analyse de la source glottique: une étude combinatoire par vidéoendoscopie ultra-rapide et électroglottographie

Sevasti-Zoi Karakozoglou

Rapport - Master 2 Recherche Informatique

Université Paris-Sud XI, Département d'Informatique

Université de Crète, Département d'Informatique

LIMSI-CNRS

Groupe Audio-Acoustique

## Résumé

La voix est notre outil de communication principal. Cet instrument humain est important pour de nombreux professionnels, tels que les chanteurs et les professeurs. Avec l'apparition de divers troubles de la voix, les études cliniques imposent d'évaluer le fonctionnement de la voix. Les techniques d'investigation cherchent à fournir des informations complètes sur le mécanisme de production de la voix et en particulier sur le comportement vibratoire des plis vocaux et sur les caractéristiques acoustiques correspondantes.

Les plis vocaux sont composés de replis jumeaux de la muqueuse, étirés horizontalement à travers le larynx. Leurs vibrations module le flux d'air expulsé des poumons pendant la phonation, fonctionnant ainsi comme une excitation acoustique, c'est à dire la source de la parole. L'électroglottographie est une technique d'investigation non-invasive qui mesure l'impédance lié à la section de l'aire glottique, zone située entre les plis vocaux. L'observation directe des vibrations de plis vocaux ne peut qu'être atteinte avec les techniques d'imagerie invasives, en enregistrant les vibrations des plis-vocaux vues du dessus du larynx.

La vidéoendoscopie ultra-rapide est l'approche la plus prometteuse pour évaluer directement les vibrations des plis vocaux. La segmentation automatique de l'aire glottique, à partir des données acquises, est un défi majeur. Dans ce travail, nous abordons ce problème en proposant un cadre de contours actifs au niveau local pour la segmentation de la glotte. La méthode pourvoit la localisation de la glotte avec une grande précision, tout en permettant de découper et de fusionner le contour selon le modèle vibratoire. La méthode est entièrement automatique. La sélection des paramètres est effectuée automatiquement; elle est fondée sur des statistiques et observations empiriques, toutes deux réalisées sur une grande variété de séquences vidéos.

La vérification manuelle a présenté des modifications de  $\pm 10$  pixels, qui correspondent à moins de 1% de l'aire glottique moyenne. La quantité immense de données qui doit être évaluée, à la fois quantitativement et qualitativement, nécessite de réduire le nombre des dimensions de l'information spatio-temporelle ainsi que de représenter de manière efficace, compacte et sans perte les données ultra-rapides. Pour cette raison, des représentations à une et à deux dimensions ont été utilisées. Une comparaison a été effectuée sur les paramètres glottiques (fréquence fondamentale, quotient ouvert) estimés sur le signal électroglottographique et le signal de l'aire glottique, en utilisant le signal électroglottographique comme référence.

Superviseur

**Christophe d'Alessandro**

**Directeur de Recherche, LIMSI-CNRS**

**Groupe Audio - Acoustique**





# Ανάλυση γλωττίδας συνδυάζοντας στροβοσκόπηση υψηλής ταχύτητας και ηλεκτρογλωττογραφία

Σεβαστή-Ζωή Καρακώζογλου

Μεταπτυχιακή Εργασία

Πανεπιστήμιο Paris-Sud XI, Τμήμα Επιστήμης Υπολογιστών

Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών

LIMSI-CNRS

Audio & Acoustics Group

## Περίληψη

Η φωνή είναι ένα πολύτιμο εργαλείο επικοινωνίας και η αξία της είναι ανεκτίμητη τόσο για τις κοινωνικές όσο για τις επαγγελματικές μας δραστηριότητες. Οι κλινικές μελέτες υπαγορεύουν την αναγκαιότητα αξιολόγησης της λειτουργίας φωνής. Οι διαγνωστικές τεχνικές επιδιώκουν να παρέχουν εκτενείς πληροφορίες σχετικά με το μηχανισμό παραγωγής φωνής και, πιο συγκεκριμένα, για το μηχανισμό κίνησης των φωνητικών χορδών και τα αντίστοιχα ακουστικά χαρακτηριστικά.

Οι φωνητικές χορδές αποτελούνται από δύο πτυχές του βλεννογόνου και βρίσκονται στο λάρυγγα. Η φωνή παράγεται από τις δονήσεις των φωνητικών χορδών, ελέγχοντας τη διέλευση του αέρα που εκπνέεται από τους πνεύμονες. Η ηλεκτρογλωττογραφία είναι μια μη επεμβατική διαγνωστική τεχνική που μετρά την αντίσταση που σχετίζεται με τη γλωττίδα, την περιοχή μεταξύ των φωνητικών χορδών. Άμεση παρατήρηση των δονήσεων των φωνητικών χορδών μπορεί όμως να επιτευχθεί μόνο με επεμβατικές διαγνωστικές τεχνικές.

Η υψηλής ταχύτητας βιντεοστροβοσκόπηση αποτελεί την καλύτερη διαγνωστική εξέταση για τον έλεγχο της φώνησης και της παθολογίας του λάρυγγα. Η τμηματοποίηση της γλωττίδας αποτελεί τη σημαντικότερη πρόκληση. Σε αυτήν την εργασία προτείνεται μια μέθοδος που βασίζεται στα τοπικά ενεργά περιγράμματα (local-based active contours). Η μέθοδος προβλέπει τον εντοπισμό της γλωττίδας με μεγάλη ακρίβεια και επιτρέπει το περίγραμμα να σπάσει και να συγχωνευτεί, σύμφωνα με το μοτίβο δονήσεων των φωνητικών χορδών, ανιχνεύοντας τη γλωττίδα με πολύ καλή ακρίβεια. Η μέθοδος είναι αυτόματη και η επιλογή των παραμέτρων γίνεται με βάση στατιστικά στοιχεία και εμπειρικές παρατηρήσεις σε μεγάλο όγκο δεδομένων.

Η ακρίβεια της τμηματοποίησης αξιολογήθηκε από χρήστες με χρήση κατάλληλου διαδραστικού εργαλείου. Η τεράστια ποσότητα των δεδομένων που πρέπει να αξιολογηθούν τόσο ποσοτικά όσο και ποιοτικά, υπαγορεύει την ανάγκη για μείωση των διαστάσεων της χωροχρονικής πληροφορίας και την αποτελεσματική απεικόνιση των δεδομένων υψηλής ταχύτητας μέσα σε ένα συμπαγές, εύχρηστο και χωρίς απώλειες τρόπο. Μονοδιάστατες και δισδιάστατες απεικονίσεις χρησιμοποιήθηκαν για αυτόν τον λόγο. Χρησιμοποιώντας το ηλεκτρογλωττογραφικό σήμα ως σήμα αναφοράς, εκτιμήθηκαν ακουστικοί παράμετροι από το ηλεκτρογλωττογραφικό σήμα και από το σήμα του εμβადού της γλωττίδας.

Επόπτης

**Christophe d'Alessandro**

**Research Director, LIMSI-CNRS**

**Audio & Acoustics Group**



## *Acknowledgements*

*I would like to thank my supervisor Christophe d'Alessandro for welcoming me and advising me during my internship in Audio & Acoustics Group. I would also like to thank Yannis Stylianou for initially suggesting this most interesting collaboration. Last, but not least, I would like to express my gratitude to Nathalie Henrich, without whose guidance and generous help this work would not be complete.*



# Contents

<b>Acknowledgements</b>	<b>xi</b>
<b>List of Figures</b>	<b>xv</b>
<b>List of Tables</b>	<b>xvii</b>
<b>Abbreviations</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 State-of-the-art in voice function assessment</b>	<b>3</b>
2.1 Principles of voice production . . . . .	3
2.1.1 Vocal folds . . . . .	4
2.1.2 Voice production features . . . . .	6
2.2 Investigation techniques . . . . .	7
2.2.1 Non-invasive investigation techniques: Electroglottogram . . . . .	7
2.2.2 Invasive investigation techniques: towards the direct observation of vocal folds . . . . .	9
2.2.2.1 Former imaging techniques . . . . .	9
2.2.2.2 High-speed videoendoscopy . . . . .	11
2.3 The need of comparative studies . . . . .	11
2.3.1 Previous studies combining different investigation techniques . . . . .	11
2.3.2 The challenge of high-speed videoendoscopy . . . . .	13
<b>3 High-speed video processing</b>	<b>15</b>
3.1 Segmentation methods . . . . .	15
3.2 Segmenting high-speed video sequences . . . . .	17
3.2.1 State-of-the-art in glottis segmentation . . . . .	18
3.3 Glottis segmentation: a new method . . . . .	19
3.3.1 Overview of the proposed method . . . . .	19
3.3.2 Extraction of landmark frames . . . . .	19
3.3.3 Glottis localization Part I: Size reduction of the video sequence . . . . .	19
3.3.4 Contrast enhancement of the video sequences . . . . .	20
3.3.5 Glottis localization Part II: glottis in a box . . . . .	21
3.3.6 Segmentation of landmark frames . . . . .	21
3.3.6.1 Curve Initialization . . . . .	21
3.3.6.2 Segmentation by local-based active contours . . . . .	23

---

3.3.7	Segmentation propagation . . . . .	26
3.3.8	Post-processing of the segmentation results . . . . .	26
3.4	Using digital kymographic sequences for glottis segmentation . . . . .	26
3.5	Contribution . . . . .	27
3.6	Representation of segmentation data . . . . .	27
3.6.1	One-dimensional representation . . . . .	28
3.6.2	Two-dimensional representation . . . . .	28
3.6.2.1	Computation of Phonovibrograms (PVG) . . . . .	29
3.6.2.2	Computation of Glottovibrograms (GVG) . . . . .	31
<b>4</b>	<b>Materials &amp; Methods</b>	<b>35</b>
4.1	Database Presentation . . . . .	35
4.1.1	Data acquisition . . . . .	35
4.1.2	Recordings . . . . .	36
4.1.3	Synchronization issues . . . . .	36
4.2	Automatic segmentation . . . . .	40
4.3	Manual evaluation of the segmentation results . . . . .	40
<b>5</b>	<b>Results</b>	<b>43</b>
5.1	High-speed video segmentation analysis . . . . .	43
5.1.1	Subjective evaluation . . . . .	43
5.1.2	Comparison between manual and automatic segmentation . . . . .	43
5.1.3	Segmentation comparison on HSV and DKG sequences . . . . .	47
5.2	Comparison on glottal parameters estimated from EGG and HSV data . . . . .	48
5.2.1	Fundamental frequency estimation . . . . .	48
5.2.2	Glottal closing and opening instants estimation . . . . .	50
5.2.3	Open quotient estimation . . . . .	52
5.3	Visualization of DEGG, GLA and GVG signals . . . . .	52
<b>6</b>	<b>Conclusions</b>	<b>57</b>
	<b>Bibliography</b>	<b>59</b>

# List of Figures

2.1	Overview of human voice production mechanism . . . . .	3
2.2	A simplified scheme of voice production . . . . .	4
2.3	Vocal folds schematic representation . . . . .	5
2.4	Glottal airflow velocity schematic representation . . . . .	6
2.5	Principle of an electroglottogram . . . . .	7
2.6	The phases of the EGG and DEGG in comparison with the glottal flow . . . . .	8
2.7	Peaks in EGG and DEGG signal . . . . .	9
2.8	Diagram of imaging equipment with a solid endoscope . . . . .	10
2.9	Example of a videokymographic image . . . . .	10
2.10	Example of a high-speed sequence . . . . .	11
3.1	Size reduction by glottis localization. . . . .	20
3.2	Example of the pre-processing on a random image . . . . .	20
3.3	Glottis localization in three consecutive landmark frames. . . . .	21
3.4	The process of curve initialization. . . . .	22
3.5	Example of curve evolution . . . . .	25
3.6	From the HSV sequence to the DKG sequence . . . . .	27
3.7	Example of GLA signal . . . . .	28
3.8	Vocal-fold edge separation . . . . .	29
3.9	Linking glottal endings with the detected contours. . . . .	30
3.10	PVG of a video sequence. . . . .	31
3.11	Linked contours with corresponding glottal axis . . . . .	31
3.12	Glottovibrogram of a video sequence . . . . .	32
3.13	Comparison between PVG and GVG (part I) . . . . .	33
3.14	Comparison between PVG and GVG (part II) . . . . .	33
3.15	GVG with maximum speed profile . . . . .	34
4.1	Examination by high-speed videoendoscopy (UKE) . . . . .	35
4.2	Screenshot from the interface tool . . . . .	40
5.1	Video and segmentation subjective assessment . . . . .	44
5.2	Differences between automatic and manual segmentation per high-speed video sequence. . . . .	44
5.3	Absolute differences between automatic and manual segmentation per high-speed video sequence. . . . .	45
5.4	Segmentation errors and manual evaluation (part I) . . . . .	45
5.5	Segmentation errors and manual evaluation (part II) . . . . .	46
5.6	Percentage of changed images per sequence during annotation . . . . .	46

---

5.7	Error of segmentation of glottal closing and opening instants per high-speed video sequence . . . . .	47
5.8	Absolute error of segmentation of glottal closing and opening instants per high-speed video sequence . . . . .	48
5.9	Kymographic (DKG) images from two sequences. . . . .	48
5.10	Segmentation results in sequence of frames by DKG and HSV processing . . . . .	49
5.11	An example of difficulty in $F_0$ estimation . . . . .	50
5.12	An example of difference in $F_0$ estimation . . . . .	50
5.13	Error of alignment of glottal closing instants. . . . .	51
5.14	Error of alignment of glottal opening instants. . . . .	51
5.15	Open quotient estimation per high-speed video sequence . . . . .	52
5.16	Synchronous representation of DEGG, GLA and GVG signals ( <i>HH_SEQ_0057a</i> )	53
5.17	DEGG, GLA and GVG signals ( <i>HH_SEQ_0058</i> ) . . . . .	54
5.18	DEGG, GLA and GVG signals ( <i>HH_SEQ_0059</i> ) . . . . .	55
5.19	DEGG, GLA and GVG signals ( <i>HH_SEQ_0069</i> ) . . . . .	56



# List of Tables

4.1	Used recordings from the UKE database (spoken samples) . . . . .	37
4.2	Used recordings from the UKE database (sung samples, part I) . . . . .	38
4.3	Used recordings from the UKE database (sung samples, part II) . . . . .	39
5.1	Cases in which $F_0$ estimation indicates gross error. . . . .	49



# Abbreviations

<b>EGG</b>	Electroglottogram
<b>CDD</b>	Charged couple device
<b>DECOM</b>	DEgg Correlation-based Open quotient Measurement
<b>DEGG</b>	Derivative of the electroglottogram
<b>DGLA</b>	Derivative of the glottal area waveform
<b>DKG</b>	Digital kymography
<b>GCI</b>	Glottal closing instant
<b>GLA</b>	Glottal area waveform
<b>GOI</b>	Glottal opening instant
<b>GVG</b>	Glottal Vibrogram
<b>HSV</b>	High-speed videoendoscopy
<b>PGG</b>	Phottoglottography
<b>PVG</b>	Phonovibrogram
<b>SRG</b>	Seeded region growing



# Chapter 1

## Introduction

Voice is our primary communication tool. This human instrument is an important tool for many professionals such as singers and teachers. With the appearance of various voice disorders, clinical studies dictate the necessity of voice function assessment. Investigation techniques seek to provide for comprehensive information on the voice production mechanism and especially the interaction of the glottal source with the larynx.

A relationship between the glottal source and the vocal tract is believed to exist (Rothenberg, 1981a, 1981b). The vocal folds, or vocal cords, consist of two muscles and are located within the larynx at the top of the trachea. Their vibration modulates the air flows expelled from the lungs during phonation (Titze 1988, 1998). In (Rothenberg, 1981a) a preliminary study investigated relations between the glottal air flow and the vocal fold contact area. The relationship between the glottal source and the vocal fold vibrations is still under investigation.

Investigation techniques are being used for the examination of the vocal-fold vibrations. Electrolottogram is a non-invasive technique which measures the impedance related to the glottal area, i.e. the area between the vocal cords. This assumption requires direct measurements of the laryngeal dynamics in order to be validated (Childers et al. 1990, Henrich et al., 2004, Golla et al., 2009).

Imaging techniques seek to complement the information from the electroglottography by providing high quality images directly recorded from the posterior part of the larynx (Gerratt et al., 1991, Kiritani et al., 1990, Svec and Schutte, 1996, Farnsworth, 1940) The vocal-fold vibrations can be recorded directly with high-speed imaging techniques, especially with high-speed videoendoscopy, without loss of information. The vast amount of data from the high-speed imaging techniques requires further processing in order to track the vocal-fold motion. Efficient data representation techniques must be used in order to avoid a time-consuming and impractical navigation through hours of high-speed videos (Demeyer et al., 2009, Lohscheller et al., 2008,

Moukalled et al., 2009).

The presented work aims to contribute to glottal source analysis. We propose a new segmentation method to successfully tract the glottal area using data from high-speed videondoscopy recordings. This method presents a precise way of localizing the glottal area. The method ensures the correct segmentation and the tracking of the glottal area, allowing the contour to split and merge, thus following the vocal-fold vibration. We seek to represent the amount of data in a efficient and practical visualization. The final step of this work consists of comparing findings from different investigation techniques in order to clarify the underlying dynamics.

The remainder of this thesis is structured as follows. Chapter 2 deals with state-of-the-art investigation and processing methods in voice function assessment. Chapter 3 deals with high-speed video processing. Materials and methods used during this work are presented in chapter 4. Chapter 5 presents the results. Finally, concluding remarks are presented in Chapter 6.

## Chapter 2

# State-of-the-art in voice function assessment

### 2.1 Principles of voice production

The human voice production mechanism can be roughly divided into three parts: the lungs, the larynx and the vocal tract (figure 2.1) (Quatieri, 2001). The lungs function as a source of air flow and pressure. When voiced speech is uttered, the vocal folds vibrate periodically and modulate the air flow expelled from the lungs into a sequence of air puffs, thus functioning as an acoustic excitation and consequently, the source of voiced speech. The larynx is a major

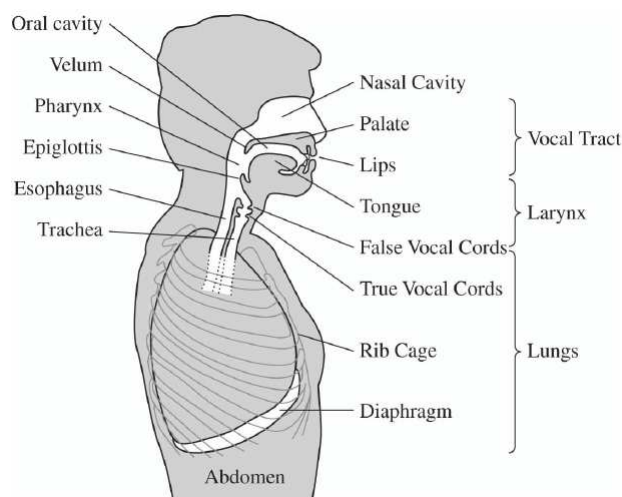


FIGURE 2.1: Overview of human voice production mechanism (Quatieri, 2001)

(but not the only) source of sound in speech, generating sound through the periodical opening and closing of the vocal folds. To oscillate, the vocal folds are brought near enough together so

that air pressure builds up beneath the larynx. The folds are pushed apart by this increased subglottal pressure, with the inferior part of each fold leading the superior part. Under the correct conditions, this oscillation pattern will sustain itself. Sound is generated in the larynx by chopping up a steady flow of air into little puffs of sound waves. The vocal tract is a set of cavities above the larynx and functions as an acoustic filter that shapes the spectrum of the sound. The sound is radiated to the surrounding air at the lips and nostrils. The perceived pitch of a person's voice is determined by a number of different factors, not least of which is the fundamental frequency of the sound generated by the larynx (Titze, 1998). A schematic representation of voice production is shown in figure 2.2.

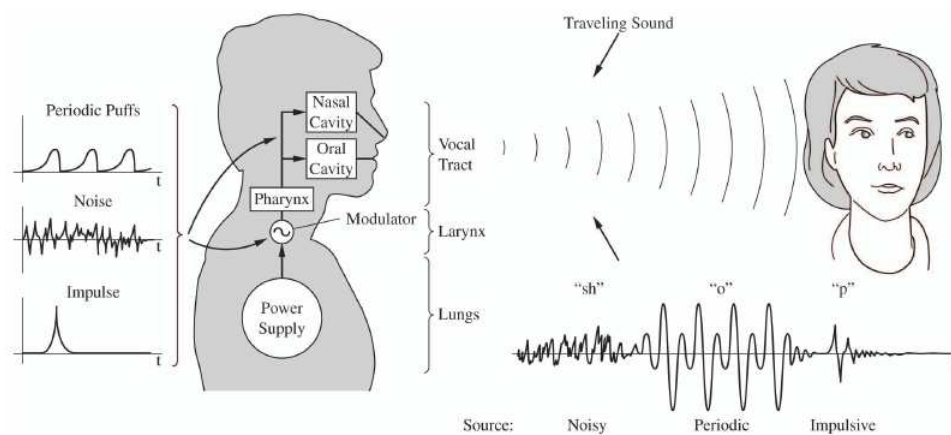


FIGURE 2.2: A simplified scheme of voice production ( (Quatieri, 2001)).

### 2.1.1 Vocal folds

The vocal folds are composed of twin infoldings of mucous membrane stretched horizontally across the larynx. They are composed of several layers with different stiffness properties. The topmost layer, the epithelium, is a cover layer on top of mucosal tissue, called lamina propria, which can be divided into three layers. The stiffness of the mucosal layers increases with depth. The innermost layer of the vocal folds is an elastic muscle (musculus thyroarytenoideus). They are attached posteriorly to the arytenoid cartilages, and anteriorly to the thyroid cartilage. Their outer edges are attached to muscle in the larynx while their inner edges margins are free. They are constructed from epithelium, but they have a ligament and a muscle, namely the vocalis muscle which tightens the front part of the ligament near to the thyroid cartilage. They are flat triangular bands and are pearly white in color. Above both sides of the vocal folds are the “ventricular folds” or “false vocal folds”. The vocal folds are represented in figure 2.3. The space between the vocal folds is called the glottis or glottal area. The relationship between projected



glottal area and flow conductance is likely to be significant in determining voice quality under some conditions (Negus, 2009, Rothenberg, 1981a, Stevens and Weismer, 2001, Titze, 1988).

