LARYNGEAL MECHANISMS AND VOCAL FOLDS FUNCTION IN ADDUCTOR LARYNGEAL DYSTONIA DURING CONNECTED SPEECH

By

Ahmed Yousef

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Communicative Sciences and Disorders – Doctor of Philosophy
Mechanical Engineering – Dual Major

2023

ABSTRACT

Adductor laryngeal dystonia (AdLD) is a neurological voice disorder that disrupts laryngeal muscle control during running speech. Diagnosis of AdLD is challenging because of the limited scientific consensus on accurate diagnostic criteria as it can mimic voice features of other voice disorders. The use of laryngeal high-speed videoendoscopy (HSV) as a powerful tool to capture the detailed vocal fold (VF) vibrations has been almost nonexistent to study AdLD and limited to sustained phonation, not connected speech in which AdLD's symptoms manifest. The present dissertation aims to address the previous literature gap using HSV and provide, for the first time, quantitative analysis for the impaired vocal function in AdLD during connected speech. To accomplish this, HSV recordings were collected from vocally normal adults and AdLD patients during connected speech. Five different studies were implemented in order to analyze and extract clinically relevant information from these recordings.

The first study investigated the differences between AdLD and normal controls based on evaluating running speech durations in HSV over which VFs were visually obstructed by excessive movements of laryngeal tissues. To facilitate these analyses, a deep learning tool was developed to automatically classify HSV frames in terms of detecting visual obstructions in the VF images. The second study provided a new image segmentation tool for detecting VF edges during running speech in HSV. This tool was developed using a unique combination of the active contour modeling method and a machine-learning based method (k-means clustering) to segment VF edges in HSV kymograms. The third study developed a quantitative representation of VF dynamics in AdLD in running speech using HSV. A deep learning technique was used based on the tool developed in study two to segment the glottal area/edges and extract the glottal area waveform from the HSV recordings for analysis. The fourth study analyzed the pathological vocal function of AdLD during phonation onset and offset in connected speech using HSV. An automated approach was developed and validated with manual analysis to measure and compare the glottal attack and offset times between AdLD group and normal controls. Study five presented a one-mass lumped model that can estimate glottal area waveform and biomechanical characteristics of VFs based on HSV data.

The results of study one showed the accurate detection of the visual obstructions of the VF frames – facilitating the study of laryngeal activities in AdLD. The findings revealed that AdLD group exhibited longer durations of obstructions – making this measure a potential candidate for

AdLD assessment. Also, indicating parts of connected speech that provide an unobstructed view of VFs allows for developing optimal passages for precise HSV examination and disorder-specific clinical voice assessment protocols. Study two and three demonstrated promising performance of the proposed automated tools to detect VF edges and analyze glottal area waveforms. These accurate techniques overcame the challenges involved in HSV analysis including the poor image quality during running speech and the excessive laryngeal maneuvers of AdLD. Future research should benefit from these newly developed automated tools for HSV analysis of VF vibrations in running speech to explore diagnostically relevant information in both vocally normal adults and AdLD. The findings of the fourth study revealed the accurate measurements of the glottal attack and offset times using the developed automated technique. The measurements showed significant longer attack time in AdLD and more variability of the attack and offset times in AdLD due to the irregularity of the VF vibratory behavior in this disorder. The results of this study also demonstrated an agreement with the previous findings in literature. Accordingly, glottal attack time might be a compelling measurement of the severity of AdLD, which can be further investigated in future using the developed tool with larger sample size and, even for different voice disorders. Obtaining such measures in running speech opens up new lines of research to explore the clinical significance of these measurements and address the diagnostic challenges in AdLD. In the last study on modeling, the results show the successful optimization of the developed one-mass model to closely capture the characteristics of VF vibrations observed in the HSV running speech sample. The study uncovered the potential of this simplified model to estimate biomechanical properties of VFs with minimal computational cost non-invasively – paving the path for future research to utilize this model for analyzing connected speech samples and study the impaired VF dynamics in AdLD.

This dissertation is dedicated to my beloved wife, Noura, my unwavering source of happiness, motivation, and strength. Her limitless support and patience during the pandemic's challenges that kept us apart for years is exceptional. Without her by my side, completing this PhD would not have been possible.

I would like to dedicate this dissertation to my Mum, Laila, my Dad, Dr. Mokhtar Yousef, my brother, Mohamed, my sister, Mai, and all my family members for their unconditional support, care, and belief in me.

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to the individuals who have made unique contributions to my academic growth. Above all, I am profoundly indebted to my PhD advisor, Dr. Maryam Naghibolhosseini, who offered unwavering support and invaluable guidance throughout my PhD training and made my shift from engineering to science smooth. I truly appreciate the countless number of hours she invested in offering enlightening research ideas, reviewing my work, providing thought-provoking feedback, and encouraging me to excel. Without her patience, dedication, and exceptional mentorship this research would not have been possible. I would also like to express my sincere gratitude to Dr. Mohsen Zayernouri for his significant contribution to this dissertation. His in-depth knowledge and extensive expertise in Mechanical Engineering have considerably improved the quality of this research. I am genuinely thankful for the opportunity to work alongside such a knowledgeable mentor. I would like to extend my utmost appreciation to Dr. Dimitar Deliyski, for his consistent support throughout my PhD journey and his commitment to provide an excellent research environment for my professional development. Furthermore, I am wholeheartedly grateful to my committee members, Dr. Eric Hunter and Dr. Jeff Searl, for their valuable guidance and persistent help throughout the development of my PhD.

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

AdLD	Adductor Laryngeal Dystonia
AbLD	Abductor Laryngeal Dystonia
LD	Laryngeal Dystonia
MTD	Muscle Tension Dysphonia
ET	Essential Tremor
CAPE-V	Consensus Auditory Perceptual Evaluation of Voice
HSV	High-Speed Videoendoscopy
EGG	Electroglottography
VF	Vocal Fold
GAT	Glottal Attack Time
GOT	Glottal Offset Time
ACM	Active Contour Modeling
DNN	Deep Neural Networks
Q	Research Question
H	Hypothesis
GAW	Glottal Area Waveform
SLP	Speech-Language Pathologists
fps	Frames Per Second
CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
c_k	K-Means Cluster Centroid
D	K-Means Euclidean Distance
I	Image Intensity
E	Active Contour Energy Function
E_{image}	Active Contour Internal Energy Function
E_{int}	Active Contour External Image Function

γ	Active Contour Elasticity Weight
β	Active Contour Rigidity Weight
∇I	Image Gradient
K_w	Kymogram Image Width
K_h	Kymogram Image Height
IoU	Intersection Over Union
DC	Dice Coefficient
F_1	Boundary-F_1 Score
ML	Machine Learning
m	Vocal Fold Mass
k	Vocal Fold Elasticity
c	Vocal Fold Damping Coefficient
AUC	Area Under The Curve
P_s	Subglottal Pressure
P_1	Inlet Glottis Pressure
P_2	Outlet Glottis Pressure
Q_g	Glottal Air Flowrate
d	Vocal Fold Thickness
l	Vocal Fold Length
w	Vocal Fold Width
$\mathcal{F}$	Net Vocal Fold Force
F	External Vocal Fold Force
P_B	Bernoulli Pressure
ρ	Air Density
μ	Coefficient of Air Viscosity
A_g	Glottal Area
A_{g0}	Initial Glottal Area
X_c	Critical Vocal Fold Displacement
$\overline{P}_s$	Typical Subglottal Pressure
P_{Smax}	Maximum Built-Up Pressure
t_c	Vocal Fold Closure Time

c'	Vocal Fold Closure Damping Coefficient
Δt	Time Step
K_1	Initial Slope Estimate
K_2	Second Slope Estimate
K_3	Third Slope Estimate
K_4	Fourth Slope Estimate
$x_\circ$	Initial Displacement
$V_\circ$	Initial Velocity
α	Scaling Factor
Obj	Objective Function
A_{Model}	Simulated Glottal Area
A_{HSV}	Experimental Glottal Area
q	Optimizing Parameters Vector
PSO	Particle Swarm Optimization
N	Number of Swarm Particles
J	Total Iteration Number
*	Optimum Value
v_i	Swarm Particle’s Velocity
q_i	Swarm Particle’s Position
p_b	Best Swarm Particle Position
g_b	Best Swarm Global Position
W	Swarm Particle Inertia Weight
Z_1	Swarm Particle Cognitive Parameter
Z_2	Swarm Particle Social Parameter

CHAPTER 1: INTRODUCTION

1.1. Voice Production and Assessment

Divulging the mastery behind speech production has been a desire for scientists. This desire emerged about a century ago [1] when scientists aimed to understand the governing physics of phonation and voice production. Human voice production process works through an energy conversion; the aerodynamic energy generated by the lungs is converted into the acoustic energy and sound in the vocal tract; this conversion happens when the vocal folds (VFs) vibrate and, appropriately modulate the glottal airstream [2, 3]. Different theories were proposed to better understand the voice production mechanisms and interpret the complex interaction between the aerodynamics of glottal airflow, vibration of VFs, and the acoustic output of the vocal tract [4, 5, 6, 2, 7]. One of the well-established theories is the Aerodynamic-Myoelastic Theory which offers a foundation for understanding human voice production. It states that the vibratory motion of the VFs during phonation are produced by a combination of both the aerodynamic forces of the airflow and the VF tissue dynamics [2, 3]. Other theories were recently developed like the nonlinear source-filter theory proposed by Titze [8]. The subglottal system below the larynx was defined as a sound source (source of energy), which helps in sustaining the VF vibration in phonation [9]. The vocal tract was considered as a filter that convolved with the source to generate the sound [10, 11, 12, 13].

Understanding these underlying mechanisms of voice production and particularly, the vibratory behavior of VF as a vital component in the larynx helps in providing better healthcare, medical diagnosis and treatments for individuals who suffer from voice problems and degraded voice quality. This cannot only enhance individuals’ quality of life and social well-being but also individuals’ work productivity and health care cost. Therefore, several tools and methods were developed to obtain a better assessment of the VF and the overall voice quality. One assessment approach is through analyzing the output aerodynamic signal (glottal airflow) from the phonatory system. Several measures can be obtained from the change in the glottal airflow due to the vibratory abduction–adduction movement of the VF [14]. Open quotient is an example of the aerodynamic measures and is defined as the portion of the vibratory cycle with an open glottis. A large open quotient value relates to a breathier voice quality whereas a small value associates with a more pressed quality [15]. Analyzing the output acoustic signal is another method for voice assessment. This can be done through either objective acoustic measurements or subjective perceptual assessments. Acoustic measures are generated using signal processing methods and can provide a quantitative tool for assessment of voice quality. These objective measures can be divided into three different categories: (1) perturbation measures such as jitter and shimmer [16]; (2) noise measures such as signal to noise ratio and harmonic to noise ratio [17]; (3) spectral/cepstral measures such as cepstral peak prominence and Mel-frequency cepstral coefficients [18]. The second way of analyzing the acoustic signals is through the auditory perceptual evaluation, which is the most commonly used approach in voice clinics and mainly depends on the level of expertise of the evaluator. Evaluation tools have been developed as standardized scales in order to reduce the possible variability and inconsistencies in the perceptual evaluation of voice disorders. One of the most recent effective standard scales is the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) [19]. CAPE-V enables the analysis and assessment of different voice features, namely, severity, roughness, breathiness, strain, pitch, and loudness. The CAPE-V rating is done using a visual analog scale on a 100-mm line and has standard vocal tasks to assess the voice quality. Additionally, the CAPE-V form includes an ordinary scale of moderate, mild, and severe to make a perceptual judgment of the voice. It was found that this rating procedure is consistent and reliable among the raters [20].

Another assessment approach is performed using imaging techniques to directly visualize the activities and the different configurations of the larynx and, particularly, VFs for a reliable assessment of the voice. The most common modalities for laryngeal imaging are electroglottography (EGG), videostroboscopy, and high-speed videoendoscopy (HSV). EGG is a voice assessment technique that is used to analyze the contact of the VFs with large sampling rates [21]. The EGG principle depends on the variation in the electrical conductivity between tissue and air. That is, two or more electrodes are placed on both sides of the larynx and a high frequency and low voltage electric current is fed between the electrodes. During the VFs vibration, the contact area between the VFs changes, hence, the electrical impedance between the electrodes varies. This variation in the impedance is reflected in the EGG output [21]. Several characteristic points of VF vibration can be obtained from EGG such as at the beginning of opening the upper margin of VF, at the complete closure of the lower margin, and at full VF contact. Different measures can be generated based on these extracted characteristic points, such as the contact quotient (ratio between the contact phase and total time of the vibratory cycle), open quotient (ratio between the open phase and total time), and speed quotient (ratio between opening and closing time) [22]. As an example of how these measures can be related to the assessment of voice quality, a breathy voice quality is associated with smaller contact quotient [22].

The current laryngeal imaging technique that is widely used in clinics for voice assessment is videostroboscopy [23, 24, 25]. Using an endoscope coupled with a stroboscopic light in videostroboscopy, video recordings of the laryngeal structures can be obtained which allows to visually assess laryngeal tissue health and VF vibrations [23, 24, 25]. Videostroboscopy can only capture stationary phonation events during periodic VF vibrations. Although videostroboscopy is used during connected speech, where most of voice disorders reveal themselves, it can only capture gross laryngeal adjustments. That is, the functional assessment of VFs vibration using videostroboscopy is limited to sustained vocalizations only [26, 27, 28, 29]. Due to the low sampling rate of the camera (resulting in a low temporal resolution), videostroboscopy is incapable of capturing the cycle-to-cycle and intra-cycle details of VFs vibration, which is critical when those vibrations are aperiodic – a common occurrence in voice disorders [30, 31]. The recent advancement of coupling flexible fiberoptic endoscopes with laryngeal high-speed videos serves to overcome the previous limitations of stroboscopy by offering high recording frame rates (thousands of frames per second) and capturing the true VF vibrations [30, 31, 32, 33]. Using HSV allows the visualization and analysis of the detailed pathological phonatory events in voice disorders during running speech [34, 35, 36, 37, 38, 39, 40] such as the true VF oscillations (cycle to cycle) [41, 42, 43, 44], phonation onsets and offsets [45, 46, 47, 48], voice breaks [36], and singing [33]. HSV will be revisited and discussed in detail later in this chapter.

Researchers have been using these different assessment modalities and approaches to study and analyze the different voice disorders. Among these modalities, imaging techniques (particularly the advanced HSV) can provide accurate information regarding the underlaying mechanisms of voice production, vocal function, and their dynamics. The high capabilities of imaging techniques allows researchers to study laryngeal dynamics and VF function in dysphonic voices [49, 50, 51, 32]. However, there is a huge gap in literature in terms of studying voice disorders using HSV, especially in neurological voice disorders whose symptoms mostly appear during connected speech. One of these neurological voice disorders that has not been well documented in literature is laryngeal dystonia (LD) which will be discussed in detail in the following subsection.

1.2. Adductor Laryngeal Dystonia (AdLD)

LD is a neurogenic, chronic voice disorder that causes the intrinsic laryngeal muscles to contract, or spasm, involuntarily during phonation [52]. LD affects an estimated 1 per 100,000 people (with a prevalence of 35,000-50,000 cases in the United States) [53]. There is a female predominance (79% of the patients are women) [53]. The average age on the onset of LD ranges from 40 to 50 years old [54]. The patients typically report a sudden onset of symptoms, which gradually progress until become severe within few months or few years [55]. As a chronic voice disorder, LD affects the daily communication of the patients and leads to social isolation and occupational disability [56]. The LD etiology remains elusive. However, recent studies demonstrated some association between the LD development and genetic, environmental, and familial factors [55, 57]. Although most LD cases have focal laryngeal dystonia [54], LD symptoms may appear in patients with other neuromuscular disorders; for example, around 25% of patients with essential tremor suffer from LD [58]. Other scholars hypothesized that the pathophysiology of LD may arise from an increase in brain plasticity, sensory abnormalities, and a reduced inhabitation of intracortical processes [59].

LD is characterized as a task-specific dystonia where it only occurs during connected speech and its severity relies on the demands of the vocal task [57]. It was reported that LD signs are more likely to appear during connected speech than sustained/prolonged vowels [60]. This is due to the increased motor complexity of running speech compared to sustained phonation, which provokes more sever laryngeal spasms and higher strain. The complexity stems from the rapid transitions during running speech which requires switching between voiced and nonvoiced sounds whereas no such transitions exist in sustained vowels [60]. LD is typically divided into three subtypes: adductor LD (AdLD), abductor LD (AbLD), and mixed. Patients with AdLD experience spasmodic overclosure of the VFs during phonation, particularly when the VFs are approximating, leading to excessive phonatory breaks and a strained voice quality with cessation of airflow [55]. In contrast, patients with AbLD exhibit excessive involuntary opening of the VFs during phonation – leading to a transient breathy voice quality with excessive escape of airflow [52, 61]. Additionally, some clinicians recognize patients who suffer from both conditions, mixed LD [61]. Since AdLD is the most common form of LD with 80% of all LD Patients [55], it will be investigated in this dissertation.

Diagnosis of AdLD is challenging because AdLD can coexist with other neuromuscular disorders that have similar voice symptoms [62]. Although current diagnosis of AdLD mostly relies on auditory–perceptual features [63], other functional voice disorders such as muscle tension dysphonia (MTD) can mimic the voice characteristics of AdLD – resulting in diagnostic confusion [64]. MTD patients can have hypercontraction in the laryngeal muscles and, hence, a strained voice. Further, MTD cases may normally cough, cry, and sing similar to AdLD [62]. There are no diagnostic criteria in the current clinical practice to differentiate between AdLD and MTD even for experienced clinicians. Given that the treatments of MTD and LD is completely different, misdiagnosis can lead to inappropriate/needless surgical or medical interventions [65]. Hence, researchers tried to differentiate between the two disorders. In this regard, they found that AdLD is a "task dependent" dystonia (Less severe during sustained vowels than running speech), yet MTD is not (equally severe regardless of the vocal task) [66]. Essential tremor (ET) might also be mistaken for AdLD where above 25% of ET suffer from laryngeal tremor. This is because laryngeal muscular tremors can mimic glottic stops as in AdLD. But misdiagnosis can be avoided knowing that the tremor with ET is present in sustained phonation while it is not present with AdLD [62].

Similar to the challenges in diagnosing AdLD, ineffective treatments of AdLD can also occur. That is, the difficulty in the differential diagnosis of AdLD may lead to needless treatments and surgical interventions because, for example, if LD is misdiagnosed with MTD (the treatments of MTD and LD are completely different) [67, 68]. The main treatment for AdLD is botulinum toxin injection into the affected muscle(s): thyroarytenoid, interarytenoid, and lateral cricoarytenoid [62]. The injection is effective and provides temporary relief; usually, it is done every 3-4 months. Studies showed that this treatment enhances the acoustic/aerodynamic measurements and the voice quality of AdLD patients [69]. However, side effects may exist after the injection such as incomplete glottic closure in AdLD. Another treatment option for AdLD is provided through a surgery which includes denervation and reinnervation of the recurrent laryngeal nerve, recurrent laryngeal nerve sectioning, type II thyroplasty, and thyroarytenoid muscle neuromyectomy [70]. Voice therapy is also considered as a complementary treatment, provided by speech-language pathologists that can help mitigating AdLD symptoms. For example, some studies show that when voice therapy is provided after the Botox injection, it gives patients a longer time with alleviated voice symptoms before needing to repeat the injection [71].

1.3. High-speed Videoendoscopy (HSV)

The voice production in AdLD has been studied using different assessment tools that we discussed earlier in this document such as the acoustic analysis [72, 73], fiberoptic laryngoscopy [74] and aerodynamic measurements [64]; yet the pathophysiology and differential diagnosis of AdLD is still not fully understood. Despite the use of these different assessment tools, the use of HSV has not been well investigated in literature. Laryngeal imaging tools can be used to observe and diagnose the impaired voice production function in AdLD during connected speech. This is because the AdLD symptoms are mostly reveal themselves during running speech [75, 76, 77, 78]. HSV is a powerful tool that can offer high frame rates and temporal resolution [30, 31, 32, 33]. The main advantage of HSV resides in the ability to visualize both periodic and aperiodic VF movements that otherwise would not be feasible with videostroboscopy [49, 41]. Such capability makes HSV viable for examining the variations between and within vibratory cycles of VFs which are associated with their aperiodic motions in AdLD during running speech [30].

This aforementioned capability provides the opportunity to develop new tools to objectively analyze the entire vibratory cycles during phonation in AdLD. Hence, the potential of using HSV has been investigated in previous studies as a promising tool to understand the underlying voice production mechanisms in dysphonic voices [49, 79, 32]. The clinical assessment of VFs vibration using videoendoscopic images is commonly performed subjectively with visual inspection of the data. However, employing efficient quantitative methods for voice analysis using HSV would be valuable for clinical voice examination. Hence, extracting useful, quantitative measurements of the dynamic motion of the VFs in HSV recordings could allow the clinicians to obtain clinically relevant characteristics of the VF oscillations during connected speech.

The measurements and features of the VF vibrations during either sustained phonation or running speech can be extracted by visually analyzing HSV recordings. During the sustained phonation (e.g., the production of the vowel /i/) and steady-state VF oscillations, features such as periodicity, VFs symmetry, glottal closure, and information about the mucosal wave and its aggregation can be obtained from HSV [32]. In addition, HSV is a unique tool to study and analyze aperiodic speech and asymmetric vibrations of the VFs, which is common in voice disorders that cause perturbed periodicity in VF vibrations. This aperiodicity cannot be analyzed using videolaryngoscopy. Hence, HSV is a powerful approach to tackle this problem as it can be used to assess the most transient VF vibratory behaviors regardless of the periodicity of the vibrations

[50]. So, visual information about the phonatory breaks, laryngeal spasms, onset and offset of phonation, and any other laryngeal movements that involve rapid maneuvers can be obtained from HSV data for clinical examinations in future.