The vocal-folds movement is controlled via the vagus nerve, so as to remain open during inhalation, closed when holding one's breath and vibrating for speech or singing. Vibration occurs mainly in the mucosal part of the tissue. Observations of vocal fold vibration using several techniques have revealed that the lower and upper portions of the vocal folds do not oscillate in phase. A wave propagates from the lower margins of the vocal folds towards the upper margins, which creates a wave-like motion traversing upwards in the vocal fold cover layer. This phenomenon is referred to as the mucosal wave (Story, 2002). Furthermore, opening and closure do not often occur simultaneously along the entire length of the vocal folds in the horizontal plane either. Instead, opening and closure often proceed from one end to the other in a zipper-like motion.

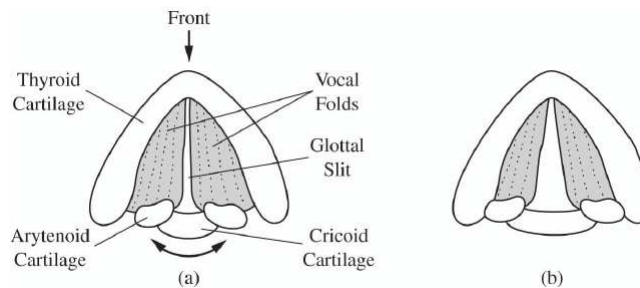


FIGURE 2.3: Vocal folds schematic representation from a top-view positions while voicing (a) and breathing (b) (Stevens and Weismer, 2001).

Men and women have different vocal fold sizes. Adult male voices are usually lower pitched and have larger folds. The male vocal folds are between  $17\text{mm}$  and  $25\text{mm}$  in length. The female vocal folds are between  $12.5\text{mm}$  and  $17.5\text{mm}$  in length. Vocal-fold size affects the pitch of the uttered voice. The frequency of glottal vibration determines the fundamental frequency of speech, which is commonly denoted by  $F_0$ . The average speaking  $F_0$  is approximately  $120\text{Hz}$  for men,  $200\text{Hz}$  for women and even higher for children. The range of variation is large: fundamental frequencies well below  $100\text{Hz}$  are not uncommon for men, while tenor singers may reach frequencies above  $600\text{Hz}$  and counter tenors much higher frequencies. Women's lowest fundamental frequencies are below  $150\text{Hz}$  (under vocal-fry register, a woman could speak at even  $50\text{Hz}$ ), while the upper limit of a soprano's singing range may exceed  $1300\text{Hz}$ .

The vocal folds can vibrate in different configurations that differ in the length and thickness of the vocal folds, as well as the muscular tensions involved. These modes are called registers or laryngeal mechanisms (Henrich, 2006). The main laryngeal mechanisms in speech and singing are laryngeal mechanism M1, corresponding to the chest or modal register, and laryngeal mechanism M2, corresponding to falsetto register. Men usually produce speech in laryngeal

mechanism M1. The vocal folds are thick and vibrate along their entire length, the glottis is tightly closed during each cycle, and there is a vertical phase difference in vibration. In M2, the vocal folds are thin, the glottis is not necessarily completely closed, and there is no vertical phase difference. Higher fundamental frequencies can be obtained in M2 rather than in M1.

### 2.1.2 Voice production features

As in any phonetic description, the description of voice needs appropriate, unambiguous and distinctive labels. Speech production models are usually described as source/filter models. Parameters dealing with voice quality, vocal effort and prosodic variations can be extracted from these models. Other than perceptual parameters, such as hoarseness, roughness and breathiness, model parameters include time-domain features of the glottal flow signal. The glottal flow signal is represented in figure 2.4. This signal can be characterized by a set time-domain features (Doval et al., 2006). The fundamental frequency is one of the primary features. The maximum excitation  $E$  of the signal is also important and corresponds to the main speech waveform peak in time domain. The open quotient ( $O_q$ ) is defined as the ratio between the open phase duration and the fundamental period and controls the relative duration of the glottal flow pulse. The asymmetry coefficient  $\alpha_m$  is defined as the ratio between the open phase duration and the fundamental period and defines the asymmetry of the flow. Finally, the return phase quotient  $Q_a$  is defined as the ratio between the opening phase and the open phase durations and is the relative duration of the return phase.

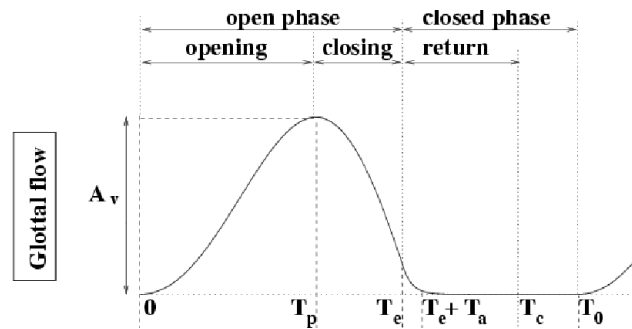


FIGURE 2.4: Glottal airflow velocity schematic representation (Doval et al., 2006).

Many of these parameters can be analyzed by investigation techniques, such as electroglottography and high-speed videoendoscopy, without the need of full modelization. One of the purposes of this work is to estimate these parameters from the presented methods.

## 2.2 Investigation techniques

Analyzing the vocal-fold vibrations is a challenging study since the larynx is not easily accessible. There are though several investigation techniques that allow for tracking of laryngeal activities. The objective goals are tracking of vibratory behaviour and/or corresponding acoustic features. Non-invasive techniques that do not require clinical operations are preferable, because they do not interfere with running speech and they do not require special precautions for the placement of the measurement equipments, thus requiring medical supervision. This comes with an important trade-off, the indirect observation of the larynx. In the following sections, an overview of non-invasive and invasive techniques will be presented.

### 2.2.1 Non-invasive investigation techniques: Electroglottogram

An electroglottograph (EGG) is a non-invasive device that indexes the contact area between the two vocal folds. A small, high-frequency current is passed between two electrodes that are secured around the neck at the level of the larynx. The opening and closing of the vocal folds causes variation in the electrical impedance of the current.

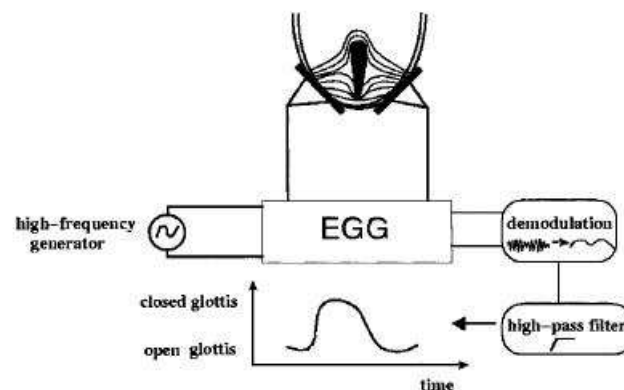


FIGURE 2.5: Principle of an electroglottogram, using the vocal fold contact area convention in which the EGG signal is represented as a function of vocal fold contact (Henrich et al., 2004).

While human tissue is a fairly good conductor of electricity, air is not. During phonation, the vocal folds are periodically separated by the glottis. As the vocal folds move apart, the glottis opens and the electrical impedance across the larynx is increased. When the vocal folds come closer together, the size of the glottis decreases, thereby decreasing the electrical impedance across the larynx. These changes in impedance are then displayed in a signal onscreen.

The EGG signal depicts the impedance variation as a function of time. The EGG signal gives a measure of vocal-fold contact area and nothing about the glottal area in the open phase. Furthermore, skin moistness and movements of the larynx during recording affect the impedance

measures. The built-in distortion levels in the EGG devices also introduce noise. However, EGG signals provide with significant information about the vocal fold behaviour during phonation. Its interpretation remains under ongoing studies.

During a vocal fold vibratory cycle, the corresponding EGG signal can be described as followed, related to the glottal air flow and physiological events, as described in Henrich et al., 2004 (Figure 2.6).

- 1 – 3 Closing phase. The lower margins of the vocal folds begin to close (1 to 2), then propagating to the upper margins. The maximum slope occurs in 2.
- 3 – 4 Closed phase. The vocal folds are fully closed.
- 4 – 6 Opening phase. The vocal folds begin to open from their lower part. The instant of maximum slope occurs in 5.
- 6 – 1 Open phase. The vocal folds are apart, corresponding to a rather flat signal.

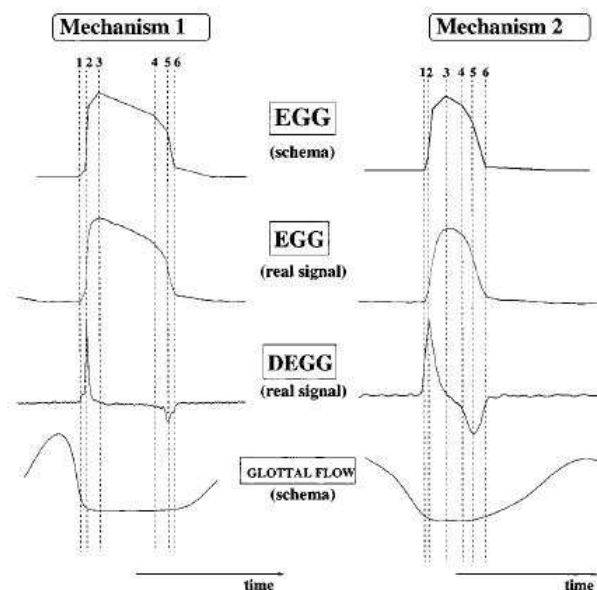


FIGURE 2.6: The phases of the EGG and DEGG in comparison with the glottal flow (Henrich et al., 2004).

Peaks present in derivative of the electroglottographic signal (DEGG) may be considered as reliable indicators of glottal opening and closing instants (Henrich et al., 2004); the latter being defined by reference to the glottal flow, as the instants when the flow starts to increase greatly from the baseline (opening) and decrease greatly to the baseline (closing).

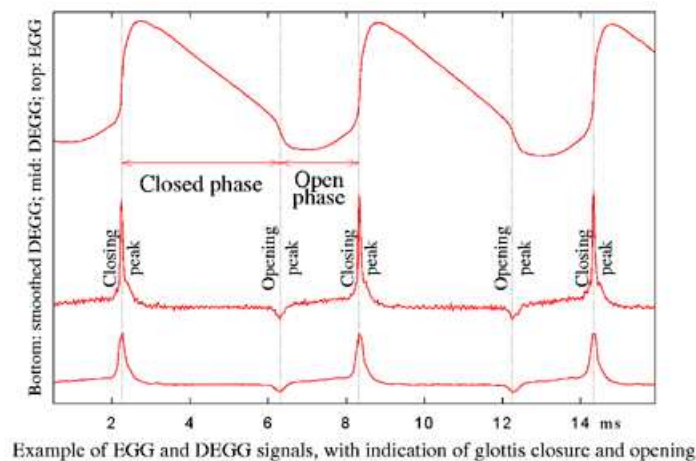


FIGURE 2.7: This figure illustrates the presence of alternating positive and negative peaks on the derivative of the electroglottographic signal. The positive peak on the DEGG signal corresponds to the glottal closing instant (GCI), i.e. the instant when the vocal fold contact area increases with greatest velocity. The negative peak, which is the point at which the EGG signal falls most steeply, corresponds to the glottal opening instant (GOI), i.e. the instant when the vocal fold contact area decreases with greatest velocity (<http://voiceresearch.free.fr/egg>).

## 2.2.2 Invasive investigation techniques: towards the direct observation of vocal folds

Signals other than the speech and EGG signal can be acquired with several imaging techniques. The main interest is focused on detecting the glottal area and the geometry of the vocal folds vibrations.

### 2.2.2.1 Former imaging techniques

Timcke et al., 1958, were among the first to use a laryngoscope, for direct observation and with the use of a recording system in order to examine vocal-folds vibrations.

Photoglottography (PGG) consists of using a light source, below the vocal folds, and monitoring the light amount proportional to the glottal opening. It can provide information regarding glottal activity, such as opening and closing instants (Gerratt et al., 1991). It is an invasive method and its clinical use remains limited, especially due to the fact that no accurate measurements can be acquired.

Laryngeal stroboscopy offers an efficient way to observe the anatomy and to some extent the behavior of the larynx. It is used to obtain a video sequence of the vocal fold vibration. A flashing light is used to illuminate the glottis with a frequency a bit lower than the frequency of the vibration, so that each flash occurs at a slightly later phase of the vibration period than the

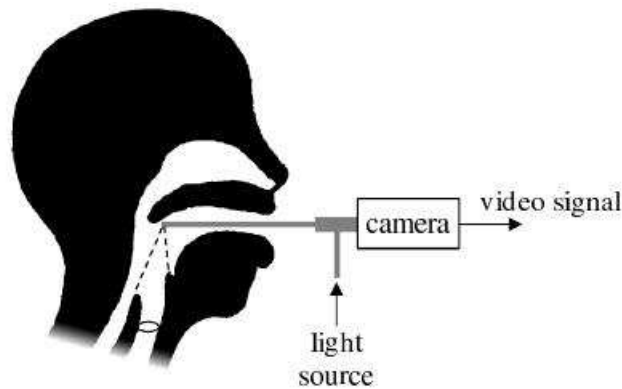


FIGURE 2.8: Diagram of imaging equipment with a solid endoscope

previous one. The frequency, however, of vocal fold vibration imposes serious limitations and only regular vocal fold vibration can be captured with high quality video (Kiritani et al., 1990).

Videokymography (VKG) is a technique which provides visual information of the vocal fold vibration at much higher temporal resolution than the stroboscopy. The system uses a modified video camera able to work in high-speed mode (nearly 8000 images/sec). The camera selects one horizontal line, transverse to the glottis, from the laryngeal image. The resulting successive lines are constructing a single image, filling it from top to bottom. This visualization provides information on frequency, amplitude, left-right asymmetries and the phases of the glottal cycle. The method was first introduced by Svec and Schutte, 1996. Up to recent, only qualitative con-

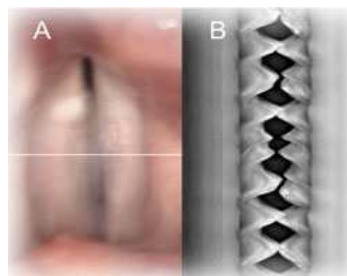


FIGURE 2.9: Example of a videokymographic image. One line, perpendicular to the glottal axis, is chosen and this area is recorded. The successive lines form the kymographic image (<http://www.kymography.com/>).

clusions could be mostly drawn from a VKG. Moukalled et al., 2009, presented a segmentation method of the vocal fold edges from the digital kymographic (DKG) sequence. The important difference between VKG and DKG is that DKG is based on the recording of full images at high frequency and then transforming them into corresponding kymographic images, while VKG is produced by recording one single line through time. An example of VKG is shown in figure 2.9.

### 2.2.2.2 High-speed videoendoscopy

The most powerful way of precisely capturing the vocal fold vibrations as an entity is the high-speed videoendoscopy (HSV). It consists of an imaging system capable of capturing full images at very high frame rate. A rigid or flexible endoscope can be placed in the larynx, equipped

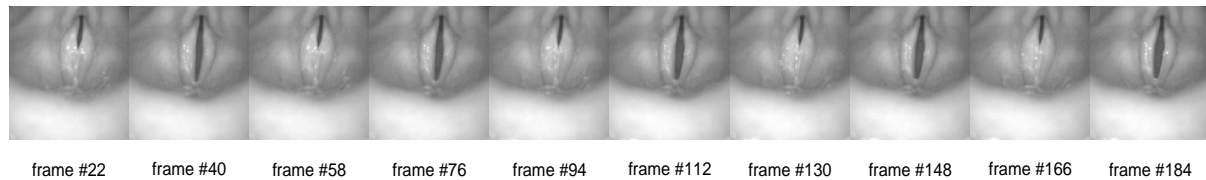


FIGURE 2.10: Example of a high-speed sequence, recorded from a male speaker.

with a camera system in order to record from a top-view position. High-speed cinematography was first developed at the Bell Laboratories in the 1930s and described by Farnsworth, 1940. However, only recently has the technology advanced so as to have high image analysis and high frame rate. Temporal and spatial resolution remain a limiting factor. There is a trade-off between image resolution and frame rate due to limited data transfer speed. The light source is another limiting factor, since it is being fed to the larynx through an endoscope with a small diameter, and the power of the light is limited.

HSV systems normally provide with image resolution from 100 to 300 pixels in each direction and frame rates up to 10000 frames per second. Due to the huge amount of memory needed to store the video sequence, the recording is usually restricted to a few seconds. Gray-scale images are usually recorded, since color cameras imply bigger amount of light. This imaging technique allows for full investigation of the vocal fold dynamics.

## 2.3 The need of comparative studies

Full investigation of the laryngeal dynamics is an ongoing study. Each of the previously described investigation techniques provide for complimentary information. Only by combining them we can achieve better understanding of the underlying phenomena.

### 2.3.1 Previous studies combining different investigation techniques

There has been an extensive interest in the research community to investigate the findings from electroglottography by identifying EGG features and correlating them to specific aspects of laryngeal physiology. Imaging methods, such as laryngeal stroboscopy, ultra high-speed laryngeal films, photoglottography, and more recently high-speed videoendoscopy, are all used

in order to capture the glottal area waveform and the geometry of vocal fold vibrations. Each study has tried to either validate or interpret the EGG features by direct observation of the imaging findings or by analysis. The list is by no means complete. Effort has been made to include the most interesting studies.

In (Rothenberg, 1981a), EGG was recorded simultaneously with oral airflow, which was then inverse-filtered to get an estimate of the glottal flow. As a result, a seven-stage model of the EGG waveform was presented.

Baer et al., 1983, compared signals from synchronized high-speed filming, acoustic recording, photoglottography and electroglottography. The results indicated that PGG and film measurements give essentially the same information for peak glottal opening and glottal closure and that the EGG signal appears to reliably indicate vocal fold contact.

Childers et al., 1990, recorded simultaneously EGG, speech signal and ultra high-speed laryngeal films. They performed point to point comparisons between the EGG and glottal area signals, as well as the estimates of open quotient and relative average perturbation. The goal of the research was to investigate the validity of the electroglottography as an analysis method and to relate the phases of their signals to laryngeal events.

Hertegard et al., 1995, used simultaneously stroboscopy, flow glottography and electroglottography. The goal was to examine the variations in glottal area and the vibratory patterns during different modes of phonation.

Larsson et al., 2000, presented a combined high-speed-acoustic-kymographic analysis package to examine different voice qualities.

Schutte and Miller, 2001, showed how kymography can avoid misinterpretations of the EGG signal and reported substantial agreement between EGG features and kymography.

Granqvist et al., 2003, examined the relationship between vocal fold vibrations and glottal flow by comparing various synchronized recordings.

Henrich et al., 2004, indicated the validity of DEGG signal as indicators of glottal closing and opening instants by comparing simultaneously recorded EGG and high-speed recording.

Degottex et al., 2008, presented a comparative study of EGG signals and videoendoscopic images of various phonatory conditions.

Golla et al., 2009, used markers to break the EGG signal into phases in order to compare it with the vibratory behaviour seen on digital kymography. They examined the findings for different registers.



### 2.3.2 The challenge of high-speed videoendoscopy

Although high-speed videoendoscopy is the most promising approach to directly examine vocal-fold vibrations, there is need for objective evaluation of the vast amount of the given data, both quantitative and qualitative. The large amount of data represented is prohibitive; a *2sec* video sequence recorded at *4000fps* requires approximately *512MB* for storage. Even a small database will require a lot of *GB* for storage. If a video sequence is recorded at *4000fps*, it cannot be viewed for evaluation at this speed, so, if we choose to view it at *15fps*, we will need more than *7min* to observe the *2sec* sequence. It is impossible to rely solely on visual inspection of hours of video.

Therefore, we need to represent our data efficiently in a compact and handy form. It is very important to reduce the dimensionality. Spatio-temporal information must be represented without loss of information. In the following part, efficient methods for evaluating and representing vocal-fold dynamics will be presented.



## Chapter 3

# High-speed video processing

The goal of computer vision is to model and automate the process of visual recognition, a term we interpret broadly as “perceiving distinctions between objects with important differences between them”. In computer vision, segmentation refers to the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze (Forsyth and Ponce, 2002). Image segmentation is typically used to locate objects and boundaries<sup>1</sup> in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. The outcome of segmentation is a logical mask, labelling the background zero (non-object) and the object of interest one. The details of the representation and exploitation of the visual characteristics abide to quite general desirable features, but depend heavily on each task. Although an analogy to video segmentation exists, the use of stand-alone image segmentation techniques is neither trivial nor guaranteed to succeed. In video segmentation, we seek to exploit consistency in topology, shape and representation throughout the sequence.

### 3.1 Segmentation methods

In order to better understand the algorithms used for glottis segmentation, it is important to present an overview of certain segmentation methods. The list is far from complete, but intends to give a better insight on the methods used in the literature for the present task.

**Clustering methods** The most simple clustering method is the k-means algorithm, an iterative technique to partition an image into k clusters (MacQueen, 1966). The algorithm

---

<sup>1</sup>In terms of perceptual semantics, an object is defined by its boundaries and boundaries define the object’s outline.

assumes that the variance between the pixels and the clusters center pixel is an appropriate measure of the cluster scatter. It also implies prior knowledge of the number  $k$  of clusters, which is also its main drawback.

**Region-based methods** This category consists of two sets of methods, the thresholding and the region-growing methods.

**Thresholding methods** Thresholding or histogram-based methods (Kohler, 1981, Haralick and Shapiro, 1985) can be very efficient compared to other segmentation techniques under circumstances, because they typically require only one pass through the pixels. A histogram is computed from the pixels color or intensity information. The peaks and valleys in the histogram are used to classify the clusters in the image. It is assumed that the histogram is at least bimodal, that is to have two peaks, which is not always the case.

In images with low contrast and objects with heterogeneous profile we cannot use histogram-based methods as the segmentation result is likely to converge poorly.

**Region-growing methods** Seeded region-growing method (SRG) (Adams and Bischof, 1994, Mehnert and Jackway, 1997) is a method which examines neighboring pixels of an initial set of seed points and determines whether the neighboring pixels should be added to the region. Seed points selection is object-dependent and should be done accordingly to the properties we seek in an object. The regions are then grown from these seed points to adjacent points depending on a region membership criterion. The criterion of similarity or homogeneity is task-dependent and usually considers intensity, color, texture and shape. An area threshold is essential and should be combined with conversion criteria. With robust criteria and relatively clear edges the algorithm converges to the regions that have the same properties we define.