However, using HSV in voice clinics remains a daunting task for clinicians since the data obtained through HSV are difficult to analyze with visual inspection; a short HSV recording can yield thousands of frames needing to be assessed, which is a time-consuming process [80, 81]. The semiautomated and fully automated objective analysis and measurements of HSV can overcome this challenge. Some of these measures include the closed/open quotient, left-right phase and amplitude asymmetry, axis shifts during closure, period/glottal width irregularity, glottal attack time (GAT) and glottal offset time (GOT) [32, 82]. Among these objective measures, GAT and GOT will be discussed as it will be included in the present dissertation analysis. GAT is closely related to the onset of VF vibration and the sound generation while GOT is associated with the offset of VF vibration and the end of a phonation. Both measures are critical factors to study the pathophysiology of voice disorders. Previously, these measures were manually extracted through the visual analysis of HSV data from vocally normal individual and patients with neurogenic voice disorders (LD and unilateral VF paralysis) during connected speech [83]. This work included a small sample size but emphasized the importance of measuring GAT and GOT and how they are critical for the voice characterization of neurogenic voice disorders. However, these two measures were extracted manually by visual raters, which was a time-consuming process resulting in analyzing a limited number of participants. This emphasizes the importance of developing these measures using automated techniques which will be done in this dissertation work.

1.4. Biomechanical Characteristics of Vocal Folds in AdLD

In the previous section, we discussed about the high capability of an advanced imaging technique as in the HSV technology and how it can be used to obtain a variety of useful clinical measures through analyzing the video recordings. In this section, a different set of indirect measures that can also be obtained from HSV will be discussed. These indirect measures cannot be directly obtained from visually analyzing the HSV frames and videos because they need a model to be coupled with the HSV analysis. These indirect measures are closely associated with the biomechanical properties and behavior of VFs movement, and they can be obtained by designing biomechanical models based on HSV. Examples of these indirect features include VF masses, the stiffness and elasticity properties of the VF tissues as well as information about the subglottal pressure.

The main advantage that these models provide is that they can provide measures that cannot be directly extracted from HSV or any other traditional recording approaches like EGG [84, 85, 86, 87]. These indirect biomechanical parameters are essential to understand the underlying physics of phonation [88, 89, 90, 91]. These parameters could be estimated using biomechanical models through an inverse problem to infer VF tissue properties in a non-invasive way. The inverse problems to predict the parameters of VFs vibration during phonation have been first introduced around 20 years ago [92]. In the inverse problems, the model parameters are optimized so that the model can generate a behavior similar to the experimental observations obtained from, e.g., HSV video data. If the optimization succeeds, the model can infer the biomechanical parameters of VFs [93].

The lumped element models are commonly used in this inverse analysis [94] to obtain the biomechanical characteristics of VFs because they are simple and can simulate different VF vibration behaviors with a few parameters. These models are designed such that the VF tissues are described by a small number of discrete, rigid masses. The neighboring masses are coupled by springs and interact with the external aerodynamic loading and/or acoustical loading [95]. The model parameters are specified based on each model component: masses (VF tissue), springs (to impose the elasticity effect), and dampers (to damp the motion). These models can be designed in different configurations: one- [96, 97, 98, 99], two- [94], three- [100], and multi-mass models [96, 97, 98]. The main advantage of the one-mass model is that it has a simple structure and minimal number of control parameters. Despite its simplicity, the model can still capture the characteristic vibratory features of the self-sustained oscillation of the VFs. These features make this simple model a compelling choice to reduce the computational coset, particularly in the real-time tasks [96, 97, 98, 99]. The two-mass model is the widely used model that can mimic the underlying mechanism of VF dynamics such as phase shifts of the upper and lower VF edges. This model was frequently used in the inverse analysis during the sustained phonatory speech as it captures self-sustained vibrations, asymmetric VF vibrations, and nonlinear VF dynamics [101, 102].

Different modeling works have used the HSV as an experimental data to build their models through the inverse analysis technique in humans to predict different measures. One of these models was done by Döllinger et al. [92] using the two-mass model. They used Nelder–Mead algorithm for optimizing the model with HSV of two human subjects. From the HSV, the medial VF edges were extracted with time using image processing techniques. The extracted trajectory of the VF edges (which refers to the change in the spatial location of VF edges across the recording time) was used to optimize and estimate the model parameters that yield a close trajectory against the experimental data. They predicted VF masses/stiffnesses and the subglottal pressure for each subject. Several studies tackled the inverse problem to quantify the asymmetry in both VF properties and VF vibration patterns [93, 103, 104, 105, 84, 106, 92]. These studies used HSV data from vocally normal participants [92] and patients with unilateral VF paralysis [104] and functional dysphonia [92], as well as from animals [93, 103]. Noting that the VF edges were extracted from the previous HSV data at the medial cross-section during vibration [84]. Only one study [106] used both HSV and EGG; they utilized HSV to extract VF edges whereas EGG to obtain three characteristic points of VF vibration: at the beginning of opening of the upper margins of the VFs, at the complete closure of the lower margins, and at the full VF contact. The prior studies used the same two-mass model with slight changes like using vertical coupling between masses [104], utilizing variable spring stiffnesses [93, 103], or including the collision force between the VFs [93, 103]. The common optimized parameters in these studies were the masses and their displacements, stiffness of the springs, stiffness during collision, damping coefficients, glottal length, thickness of masses, areas between masses, and the subglottal pressure [93, 103, 104, 105, 84, 106, 92]. Particle swarm technique and genetic algorithm were mostly utilized for optimization [93, 103].

The biomechanical characteristics for the pathological phonation of different disorders including LD can be modeled using the lumped-element models. For example, by including the anterior-posterior variations in a 3D model, incomplete glottal closure can be simulated [107, 108]. The VF vibratory characteristics of LD associated with different voice qualities, ranging from pressed as in AdLD to a breathy voice as in AbLD, can also be modeled as well [109]. The unilateral laryngeal nerve paralysis represents a phase lag between the healthy and damaged VFs leading to asymmetries in VF tissues. This disorder can be simulated by decreasing the mass and increasing all the stiffnesses of the damaged VFs [95, 110, 111]. Polyps and nodules, which refer to geometric abnormalities can be mimicked too [112] by adding an extra mass to the affected VF, altering the stiffness/damping coefficients, modifying the collision force, and altering the subglottal pressure loading [113, 114]. The Parkinson’s disease, which causes breathiness, vocal tremor, and an incomplete glottal closure [115, 116], is another disorder that lumped models can emulate [117, 118] through using time-varying model parameters and increasing the springs stiffness [95]. Despite the significant potential of the inverse analysis, attempts are almost nonexistent to extract the biomechanical characteristics of the impaired VF vibrations during running speech. That is, a major knowledge gap exists in terms of analyzing connected speech based on the biomechanical characteristics of VFs using biomechanical models. Prior models were not able to extract indirect biomechanical parameters during running speech; they were just focused on sustained phonation. This is due to the lack of the experimental data collected in connected speech, particularly the imaging data (HSV), that are necessary to design/validate those models and generate the various biomechanical characteristics of VF vibrations. Another reason for the lack of previous models is due to the complexity of simulating running speech since the phonatory events are transitory and convoluted. This high complexity in running speech requires building biomechanical models to simulate the phonation with myriad parameters that need to be optimized which is generally not favorable as it will considerably increase the computational cost, particularly in case of using a complicated model for the inverse analysis problem. This difficulty resulted in a lack of knowledge in terms of studying connected speech and AdLD symptoms since these symptoms only elicited during connected speech which, as mentioned, are very complicated to be simulated.

Therefore, as we discussed previously, it is crucial to develop techniques to automatically analyze VF vibrations in order to be able to generate the analysis we mentioned earlier. Obtaining these measures automatically from the HSV video recordings requires the automated segmentation of VF edges and the glottal area. These automated segmentations will determine the location and shape of the vibrating VFs during speech. The objective representation of the edges of the VFs will allow the development of the different HSV-based measures. In the next section, the different automated methods and approaches that were implemented in literature to spatially capture the VF vibrations will be discussed.

1.5. Automated HSV Analysis

Considering the large number of images generated from HSV recordings, visual analysis is not a practical solution. This emphasizes the need for automated techniques that can analyze and process the HSV videos in order to obtain useful measurements and information about the VFs function during speech. Segmentation is a fundamental step to analyze HSV video sequences and a building block needed to extract such measurements from the video data. Segmentation can be classified into three different types: temporal segmentation, spatial segmentation, and spatiotemporal segmentation. Temporal segmentation is a process of dividing a video sequence into well-defined time segments (short sequences) in order to extract useful temporal information from the video. An example of this type of segmentation would be the instances in a HSV sequence during which VFs are visually unobstructed or would be the time segments during which a phonation or VFs vibration occurs. Another type of video analysis occurs spatially which is called the spatial segmentation. This technique identifies the region of interest (e.g., a moving object like VFs) in the frames across the video sequence and provides its spatial location and structure for further analysis. For example, this technique can be used to identify the edges of the vibrating VFs and highlight the spatial location of the glottal area in the different HSV frames. This would facilitate the downstream analysis of VF vibrations and the development of the HSV-based measures. The third type of segmentation is spatiotemporal combines the previous two segmentation techniques such that the segmentation is performed on image sequences or video data in both the time and space domains.

In literature, there are two main methods to apply segmentation on HSV images: by either the image processing algorithms or the machine/deep learning techniques. The first method includes the classical image processing techniques for extracting useful spatial and temporal information/features from images. In terms of temporal segmentation, a previous study on using HSV in connected speech was proposed, which developed an automated temporal segmentation method using a statistical-based image processing algorithm. This algorithm was able to extract the timestamps for the onsets and offsets of vocalizations and epiglottic obstructions of the VFs [119]. For spatial segmentation, several studies developed and applied the traditional image processing techniques for spatial segmentation of the VF edges in HSV during sustained phonation [120, 121, 41, 122, 123]. The main approaches for extracting the VF edges from the HSV data are the region growing [121, 124, 125], histogram thresholding [41, 126], level sets [127], and active contour modeling (ACM) approaches [122, 123]. These image processing methods were used for HSV analysis in isolated sustained vowels. Most of the developed methods are not fully automated and require visual inspection of the data and some manual analysis [122, 124, 127, 128, 129]. Region growing and histogram thresholding are both vulnerable to image noise and homogeneity [120]. The level set method can accurately estimate the glottal cycle only when the VFs are closed, and is also prone to noise [127]. The ACM approach, however, is less sensitive to noise and intensity inhomogeneity in images, can be initialized anywhere, even across boundaries, and efficiently preserves global line shapes [130, 131]. Hence, this approach is an alternative promising technique for spatial segmentation of the glottal boundaries [132]. The ACM method can be used to dynamically locate the contour of the desired image features, such as the edges of the glottis. The active contour (aka snake) is a spline, which deforms based on certain energy minimization rules to capture the glottal edges. The ACM approach has previously been employed to detect the glottal edges i) spatially in each HSV frame: this method is based on using closed loop snakes for each individual HSV frame [121, 122]; ii) temporally in HSV kymograms: this method uses two open curve snakes for detection of the right and left VFs [123, 133]; and iii) spatio-temporally: the open curve snakes are used for glottal edge detection in the HSV kymograms across different cross sections of the VFs and the extracted edges are then registered back to each HSV frame [123]. The existing studies for spatial segmentation of glottal edges were performed on HSV data obtained during sustained vowels and not in connected speech.

The second method that was used in literature for both spatial and temporal segmentation in HSV data is the machine and deep learning technique. Using this advanced method, we can efficiently classify and cluster similar structures and/or discover hidden patterns in a sequence of image data with a low computational cost. In literature, deep learning was used as a tool for temporal segmentation and, particularly, laryngoscopic image classification. Deep learning is a subcategory of machine learning and utilizes deep neural networks (DNN) that can learn features from known/labeled image data (training data) and make predictions on new image data, based on the learned features [134]. Deep learning has shown promising performance in a variety of diagnostic tasks from medical images [135]. There is an immense potential for using deep learning techniques to analyze laryngeal images. Accordingly, several recent studies have applied deep learning on laryngoscopic videos to automatically select frames that display sufficient diagnostic information allowing clinicians to find abnormalities in a timely manner, yet these models were vulnerable to overfitting due to the limited size of the training dataset (i.e., only a few hundred images) [136, 137, 138]. Others used deep learning as a classifier to recognize laryngeal pathology (such as polyps, leukoplakia, vocal nodules, and cancer) based on thousands of laryngoscopic images as a larger training dataset [139, 140, 141]. However, none of the previous studies were conducted using HSV in connected speech for frame selection/classification, which is very important in studying voice disorders as they mostly reveal themselves in running speech. Additionally, in terms of the type of voice disorder, this literature did not investigate AdLD. HSV in connected speech imposes lower image quality than in sustained phonation and exhibits excessive laryngeal maneuvers across frames, which urges the need to develop even more efficient methods that can deal with these challenging conditions [81]. In addition to the temporal segmentation and image classification tasks, deep learning (specifically the DNN) has also been utilized for the spatial segmentation task in order to capture VF edges and the glottal area in HSV data. Five recent deep learning techniques based on DNN have successfully segmented the glottis and VF edges in HSV recordings with satisfying accuracy [142, 143, 144, 145, 146, 147]. The HSV datasets in these studies, however, were recorded during production of sustained phonation using a rigid HSV with high image quality. Also, these approaches used DNN for the spatial segmentation task and, hence, required manual labelling/annotation of the glottal edges/area in HSV frames to train the neural networks. Keeping in view that these previous studies for spatial segmentation used only sustained phonation as their data set, expanding this to connected speech using a flexible HSV system is an important next step.

1.6. Research Gaps, Questions and Hypotheses

The current diagnosis of AdLD is predominantly based on only auditory–perceptual assessment which causes diagnostic confusion as AdLD symptoms can mimic other disorders such as MTD [67, 68]. This is because the auditory-perceptual evaluation does not provide enough information that would help in differentiating between the two disorders – leading to needless surgical interventions or treatments [148, 149]. Therefore, there is a need for other effective assessment tools like the imaging techniques which can provide more information about the impaired laryngeal mechanisms and vocal function in AdLD and, eventually, enhance its differential diagnosis [150]. Since the previous studies found that the symptoms of AdLD typically appear in connected speech, not in a sustained phonation context [66, 151], the most appropriate powerful imaging technique to study AdLD is the HSV tool. HSV allows to visualize and analyze the detailed VF vibrations as well as the various phonation events in AdLD during connected speech [119, 35, 37]. This, in turn, could lead to a more accurate diagnosis of AdLD. However, most of the previous studies neither used HSV to study connected speech nor used HSV to investigate AdLD. Moreover, given the massive number of frames present in the HSV recordings which, definitely, needs an automated analysis, there is, however, a lack of effective automated tools (the temporal and spatial segmentation) for HSV analysis in connected speech. The reason for this is that using a transnasal flexible endoscope with a high-speed camera to record connected speech imposes challenges to the available automated image processing tools in term of image quality and excessive laryngeal maneuvers. The present dissertation aims to fill these gaps in research. Several measures can be directly extracted from analyzing HSV as mentioned before such as GAT and GOT. These measures can lead to deeper insights on the physiological changes in the impaired voice production of AdLD patients. On top of that, HSV can also be utilized to construct individual-specific models by which different biomechanical measurements can be generated to study AdLD in connected speech such as the elasticity and viscosity of the VF tissues [93]. These biomechanical measures can be generated using modeling techniques that have not been discussed before in literature to study the vocal function in running speech. A summary of the research gaps, research questions (Q), and hypotheses (H) are listed as follows:

Research Gap 1:

Current standard passages used for perceptual voice assessment of, e.g., AdLD exhibit difficulty to clearly visualize VFs when using HSV during connected speech. This urges suggestions for new speech tasks that would best allow for a better HSV assessment in these populations. This can be done by introducing automated approaches to detect the visual obstructions of the VFs in HSV and by which new assessment passages can be tested and optimized. However, there is a lack of effective automated techniques that can be used for image/frame classification and temporal segmentation of HSV during connected speech. Several studies proposed temporal segmentation techniques and image classification procedures to automatically detect laryngoscopic images that display sufficient diagnostic information. However, none of these studies were performed using HSV in connected speech, which is very important in studying voice disorders such as AdLD as they mostly reveal themselves in running speech. Only one study by our research lab was conducted for the temporal segmentation task but it only used a single HSV recording of a vocally normal individual in connected speech [119]. This emphasizes the need to develop reliable methods to temporally analyze connected speech in HSV. **Aim 1 is to build an automated tool to classify HSV frames by detecting the instances during which the image of the VFs is optically obstructed.** This tool will be able to automatically detect the time segments (HSV frames) that display an unobstructed clear view of VFs during a token of speech. This aim will address the following research questions.

Q1.1: *Can DNN accurately classify HSV frames in AdLD during connected speech regardless of the excessive laryngeal maneuvers?*

H1.1: DNN can accurately classify HSV frames based on whether these frames display an obstructed view of the VFs.

Q1.2: *Does the presence of AdLD affect the durations over which VFs are visually obstructed in HSV during running speech?*

H1.2: The duration of the visual obstruction of the VFs will be longer in AdLD versus normal controls during connected speech.

Research Gap 2:

Another lack lies in developing automated spatial segmentation methods amenable to HSV analysis during connected speech. Successful spatial segmentation in HSV data would lead to a precise detection/localization of VF edges during vibration in connected speech. This can be used to assess and evaluate the VF behavior during running speech. Yet, the existing studies for spatial segmentation of VF edges were performed on HSV data obtained during sustained vowels. The utilized image processing tools in these studies are ineffective in terms of being applied to connected speech with more challenging conditions (poor image quality and excessive laryngeal movements).

Aim 2 is to spatially represent VF edges in connected speech through developing a robust automated image segmentation tool considering the poor image quality in the transnasal HSV.

Q2: C*an VF edges be accurately and robustly segmented in HSV data during running speech in the presence of image noise?*

H2.1: The dark glottal area can be successfully silhouetted against the brighter surrounding VF tissues.

H2.2: ACM can accurately segment VF edges in HSV data with excessive image noise during VF vibrations.

H2.3: A clustering technique can be combined with ACM to build a hybrid method improving the edge segmentation accuracy of ACM during vocalization and when VFs are not vibrating.

Research Gap 3:

Although previous studies examined AdLD disorder in running speech using acoustic analysis, laryngoscopy, and aerodynamic measurements [72, 73, 64, 74], utilizing an advanced imaging technique like HSV has not been well investigated. Limited research was found on HSV use in studying AdLD, which was mainly conducted on sustained phonation, not running speech. Therefore, there is a need for further investigation into analyzing HSV during connected speech tasks to gain a more complete understanding of the impaired vocal function in AdLD. To do so, it is critical to develop automated approaches to provide analytical representation of the VF dynamics in AdLD during connected speech. Developing these approaches – though challenging due to the excessive laryngeal maneuvers existed in AdLD – will represent a considerable contribution to the existing literature in terms of offering a distinctive quantitative portrayal of the impaired behavior of VF vibrations. This can provide unique insights into the underlying VF dynamics in AdLD

Aim 3 is to provide quantitative representation of VF dynamics in AdLD during connected speech by extracting glottal area waveform (GAW) and glottal edges from HSV.

Q3: *Can the GAW be automatically extracted given the inferior image quality in the fiberoptic HSV and the excessive laryngeal movements in AdLD during running speech?*

H3.1: The hybrid method can be used as an automated labeling tool to train a robust DNN on detecting the glottal area in HSV during running speech.

H3.2: This trained DNN will be successfully implemented for the automated extraction of the GAW in AdLD and normal controls even with its challenging image conditions.

H3.3: The glottal midline along with the left and right VF edges can be successfully captured based on the segmented glottal area.

Research Gap 4:

Evaluating and investigating the pathological vocal function during phonation onset and offset of AdLD has been almost non-existent using HSV in running speech [152, 46]. Previous research showed that voicing onset/offset times might contribute to assess AdLD symptoms using acoustic analysis in running speech [72, 73], not HSV in connected speech. Thus, developing automated method for extracting quantitative measures of VF behavior at the onset and offset of phonation in HSV is important to understand the impaired vocal function in AdLD.

Aim 4 is to automatically analyze phonation onset and offset from HSV and measure glottal attack and offset times in AdLD and normal controls. Those two measures are critical factors to study the pathophysiology of voice disorders and, particularly, AdLD [153]. This aim will answer the following research question.

Q4: Are the glottal attack and offset times different between AdLD and normal controls?

H4.1: An automated algorithm can be developed to measure GAT and GOT with comparable accuracy to visual measurements.

H4.2: GAT and GOT will be significantly higher in AdLD versus normal controls.

H4.3: GAT and GOT will show more variability in AdLD subjects.