However, region-growing methods are affected by noise and variation in intensity and they cannot easily distinguish the shading of real images, while being time-consuming.

**Model-based models** The main idea is that objects of interest have some kind of repetitive form of representation. By modelling the properties, we can use this knowledge as a prior to segment the image. State of the art methods for knowledge-based include active shape and appearance models (Cootes et al., 1995), level-set methods (Sethian, 1999, Paragios and Deriche, 2002) and active contours models.

Active contour models, also known as Snakes are mainly used to dynamically locate the contour of an object. A snake is an energy-minimizing spline guided by external constraint forces and influenced by image forces that pull it towards desired features, such as lines and edges (Kass et al., 1988). By choosing properly an initial contour near the object of interest, the model will converge to the desired solution. An energy functional is associated with the curve in terms of its shape and distance from desired image features.

The features must be wisely defined in a way that the final position of the contour will have a minimum energy. The problem of object detection is thus treated as an energy minimization problem.

There are two main categories of active contours, edge-based and region-based.

**Edge-based models** Edge-based active contour models utilize image gradients to identify object boundaries. They are very sensitive to image noise and depend on initial curve placement. There is no need for global constraints on the entire image and the snake can converge to the correct object in certain cases (Caselles et al., 1997, Kichenassamy et al., 1996).

**Region-based models** Region-based models model the foreground and the background statistically and seek for an energy optimum where the model best fits the image (Chan and Vese, 2001, Yezzi et al., 1999). Features used for this task may include intensity information, texture and known distributions. These models are more robust against initial curve placement and image noise than edge-based models. However, they may fail when used to segment heterogeneous objects, due to the use of global statistics.

## 3.2 Segmenting high-speed video sequences

The choice of segmentation method is not trivial. It depends on the quality of the used data, the features profile of the object of interest and the computational demands. Following the modelling of our data which will point out the most suitable segmentation method, additional constraints must be applied. That is, ready-made recipes do not exist.

In the case of high-speed videoendoscopy, and for the purposes of the present work, we wish to identify the glottis, i.e. the area between the vocal folds. In terms of computer vision, the glottal area is the foreground object and the rest of the image is the background. Perceptually, the glottal area is well defined by direct visual inspection. However, the nature of high-speed recordings and the vibration pattern implies further constraints in the segmentation process. Low image quality, poor lighting conditions usually degrade the final video sequences. Light reflections on the laryngeal tissues, enhanced by the epithelium, which is extremely reflective due to hydration, as well as saliva and mucus may create regions where the glottal area is not distinctive. The above circumstances lead us to conclude that the glottal area should be treated as an object with heterogeneous feature profile, which is not always present.

In the following section, previous works on glottis segmentation will be presented.

### 3.2.1 State-of-the-art in glottis segmentation

There have been a few methods in the literature which deal with glottis segmentation, mainly due to the recent development of robust high-speed videoendoscopy systems. Two segmentation methods have been used for this purpose: active contours and region-growing methods.

Marendic et al., 2001, were the first to present an active contour algorithm for vocal fold extraction from high-speed data. The vocal-fold vibration pattern was taken into consideration to reinitialize the algorithm. They applied their algorithm in one sequence, using parameters chosen empirically.

Allin et al., 2004, used a snake based segmentation for the medial edges of the vocal folds. Although they used data from a stroboscopic system, their approach is interesting because they use the Fischer linear to achieve a coarse segmentation of the sequence. The method demands the training of a color classifier for each sequence from more than one frame. Then, they use active contours to refine the result.

The third method that uses active contours is presented in (Moukalled et al., 2009). They employ a pair of open-curve snakes on the digital kymographic sequences. The method requires from the user to define the posterior and anterior points in an image and verify the segmentation result in one DKG frame, before it propagates to the rest of the sequence. The segmentation results are applied to the HSV sequences, once the segmentation is completed.

Yan et al., 2006, proposed the use of a region-growing method. The initial region of interest needs to be defined manually and the seed points are computed by assuming Rayleigh distribution on intensity. They do not take advantage of the vibration pattern neither present special constraints for frames depicting closed glottis.

Lohscheller et al., 2007, have presented a method where the user defines one seed point in an image of his choice, which is used for the region-growing method. The initial segmentation must first be verified so as to proceed with the rest of the sequence. They use threshold criteria for each horizontal line and reiterate the seeding procedure from the segmentation results to compensate for potential glottal drifts. The user has to supervise the procedure and select some parameters by visual inspection.

Finally, the last method using a seeded region-growing method is proposed by Demeyer et al., 2009. The seed point for the region-growing method is the maximal response of a laplacian of gaussian filter. They use intensity as the only homogeneity criterion and due to the uncertainty of size estimation, the size can be underestimated. They apply this method to periodical frames, where the glottis is supposed to be maximal. The region-growing results are propagated to the rest of the sequence using a level set method. Parameters are chosen empirically.

### 3.3 Glottis segmentation: a new method

#### 3.3.1 Overview of the proposed method

The method we are presenting in this work seeks to enhance the power of existing segmentation methods by proposing a set of steps which facilitate the segmentation process. It consists of an automatic method which does not require user intervention. It is an active contours based segmentation method taking into consideration the vocal-fold vibration pattern and the variability of high-speed recordings. Internal parameters are computed automatically for each treated sequence, while the active contours parameters are fixed for all sequences.

#### 3.3.2 Extraction of landmark frames

The open glottis is the darkest region and the one that evolves mostly within the sequence. Based on this fact, we extract the frames with minimum sum of pixel intensities. These frames are called landmark frames (equation 3.1). The landmark frames represent the open states of the glottal cycles within the sequence under consideration. To ensure that all selected landmark frames will represent maximal glottal areas, in cases where the intensity range is narrow, we check the found indices and suppress those that correspond to high overall mean intensities.

$$I_{landmark} = \min_{i=1..k} \left( \sum_x \sum_y I_i(x, y) \right) \quad (3.1)$$

Each of these frames will be used as reference for the segmentation propagation within each glottal cycle, as it will be presented in the following sections.

#### 3.3.3 Glottis localization Part I: Size reduction of the video sequence

Since our region of interest covers only a part of the entire image, there is no need to process the entire image for localization and computational reasons. The size of the video sequences used in this work is  $256 \times 256$  pixels and the glottal area usually covers less than 25% of the entire image size. More details on the database are presented in chapter 4.

The procedure is an edge-based morphological processing of a landmark frame. The idea is to find a large, nearly vertically oriented area. We apply a Sobel filter for detection of strong edges in vertical direction. Then, we perform a morphological closing on the gradient map, so as to connect small related regions and we detect these regions by connected component analysis. We choose the object with the largest area and vertical orientation. Around the selected area, we compute a rectangle surrounding it, called the bounding box. The final bounding box is

expanded so as to compensate for the glottal drift and/or movements of the endoscope. This step allows us to reduce the amount of data to be processed and treat larger video sequences. The coordinates of the cropped rectangle are stored and once the segmentation is performed, we can go back to the initial sequence.

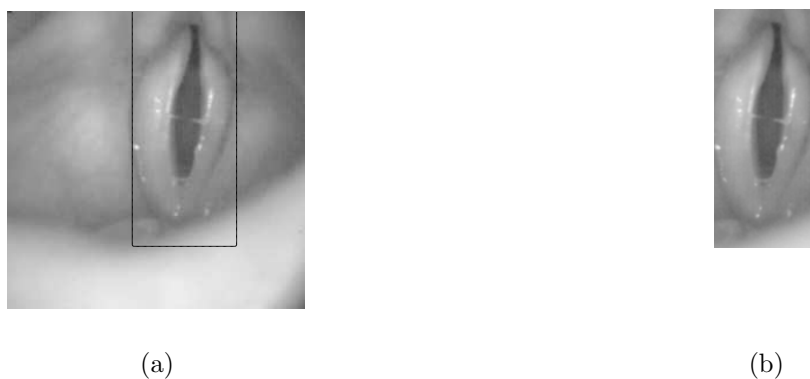


FIGURE 3.1: Size reduction by glottis localization. (a) The bounding box on the original image ( $256 \times 256$  pixels). (b) The cropped image ( $202 \times 89$  pixels). From the cropped video sequence, we can go back to the original sequence without any loss.

### 3.3.4 Contrast enhancement of the video sequences

The video sequence was enhanced so as to improve the contrast. In order to improve the local contrast in the images, bringing out more detail in the glottal area while avoiding significant noise introduction, the contrast limited adaptive histogram equalization algorithm (CLAHE) was used (Zuiderveld, 1994). It consists of a generalization of adaptive histogram equalization and computes several histograms, each corresponding to a distinct section of the image, and uses them to redistribute the lightness values of the image, thus compensating for noise amplification. The enhanced video sequence is used for the following steps of the algorithm.

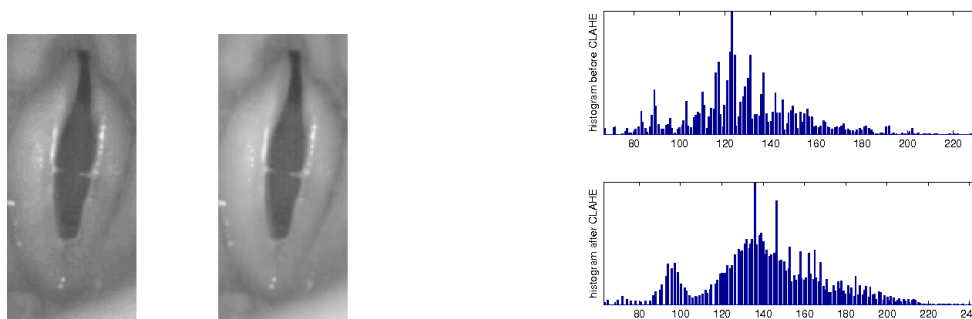


FIGURE 3.2: Example of the pre-processing on a random image: from left to right, the original image and the image after the CLAHE enhancement. The contrast is higher, as proved by the histogram change depicted on the right.



### 3.3.5 Glottis localization Part II: glottis in a box

It is necessary to localize with some precision the object to segment. The first step is to define an area in which to search for the object of interest. We apply the same processing as described in subsection 3.3.3 to get a tighter bounding box surrounding the glottal region in every landmark frame.

Once the bounding boxes are computed, they need to be propagated to the rest of the frames. It is also necessary to define whether the glottis exists in all images or not. In an active contours framework, the algorithm may evolve in excrescences if further constraints are not applied. This is solved with two techniques. Firstly, we take into consideration the pixel intensities of each image. If the image's minimum value remains over a global threshold, the median of the pixels intensities of the entire sequence, it is assumed that there is no glottal opening present. Then, the bounding box is computed. If it is centered far from the landmark's bounding box or if it does not exist, it is assumed that there is no glottal opening as well. When these two conditions coincide, we can exclude the image from further processing.



FIGURE 3.3: Glottis localization in three consecutive landmark frames.

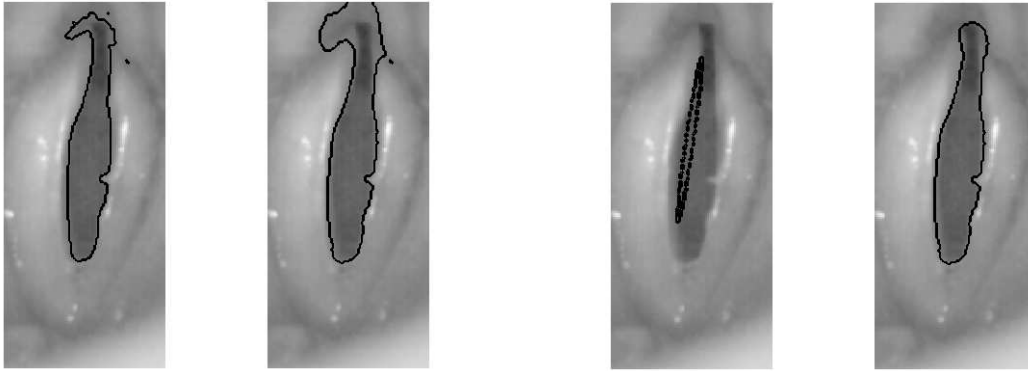
So far, we have an estimation of the duration of glottal cycles, the position where the glottal area lies, a loose maximum threshold for its area and its scale. This knowledge will help us to begin the segmentation process.

### 3.3.6 Segmentation of landmark frames

#### 3.3.6.1 Curve Initialization

Curve initialization is of major importance in active contours methods. We need to define an initial mapping of the object of interest with the greatest accuracy we can achieve in order to help the curve conversion.

However, the glottis is an object with heterogeneous feature profile. Even though it is always darker than the surrounding tissues, there might be regions where local statistics do not provide substantial similarity criteria. We therefore propose the use of two methods for curve initialization.



Curve initialization by thresholding; the initial curve and the final curve

Curve initialization by ellipse; the initial curve and the final curve

FIGURE 3.4: The process of curve initialization. For a landmark frame, we compute two initial curves, the threshold-based and the ellipse. By comparing the region at which the segmentation algorithm converges, we manage to successfully segment the landmark frames. While the threshold-based curve seems to capture most of the glottal area, the active contours converge to the glottal area and to an irrelevant area. The ellipse curve, which is all inside the glottal area, provides with better initialisation statistics and thus a better segmentation result.

Finally, this is the curve that will be used for propagation.

The first method consists of finding the intensity threshold within the bounding box of each landmark frame. In cases of high contrast, the glottal region is much darker and relatively homogeneous and a threshold is sufficient for initial discrimination. The threshold is found by choosing it to be in the valley of a smoothed bimodal histogram. This method is also called the mode method (Glasbey, 1993). However, the assumption of the bimodal histogram is not always valid, as it depends on the statistics of the image. To address this problem we propose the use of a localization-based map.

Based on the bounding box computed on a previous step, we compute an ellipse whose center is located on the center of the bounding box and size proportional to the bounding box's size. Its orientation is based on the orientation of the glottis computed during the localization. This ellipse-shaped mask covers the glottal area and points out with good accuracy the area to which the active contours should converge.

The contour of the landmark frames are estimated with these two initial maps. A comparison of the computed contours is made based on the size of area they cover and the maximal separability of the object to the background in terms of intensity. The contour which best fits the above criteria will be used for the propagation of the segmentation to the rest of the sequence.

### 3.3.6.2 Segmentation by local-based active contours

The segmentation method used in this work is based on the framework proposed by Lankton and Tannenbaum, 2008, called local region-based framework for guiding active contours. The idea is to allow the foreground and background to be modeled in terms of smaller local regions, since foreground and background regions cannot be always represented with global statistics. This framework allows for correct conversion in cases of inhomogeneity, common in medical images.

By analysing the local regions we wish to construct a set of local energies at each point along the curve. Each point along the curve is considered separately and moves to minimize the energy computed in its own local region. To compute the energies, local neighborhoods are split into local interior and local exterior by the evolving curve. The energy minimization is performed by fitting a model to each local region.

Let  $I$  be an image and  $C$  a closed contour represented as the zero level set of a signed distance function <sup>2</sup>  $\phi$ . The interior of  $C$  is specified by the approximation of the smoothed Heaviside function, given in equation 3.2. The exterior of  $C$  is defined as  $(1 - \mathcal{H}\phi(x))$ .

$$\mathcal{H}\phi(x) \begin{cases} 1, & \phi(x) < -\epsilon \\ 0, & \phi(x) > \epsilon \\ \frac{1}{2} \left\{ 1 + \frac{\phi}{\epsilon} + \frac{1}{\pi} \sin \left( \frac{\pi\phi(x)}{\epsilon} \right) \right\}, & \text{otherwise} \end{cases} \quad (3.2)$$

The area just around the curve is given by the derivative of the  $\mathcal{H}\phi(x)$ , a smoothed version of the Dirac delta

$$\delta\phi(x) \begin{cases} 1, & \phi(x) = 0 \\ 0, & |\phi(x)| > \epsilon \\ \frac{1}{2\epsilon} \left\{ 1 + \cos \left( \frac{\pi\phi(x)}{\epsilon} \right) \right\}, & \text{otherwise} \end{cases} \quad (3.3)$$

A point in the image  $I$  is represented by  $x$  and  $y$  coordinates, which are independent spatial variables. Equations 3.2, 3.3 were presented with  $x$  only, for simplicity reasons. For masking local regions, we will use the term  $\mathcal{B}(x, y)$ . This term represent the neighborhood of the point, both interior and exterior, in which the energies are computed.

$$\mathcal{B}(x, y) \begin{cases} 1, & \|x - y\| < r \\ 0, & \text{otherwise} \end{cases} \quad (3.4)$$

---

<sup>2</sup>In two dimensions, the level set method represents a closed curve  $C$  using an auxiliary function  $\phi$ , the level set function, as the zero level set of  $\phi$ , that is  $C = \{(x, y) \mid \phi(x, y) = 0\}$ .  $\phi$  is assumed to take negative values inside the region and positive values outside (Sethian, 1999).

The radius parameter  $r$  in equation 3.4 controls the locality of the segmentation results. It is task-dependent, depends on the scale of the object of interest and the proximity of the surrounding clutter and controls the smoothness of local statistics along the curve. The smaller the object is, the smaller should the value of  $r$  in pixels be. It is chosen to be one-third of the width of the bounding box for each landmark frame.

The energy functional is given in equation 3.5:

$$E(\phi) = \underbrace{\int_{\Omega_x} \delta\phi(x)}_A \underbrace{\left( \int_{\Omega_y} \mathcal{B}(x, y) F(I(y), \phi(y)) dy \right)}_B dx + \lambda \underbrace{\int_{\Omega_x} \delta\phi(x) \|\nabla\phi(x)\| dx}_C \quad (3.5)$$

This formulation allows the computation only in the proximity of the points. The three terms of the formula contribute as following:

- A This term allows the computation of the energy just around the curve, allowing us to ignore inhomogeneity far from the region of interest. It allows the curve to split and merge, using only the contribution of the neighborhood's statistics.
- B For every point  $x$  chosen by  $\delta\phi(x)$ , the mask  $\mathcal{B}(x, y)$  ensures that  $F$  will operate only on local image information about  $x$ . Thus, the terms  $A$  and  $B$  compute the sum of  $F$  values for every  $\mathcal{B}(x, y)$  neighborhood around the zero level set.
- C This term ensures the curve smoothness weighted by a parameter  $\lambda$ .

In the present framework, any energy can be used as  $F$ , giving local adherence to a given model at each point along the contour. We choose to use the Chan-Vese model (Chan and Vese, 2001), which is a constant intensity model and its formula is given in equation 3.6. It models the foreground and background as constant intensities represented by their means,  $u$  and  $v$ .

$$E = \int_{\Omega_y} (\mathcal{H}\phi(y)(I(y) - u)^2 + (1 - \mathcal{H}(y))(I(y) - v)^2) dy \quad (3.6)$$

The mean intensities of the interior and exterior regions,  $u$  and  $v$  respectively, are given by the following equations:

$$u = \frac{\int_{\Omega_y} \mathcal{H}\phi(y) I(y) dy}{\int_{\Omega_y} \mathcal{H}\phi(y) dy} \quad (3.7)$$

$$v = \frac{\int_{\Omega_y} (1 - \mathcal{H}\phi(y)) I(y) dy}{\int_{\Omega_y} (1 - \mathcal{H}\phi(y)) dy} \quad (3.8)$$

while the localized formulas are given by the following equations:

$$u_x = \frac{\int_{\Omega_y} \mathcal{B}(x, y) \mathcal{H}\phi(y) I(y) dy}{\int_{\Omega_y} \mathcal{B}(x, y) \mathcal{H}\phi(y) dy} \quad (3.9)$$

$$v_x = \frac{\int_{\Omega_y} \mathcal{B}(x, y)(1 - \mathcal{H}\phi(y))I(y)dy}{\int_{\Omega_y} \mathcal{B}(x, y)(1 - \mathcal{H}\phi(y))dy} \quad (3.10)$$

The energy functional  $F$  is formed as following by replacing the local statistics from equations 3.9, 3.10:

$$F = \mathcal{H}\phi(y)(I(y) - u_x)^2 + (1 - \mathcal{H}\phi(y))(I(y) - v_x)^2 \quad (3.11)$$

Thus, by replacing 3.11 into the equation 3.5, we have the localized energy. The evolution equation for  $\phi$  is obtained by taking the derivative of  $F$  with respect to  $\phi(y)$  (equation 3.12).

$$\nabla_{\phi(y)} F = \delta\phi(y)((I(y) - u_x)^2 - (I(y) - v_x)^2) \quad (3.12)$$

The curvature<sup>3</sup> flow is computed by inserting the derivative of  $F$  into equation 3.5:

$$\frac{\partial\phi}{\partial t}(x) = \delta\phi(x) \int_{\Omega_y} \mathcal{B}(x, y)\delta\phi(y)((I(y) - u_x)^2 - (I(y) - v_x)^2)dy + \lambda\delta\phi(x)div\left(\frac{\nabla\phi(x)}{|\nabla\phi(x)|}\right) \quad (3.13)$$

Within this framework, we compute the minimum energy of the flow. The minimum is obtained when each point on the curve has moved such that the local interior and exterior about every point along the curve is best approximated by local means  $u_x$  and  $v_x$ . The algorithm begins by computing the statistics on the original curve. The original curve is extracted from the initial mask and the algorithm runs until convergence; until the contour no longer changes, without surpassing 150 iterations. In figure 3.5 we can see an example of segmentation on a single image.

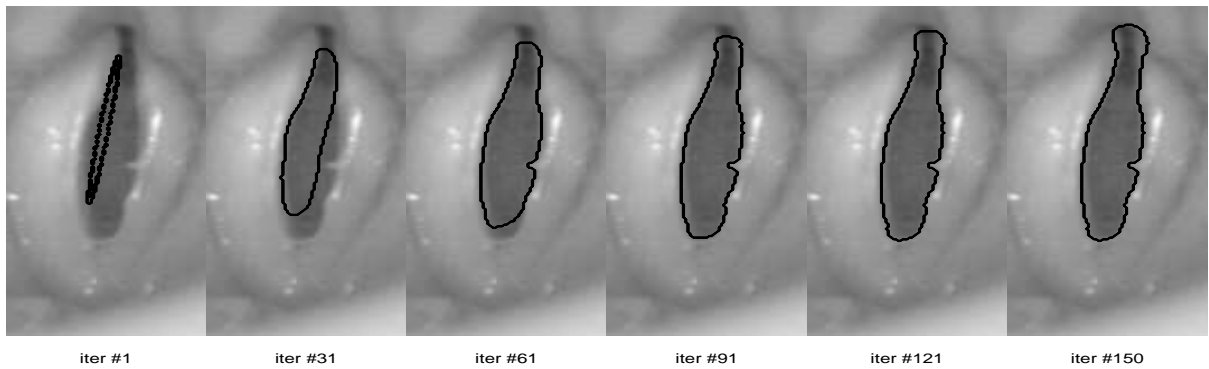


FIGURE 3.5: Curve evolution in a landmark frame. From the ellipse-shaped curve (iteration # 1), the curve evolves until it converges to the medial vocal-fold edges (iteration # 150). Curves each 30 iterations are shown.