Research Gap 5:

A major knowledge gap exists in studying the biomechanical characteristics of the VFs in connected speech. Previous studies showed the successful development of the biomechanical measures using inverse analysis of HSV data through model-based methods. Yet these papers studied the biomechanical behavior of VFs neither in connected speech nor in AdLD disorders. These modeling studies used HSV data recorded during prolonged vowels, which are impractical for analyzing the pathological voice function as it mainly appears in running speech. Moreover, complex lumped models with multiple masses have been intensively utilized in literature. Given the complexity of analyzing connected speech, using simplified models becomes essential for optimizing simulations. However, simple models, like one mass, have not been explored before to conduct inverse analysis of HSV data. This model offers a simple structure, and with minimal control parameters, it can still represent the self-sustained VF vibration. Such advantage besides its low computational demands makes the one-mass model a compelling candidate for the present inverse analysis problem to capture the prominent features of VFs.

Aim 5 is to develop a simplified model-based approach that can determine the biomechanical characteristics of VF including VF mass, elasticity and viscosity based on the HSV running speech sample. The model parameters are obtained/optimized based on inverse analysis of HSV data using the dynamic vibration of VF extracted form HSV data.

Q5: *Can a simplified one-mass model be optimized to accurately match the vibratory behavior of VF extracted from HSV?*

H5.1: A simplified one-mass model can successfully simulate both the vibratory and closure phases of VF motion.

H5.2: The particle swarm optimization technique will enable accurate optimization of the model to predict the experimental glottal area waveform.

H5.3: The optimized model parameters, obtained through inverse analysis of HSV data, can estimate the VF mass, elasticity, and viscosity indices.

1.7. Dissertation Structure

Chapter 2 focuses on developing the required methodology to answer the aforementioned research questions. A description of the study subjects and the clinical HSV data acquisition is included with two different datasets (color and monochrome HSV data). The methodology used to generate the HSV-based measures and the conducted analysis are profoundly elaborated in this chapter. This methodology consists of five studies including various techniques and different analysis that are discussed in Chapter 2 in detail under a separate section. Here is a brief summary regarding the different approaches and analysis performed in each study. Study I: Obstruction detection of the VF view is a method for extracting the frames that display an obstructed view of VFs. This method includes a developed deep learning framework to conduct this task. Study II: Image segmentation of VF edges is an algorithm developed for capturing VF edges in HSV-based extracted kymograms. Study III: A deep learning approach for analytical representation of the VF dynamics and glottal area aims at providing analytical representation of the dynamic movement of the VFs. Study IV: Automated measurements of GAT and GOT which are extracted and analyzed from the video recordings to assess the impaired VF dynamics in AdLD during the phonation onset and offset. Study V: Developing and optimizing a lumped modeling technique to simulate the oscillatory characteristics of the VFs and optimizing it using the experimental HSV data. Chapter 3 presents and describes the results corresponding to each developed methodology and technique in the dissertation and is organized similar to Chapter 2 for clarity. Chapter 4 discusses the results of the various developed methods and the different performed analysis. Finally, Chapter 5 presents a conclusion of each study including a summary of the findings.

CHAPTER 2: METHODOLOGICAL APPROACH

2.1. Research Design

This chapter provides information about the methodologies required to pursue the target aims of this dissertation, discussed in the previous chapter. The research design, methods, and analysis of each corresponding aim will be presented. In addition, information about the study subjects and the data acquisition are included at the beginning of this chapter. There are two different sets of HSV data that are used and analyzed in this dissertation study. Each dataset is utilized to pursue specific aims and address their related research questions and hypotheses. The first dataset was recorded during running speech from a vocally normal speaker using a flexible HSV with a color high-speed camera. This recording had an inferior noisy image quality with dim lighting. This challenging HSV data will be used to fulfil Aim 2 and a part of Aim 3. The methods and the image processing techniques related to these aims are mainly designed and implemented using this challenging recording in order to demonstrate the robustness of the introduced approaches. The successful application of the proposed methods on such challenging image conditions facilitates the implementation to less challenging monochromatic HSV images since a monochrome camera provides a higher sensitivity and dynamic range with better pixel representation. Monochromatic HSV recordings make up the second HSV dataset which are utilized to execute the remaining research aims in this dissertation. The second set includes HSV recordings obtained from both vocally normal participants and patients with AdLD using a monochrome camera; each subject's HSV data were recorded during running speech.

This chapter includes a description of each methodology used to analyze the HSV data in order to address the research questions and aims of the present dissertation. Different methods and approaches are introduced for the data analysis and discussed under different sections. The first study was developed as a temporal (classification) technique which will pursue Aim 1. In this approach, an effective tool was introduced using DNN to analyze the monochrome HSV dataset for both vocally normal adults and patients with AdLD. The tool was designed such that it can automatically recognize and classify the time segments in HSV recordings that displayed an obstructed view of VFs. Information regarding the DNN structure, training and evaluation are included in this section. This temporal method allows to analyze the differences between the normal controls and AdLD in terms of the durations of VF obstructions.

The second study aims at developing a spatial segmentation method that can successfully detect VF edges in HSV-based kymograms after a series of image processing steps to preprocess the color HSV data. These analyses fulfill Aims 2. The developed approaches introduced improved image processing techniques that were designed as a combination of the ACM method and a clustering technique. This hybrid method was developed based on the color HSV dataset. The design and the implementation of this method will be discussed in detail in this chapter. Aim 3 was implemented using another DNN method. The spatial segmentation technique, developed in Aim 2, was used to train a robust DNN that can perform spatial segmentation to analyze the GAW and provide an analytic representation of the vibratory characteristics of VF in Aim 3. Information about how this DNN method for spatial segmentation was built, trained, and evaluated are included. This developed DNN tool was then applied to the monochrome HSV data to analyze the vibratory characteristics of VF in AdLD patients by automatically segmenting the glottal area/edges within the HSV recordings; this implementation is also discussed in detail in this chapter. The spatial segmentation DNN tool facilitated the development and analyses of the HSV-based measurements, which is discussed as the fourth study in this dissertation: particularly the generating the GAT and GOT measures. In the fifth study, a lumped model was developed, simulated and optimized in combination of the extracted results of the spatial segmentation of the VFs and the glottal area. This model pursued Aim 5, which is discussed in the last section of this chapter.

2.2. Study Subjects

The details of two subject pools (subject pool one and subject pool two) that are included in the present dissertation are discussed in this section. The main difference between the two subject pools is that the subject pool one included a vocally normal speaker while the subject pool two contained both vocally normal speakers and AdLD patients. Also, the two subject pools were recruited and examined at two different locations. Further information about each subject pool is included in the next two subsections.

2.2.1. Subject Pool One

The first HSV dataset (subject pool one) was obtained from a vocally normal female (38 years of age) who did not have any history of voice disorder. The following inclusion criteria were used for the vocally normal participant: 1) at least 18 years of age and proficiency in reading English written text; 2) no prior history of intubation injury or airway/laryngeal surgery; 3) normal hearing;

4) absence of structural abnormalities including lesions and/or VF paralysis/paresis, 5) absence of perceptual symptoms of the classical dysarthria; 6) cognitively intact and able to undergo the flexible HSV protocol. The subject was recorded during running speech while reading the "Rainbow Passage." The examination was done at the Center for Pediatric Voice Disorders, Cincinnati Children's Hospital Medical Center and was approved by the Institutional Review Board. This subject pool is used to pursue Aim 2 and part of Aim 3.

2.2.2. Subject Pool Two

Monochrome data from 9 participants within the age range of 35–76 years old were collected. The current study population includes 4 vocally normal participants (3 female and 1 male) without history of voice disorder and 5 patients with AdLD (4 female and 1 male). The data collection was conducted at the Mayo Clinic in Scottsdale, AZ, and approved by the Institutional Review Board. The goal of this data collection is to use the resulted HSV dataset from the second subject pool to fulfill Aims 1, 3, and 4.

The inclusion criteria for the normal participants were similar to the criteria mentioned in the previous subsection with the first dataset of subject pool one. The inclusion criteria for the AdLD subjects were [154, 155]: 1) at least 18 years of age and proficiency in reading English written text; 2) no prior history of intubation injury or airway/laryngeal surgery; 3) normal hearing; 4) absence of structural abnormalities including lesions and/or VF paralysis/paresis; 5) absence of perceptual symptoms of the classical dysarthria; 6) auditory-perceptual characteristics consistent with the disorder (evidence of phonatory breaks on voiced sounds and a strained-strangled quality, and no obvious tremor during phonation); 7) occasional moments of normal sounding voice; 8) improved voice for non-speech vocalizations; 9) improved voice quality for phonation at higher pitches; 10) cognitively intact and able to undergo the flexible HSV protocol; 11) not stimulable for voice change with facilitation techniques (e.g., distraction, improving breath and voice coordination and forward resonance, manual manipulation, etc.); 12) no recent Botox treatment or surgical treatment for AdLD. A summary of the study inclusion criteria for both groups of subjects is included in Table 2.1.

Table 2.1. Study inclusion criteria for AdLD patients.

Inclusion Criteria	**AdLD group**	**Non-pathological group**
At least 18 years of age and proficiency in reading English written text.	X	X
No prior history of intubation injury or airway/laryngeal surgery.	X	X
Normal hearing.	X	X
Absence of structural abnormalities including lesions and/or VF paralysis/paresis.	X	X
Absence of perceptual symptoms of the classical dysarthria.	X	X
Cognitively intact and able to undergo the flexible HSV protocol.	X	X
Auditory-perceptual characteristics consistent with the disorder (evidence of phonatory breaks on voiced sounds and a strained-strangled quality, and no obvious tremor during phonation).	X	
Occasional moments of normal sounding voice.	X	
Improved voice for non-speech vocalizations.	X	
Improved voice quality for phonation at higher pitches.	X	
Not stimulable for voice change with facilitation techniques (e.g., distraction, improving breath and voice coordination and forward resonance, manual manipulation).	X	
No recent Botox treatment or surgical treatment for AdLD.	X	

Furthermore, all vocally normal participants were screened by a voice specialized SLP (15+ years' experience), prior to consent, based on the above inclusion criteria. All participants with AdLD were diagnosed through the consensus of a voice-specialized SLP (5 years of experience) and a fellowship-trained laryngologist (15+ years of experience) based on the mentioned inclusion criteria. Also, in order to obtain an accurate diagnosis of AdLD, subjects were identified from the treatment-seeking population at the Mayo-AZ Otolaryngology – Head & Neck Clinic. All participants were seen by both a speech-language pathologist and a laryngologist. A voice evaluation and laryngoscopy were completed during sustained phonation and connected speech.

Additionally, stimulability tasks were completed. Post full voice evaluation, a diagnosis of AdLD was assigned via multidisciplinary consensus involving the laryngologist and one of 3 speech-language pathologists specialized in voice disorders. All participants with the diagnosis of AdLD had no evidence of tremor via consensus between the treating speech-language pathologists and laryngologist.

2.3. Data acquisition

This section explains how the data acquisition was performed. The two subject pools included in the present work were examined and recorded differently. The difference mainly lied in the way of collecting the experimental data such that different HSV systems and setups were used to obtain the video recordings from the participants. An HSV system with a color high-speed camera was utilized to collect recordings from the subject pool one whereas a monochrome high-speed camera was used to collect the HSV data from the subject pool two. In addition, each camera had different spatial resolution. A detailed description of each HSV setup is discussed in the following two subsections.

2.3.1. Subject Pool One (Color HSV System)

A custom-built color HSV system with 4,000 frames per second (fps) and 249 µs integration time was used for the data acquisition from the subject pool one (a vocally normal speaker). The recording length was 29.14 s (116,543 frames in total) with HSV image resolution of 256x256 pixels. The recording system included a FASTCAM SA-Z color high-speed camera (Photron Inc., San Diego, CA) coupled with a 3.6-mm Olympus ENF-GP Fiber Rhinolaryngoscope (Olympus Corporation, Tokyo, Japan), and a 300-W xenon light source, model 7152A (PENTAX Medical Company, Montvale, NJ). The camera had a 12-bit color image sensor with sensitivity of ISO 20,000 and 64 GB of cache memory divided into two 32-GB partitions. The selected zoom lens adapter had a focal distance of 45 mm in order to provide the optimal pixel representation and dynamic range. The distance of the endoscope to the VFs was selected to ensure that despite the active maneuvers of the larynx during connected speech, the VFs always fall within the field of view of the endoscope during the recording. The recorded HSV sequence was saved as an uncompressed 24-bit RGB AVI file and then analyzed. The camera used for the data collection had a native resolution of 1,024x1,024 pixels at 20,000 fps. However, for the purposes of the study, the resolution was set to 256x256 pixels, which at the chosen speed of 4,000 fps provided for up to 30 seconds per partition to record the reading of the Rainbow Passage. The selected frame rate was shown to be clinically acceptable for voice assessment [42]. Moreover, using 256x256 pixels provided the optimal balance between the image resolution, camera frame rate, duration of the recording, and the light sensitivity necessary for this data collection.

2.3.2. Subject Pool Two (Monochrome HSV)

A different HSV system was used to collect the HSV video data from the second subject pool. For this second HSV dataset, the experimental setup included a Photron FASTCAM mini AX200 high-speed monochrome camera (Photron Inc., San Diego, CA). The camera was coupled with a flexible nasolaryngoscope. The video recordings were collected at spatial resolution of 256 x 224 pixels with a rate of 4,000 frames per second (fps). This chosen spatial resolution was appropriate for the present dataset to be analyzed due to the reasons mentioned in the previous HSV system (in the previous subsection). The recording procedure consisted of several connected speech samples. As such, in the same recording session, each subject was asked to complete reading the six CAPE-V sentences, and a reading of part of the "Rainbow Passage" (the first six sentences). The full length of HSV recordings varied between 50 to 100 s among participants. The HSV files (up to 32GB each) were stored on a computer connected to the HSV camera as mraw files (the file format of the Photron camera used in this study), then transferred to the laboratory data server after de-identification.

2.4. Study I: Automated Detection of Vocal Fold Image Obstructions

The purpose of this study was intended to fulfill Aim 1 by addressing the following:

Q1.1: Can DNN accurately classify HSV frames in AdLD during connected speech regardless of the excessive laryngeal maneuvers?

H1.1: DNN can accurately classify HSV frames based on whether these frames display an obstructed view of the VFs.

Q1.2: Does the presence of AdLD affect the durations over which VFs are visually obstructed in HSV during running speech?

H1.2: The duration of the visual obstruction of the VFs will be longer in AdLD versus normal controls during connected speech.

This study investigates a new approach to temporally segment the HSV recordings. This approach was utilized and introduced in order to analyze the monochrome HSV datasets [152] and was thoroughly discussed in this section. This method was developed using a deep learning framework where a DNN was designed as a classifier. The main objective of this temporal method was to automatically classify the video frames from the HSV recordings into either a frame with a clear display of VF view or a frame with an obstructed view of VFs. This proposed method fulfilled Aim 1 and its hypothesis using the vocally normal and AdLD subjects from the second sample pool (the monochrome dataset). As this approach was designed using a classifying DNN, the structure, training, implementation, and evaluation of this classifying network are discussed in the following subsection. The results of implementing this approach are also discussed at the beginning of the results chapter.

2.4.1. Convolutional Neural Network (CNN) Architecture

The deep-learning method we propose for classification is based on convolutional neural networks (CNN), which is a well-known technique with a promising performance for image classification [156, 157]. The network architecture is built using 64-bit MATLAB R2019b (MathWorks Inc., Natick, MA) as a powerful platform for CNN implementation. Figure 2.1 illustrates the main network architecture, which is mainly designed through 10 layers of 3x3 unpadded convolutions. Each convolution is followed by batch normalization to accelerate the training of the neural network and alleviate the sensitivity of the network to its initialization. A nonlinearity term is then added by including a rectified linear unit (ReLU) to accelerate the computations and improve the network performance. ReLU converts values below zero in the convolution input to a value of zero while keeping the values above zero the same. After each pair of two successive convolution-batch normalization and ReLU layers, a down sampling step is applied using a 2x2 max pooling with a strid of 2 where the number of feature maps are doubled. The dimensions of the feature maps corresponding to the different convolutional layers are illustrated in Figure 2.1 at each down sampling stage. The last layer is a sigmoid layer to classify whether the input frame contained the VFs or not (2 classes: unobstructed VF or obstructed VF). The input of the network-based classifier is the HSV frames (256x224 pixels) and are classified as images with unobstructed or obstructed VF view in the classifier output.

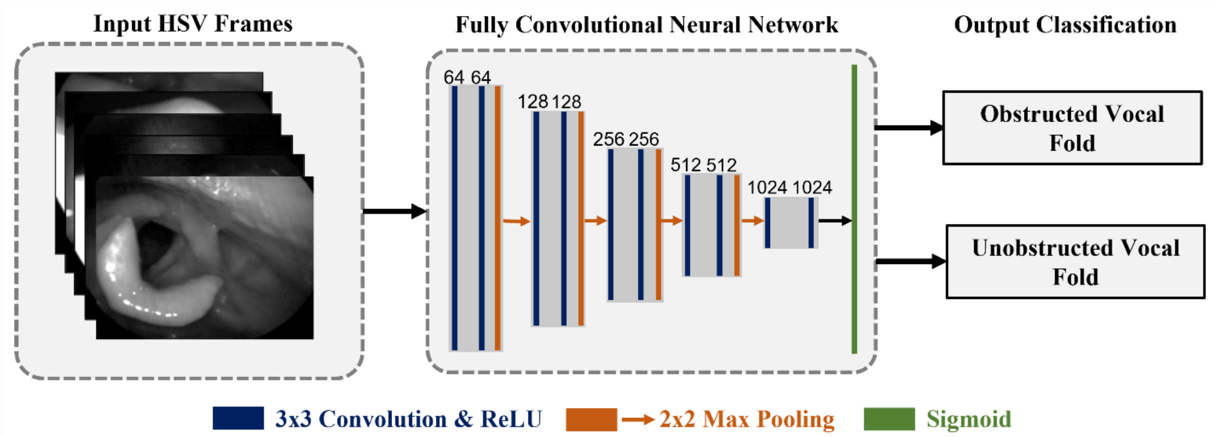


Figure 2.1. A Schematic diagram for the automated deep learning approach, developed in this work. The HSV video frames serve as the input to the automated classifier. The detailed structure of the convolutional neural network is illustrated. The input frames are processed through several layers of 3x3 convolutions combined with rectified linear unit (ReLU) layers (in dark blue), followed by multiple 2x2 max pooling layers (in orange). The last layer includes a sigmoid layer (in green). The dimensions of the feature maps corresponding to the different convolutional layers are also included in the figure. The neural network classifies each frame into two classes as a classification output: either a frame with unobstructed VF(s) or a frame with obstructed VF(s).

2.4.2. Classification CNN Training

The training dataset is constructed using the monochrome HSV data (subject pool two). That is, a total of 11,800 HSV frames is manually selected and extracted from six monochrome HSV recordings of three normal participants and three AdLD patients – creating an image dataset by which the network is trained and validated. These chosen frames are labeled by a rater into two classes: 5,900 images with unobstructed VF view and 5,900 images with obstructed VF view. For the first class, the frames are selected such that all different phonatory gestures that may occur during running speech are represented. For example, we select frames during sustained VF vibration, pre-phonatory adjustments, phonation onset/offset, and no VF vibration. For the second class, we consider frames that displayed various types of VF obstructions due to, e.g., epiglottis movements, the movements of the left/right arytenoid cartilages, laryngeal constriction, false VF movements, or a combination of any of these various obstruction types. Generally, an obstruction is defined such that if more than 50% of VF(s) was obstructed, the frame would be classified as a frame with VF obstruction.

The total of 11,800 training frames are divided into two independent datasets, where 10,620 frames (90%) form a training dataset, and 1,180 frames (10%) are assigned to a validation dataset. The validation dataset is created to tune the network parameters and avoid any overfitting that may occur during training. Adam optimizer, a stochastic gradient descent optimizer, is used for training (refer to [158], for the complete description of implementing Adam optimizer). The training is performed with a batch size of 16 (16 images processed in each training iteration) for a maximum of 20 epochs (the maximum number of training iterations). The outcome of the trained network return one of two labels for each frame in the video (either "Unobstructed VF" or "Obstructed VF").

2.4.3. Classification CNN Evaluation

An additional 2,250 HSV frames are manually extracted and labeled from the HSV recording of an AdLD participant (selected from the second subject pool as well) who is not included in the training dataset. Those labeled frames constitute a testing dataset, which is important to verify the generalization capability of the trained network on a set of new images. Also, the robustness and stability of the trained network are thoroughly evaluated through the comparison of automated and visual classifications for the entire HSV recordings of two participants. A rater visually analyzes the recordings of a vocally normal participant (264,400 frames) and an AdLD patient (399,384 frames) and determines the timestamps (frame numbers) between the beginning and ending of each obstruction.