<sup>3</sup>Intuitively, curvature is the amount by which a geometric object deviates from being flat, or straight in the case of a line, but this is defined in different ways depending on the context (Struik, 1988).

### 3.3.7 Segmentation propagation

Once the segmentation on all landmark frames is computed, we continue by segmenting the rest of the frames. To ensure temporal consistency, the segmentation is propagated by using as initial mask for the  $k^{th}$  frame the segmentation result from the  $(k - 1)^{th}$  or the  $(k + 1)^{th}$  frame, depending the position of the landmark frame.

### 3.3.8 Post-processing of the segmentation results

There may be cases where the segmentation may converge to inconsistent regions. We keep contours that are present within the limits of the bounding boxes in order to suppress undesired contours. To ensure maximal separability in terms of local statistics, we compare the mean intensity of the segmented regions with respect to its neighborhood. If the mean intensity is significantly bigger than the surroundings or the rest of the regions within the same image, if any, we exclude this region, as it will most likely have not captured glottal area. Furthermore, for each segmented object whose histogram is bimodal with a high intensity threshold, we apply threshold segmentation on it to capture the homogeneous region with low intensity. The segmentation matrix was finally smoothed so as to exclude holes in the found regions.

## 3.4 Using digital kymographic sequences for glottis segmentation

The above described procedure can be used for glottis segmentation using the digital kymographic (DKG) sequences. The DKG sequences can be acquired by simply reshaping the video sequence matrix comprised of  $N$  frames along the y-axes, as following:

$$\begin{aligned} I_{HSV} &= \{I_i(x, y) | x = 0..255, y = 0..255, i = 1..N\} \Rightarrow \\ I_{DKG} &= \{I_x(y, i) | y = 0..255, i = 1..N, x = 0..255\} \end{aligned} \quad (3.14)$$

The segmentation method can be applied to the transformed video sequences with small changes. The object localization used information on the nature of the area of interest, which means that we search for an horizontally-aligned area, comprised of possibly more than one small regions. Additional tuning may be required. The segmentation proceeds as described in the previous sections and the final segmentation is then applied on the HSV sequences.

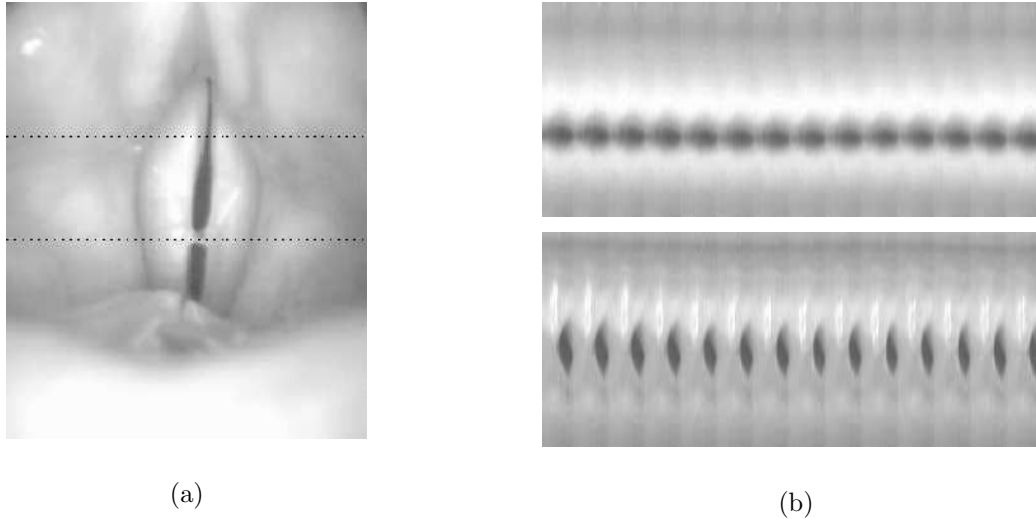


FIGURE 3.6: Example of acquiring the DKG sequence from the corresponding HSV sequence. The left image shows a frame from the HSV sequence. The dotted lines correspond to the index kymographic images on the right.

### 3.5 Contribution

In the present work, we make the following contributions. First, we provide a precise way of localizing the glottal area. The proposed scheme provides a tight mask surrounding the object of interest, information on the maximum area threshold and information on the presence or absence of glottal area. Second, we propose the use of local-based active contours, so that objects with heterogeneous statistics can be segmented when global energies fail to successfully converge. The method allows the contour to split and merge, thus capturing  $n$  objects, which is often the case in pathological vocal-fold vibrations. Furthermore, the method is fully automatic and it does not require human intervention. Parameter selection is performed automatically, based on the sequence statistics, and when necessary, parameters were chosen empirically by studying a wide range of high-speed sequences (refer to chapter 5 for a detailed presentation of the database). Finally, the proposed method is not data-dependent. The use of different steps to be applied on a video sequence, allows the use of it on different data sets. In chapter 5 we will present the segmentation results on DKG sequences, as validation to the segmentation discriminative power.

### 3.6 Representation of segmentation data

It is essential to represent the segmentation data in a compact and handy form. Along with the video sequences with the detected contours, recorded at 15 fps for better visualization, we will present the following representations.

### 3.6.1 One-dimensional representation

The sum of pixels of the segmented regions for every image gives the glottal area waveform (GLA). Let us consider  $I^s$  the segmentation matrix, having same size as the  $I$  video sequence. The segmentation matrix is a logical matrix, where 1 is assigned to pixels belonging to the glottal area and 0 is assigned to pixels belonging to the background. GLA is formulated as follows:

$$GLA(i) = \sum_x \sum_y I_i^s(x, y), \quad i = 1..k \quad (3.15)$$

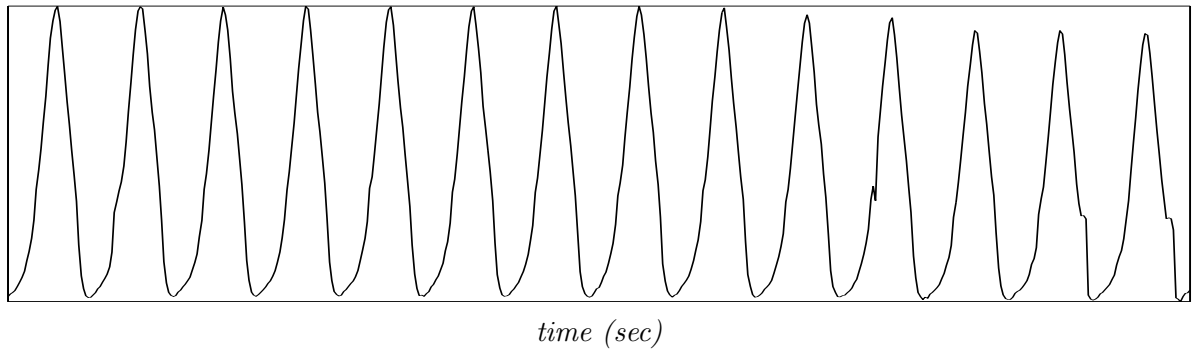


FIGURE 3.7: Example of glottal area waveform signal (video recorded at 4000fps)

In figure 3.7 we see a typical GLA signal. The valleys correspond to the instants where the glottal area is maximally closed, while the peaks correspond to the instants where the glottal area is maximally open. With this representation, we track the area of the glottis through time. A detailed work on the use of GLA for evaluation of vocal-fold vibratory characteristics and its comparison with the EGG will be presented in chapter 5. Since the calibration parameters of the recording procedure are unknown, the GLA represents the relative glottal area in pixels and is normalized within the interval  $[0, 1]$ . In the case where calibration parameters would be known, the actual glottis size could be computed.

### 3.6.2 Two-dimensional representation

No matter how representative the GLA signal can be in terms of global area evolution, all information on the vocal-folds geometry is omitted. The first attempt of glottal shape representation was made by Westphal and Childers, 1983. In (Lohscheller et al., 2008) the Phonovibrogram (PVG) is introduced, which is a further development of spatio-temporal plots of vocal-fold vibrations (Neubauer et al., 2001). The PVG is a 2-D diagram of vocal fold vibrations. This representation transforms vocal-fold movements into well-defined geometric objects, thus allowing direct assessment of the vocal-fold dynamics of an entire video sequence in a single image.



### 3.6.2.1 Computation of Phonovibrograms (PVG)

Let us consider the segmentation matrix  $I^s(x, y, i)$  (equation 3.16). As stated before, it consists of a logical matrix, defining foreground (the glottal area) as true and background (everything else) as false. We will call each segmented region within an image  $A_j, j = 1..k$ .

$$I^s(x, y, i) = \begin{cases} 1, & \text{pixel} \in \text{glottal area} \\ 0, & \text{background} \end{cases} \quad (3.16)$$

For each region  $A_j$  in the entire sequence, we compute the corresponding linear regression

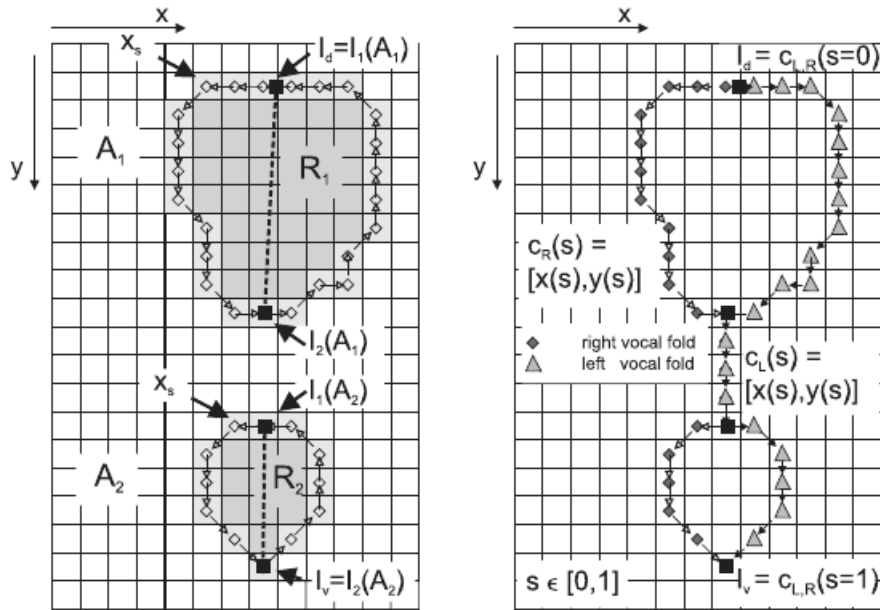


FIGURE 3.8: Vocal-fold edge separation. By defining linear regression lines  $R_i(x, y)$  for every fragment  $A_j$ , the left and right vocal-fold edges are delineated within an image  $I(x, y)$  (Lohscheller et al., 2007).

line  $R_i(x, y)$ , which corresponds to the symmetry axis of each region  $A_j$ . The regression line indicated the orientation of each region. The intersection points between the regression line  $R_i(x, y)$  and the contour of  $A_j$  define points  $I_{1,2}(A_j)$ , which are the anterior and posterior points of the region. For each image, we define as dorsal ( $I_d$ ) and ventral ( $I_v$ ) points of the entire glottal area, the  $I_{1,2}(A_j), j = 1..m$  points with the minimum and maximum y-coordinate respectively. The dorsal and ventral points for every image correspond to the posterior and anterior points of the left and right vocal-fold edges,  $c_{L/R}(s) = [x(s), y(s)]$ , where  $s \in [0, 1]$  is the parametric domain of  $c_{L/R}(s)$ . In figure 3.8 we can see a scheme of the above described procedure.

In order to relate the segmented contours, we need to create a continuous representation of vocal-fold vibrations. We need to link the dorsal and ventral points of all images within a sequence.

For doing this, we identify the frames where the glottal area is maximal, thus defining the glottal cycles. It is assumed that within a cycle, the dorsal and ventral points don't change dramatically and that they can be approximated for the intermediate frames by linear approximation with the following formulation.

Let us consider  $D(T_O) = c_{L/R}(s = 0, t_i)$  the most dorsal points and  $V(T_O) = c_{L/R}(s = 1, t_i)$  the most ventral points within the open states  $t_i = T_O$ . For all intermediate points we have:

$$\begin{aligned} D(t_i) &= D(T_O) + \frac{D(T_{O+1}) - D(T_O)}{T_{O+1} - T_O}(t_i - T_O), & T_O < t_i < T_{O+1} \\ V(t_i) &= V(T_O) + \frac{V(T_{O+1}) - V(T_O)}{T_{O+1} - T_O}(t_i - T_O), & T_O < t_i < T_{O+1} \end{aligned} \quad (3.17)$$

By connecting the  $c_{L/R}(s, t_i)$  to the corresponding  $D(t_i)$ ,  $V(t_i)$  positions, we have the continuous representation of the parts of vocal fold edges which are closed, and therefore, undetected from the segmentation methods, as showed in figure 3.9. The glottal main axis  $g(m, t)$  is therefore the connection line between  $D(t_i)$  and  $V(t_i)$ .

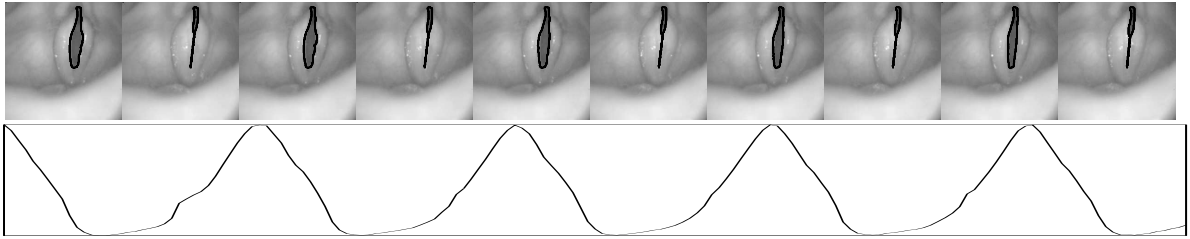


FIGURE 3.9: Linking glottal endings with the detected contours. Within a video sequence, 10 frames with the corresponding segmentation are presented. From the open state to the instant where the glottal area is minimal within its glottal cycle, the contour is linked to the most dorsal and ventral points.

For a continuous representation of the vocal-folds, the glottal axis  $g(m, t)$  and the vocal-fold edges  $c_{L,R}(m, t)$  are equidistantly sampled with  $m \in [0, M]$ . For each image we define the deflections of vocal-fold edges perpendicular to the glottal axis as the absolute values of the distances:

$$\delta^{l,r}(m, t) = \|g(m, t) - c_{L,R}(m, t)\|_2, \quad \forall m \quad (3.18)$$

The vocal-fold edges are considered to be longitudinally split, with the left vocal fold turned  $180^\circ$  around the posterior ending  $D$ , so that the length of the  $\delta^{l,r}(m, t)$  is  $2M + 1$ , where  $M$  is the sampling size. The distances  $\delta^{l,r}(m, t)$  for the entire video sequence are stored in a matrix  $\mathcal{D}_{L,R}^{(2M+1) \times (T+1)}$ . The horizontal centerline  $\mathcal{D}(m = 0, t)$  holds the dorsal endings of the vocal folds.

In order to visualize the distance matrix, each matrix element is color coded and normalized within the interval  $[0, 1]$ , with 0 (black) representing zero distance and 1 (white) representing maximum distance.

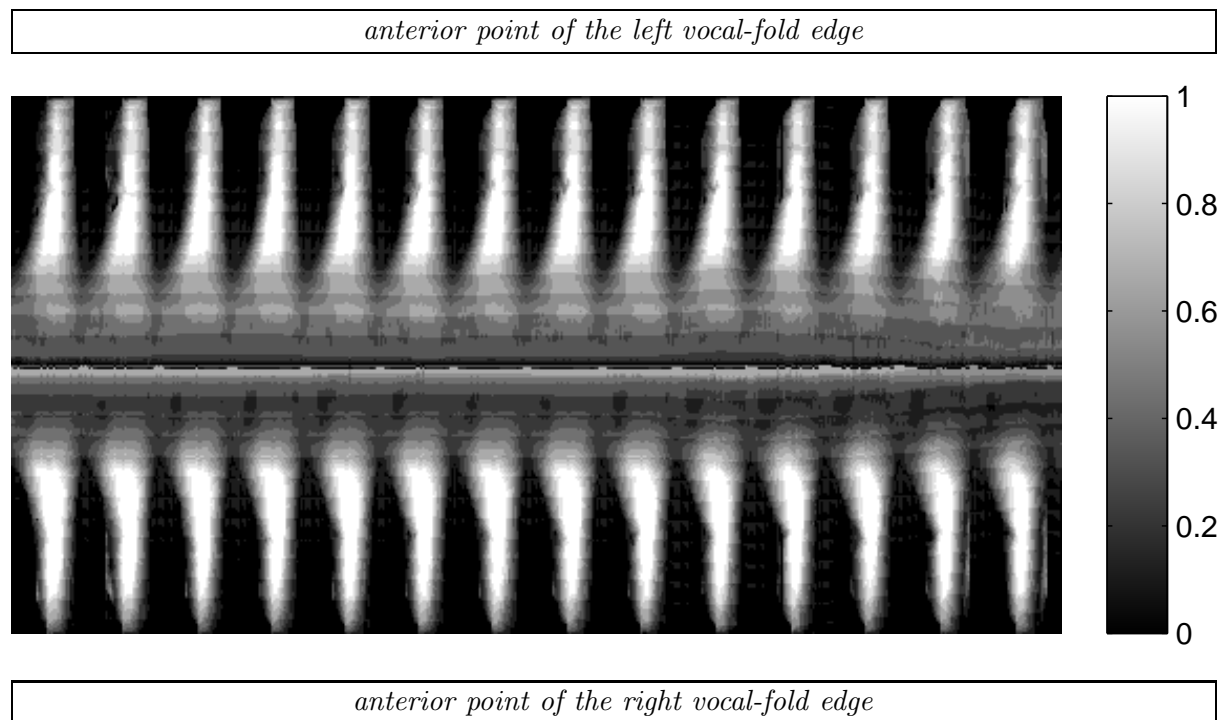


FIGURE 3.10: PVG of a video sequence.

### 3.6.2.2 Computation of Glottovibrograms (GVG)

However, the previously described PVG visualizes the deflections of the medial vocal-fold edges from the glottal axis. In order to visualize the deflection between the medial vocal-fold edges we present the following transformation of the above method.

Instead of measuring the distance between the glottal symmetry axis and the vocal-fold contours, we propose to measure the distance between the vocal-fold contours themselves, that is the distance of the points that are found across the glottal axis, on the perpendicular to the glottal axis line.

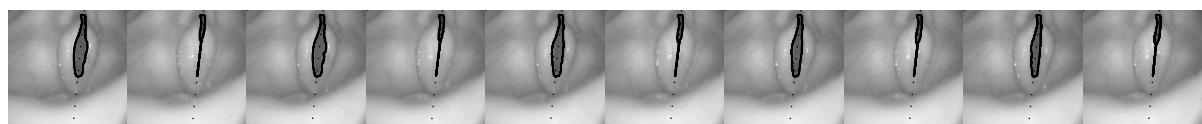


FIGURE 3.11: Linked contours with corresponding glottal axis. For the same frames presented in figure 3.9, the glottal axis is depicted as well. The distances  $\delta(m, t)$  are computed among points perpendicular to the glottal axis for each image.

$$\delta^{gl}(m, t) = \|c_L(m, t) - c_R(m, t)\|_2, \quad \forall m \quad (3.19)$$

The distances  $\delta^{gl}(m, t)$  are stored in a matrix  $\mathcal{D}^{(2M+1) \times (T+1)}$  and treated in the same way as presented before.

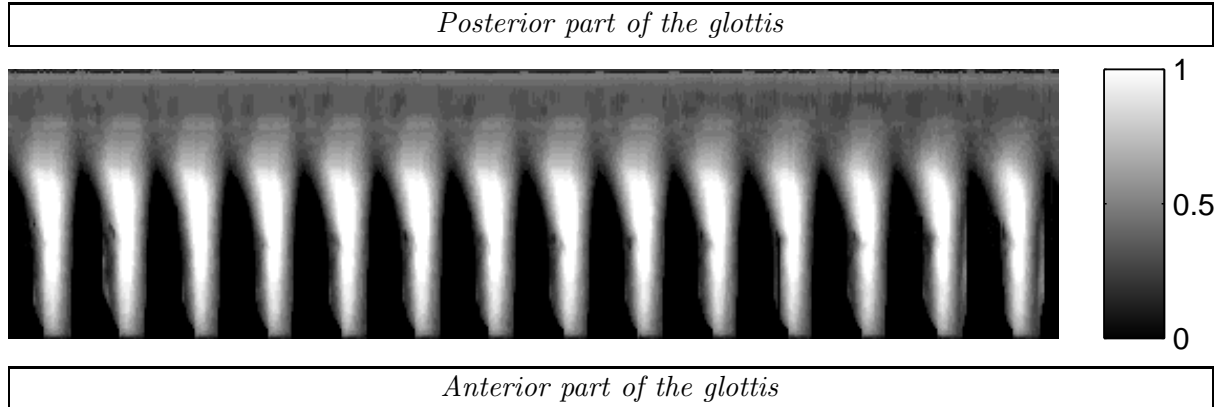


FIGURE 3.12: Glottovibrogram of a video sequence. Black corresponds to zero distance between the vocal-fold edges and white corresponds to the maximum observed distance.