We use the confusion matrix as a quantitative measurement to assess the performance of the trained network on the testing dataset and the two testing videos. The matrix demonstrates how accurate the trained network recognizes VF obstructions in the frames. From the confusion matrix, we use different metrics to evaluate the performance: class sensitivity, specificity, precision, F1-score, and the overall network accuracy. The sensitivity and specificity are the measures of the network ability to correctly predict the output based on the actual frame label as follows:

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP + FN}}, \qquad (2.1)$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN + FP}}. \qquad (2.2)$$

TP (true positive) is the number of frames for which the network correctly predicts the positive class (frames with an unobstructed view). TN (true negative) is the number of frames for which the trained network correctly predicts the negative class (frames with an obstructed view). In contrast, FP (false positive) and FN (false negative) are the number of incorrectly classified frames as images with unobstructed and obstructed view, respectively. The sensitivity and specificity scores are utilized to obtain the receiver operating characteristics curve (refer to [152] for details about its generation). The area under this curve is calculated as an overall evaluation of the network performance: The larger the area, the better the accuracy of the network. The precision of correctly predicting each class is calculated from the following equation:

$$Precision\ = \frac{\mathrm{TP}}{\mathrm{TP + FP}}. \qquad (2.3)$$

F1-score is computed based on the precision and sensitivity scores of each class as another way to measure accuracy as shown in equation (2.4):

$$F1\ Score\ = 2\ \mathrm{x}\ \frac{\mathrm{Precision\ x\ Sensitivity}}{\mathrm{Precision + Sensitivity}} = \frac{\mathrm{2TP}}{\mathrm{2TP + FP + FN}}, \qquad (2.4)$$

where a high F1-score means that the network has low number of incorrect frame classifications for both classes. Also, the overall network accuracy when applied to the testing frames/videos is determined from equation (2.5):

$$Accuracy = \frac{\mathrm{TP + TN}}{\mathrm{TP + TN + FP + FN}}. \qquad (2.5)$$

After the evaluation, the network is applied to the monochrome recordings of both the vocally normal subjects and the AdLD patients. The network defined the durations and exacted the frames during which the VF was visually obstructed. Hence, these durations are compared in order to investigate any noticeable differences between the AdLD and the normal controls – seeking to address H1.2.

2.5. Study II: Image Segmentation of Vocal Fold Edges

This study is focused to address Aim 2 by fulfilling the following:

Q2: Can VF edges be accurately and robustly segmented in HSV data during running speech in the presence of image noise?

H2.1: The dark glottal area can be successfully silhouetted against the brighter surrounding VF tissues.

H2.2: ACM can accurately segment VF edges in HSV data with excessive image noise during VF vibrations.

H2.3: A clustering technique can be combined with ACM to build a hybrid method improving the edge segmentation accuracy of ACM during vocalization and when VFs are not vibrating.

Spatial segmentation is the second study in the present dissertation. The goal of this study is to spatially segment the VF edges in HSV-based kymograms using the color HSV data (subject pool 1). In order to achieve this, several image processing techniques are considered to facilitate the downstream spatial segmentation task in this work. A preprocessing step including a temporal segmentation technique is required to provide the vocalized time segments in the HSV data that only display VF vibrations to which the spatial technique can be effectively applied. The spatial segmentation approach is a combination of two image processing steps for spatial segmentation: a clustering technique and an ACM. This hybrid approach is applied to the color HSV recording to segment the VF edges after preprocessing the video. The proposed hybrid image segmentation method aimed at fulfilling Aim 2. The details of each step of developing this technique are discussed under the following subsections. The proposed image segmentation tool is developed in order to localize and segment the edges of VFs across HSV frames in HSV-based kymogrames. This tool consists of several integrated algorithms, which can be divided into two main stages as shown in Figure 2.2: a data preprocessing step (including temporal segmentation, motion compensation, and kymograms extraction) and a machine-learning based image segmentation step (including a clustering and an ACM). Each step will be discussed in detail in the following subsections. All the algorithms are developed and implemented using 64-bit MATLAB (MathWorks Inc., Natick, MA).

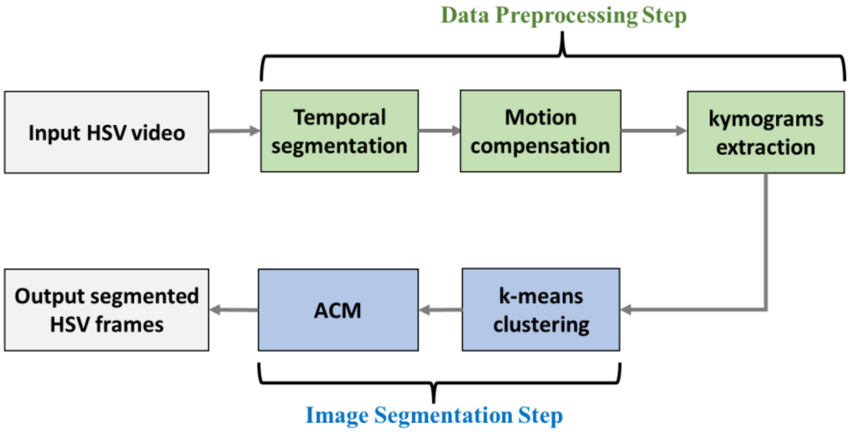


Figure 2.2. Workflow-chart of the spatial segmentation tool. The gray boxes indicate the input (HSV video data) and the output (HSV frames with segmented VF edges), the green boxes show the data preprocessing steps, and the blue boxes represent the image segmentation steps.

2.5.1. Data Preprocessing

Several preprocessing steps are applied to the color HSV data before proceeding with the proposed spatial segmentation approach as shown in Figure 2.2. These data preprocessing steps were developed in a previous study [119] and were applied to the present HSV recording. The main goal of this algorithm is to preprocess the HSV video to automatically extract timestamps of the vocalized segments (the vibratory onsets/offsets) before applying the image segmentation algorithm to obtain an analytic representation of the VF edges, which will be discussed later. These preprocessing steps are essential as a temporal segmentation stage preceding the spatial segmentation, which eventually provides an analytic representation of the VF edges in the images/frames. This is because the temporal method, discussed in this subsection, should first determine where in time the vibrating VFs are present before even the edges of the VFs can be localized and segmented. The main goal of these preprocessing steps is to extract the vocalized time segments in the video where the VFs are vibrating. A histogram analysis is first performed to enhance the contrast of the HSV frames by removing the saturated pixels. A gradient-based technique is then applied in order to eliminate the image noise and suppress any irrelevant movements in the image other than VF vibrations. This is applied to the images through a moving-average procedure to generate a motion window.

The goal of the motion window is to capture the spatial location of the vibrating VFs in a window across the frames. The motion compensation is mainly performed to overcome the problem with the VF misalignment due to the laryngeal maneuvers during running speech and the movements of the endoscopic tip relative to the laryngeal tissues over time. After detecting the location of the vibrating VFs across the frames, the HSV frames are cropped based on the location of the center of the motion window. The motion window is designed in an ellipse shape (see [37] for complete description of the motion window). It is applied to each HSV frame in order to remove the irrelevant noise and tissues. This step is essential for extracting HSV kymograms inside a rectangular window that encloses the vibrating VFs in each frame.

The HSV kymograms are then extracted for the frames between 25 ms before the onset of vocalization and 25 ms frames after the vocalization offset for each vocalization to ensure that the full pre-, post-, and peri-phonatory phases are included. For each vocalized segment, the first kymogram is extracted at the mid-line passing through the vibrating VFs inside the motion window. The spatial segmentation (explained in the upcoming section) is then applied to the kymogram to detect the glottal edges. The kymograms are extracted at different cross sections of the VFs between the anterior and posterior commissure. That is, the y-axis of the kymogram image represents the left-right dimension of the video frame, while the x-axis refers to time (frame number). The spatial segmentation algorithm is applied to each kymograms which will be discussed in the following subsection.

2.5.2. Machine-Learning Based Image Segmentation for Clustering

As mentioned earlier, after applying the data preprocessing step and extracting the kymograms of the HSV recording, a hybrid method for image segmentation is applied to the kymogram images. This image segmentation step consists of a combination of two techniques: a clustering technique and ACM as shown in Figure 2.2. This subsection discusses the development of the clustering technique. Selection of the right features from the HSV kymogram images is an essential step toward the successful implementation of the clustering method. In this work, three features are extracted from the kymogram images. The pixel intensities of the red and green channels in the extracted kymograms are considered as two features. This is because the regions of interest in the kymogram (glottal areas) have lower intensities (darker) than the neighborhood regions (the laryngeal tissues) – making pixel intensities a good predictor for glottal area. In addition, given the contrast between the intensity of the glottis and the surrounding regions, the third feature is the kymogram image gradient. As such, the image gradients are computed along the x- and y-axis in the kymogram with a step size of 8 pixels. Then, the overall gradient magnitude is calculated by taking the square root of the sum of squared of the pixels' gradients in the two directions. The three extracted features are then be normalized between 0-1 and input into a k-means clustering technique [159].

The k-means clustering algorithm is an unsupervised machine learning technique and is well-known for image segmentation [159]. The k-means clustering technique is based on partitioning a dataset into a certain number of disjoint clusters (groups of data). This technique requires the initialization of the number of clusters (k) and the center of each cluster (centroid). Refer to [159, 160] for the full detail of the implementation and the algorithm. In this study, for the kymogram images, the number of clusters are 2 (inside or outside the glottal area) and the initial centroids are chosen based on k-means++ algorithm, which uses a heuristic in order to initialize centroid seeds for k-means clustering (see [160] for the full detail of the algorithm). The clustering algorithm then computes the distance between the centroids and each pixel in the kymogram. The distance is calculated using the Euclidean distance as follows:

$$d = \|I(x, y) - c_k\|, \qquad (2.6)$$

where d is the Euclidean distance, I(x, y) corresponds to the intensity of the kymogram, x and y refer to the pixel coordinates, c_k is the cluster centroid, and k is the cluster number. Each pixel in the kymogram image is assigned to the nearest centroid based on the calculated distance leading to the formation of the initial clusters. Once the grouping is done, the algorithm recomputes the updated centroid of each cluster (c_k) as follows:

$$c_k = \frac{1}{k} \sum_{x \in c_k} \sum_{y \in c_k} I(x, y), \qquad (2.7)$$

where this new centroid is the data point to which the summation of the distances from all the pixels located in that cluster is minimal. This process is repeated iteratively – reshaping the clusters in the image at each iteration – until converging, when the distance between the new and original centroids does not change.

Instead of applying the clustering algorithm to the entire kymogram image for a specific vocalization, each kymogram is divided into smaller kymograms with a maximum of 50 frames to mitigate any possible impact of the image noise on the clustering accuracy. For example, when part of the kymogram has extremely bright pixels (saturated or near-saturated pixels), the clustering technique might be misguided, particularly with large number of frames. This might occur due to the movements of the epiglottis and large reflections from its surface.

After applying the clustering algorithm to each kymogram, each pixel in the kymogram is assigned to either cluster 1 (a pixel belongs to the glottal area) or cluster 2 (a background pixel). All the pixels in the same cluster have similar labels. Accordingly, a new binary labeled image of the kymogram is constructed, where each pixel has the binary value of the cluster number. Accordingly, the spatial location of the glottal edges in the kymograms – corresponding to the left and right VF – is determined by spatially annotating the boundary of the glottal area cluster (cluster 1) as the initial contours for the ACM method.