### Contribution

With this representation, the deflections of the glottal area are depicted within a single image. By computing the distance of the vocal-fold edges, we can depict the exact physiological behaviour of the vocal folds. From the posterior to the anterior part of the glottis, we can observe the shape of the vocal-fold edges and extract conclusions on the physiology of the vibration. The GVG visualization can delineate the vocal-folds vibratory pattern, when the PVG fails to be clear, due to poor detection of the glottal axis. In figures 3.13 and 3.14 two cases are shown, where the GVG presents clearer than the corresponding PVG the vibratory pattern.

By computing the derivative of the distance, we can also depict on this representation the speed profile of the vibrations. This can be done by superimposing on the GVG the visualisation of the derivative of the  $\mathcal{D}$  matrix. The velocity pattern along the muscles of the vocal-folds is interesting to be visualized. We can observe with that way how do the folds move under different mechanisms.

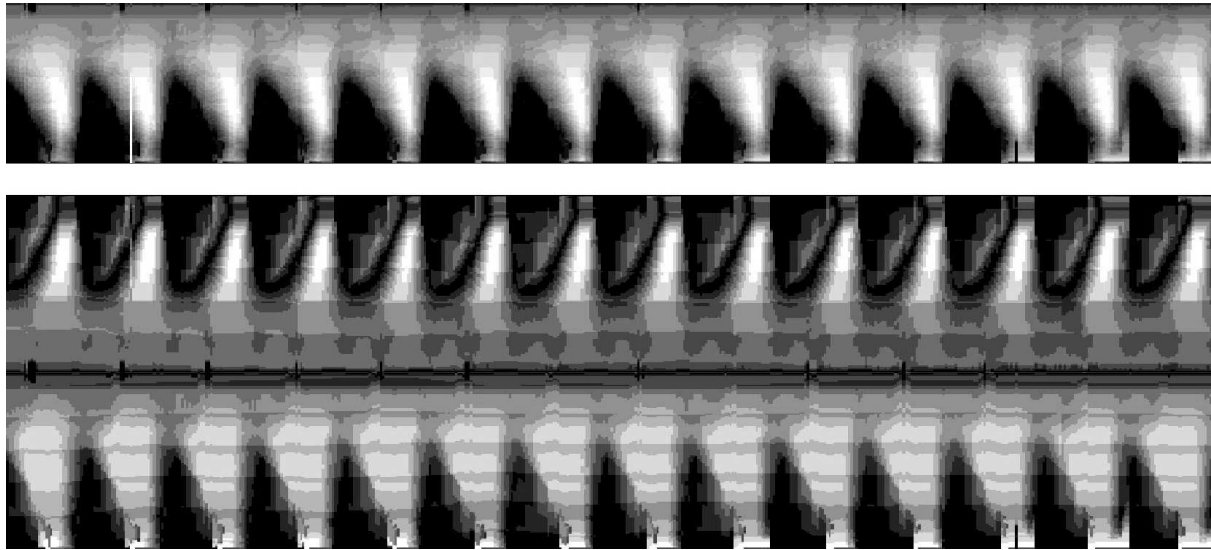


FIGURE 3.13: GVG and PVG of *HH\_SEQ\_066*. The vocal fold edges are clearer in the GVG. There are no artefacts due to computation and the closed phase is distinguishable.

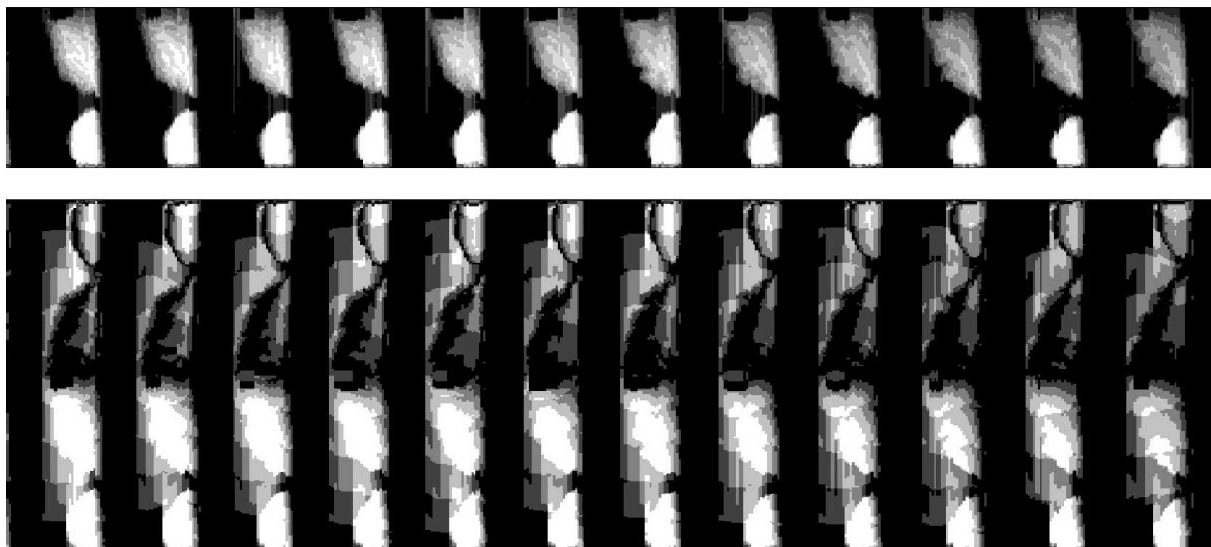


FIGURE 3.14: GVG and PVG of *HH\_SEQ\_067*. This is a case of a vibratory pattern which consists of the zipper-like movement of two distinct regions. In the PVG the left and right vocal-fold edges are not distinctive, whereas in the GVG, the two regions and their evolution are clearer. The asymmetry of the vibration is evident in the GVG; the vocal folds close faster than they open. The two glottal regions never merge completely.

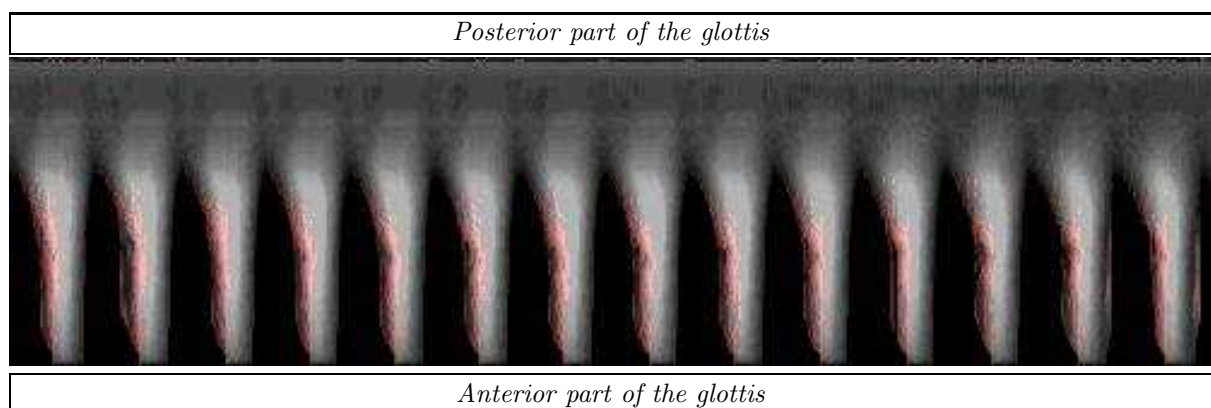


FIGURE 3.15: GVG with maximum speed profile of a video sequence (the GVG is darker due to the superimposition of the speed profile image. The red regions correspond to the points where the vocal-folds move with maximum velocity.

## Chapter 4

# Materials & Methods

### 4.1 Database Presentation

#### 4.1.1 Data acquisition

The data used during this work were taken from the UKE high-speed database, kind courtesy of Nathalie Henrich, which was recorded at the University Medical Center Hamburg-Eppendorf (UKE) in Hamburg, Germany, by the team of Pr. Hess (Frank Müller and Anna-Katharina Licht). It consists of synchronized high-speed, audio and EGG recordings of two subjects, one singer and one speaker.

For the high-speed recordings, a rigid endoscope (Wolf 90 E 60491) equipped with a continuous source of light (Wolf 5131) driven by optic fiber was used. The system was equipped with a grayscale charged coupled device (CCD) with a spatial resolution of  $256 \times 256$  pixels and sampling frequency of 2000 and 4000 *fps*. The data used were recorded at 4000 *fps*. The mouth is wide open, while the tongue is pulled between the thumb and middle finger of the doctor. Using a rigid endoscope introduced into the oral cavity of the subject limits the exploration of its possibilities in vocal production of vowels. The average duration of recordings is 4*sec*.



FIGURE 4.1: Examination by high-speed videoendoscopy (UKE)

Along with the high-speed recording, the EGG signal is acquired by an electroglottogram (Glottal Enterprises, EL-2 type (Rothenberg, 1992)). The electroglottogram is equipped with two pairs of parallel electrodes which can be adjusted for a better placement on the neck of the subject, taking into account the vertical displacement of the larynx during phonation. It delivers a current of  $10mA$  with frequency modulated at  $2MHz$ . The voltage output of the device is a few hundred millivolts and does not exceed  $1.5V$ . The electrodes are circular with a diameter of  $34mm$ . A gel is applied to their surface during their placement to facilitate contact with the skin. The EGG and audio sampling frequency is at  $44170Hz$ , directly on the medical platform. A real-time monitoring of the EGG signal is performed for each recording with an A/D oscilloscope.

### 4.1.2 Recordings

The purpose of the experiment was to compare EGG features and glottal behavior for different spoken and sung situations. Five participants were recorded with different voice qualities, pitches and transitions. The vocal-fold vibrations were recorded from a top-view position along with the EGG and audio signal.

For the purpose of this work, 60 recordings were used, taken from two male subjects. To ensure the processing of sustained phonation only, the sequences used were chosen approximately at the middle of phonation. They all comprise of 501 frames, which correspond to roughly  $125msec$  of sustained phonation. The corresponding frames and EGG signals were stored for all experiments. Tables 4.1, 4.2 and 4.3 present all recordings along with additional information; the task the subject was asked to perform, the glottal movement observed in the recording, as well as the glottovibrogram thumbnail. Complete studies on laryngeal mechanisms, mentioned in this work, can be found in (Henrich, 2006, Roubeau et al., 2009), while the phonation types and their respective characteristics can be found in (Ní Chasaide and Gobl, 1997, Trask, 1996).

### 4.1.3 Synchronization issues

There are certain synchronization issues that arise from the recording process. The medical platform records with sampling frequency  $44170Hz$ , instead of  $44100Hz$ . This introduces a delay of  $3.59e^{-5}msec$ , relative to the length of the recorded EGG signal:

$$\text{delay} = \text{length of the EGG signal} \times \left( \frac{1}{44100} - \frac{1}{44170} \right) * 1000, \quad \text{in } msec. \quad (4.1)$$

Furthermore, the different sampling frequencies ( $4000fps$  for the high-speed video,  $44170Hz$  for the EGG signal) introduce further uncertainty. For every frame captured we have  $\frac{44170}{4000} = 11.04$



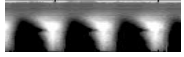
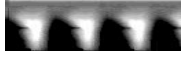
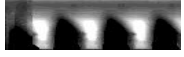
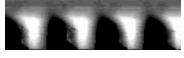


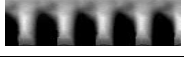





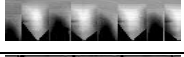
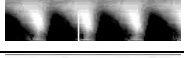





Index	Sequence Name	Description	Glottal movement	GVG thumbnail
1	<i>HH_SEQ_0057</i>	modal	anterior-to-posterior	
2	<i>HH_SEQ_0057a</i>	modal	anterior-to-posterior	
3	<i>HH_SEQ_0058</i>	modal	anterior-to-posterior	
4	<i>HH_SEQ_0058a</i>	modal	anterior-to-posterior	
5	<i>HH_SEQ_0059</i>	modal	middle-to-edges	
6	<i>HH_SEQ_0059a</i>	modal	middle-to-edges	
7	<i>HH_SEQ_0060a</i>	breathy effort	middle-to-edges	
8	<i>HH_SEQ_0061a</i>	breathy normal	anterior-to-posterior	
9	<i>HH_SEQ_0062</i>	normal to creaky	anterior-to-posterior	
10	<i>HH_SEQ_0064</i>	tensed	complex vibration	
11	<i>HH_SEQ_0064a</i>	tensed	complex vibration	
12	<i>HH_SEQ_0065</i>	modal to breathy	middle-to-edges	
13	<i>HH_SEQ_0065a</i>	modal to breathy	middle-to-edges	
14	<i>HH_SEQ_0066</i>	breathy to modal	anterior-to-posterior	
15	<i>HH_SEQ_0066a</i>	breathy to modal	anterior-to-posterior	
16	<i>HH_SEQ_0067</i>	modal to creaky	anterior-to-posterior	
17	<i>HH_SEQ_0067a</i>	modal to creaky	anterior-to-posterior	
18	<i>HH_SEQ_0068</i>	creaky to modal	anterior-to-posterior	
19	<i>HH_SEQ_0068a</i>	creaky to modal	anterior-to-posterior	

TABLE 4.1: Used recordings from the UKE database (spoken samples). The high-speed sequence indices, names are shown, as well as the used laryngeal mechanisms and the glottal vibration pattern. The last column presents a small portion of the corresponding glottovibrogram, the GVG thumbnail. When a sequence’s name holds *a* at the end, e.g. *HH\_SEQ\_0057a*, the sequence has been cropped from the same sequence *HH\_SEQ\_0057* without using intersecting frames.

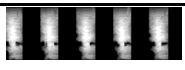
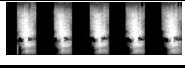
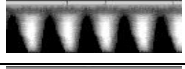
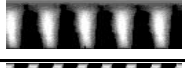

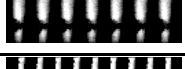

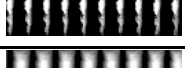
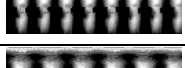

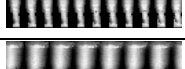
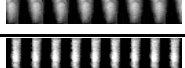
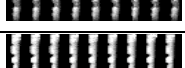
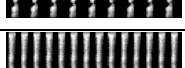
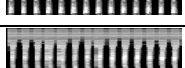



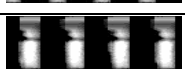
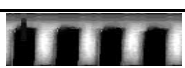
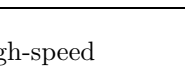
Index	Sequence Name	Description	Glottal movement	GVG thumbnail
20	<i>HH_SEQ_0069</i>	M1 D3	middle-to-edges	
21	<i>HH_SEQ_0069a</i>	M1 D3	middle-to-edges	
22	<i>HH_SEQ_0070</i>	M1 D3	anterior-to-posterior	
23	<i>HH_SEQ_0070a</i>	M1 D3	anterior-to-posterior	
24	<i>HH_SEQ_0071</i>	M1 A3 crescendo	anterior-to-posterior	
25	<i>HH_SEQ_0071a</i>	M1 A3 crescendo	anterior-to-posterior	
26	<i>HH_SEQ_0072</i>	M1 D4 crescendo	anterior-to-posterior	
27	<i>HH_SEQ_0072a</i>	M1 D4 crescendo	anterior-to-posterior	
28	<i>HH_SEQ_0073</i>	glissando M1 M2 transition	anterior-to-posterior	
29	<i>HH_SEQ_0075a</i>	glissando M1 M2 no transition	anterior-to-posterior	
30	<i>HH_SEQ_0078a</i>	glissando with transition	anterior-to-posterior	
31	<i>HH_SEQ_0079</i>	M2 A3	anterior-to-posterior	
32	<i>HH_SEQ_0083</i>	transition M1 M2 on D4	anterior-to-posterior	
33	<i>HH_SEQ_0084a</i>	Mx1	anterior-to-posterior	
34	<i>HH_SEQ_0086</i>	M2 A4	anterior-to-posterior	
35	<i>HH_SEQ_0086a</i>	M2 A4	anterior-to-posterior	
36	<i>HH_SEQ_0089</i>	breathy normal	anterior-to-posterior	
37	<i>HH_SEQ_0089a</i>	breathy normal	anterior-to-posterior	
38	<i>HH_SEQ_0090</i>	modal	anterior-to-posterior	
39	<i>HH_SEQ_0090a</i>	modal	posterior-to-anterior	
40	<i>HH_SEQ_0095</i>	modal to breathy	anterior-to-posterior	

TABLE 4.2: Used recordings from the UKE database (sung samples, part I). The high-speed sequence indices, names are shown, as well as the used laryngeal mechanisms and the glottal vibration pattern. The last column presents a small portion of the corresponding glottovibrogram, the GVG thumbnail. When a sequence’s name holds *a* at the end, e.g. *HH\_SEQ\_0069a*, the sequence has been cropped from the same sequence *HH\_SEQ\_0069* without using intersecting frames.

Index	Sequence Name	Description	Glottal movement	GVG thumbnail
41	<i>HH_SEQ_0095a</i>	modal to breathy	anterior-to-posterior	
42	<i>HH_SEQ_0096</i>	modal to creaky	medial-to-edges	
43	<i>HH_SEQ_0096a</i>	modal to creaky	medial-to-edges	
44	<i>HH_SEQ_0098</i>	$F_0$ down	posterior-to-anterior	
45	<i>HH_SEQ_0098a</i>	$F_0$ down	posterior-to-anterior	
46	<i>HH_SEQ_0099</i>	modal to breathy	posterior-to-anterior	
47	<i>HH_SEQ_0099a</i>	modal to breathy	posterior-to-anterior	
48	<i>HH_SEQ_0101</i>	$F_0$ up	posterior-to-anterior	
49	<i>HH_SEQ_0101a</i>	$F_0$ up	posterior-to-anterior	
50	<i>HH_SEQ_0102</i>	$F_0$ down	anterior-to-posterior	
51	<i>HH_SEQ_0103</i>	modal to tense, $F_0$ up	anterior-to-posterior	
52	<i>HH_SEQ_0105</i>	init final gs	posterior-to-anterior	
53	<i>HH_SEQ_0106</i>	vibrato soft to middle	posterior-to-anterior	
54	<i>HH_SEQ_0106a</i>	vibrato soft to middle	posterior-to-anterior	
55	<i>HH_SEQ_0107</i>	vibrato middle to strong	posterior-to-anterior	
56	<i>HH_SEQ_0108</i>	vibrato	posterior-to-anterior	
57	<i>HH_SEQ_0110</i>	$F_0$ down, modal to breathy	posterior-to-anterior	
58	<i>HH_SEQ_0111</i>	$F_0$ down	posterior-to-anterior	
59	<i>HH_SEQ_0112</i>	$F_0$ up	posterior-to-anterior	
60	<i>HH_SEQ_0113</i>	$F_0$ up, modal to tense	posterior-to-anterior	

TABLE 4.3: Used recordings from the UKE database (sung samples, part II). The high-speed sequence indices, names are shown, as well as the used laryngeal mechanisms and the glottal vibration pattern. The last column presents a small portion of the corresponding glottovibrogram, the GVG thumbnail. When a sequence’s name holds *a* at the end, e.g. *HH\_SEQ\_0096a*, the sequence has been cropped from the same sequence *HH\_SEQ\_0096* without using intersecting frames.

samples from the EGG signal. That means that we cannot have a direct correspondance of samples and there is a minimum time interval, under which we can only assume correspondance:

$$\begin{aligned} t_{uncertainty} &= \frac{44170}{4000} * \frac{1}{44170} * 1000 \\ t_{uncertainty} &= 0.25msec. \end{aligned} \tag{4.2}$$

## 4.2 Automatic segmentation

The automatic glottis segmentation from the high-speed video sequences, as well as from the digital kymographic sequences, was implemented in Matlab. Each sequence was processed individually; any video sequence in .bld, .avi or .mat format can be processed. The average execution time per sequence is approximately 45min.

## 4.3 Manual evaluation of the segmentation results

Automatic segmentation results have been manually verified and corrected, if needed, with the use of an interactive tool. The interface was originally presented and used in the PhD dissertations of N. Henrich (Henrich, 2001) and L. Bailly (Bailly, 2009). It consists of an interface implemented in Matlab, which interacts with the given input and treats the contour using Bezier splines (Bezier, 1972). The video and segmentation quality was subjectively evaluated by 13 participants, 7 of which were familiar with voice analysis and image processing. The connected contours, as presented during the processing of the PVG and GVG (section 3.6) were used, so as to provide the users as much information as possible on the nature of the task. Each participant was asked to assess 3 high-speed video sequences with their corresponding segmentation results. The rest of the high-speed sequences were processed by the author.

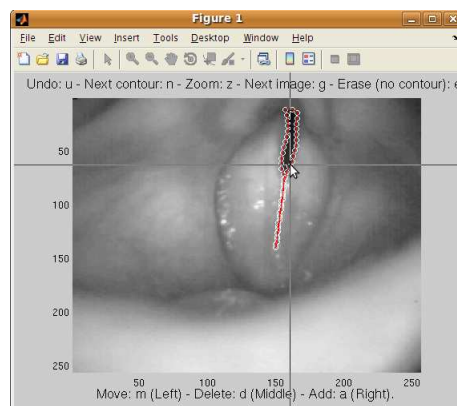


FIGURE 4.2: Screenshot from the interface tool. The user can evaluate the segmented contour for each frame individually through the interactive tool, control the contour and fix it if it does not track the glottal area.

---

For all 60 high-speed video sequences, it was requested from the participants to rate the video quality (lighting conditions, contrast relative to the discrimination of the glottal area), as well as the segmentation quality (tracking of vocal-fold movements, irrelevant excrescences). The user was presented with a sequence of frames and was evaluating consecutively each frame through an interactive window (figure 4.2). If the contour followed correctly the glottal area, the user was directing the program to the following frame. In the other case, with the use of the mouse, the user could control the contour and fix it. When the sequence processing was over, the user was to evaluate the video and segmentation quality. On average, one sequence required about 15min to be fully processed.



# Chapter 5

## Results

### 5.1 High-speed video segmentation analysis

#### 5.1.1 Subjective evaluation

The manual verification of the automatic segmentation resulted in a number of interesting findings. On a 5-point scale, with 1 representing very bad quality and 5 very good quality, the results of the subjective rating are presented in figure 5.1. The average video quality is  $4.2 \pm 0.72$  (mean value  $\pm$  standard deviation), while the average segmentation quality of all high-speed video sequences is  $4.2 \pm 1$ . In 71% of recordings, the segmentation was rated equal or higher than the video quality, while 93% was characterized as more than acceptable (average to very good).

#### 5.1.2 Comparison between manual and automatic segmentation

For a quantitative evaluation the participants were also asked to manually correct the segmentation results that did not follow the glottal area. The error of segmentation for the entire database is  $-1.8 \pm 18.8$  pixels. In figure 5.2 the error of segmentation per high-video sequence is shown. The negative mean value indicates that it is more likely for the contour to be expanded in nearby excrescences than converge before the actual vocal-fold edges. The absolute error of segmentation (by considering the absolute differences) for the entire database is  $4 \pm 18.4$  pixels. The absolute difference per sequence is shown in figure 5.3.

Most of the segmentation errors occurred in the posterior or anterior part of the glottal area. Either the active contours could converge before the actual vocal fold edges, either they include surrounding areas. As evaluated in most cases, the contour tracked very well the glottal area,

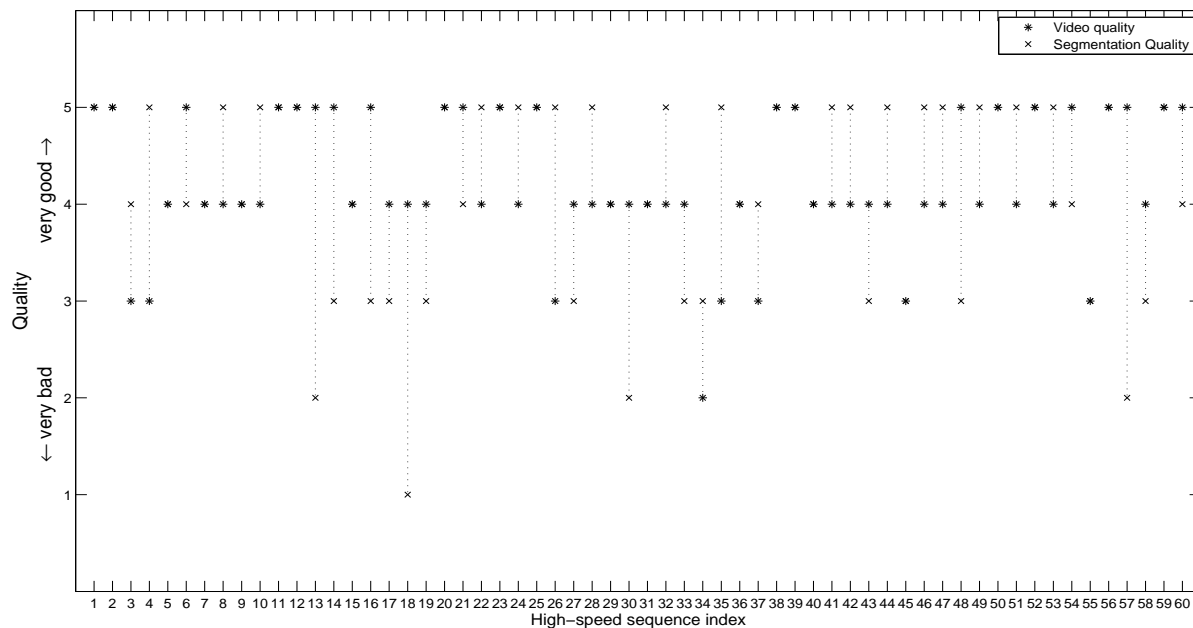


FIGURE 5.1: Video and segmentation subjective assessment of the entire database on a 5-point scale.

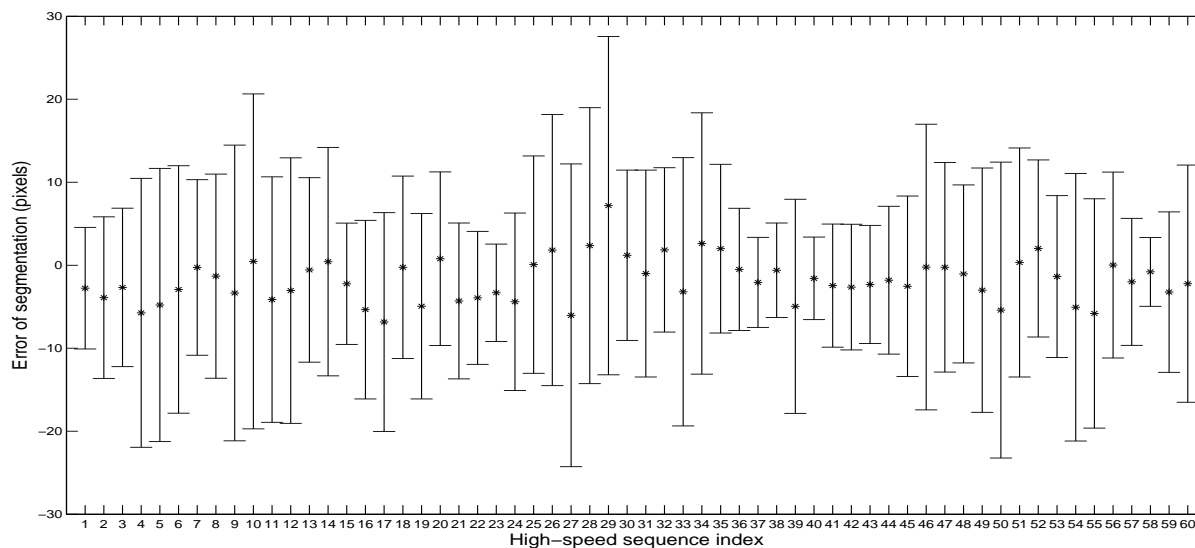


FIGURE 5.2: Differences in number of pixels between automatic and manual segmentation per high-speed video sequence. The asterisks represent the mean value of error, while the vertical bars indicate the size of standard deviation.



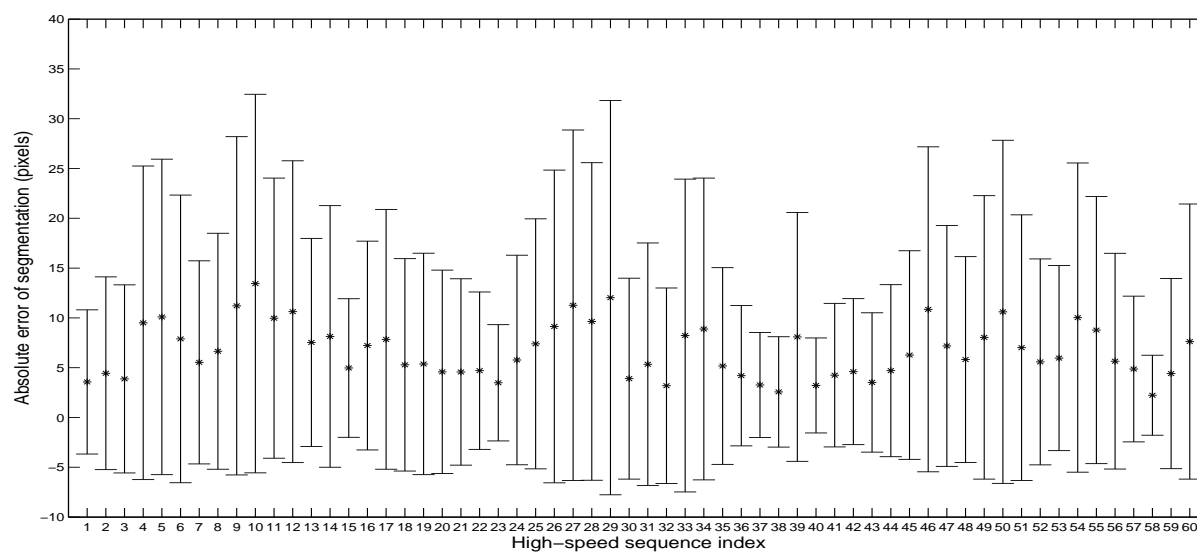
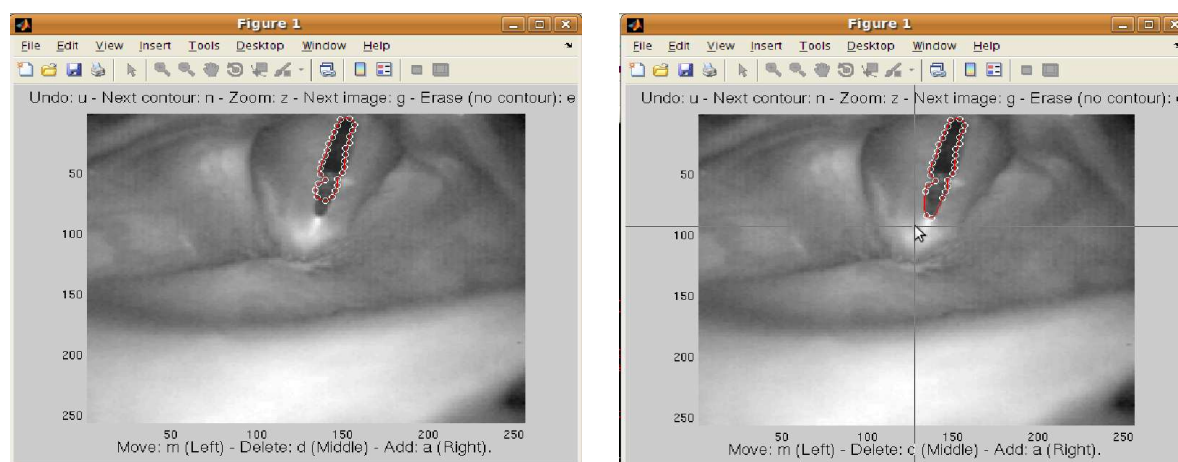


FIGURE 5.3: Absolute differences in number of pixels between automatic and manual segmentation per high-speed video sequence. The asterisks represent the mean value of error, while the vertical bars indicate the size of standard deviation.

despite the errors around the edges. Two cases where the segmentation procedure failed to track the glottal area effectively are presented in figures 5.4 and 5.5.



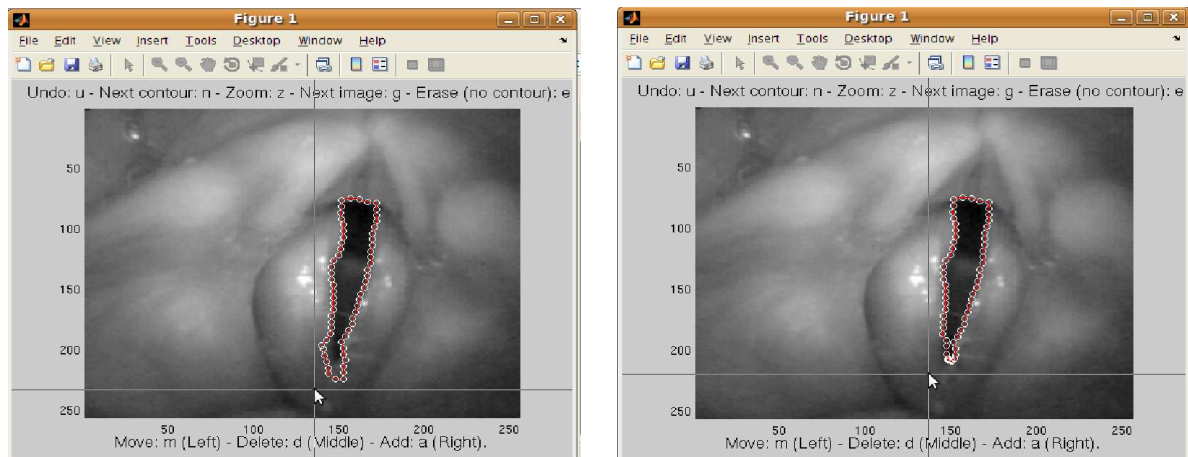
Computed contour on a image; the anterior part of the glottis has not been detected.

Corrected contour; the contour now tracks the entire glottal area

FIGURE 5.4: Segmentation errors and manual evaluation (part I). The active contours converge before the actual vocal folds, at the anterior part of the glottal area.

The amount of intra-sequence errors is also interesting. In figure 5.6, the histogram of corrections per sequence is shown.

An important aspect in evaluating the segmentation results is relative to the static phases of glottal opening and closing instants. The segmentation procedure needs to meet the demands for glottal source analysis. For that reason, the amount of error relative to the glottal instants



Computed contour on a image; the contour includes an irrelevant anterior region.

Corrected contour; the contour now tracks the entire glottal area

FIGURE 5.5: Segmentation errors and manual evaluation (part II). The contour includes excrescences at the anterior part of the glottal area due to poor local statistics, thus overestimating the actual glottal area.

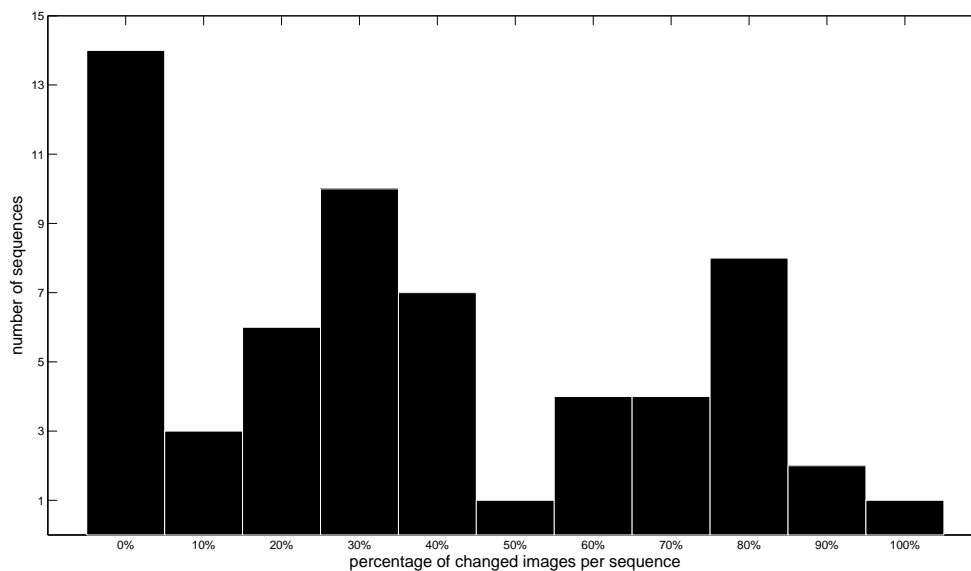


FIGURE 5.6: Percentage of images per sequence on which the glottal area contour was manually changed during annotation.

was also investigated. It is evident, as shown in figure 5.7, that there have been hardly any changes in most high-speed video sequences. In figure 5.8 the absolute error of segmentation at the glottal closing and opening instants is shown. This validates the use of data acquired for comparison with the EGG signals, as it will be presented in section 5.2.

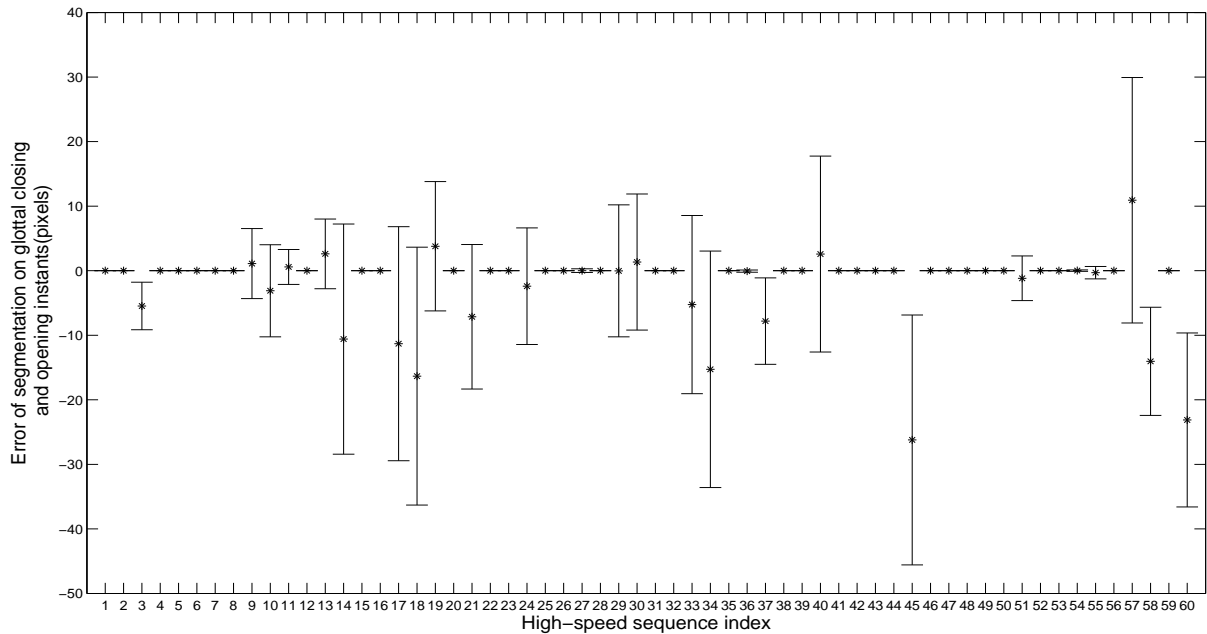


FIGURE 5.7: Error of segmentation of glottal closing and opening instants per high-speed video sequence. Frames which correspond to glottal closing and opening instants are investigated within each sequence. The asterisks represent the mean value of error, while the vertical bars indicate the size of standard deviation.

### 5.1.3 Segmentation comparison on HSV and DKG sequences

The segmentation procedure has also been used in DKG sequences. By appropriately segmenting the DKG sequences, we can propagate the results to the HSV sequence. Only a few sequences from the database have been tested. The discriminative power of this method have proven to be very good. A complete investigation, as in the case of HSV processing, is missing from this work due to lack of time.

In figure 5.10 a sequence of frames and its equivalent segmentation with the two methods is shown. Due to the nature of the treated images, the use of DKG sequences enhances the capturing power of the segmentation method, thus allowing to capturing even very small glottal regions.

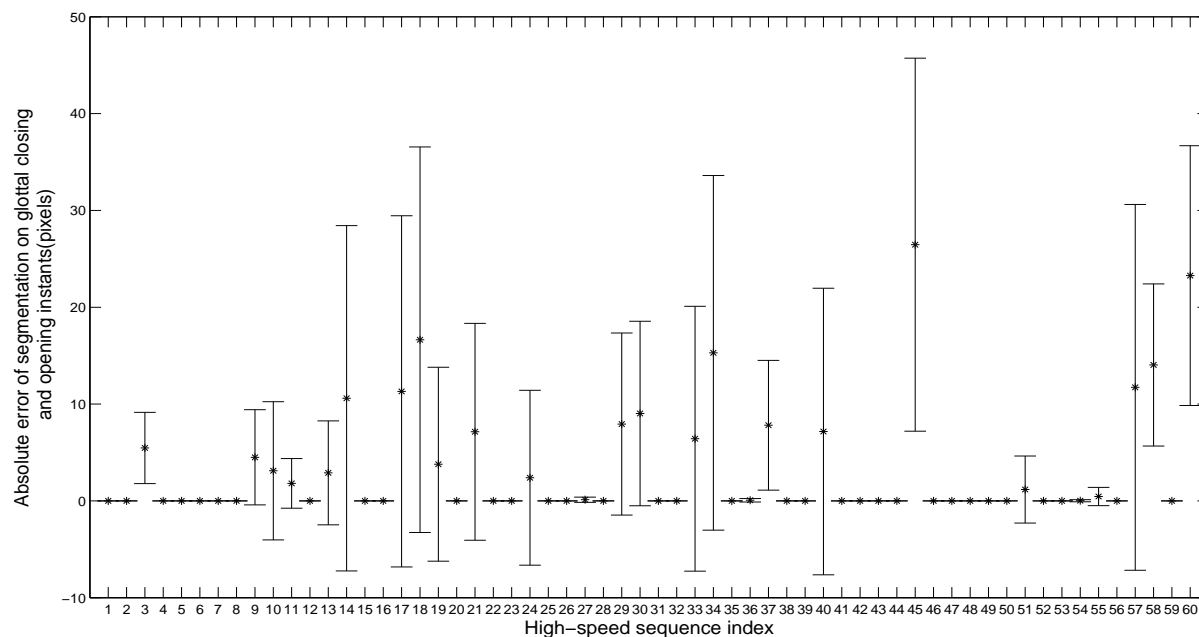


FIGURE 5.8: Absolute error of segmentation of glottal closing and opening instants per high-speed video sequence. Frames which correspond to glottal closing and opening instants are investigated within each sequence. The asterisks represent the mean value of error, while the vertical bars indicate the size of standard deviation.

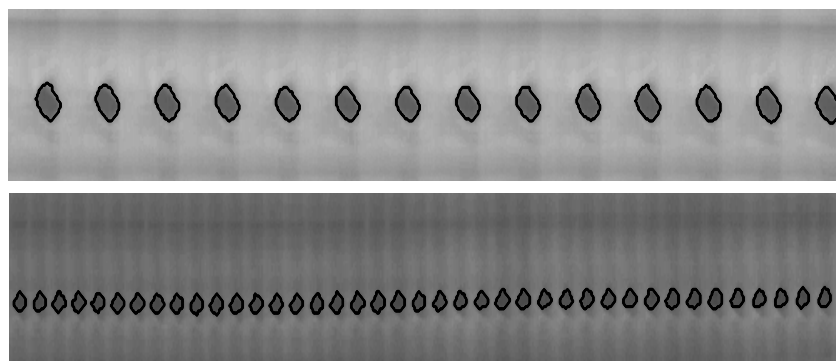


FIGURE 5.9: Kymographic (DKG) images from two sequences (from top to bottom, *HH\_SEQ\_0067a* and *HH\_SEQ\_0078a* respectively). The black splines correspond to the segmented regions.

## 5.2 Comparison on glottal parameters estimated from EGG and HSV data

### 5.2.1 Fundamental frequency estimation

The first step in analysing the descriptive power of our data processing was to compute the fundamental frequency of the EGG and GLA signals. The YIN estimator has been used for this purpose (de Cheveigné and Kawahara, 2002). The GLA signals are upsampled at  $44170Hz$  (sampling frequency of EGG and audio signals).

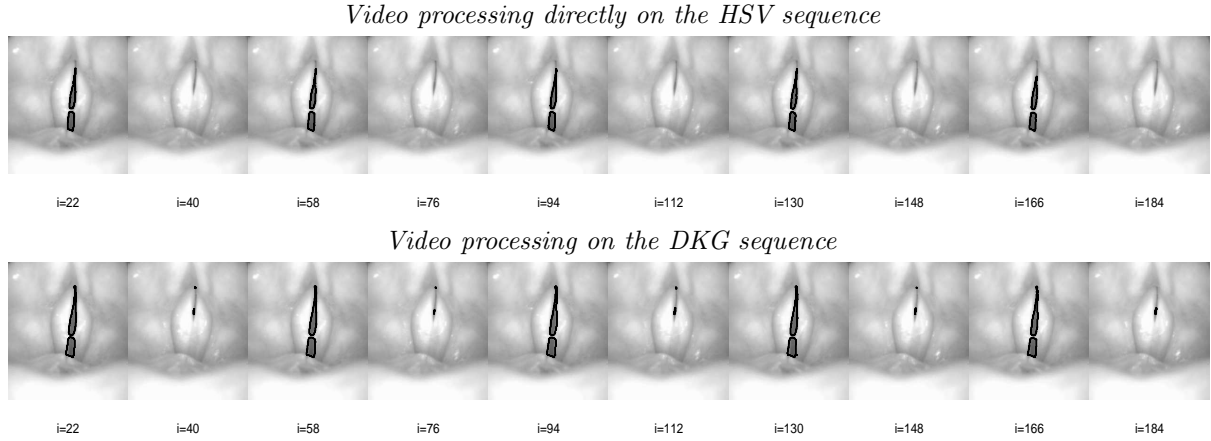


FIGURE 5.10: Segmentation results in sequence of frames (sequence (*HH\_SEQ\_0067a*)) by DKG and HSV processing.

hsv index	$F_0$ (EGG, Hz)	$F_0$ (GLA, Hz)	comments
<i>HH_SEQ_0062</i>	36	42	double peaks in both EGG & GLA signal
<i>HH_SEQ_0106</i>	368	722	high fundamental frequency
<i>HH_SEQ_0106a</i>	368	742	high fundamental frequency
<i>HH_SEQ_0107</i>	351	690	high fundamental frequency
<i>HH_SEQ_0108</i>	353	689	high fundamental frequency

TABLE 5.1: Cases in which  $F_0$  estimation indicates gross error.

In order to evaluate the accuracy of  $F_0$  computation, we used the interval difference formulation (equation 5.1). The interval between two frequencies in semitones should be lower than 0.5 so as to be considered equal. The threshold was empirically chosen, so as to ensure minimum estimation error. We have used the  $F_0$  computed from the EGG signal as the base frequency ( $f_2$ ).

$$\Delta_{f_0} = 12 \log_2 \frac{f_1}{f_2}, \quad \text{in semitones} \quad (5.1)$$

Only 5 sequences present  $\Delta_{f_0}$  greater than 0.5. For these cases, further investigation on the origin of this difference was necessary. The results of fundamental frequency estimation, for which there are significant differences, are shown in table 5.1. In the first case (*HH\_SEQ\_0062*) there are double peaks in both the EGG and the GLA signal, which cannot be processed by the frequency estimator (figure 5.11). In the other cases, we observe gross errors in estimation. The estimated  $F_0$  is nearly the double of the one estimated from the EGG signals. This is mainly due to the nature of signals, as presented in figure 5.12 for a single, yet representative case.

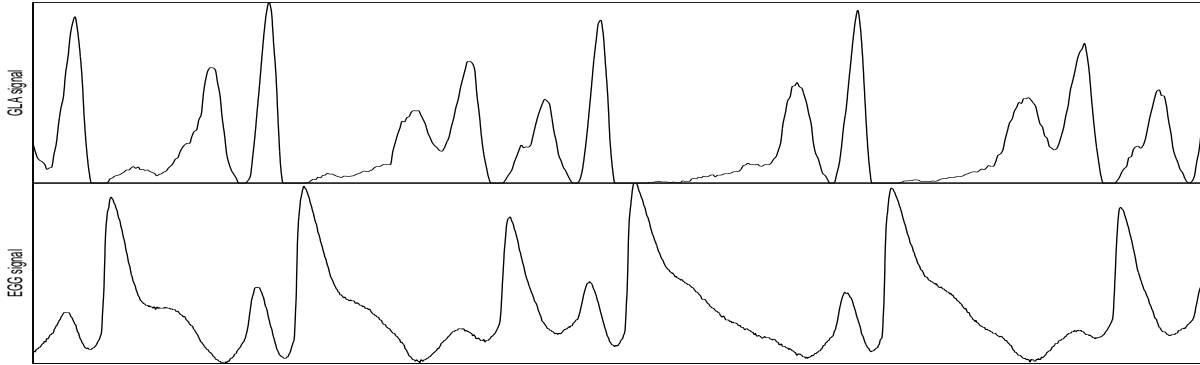


FIGURE 5.11: Difference in  $F_0$  estimation. For the sequence *HH\_SEQ\_0062*, the presence of double peaks during the vocal-fold vibration introduces noise that cannot be distinguished by the YIN estimator. The segmentation results present slight error in glottal instants ( $1.1 \pm 10.8$  pixels).

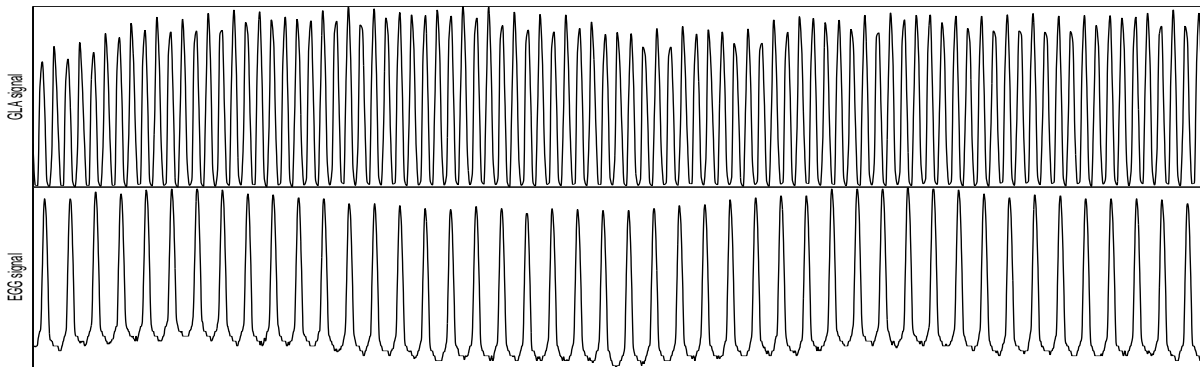


FIGURE 5.12: Difference in  $F_0$  estimation. For the sequence *HH\_SEQ\_0106*, we observe a distinct difference in frequency ( $\Delta_{F_0} = 11.67$ ), while the segmentation results present no error (error in glottal instants equal to  $0 \pm 0$  pixels).

### 5.2.2 Glottal closing and opening instants estimation

The DECOM method (Henrich et al., 2004) was used for the detection of glottal closing and opening instants on DEGG and DGLA signals. This method is applied to a four-period windowed DEGG signal which is separated into two parts: its positive part, which shows strong peaks related to GCIs, and its negative part, which shows weaker peaks related to GOIs. In figures 5.13 and 5.14 the error of alignment of glottal closing and opening instants, respectively, is presented. We can observe that in most sequences the GCIs of the DEGG signals occur before the corresponding events of the DGLA signals, while the GOIs seem to have a smaller lag. However, as previously stated in section 4.1.3, the synchronization delays due to the signal acquisition introduce an interval of uncertainty. The maximum delay of alignment for both glottal closing and opening instants is smaller than  $2.5msec$  and does not exceed the interval of  $3.59msec$ , for which there is synchronization uncertainty.

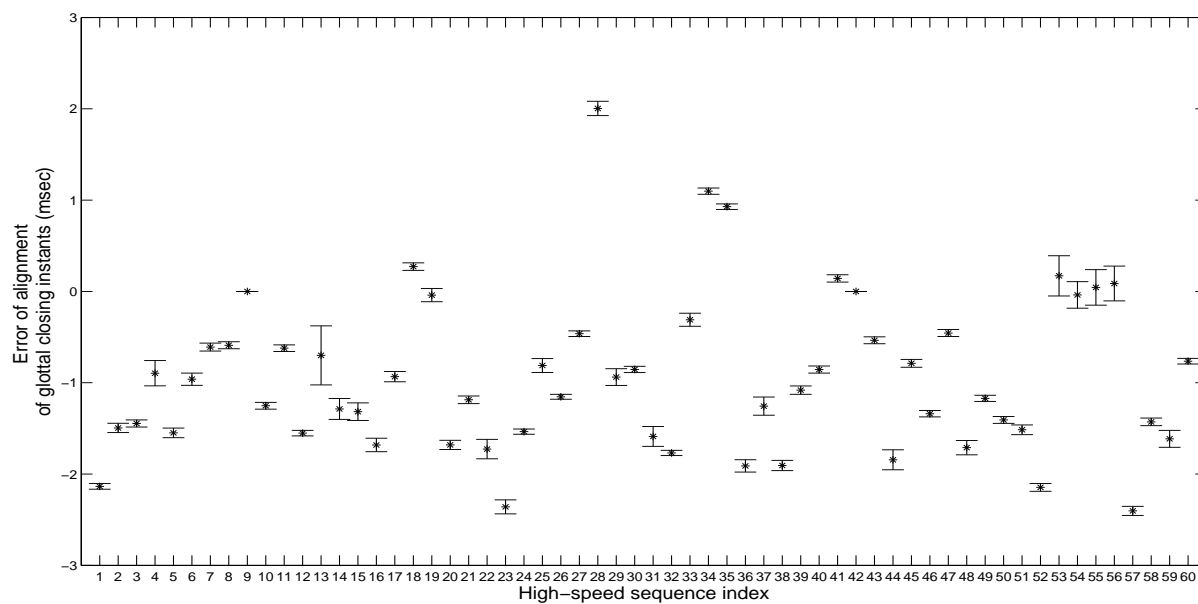


FIGURE 5.13: Error of alignment of glottal closing instants (in msec).

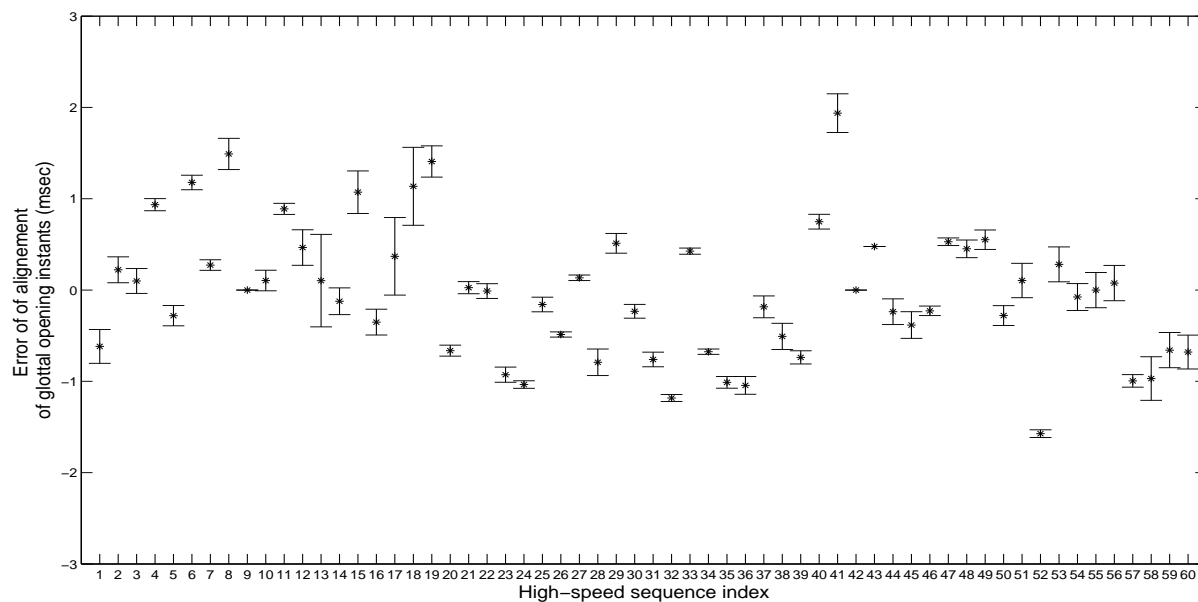


FIGURE 5.14: Error of alignment of glottal opening instants (in msec).

### 5.2.3 Open quotient estimation

There has also been an investigation on the open quotient ( $O_q$ ) estimation. The 50% threshold method on the EGG and GLA signals has been used. The results and the difference of estimation are presented in figure 5.15.

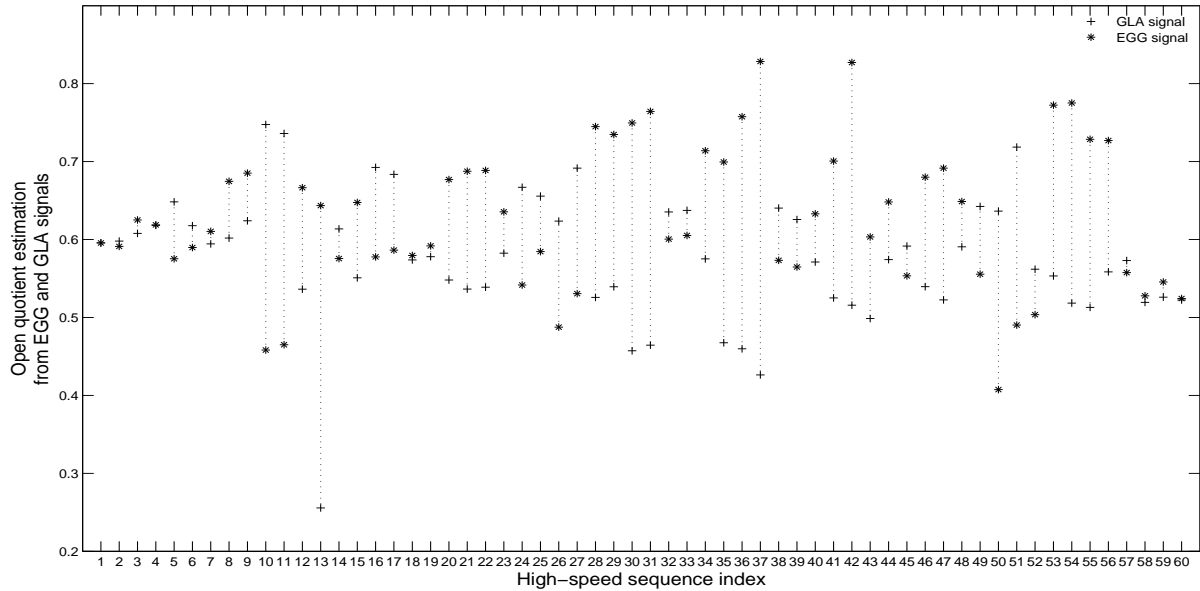


FIGURE 5.15: Open quotient estimation per high-speed video sequence. The asterisks represent the EGG values, while the crosses represent the GLA values.

Some interesting remarks can be made on the  $O_q$  estimation. First, most values fall within the eligible interval  $[0.3, 0.8]$ , which was proposed by Henrich et al. (2005) for the open quotient. Second, the difference of estimation between the two signals is related to the alignment of glottal closing and opening instants, as well as with the estimated  $F_0$ . When we observe large deviation in glottal instants alignment, we also observe a larger difference in  $O_q$  estimation.

## 5.3 Visualization of DEGG, GLA and GVG signals

An interesting way of evaluating our data is the simultaneous representation of DEGG, GLA and GVG signals. It allows the visual validation of synchronization and features estimation by also giving a complete image of the vocal-fold vibration pattern. As we can observe in the following figures (figs. 5.16, 5.17, 5.18 and 5.19), the glottal closing and opening instants estimated from the EGG signals serve as the baseline of the representation.

An interesting remark can be made from the simultaneous representation. As stated before, glottal closing instants are conceived to represent the instant the glottal area decreases with highest velocity. However, as shown in figures 5.17 and 5.18, the GCIs correspond, according



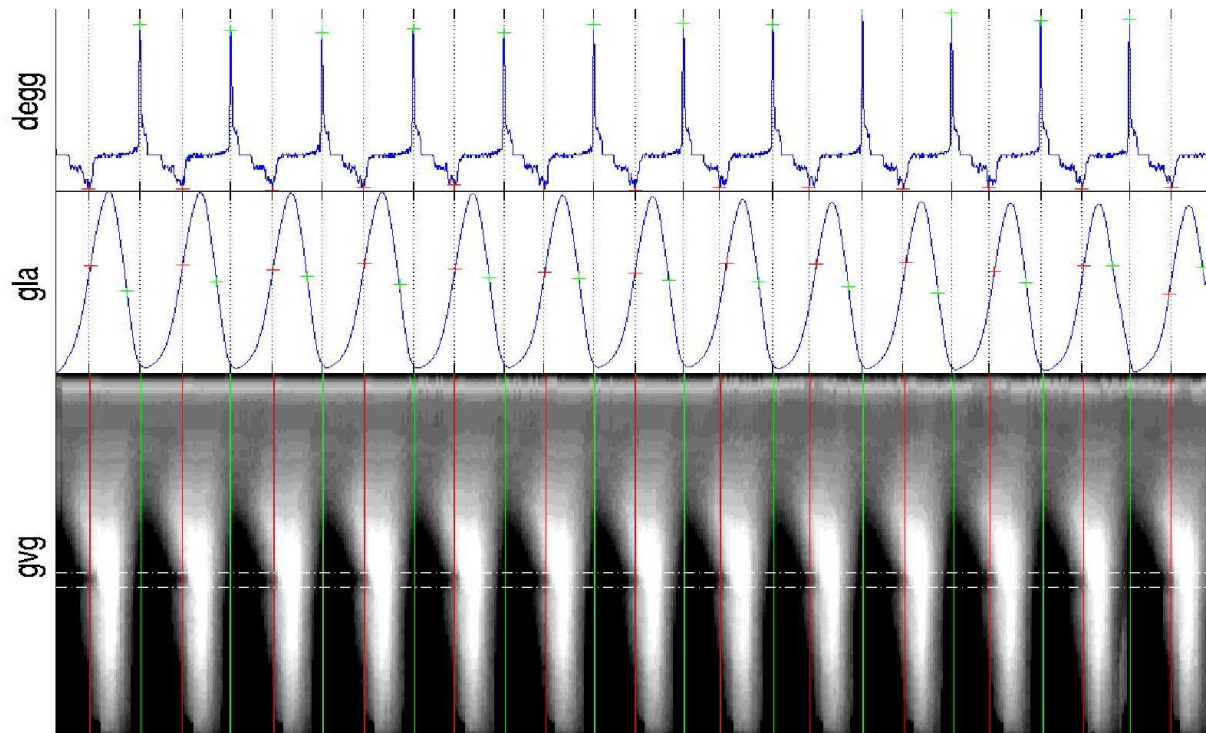


FIGURE 5.16: Synchronous representation of DEGG, GLA and GVG signals (*HH\_SEQ\_0057a*). The green crosses and lines indicate the positions of glottal closing instants, while the red ones indicate the positions of glottal opening instants, as computed from the EGG signal. In the GVG, we can observe a mucus bridge which is present in all vocal-fold vibratory cycles, highlighted in between the white dotted lines.

to the GVG to the instants where the glottal area is maximally closed. In figure 5.19, where a case of full glottal closing during cycles is presented, the closing instants fall in the middle of the closed phase.

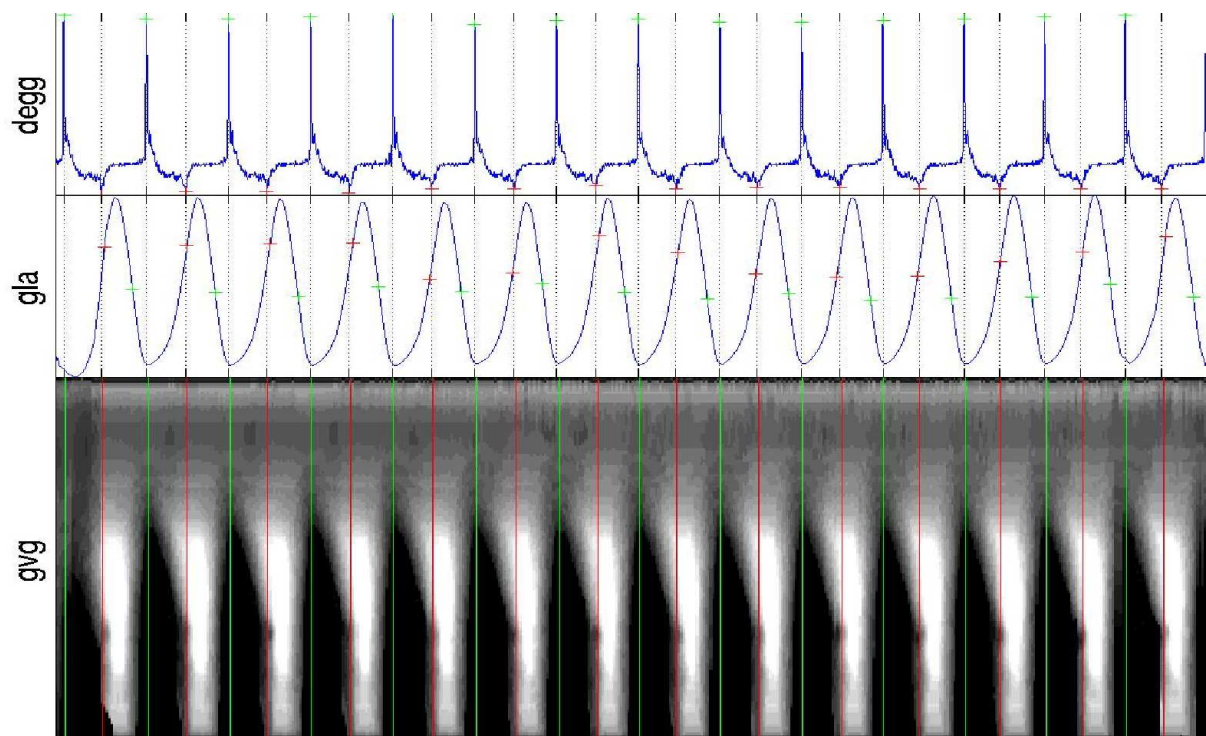


FIGURE 5.17: Synchronous representation of DEGG, GLA and GVG signals (*HH\_SEQ\_0058*). The green crosses and lines indicate the positions of glottal closing instants, while the red ones indicate the positions of glottal opening instants, as found from the EGG signal. In this case, the glottal closing instants correspond with high accuracy to the instant where the glottal area is maximally closed, while the glottal opening instants correspond to the instant where the glottal area is maximally open.

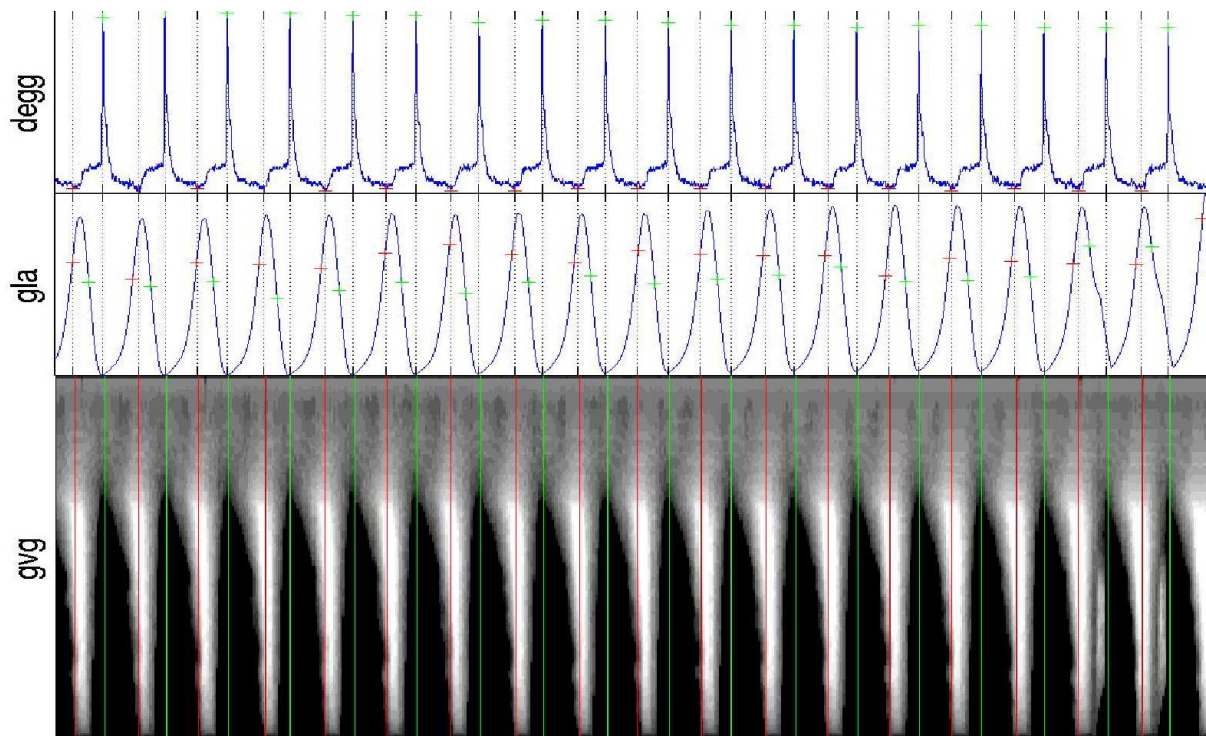


FIGURE 5.18: Synchronous representation of DEGG, GLA and GVG signals (*HH\_SEQ\_0059*). Similarly to the previous case, the glottal closing instants correspond with high accuracy to the instant where the glottal area is maximally closed, while the glottal opening instants correspond to the instant where the glottal area is maximally open.

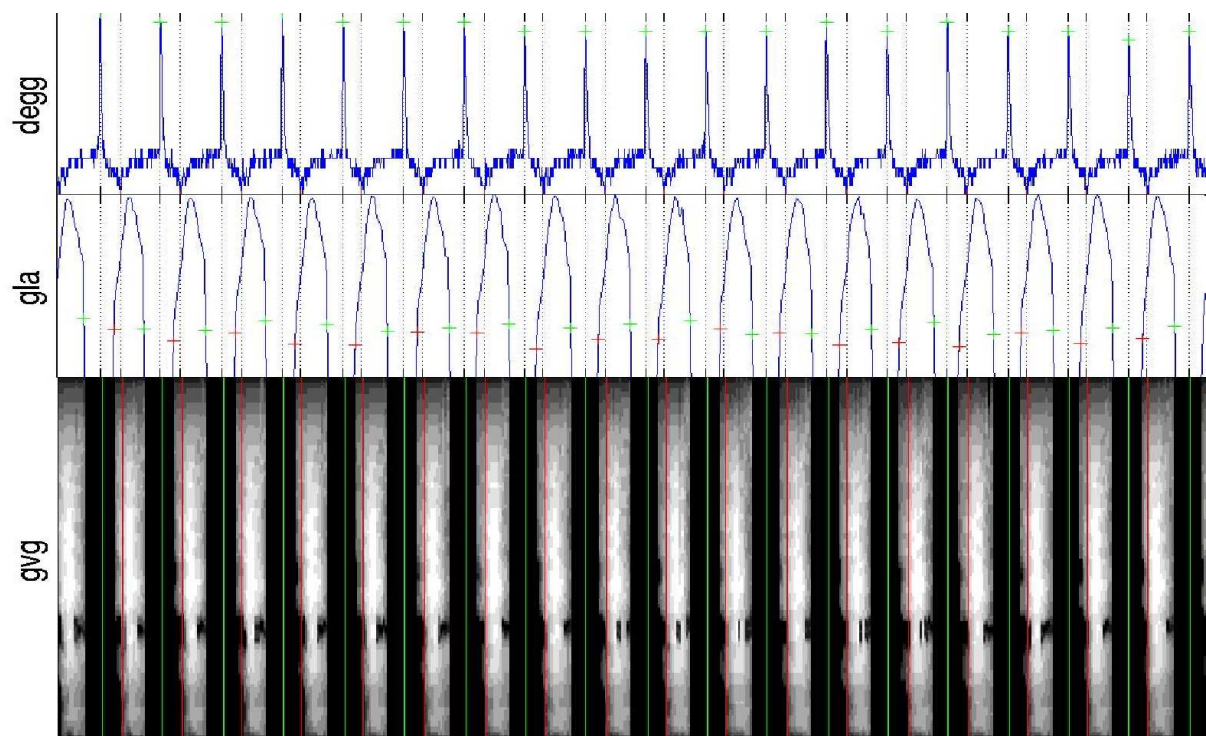


FIGURE 5.19: Synchronous representation of DEGG, GLA and GVG signals (*HH\_SEQ\_0069*). The glottal closing instants correspond with high accuracy to the instant where the glottal area is maximally closed, while the glottal opening instants correspond to the instant where the glottal area is maximally open, according to the DEGG and GLA signals. Since there is full glottal closing in all cases, the closing instants fall within the closed interval, as it is shown from the GVG. The presence of mucus bridge is also evident in the GVG.

## Chapter 6

# Conclusions

High-speed videoendoscopy is the most promising method for direct investigation of the vocal-fold vibrations. We have presented a segmentation method that tracks the glottal area. The results on the database are satisfying and indicate the discriminative power of the proposed method. The glottal area is clearly distinguished from the background. The method has also been tested on digital kymographic images, but in a much smaller scale, and the findings were satisfactory as well. A lot of work still remains to be done, however, in order to properly present findings from the DKG processing.

The proposed procedure succeeds in its task with a sequence of steps which achieve very good results. By cutting the sequences in oscillating cycles with the use of landmark frames we compensate for glottal or endoscope movements, as well as for lighting changes. We also ensure temporal consistency for a better propagation. The glottal area is tracked with high accuracy with the use of bounding masks. In order to deal with different video qualities, we introduce the use of two initial masks for the segmentation of the landmark frames. This choice guarantees the best segmentation results on the landmark frames and the correct propagation to the rest of the sequence. The use of a localized region-based active contours model ensures the correct segmentation of the glottal area. The algorithm is versatile enough to split and merge, as required by the vocal-fold evolution and geometry.

The visualizations of the segmented glottal area, phonovibrogram and the newly proposed glottovibrogram seek to effectively represent the evolution of the glottal area over time. More specifically, with the glottovibrogram we have managed to clearly represent the shape of the glottal area. The speed profile gave a better insight on the propagation of velocity. The muscles involved move with different velocity and the movement of the edges could be an interesting indicator of the corresponding dynamics. The vocal folds do not move simultaneously along their length. A posterior-to-anterior opening is observed, although there have been reported cases where the movement begins from the anterior part of the vocal folds. A velocity burst

is present when the vocal-folds begin to open. Additionally to that observation, the vocal folds move faster during the closing phase rather than the opening phase. Another behaviour of the vocal folds has also been clarified. It was conceived that partial glottal closing was a characteristic of pathological cases. Our findings prove that for normal phonations there can be partial glottal closing during vibration. This behaviour seems to be related with the used laryngeal mechanisms. In addition to the visualizations, one can always refer to the segmented sequences to clarify any behaviours.

An interesting novelty presented in this work is the simultaneous representation of the 1D and 2D signals. Our findings seem to clarify the correspondences of the glottal instants to the phases of the glottal area. Data synchronization is a bit doubtful; the sampling frequency of the medical platform is a bit higher than the usual audio sampling frequency and audio and video are recorded with different sampling frequencies. There is a minimum time interval under which we can only assume correspondance. The glottal opening instants, as estimated from the DEGG signals, seem to correspond relatively well to the instants where the vocal folds begin to open. The glottal closing instants, seem to correspond better to the instants where the glottal area is minimum.

The vocal-fold vibration pattern could be exploited in advanced segmentation methods. Possible directions in glottis segmentation should include the use of active shape models and model-based methods. The use of three-dimensional segmentation techniques, as inspired from the processing of MRI sequences could also be an interesting approach. Other future works should include the use of inverse-filtering on the GLA signal. A comparison with the glottal air flow would also be interesting.

# Bibliography

- [Adams and Bischof, 1994] Adams, R., Bischof, L., 1994 *Seeded region growing* IEEE Transactions on pattern analysis and machine intelligence, vol. 16, no.6, 641-647.
- [Allin et al., 2004] Allin, S., Galeotti, J., Stetten, G., Dailey, S.H., 2004 *Enhanced snake based segmentation of vocal folds*, IEEE International Symposium on Biomedical Imaging: Nano to Macro, 812-815.
- [Baer et al., 1983] Baer, T., Lofqvist, A., McGarr, N. 1983 *Laryngeal vibrations: A comparison between high-speed filming and glottographic techniques*. The Journal of the Acoustical Society of America, vol. 73, no. 4, 1304-1308.
- [Bailly, 2009] Bailly, L., 2009 *Interaction entre cordes vocales et bandes ventriculaires en phonation: exploration in-vivo, modélisation physique, validation in-vitro*. PhD dissertation, Université du Maine.
- [Bezier, 1972] Bezier, P., 1972 *Numerical control; mathematics and applications*. John Wiley & Sons.
- [Caselles et al., 1997] Caselles, V., Kimmel, R., Sapiro, G., 1997 *Geodesic active contours*. International journal of computer vision, vol. 22, no. 1, 61-79
- [Chan and Vese, 2001] Chan, T.F., Vese, L.A., 2001 *Active contours without edges* IEEE Transactions on image processing, vol. 10, no. 2, 266-277
- [de Cheveigné and Kawahara, 2002] de Cheveigné, A., Kawahara, H., 2002 *YIN, a fundamental frequency estimator for speech and music* The Journal of the Acoustical Society of America, vol. 111, 1917-1930
- [Childers et al. 1990] Childers, D. G., Hicks, D. M., Moore, G. .P., Eskenazi, L., Lalwani, A. L., 1990 *Electroglottography and vocal fold physiology*. J. Speech Hear. Res., vol. 33, 245-254.
- [Cootes et al., 1995] Cootes, T.F., Taylor, C.J. Cooper, D.H., Graham, J., 1995 *Active shape models-their training and application* Computer vision and image understanding, vol. 61, 38-59.

- [Degottex et al., 2008] Degottex, G., Bianco, E., Rodet, X., 2008 *Usual to particular phonatory situations studied with high-speed videoendoscopy*. ICVPB.
- [Demeyer et al., 2009] Demeyer, J., Dubuisson, T., Gosselin, B., Remacle, M., 2009 *Glottis segmentation with a high-speed glottography: a fully automatic method 3<sup>rd</sup>* Advanced Voice Function Function Assessment International Workshop.
- [Doval et al., 2006] Doval, B., d'Alessandro, C., Henrich, N., 2006 *The spectrum of glottal flow models* Acta Acustica united with Acustica, vol. 92, no. 6, 1026-1046.
- [Farnsworth, 1940] Farnsworth, D., 1940 *High-speed motion pictures of the human vocal cords* Bell Laboratories Record, vol. 18, 203-208.
- [Forsyth and Ponce, 2002] Forsyth, D.A. and Ponce, J., 2002 *Computer vision: a modern approach* Prentice Hall Professional Technical Reference.
- [Gerratt et al., 1991] Gerratt, G., Hanson, D. G., Berke, G. S., Precoda, K., 1991 *Photoglottography: A clinical synopsis*. Journal of Voice, vol. 5, no. 2, 98-105.
- [Golla et al., 2009] Golla, M., Deliyski, D., Orlikoff, R., Moukalled, H., 2009 *Objective Comparison of the Electroglottogram to synchronous high-speed images of vocal-fold contact during vibration* Manfredi C (Ed.) Proceedings of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, 6:141-144.
- [Glasbey, 1993] Glasbey, C.A., 1993 *An analysis of histogram-based thresholding algorithms* CVGIP: Graphical Models and Image Processing, vol. 55, no. 6, 532-537.
- [Granqvist et al., 2003] Granqvist, S., Hertegard, S., Larsson, H., Sundberg, J., 2003 *Simultaneous analysis of vocal fold vibration and transglottal airflow; exploring a new experimental setup* Speech, Music and Hearing, Quarterly Progress and Status Report, RIT, Stockholm, vol. 45, 35-46.
- [Haralick and Shapiro, 1985] Haralick, R.M., Shapiro, L.G., 1985 *Image segmentation techniques* Computer vision, graphics, and image processing, vol. 29, no. 1, 100-132.
- [Henrich, 2001] Henrich, N., 2001 *Etude de la source glottique en voix parlée et chantée*. PhD dissertation, Université Pierre et Marie Curie - Paris 6.
- [Henrich et al., 2004] Henrich, N., d'Alessandro, C., Doval, B., Castellengo, M., 2004 *On the use of electroglottographic signals for characterization of nonpathological phonation* The Journal of the Acoustical Society of America, vol. 115, no. 3, 1321-1332.
- [Henrich et al., 2005] Henrich, N., Doval, B., Castellengo, M., 2005 *Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency* The Journal of the Acoustical Society of America, vol. 117, no.3, 1417-1430.



- [Henrich, 2006] Henrich, N., 2006 *Mirroring the voice from Garcia to the present day: Some insights into singing voice registers* Logopedics Phoniatrics Vocology, vol. 31, no. 1, 3-14.
- [Hertegard and Gauffin, 1995] Hertegard, S., Gauffin, J., 1995 *Glottal area and vibratory patterns studied with simultaneous stroboscopy, flow glottography and electroglottography*. J. Speech Hear. Res., vol. 38, 85-100.
- [Kass et al., 1988] Kass, M., Witkin, A., Terzopoulos, D., 1988 *Snakes: Active contour models* International journal of computer vision, vol. 1, no. 4, 321-331.
- [Kichenassamy et al., 1996] Kichenassamy, S., Kumar, A., Olver, P., Tannenbaum, A., Yezzi, A., 1996 *Conformal curvature flows: from phase transitions to active vision* Archive for Rational Mechanics and Analysis, vol. 134, no. 3, 275-301.
- [Kiritani et al., 1990] Kiritani, S., Imagawa, H., Hirose, H., 1990 *Vocal cord vibration and voice source characteristics - observations by a high-speed digital image recording* ICSLP, 61-64.
- [Kohler, 1981] Kohler, R., 1981 *A segmentation system based on thresholding*, Computer Graphics and Image Processing, vol. 15, no. 4, 319-338.
- [Lankton and Tannenbaum, 2008] Lankton, S., Tannenbaum, A., 2008 *Localizing region-based active contours* IEEE Transactions on Image Processing, vol. 17, no. 11, 1-11.
- [Larsson et al., 2000] Larsson, H., Hertegard, S., Lindestad, P. A., Hammarberg, B., 2000 *Vocal fold vibrations: High-speed imagin, kymography, and Acoustic Analysis: A Preliminary Report* The Laryngoscope, vol. 100, 2117-2122.
- [Lohscheller et al., 2007] Lohscheller, J., Toy, H., Rosanowski, F., Eysholdt, U., Döllinger, M., 2007 *Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos* Medical Image Analysis, vol. 11, no. 4, 400-413.
- [Lohscheller et al., 2008] Lohscheller, J., Eysholdt, U., Toy, H., Döllinger, M., 2008 *Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics* IEEE transactions on medical imaging vol. 27, no. 3, 300-309.
- [MacQueen, 1966] MacQueen, J.B., 1966 *Some methods for classification and analysis of multivariate observations* Western Management Science Inst. University of California.
- [Marendic et al., 2001] Marendic, B., Galatsanos, N., Bless, D., 2001 *New active contour algorithm for tracking vibrating vocal folds* International Conference on Image Processing, Proceedings.

- [Mehnert and Jackway, 1997] Mehnert, A., Jackway, P., 1997 *An improved seeded region growing algorithm* Pattern Recognition Letters, vol. 18, no. 10, 1065-1071.
- [Moukalled et al., 2009] Moukalled, H. J., Deliyski, D. D, Schwarz, R. R, Wang, S., 2009 *Segmentation of laryngeal high-speed videoendoscopy in temporal domain using paired active contours* Models and Analysis of Vocal Emissions for Biomedical Applications.
- [Negus, 2009] Negus, V.E., 2009 *The comparative anatomy and physiology of the larynx* The Laryngoscope, vol. 60, no. 5, 516.
- [Neubauer et al., 2001] Neubauer, J., Mergell, P., Eysholdt, U., Herzel, H., 2001 *Spatio-temporal analysis of irregular vocal fold oscillations: Biphonation due to desynchronization of spatial modes* The Journal of the Acoustical Society of America, vol. 110, 3179-3192.
- [Ní Chasaide and Gobl, 1997] Ní Chasaide, A., Gobl, C., 1997 *Voice source variation*. The handbook of phonetic sciences, vol. 5, 427-461.
- [Paragios and Deriche, 2002] Paragios, N., Deriche, R., 2002 *Geodesic active regions and level set methods for supervised texture segmentation* International Journal of Computer Vision, vol. 46, no. 3, 223-247.
- [Quatieri, 2001] Quatieri, T.F., 2001 *Discrete-Time Speech Signal Processing* Pearson education.
- [Rothenberg, 1981a] Rothenberg, M., 1981 *Some relations between glottal air flow and vocal fold contact area*. ASHA Rep. 11, 88 - 86.
- [Rothenberg, 1981b] Rothenberg, M., 1981 *Acoustic interaction between the glottal source and the vocal tract* Vocal fold physiology, 305-323.
- [Rothenberg, 1992] Rothenberg, M., 1992 *A multichannel electroglottograph* Journal of Voice, vol. 6, no. 1, 36-43.
- [Roubeau et al., 2009] Roubeau, B., Henrich, N., Castellengo, M., 2009 *Laryngeal vibratory mechanisms: The notion of vocal register revisited*. Journal of Voice, vol. 23, 425-438.
- [Schutte and Miller, 2001] Schutte, H. K., Miller, D. G., 2001 *Measurement of closed quotient in a female singing voice by electroglottography and videokymography* Proc. of the 5<sup>th</sup> Intern. Conf. on Adv. in Quant. Laryngoscopy, Voice and Speech Research, Groningen.
- [Sethian, 1999] Sethian, J.A., 1999 *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science* Cambridge University Press.
- [Stevens and Weismer, 2001] Stevens, K.N., Weismer, G., 2001 *Acoustic phonetics* The Journal of the Acoustical Society of America, vol. 109, 17.

- [Story, 2002] Story, B. H., 2002 *An overview of the physiology, physics and modeling of the sound source for vowels* Acoustical Science and Technology, vol. 23, no. 4, 195-206.
- [Struik, 1988] Struik, D.J., 1988 *Lectures on classical differential geometry*. Dover Publications.
- [Svec and Schutte, 1996] Svec, J. G., Schutte, H. K., 1996 *Videokymography: High-speed line scanning of vocal fold vibration*. Journal of voice, vol. 10, no. 2, 201-205.
- [Timcke et al., 1958] Timcke, R., von Leden, H., Moore, P., 1958 *Laryngeal Vibrations: Measurements of the Glottic Wave: Part I. The Normal Vibratory Cycle* Archives of Otolaryngology—Head & Neck Surgery, vol. 68, no. 1.
- [Titze, 1998] Titze, I.R., Martin, D.W., 1998 *Principles of voice production*. Acoustical Society of America Journal, vol. 104, 1148.
- [Titze, 1988] Titze, I.R., 1988 *The physics of small-amplitude oscillation of the vocal folds*. Acoustical Society of America Journal, vol. 83, no. 4, 1536-1552.
- [Trask, 1996] Trask, R.L., 1996 *A dictionary of phonetics and phonology*. Burns & Oates.
- [Westphal and Childers, 1983] Westphal, L., Childers, D., 1983 *Representation of glottal shape data for signal processing* IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 31, no. 3, 766-769
- [Yan et al., 2006] Yan, Y., Chen, X., Bless, D., 2006 *Automatic tracing of vocal-fold motion from high-speed digital images*. IEEE transactions on bio-medical engineering, vol. 53, no. 7, 1394.
- [Yezzi et al., 1999] Yezzi Jr, A., Tsai, A., Willsky, A., 1999 *A statistical approach to snakes for bimodal and trimodal imagery* Proceedings of the International Conference on Computer Vision.
- [Zuiderveld, 1994] Zuiderveld, K., 1994 *Contrast limited adaptive histogram equalization* Graphics gems IV, Academic Press Professional, Inc., 474-485.
- [<http://www.kymography.com/>] Videokymography, web page. <http://www.kymography.com/>.
- [<http://voiceresearch.free.fr/egg>] Electroglottography: Open-Source software for Analysing the Electroglottographic Signal, web page. <http://voiceresearch.free.fr/egg>