2.5.3. Hybrid Method

The hybrid method for spatial segmentation is developed by combining the clustering technique, discussed above, with an ACM method. The active contour in the ACM method is a spline that deforms spatially based on an internal rule (depending on the rigidity and elasticity of the contour) and an external rule (depending on the gradient of the image) until the contour can capture the glottal edges in the image. This deformation is performed through an energy optimization process, which aims to minimize the sum of the internal and external energy functions, corresponding to the contour shape and the image gradient, respectively [132]. The glottal boundary locus resulted from the clustering method is provided to the ACM technique as the initial active contours for the right and left VFs in the kymograms. These initialized contours, estimated close to the VF edges, facilitate the efficient ACM implementation by accurately being deformed and exactly detecting the locations of VF edges in the kymograms. That is, the initial contours (also called snakes) are parametrized as a vector $v(s) = [x(s), y(s)]$, where $s \in [0,1]$. The objective energy function that needed to be minimized is defined as [132, 161]:

$$E = \int_0^1 [E_{int}(v(s)) + E_{image}(v(s))]\, ds, \qquad (2.8)$$

where E_{int} is the internal energy function and E_{image} is the external image function. The internal energy function (E_{int}) of the contours is computed by:

$$E_{int}(v(s)) = \frac{1}{2}\ [\alpha(s)\ |v'(s)|^2 + \beta(s)\ |v''(s)|^2], \qquad (2.9)$$

where $v'(s)$ and $v''(s)$ are the contours first and second derivatives, respectively; α and β are two weights included to adjust the elasticity and rigidity of the snake, respectively, which control the snake shape. The two weights α and β are set to 0.1 and 0.03, respectively.

The image energy function (E_{image}) counterbalances the internal energy and is given by:

$$E_{image}(v(s)) = -\ |\nabla I(x,y)|^2, \qquad (2.10)$$

where $\nabla I(x,y)$ is the spatial gradient of the kymogram image.

The solution for equation (2.8) is based on discretization of the energy function. The finite difference method is used to approximate the first and second derivatives in the internal energy function. The internal energy function, given in equation (2.9), can be rewritten as follows, after the discretization:

$$E_{int}(v_i) = \frac{1}{2}\ [\alpha\ |v_i - v_{i-1}|^2 + \beta\ |v_{i+1} - 2v_i + v_{i-1}|^2], \qquad (2.11)$$

where v_i refers to the ith snaxel; the snaxels are the vertices that make up the snake spline. The discretization of the image energy function, given in equation (2.10), yields:

$$E_{image}(v_i) = -\ |\nabla I(v_i)|^2. \qquad (2.12)$$

By discretizing the total energy function, given in equation (2.8), the following equation can be derived, which is considered as a dynamic-programming problem [161]:

$$E_{total} = \sum_{i=1}^{n} [E_{int}(v_i) + E_{image}(v_i)], \qquad (2.13)$$

where *n* is the total number of snaxels referring to the total number of frames included in the kymogram. The solution to the above dynamic programming problem generates a sequence of functions $\{S_i\}_{i=1}^{n-1}$ with one unknown variable v_i, where S_i is the optimum value function (see 2.14). At each S_i, a minimization is conducted over the v_i variable, where v_i is a state variable and can be assigned *m* possible values. The value of *m* refers to the number of pixels in the neighborhood of the snaxel that the algorithm searches to find the optimal v_i. In this study, the value of m is set to five.

$$\begin{aligned}
S_1(v_1) &= \underset{v_1}{min}\left[E_{int}(v_0, v_1, v_2) + E_{image}(v_1)\right],\\
S_2(v_2) &= \underset{v_2}{min}\left[S_1(v_1) + \ E_{int}(v_1, v_2, v_3) + E_{image}(v_2)\right], \qquad (2.14)\\
S_3(v_3) &= \underset{v_3}{min}\left[S_2(v_2) + \ E_{int}(v_2, v_3, v_4) + E_{image}(v_3)\right],\\
&.\\
&.\\
S_n(v_n) &= \underset{v_n}{min}\left[S_{n-1}(v_{n-1}) + \ E_{int}(v_n) + E_{image}(v_n)\right],
\end{aligned}$$

and in the general case,

$$S_i(v_i) = \underset{v_i}{min}\left[\sum_{i=1}^{i-2}\left(E_{int}(v_i)\right) + \sum_{i=1}^{i-1}\left(E_{image}(v_i)\right) + E_{int}(v_{i-1}) + E_{int}(v_i) + E_{image}(v_i)\right]. (2.15)$$

The image gradient ∇I is calculated in the vertical direction along the left-right dimension of the kymogram image in each frame with the step size of 10 pixels. The gradient is computed for each kymogram to signify the glottal edges, where the intensity changes rapidly around the initialized contours. Accordingly, the movement of each snaxel for the left and right active contours is limited to the vertical direction. The positive and negative gradients are computed next. The positive gradient of the kymogram is obtained by assigning the negative gradient pixels value of zero. The negative gradient of the kymogram is obtained by assigning the positive gradient pixels zero. The positive gradient is used in the image energy function when searching for the left VF edges while the negative gradient is used for the right VF.

To find the snakes that exactly capture the left and right VF edges, the discretized energy function in equation (2.15) is solved for i = 1, 2, 3, …, n. The minimization of the S_i functions is done recursively, and the snake is updated during each loop until the value of the snake remains almost unchanged through the minimization procedure. As such, during each loop and at each snaxel, the value of the S_i function is numerically calculated for the column-wise snaxel's five neighboring pixels. The neighboring pixel that leads to a minimum total energy is considered as the updated value of v_i. After updating all the v_i values, the next loop starts until the algorithm converges and the optimum snake spline is found. The convergence of the algorithm depends on an error function. This error function is defined as the sum of squared of the difference between the calculated snake in the current loop and the previous loop. The algorithm converges when the error became smaller than or equal to 1. The successful execution of the optimization process leads to an accurate adjustment of the initialized contours, resulted from k-means, and to a smooth segmentation of the VF edges in the kymogram images.

The hybrid method (k-means + ACM) is applied to all the kymograms at different intersections of the VFs for glottal edge detection. The detected edges in the kymograms are then registered back to the HSV frames for each vocalization to determine the spatial location of the VF edges with respect to the original HSV frames. In order to demonstrate the advantage of combining ACM with k-means, the ACM method is directly applied to detect VF edges in the kymogram images without incorporating the clustering technique. Hence, instead of using the clustering method to initialize the active contours for the ACM method, a horizontal line (an active contour) is initialized in the kymogram space – passing through the center of the glottis. The contour initialization is performed for each kymogram n_i using the first moment of inertia, denoted by $M_1(y, n_i)$, calculated as follow [82]:

$$M_1(y, n_i) = \frac{\sum_{x=1}^{K_w} \sum_{y=1}^{K_h} I(x, y, n_i) y}{\sum_{x=1}^{K_w} \sum_{y=1}^{K_h} I(x, y, n_i)}, \qquad (2.16)$$

where x and y correspond to the spatial coordinate of a pixel, $I(x, y, n_i)$ is the pixel intensity in the kymogram image, K_w is the kymogram image width, which is the number of frames in the kymogram, and K_h is the kymogram image height in pixels. Since the first moment of inertia determines the center of brightness, the kymogram is inverted to find the center of darkness (i.e., the centroid of the glottis). The resulted moment line of the kymogram image is considered as the initialized line. The ACM is then applied to the kymograms using this initialized straight line where it is deformed upward and downward until capturing the left and right VF edges. The performance of only implementing ACM to the kymogram images is evaluated in order to reveal whether the ACM is a suitable image segmentation method for localizing VF edges in HSV-based kymograms. In addition, the performance of the ACM is compared with that of the hybrid method by applying the two methods on a decent quality kymogram and a kymogram with dim lighting and a degraded quality. These different applications are selected to address the hypotheses related to Aim 2.

2.6. Study III: Deep-Learning-based Representation of Vocal Fold Dynamics

This study is conducted to pursue Aim 3 through addressing the following:

Q3: Can the GAW be automatically extracted given the inferior image quality in the fiberoptic HSV and the excessive laryngeal movements in AdLD during running speech?

H3.1: The hybrid method can be used as an automated labeling tool to train a robust DNN on detecting the glottal area in HSV during running speech.

H3.2: This trained DNN will be successfully implemented for the automated extraction of the GAW in AdLD and normal controls even with its challenging image conditions.

H3.3: The glottal midline along with the left and right VF edges can be successfully captured based on the segmented glottal area.

The successful building of the hybrid approach to capture VF edges during vibrations allows for the development of a more generalized, flexible approach that can automatically capture the VF dynamics. Since the previous hybrid technique can capture the glottal edges during sustained VF oscillations, we utilize it to build a deep learning model that can segment glottal edges/area during also the nonstationary portions of running speech such as in prephonatory adjusments and during onsets and offsets of vibration, and when there is no VF vibrations [47, 45, 46]. As such, this model is designed based on DNN and utilizes the hybrid method as an automated labeling tool for training the segmentation network in order to precisely representing VF edges in all phonatory events during connected speech. The color HSV recording is considered to implement the proposed DNN approach. Below is a description of how the hybrid method as an automated labeling tool and an explaination on the design, training, and evaluation of the DNN for spatial segmentation. After building and implementing the developed DNN model to the color HSV data, it is retrained and applied to the monochrom HSV data which is also discussed in this section. The development of a new DNN model and its implementation to the monochrom data will pursue Aim 3.

2.6.1. Automated Labelling Tool

The hybrid image segmentation tool, previously discussed, is implemented to provide an adequate estimate of the glottal area in a set of HSV frames, selected from the color HSV video, during VF vibration. This set of segmented frames form a training dataset on which a DNN is trained to accurately segment the glottal area during connected speech in different phonatory events. Hence, instead of using manual labeling to create the training data, the hybrid image segmentation method is utilized as an automated labelling tool. This is done by applying the hybrid technique to the kymograms and locating the glottal edges at different cross sections along the VF length; then, these detected edges are registered back in the HSV frames where the glottal area is identified. These segmented HSV frames during VF vibration are used as automated, labelled data for the purpose of training a segmentation network (DNN).

2.6.2. Segmentation CNN Architecture

The U-Net architecture is used, which is a fully convolutional neural network architecture. U-Net was introduced by Ronneberger and colleagues in 2015 as an image segmentation tool, particularly in the biomedical imaging field (Ronneberger et al., 2015). This network is a U-shaped network comprising of two parts: encoder and decoder. Figure 2.3 illustrates a schematic diagram of the DNN that is used in this work, which shows the proposed U-Net architecture based on the work of Ronneberger et al. The network is implemented using 64-bit MATLAB (MathWorks Inc., Natick, MA). As shown, panel (a) illustrates the general encoder-decoder design of the U-Net. Panel (b) displays the detailed structure of the network. An example of a network input (HSV frame) and output (segmentation mask) is shown in the figure in which the captured glottal area is automatically highlighted in white color. The HSV frames serve as the input. The feature maps dimensions are shown in the figure as well such that the first and the second entries represent the height and the width of the image while the third entry is the number of channels. For example, convolution (256, 256, 64) is an image with a height and a width dimension of 256 pixels along with 64 different channels.

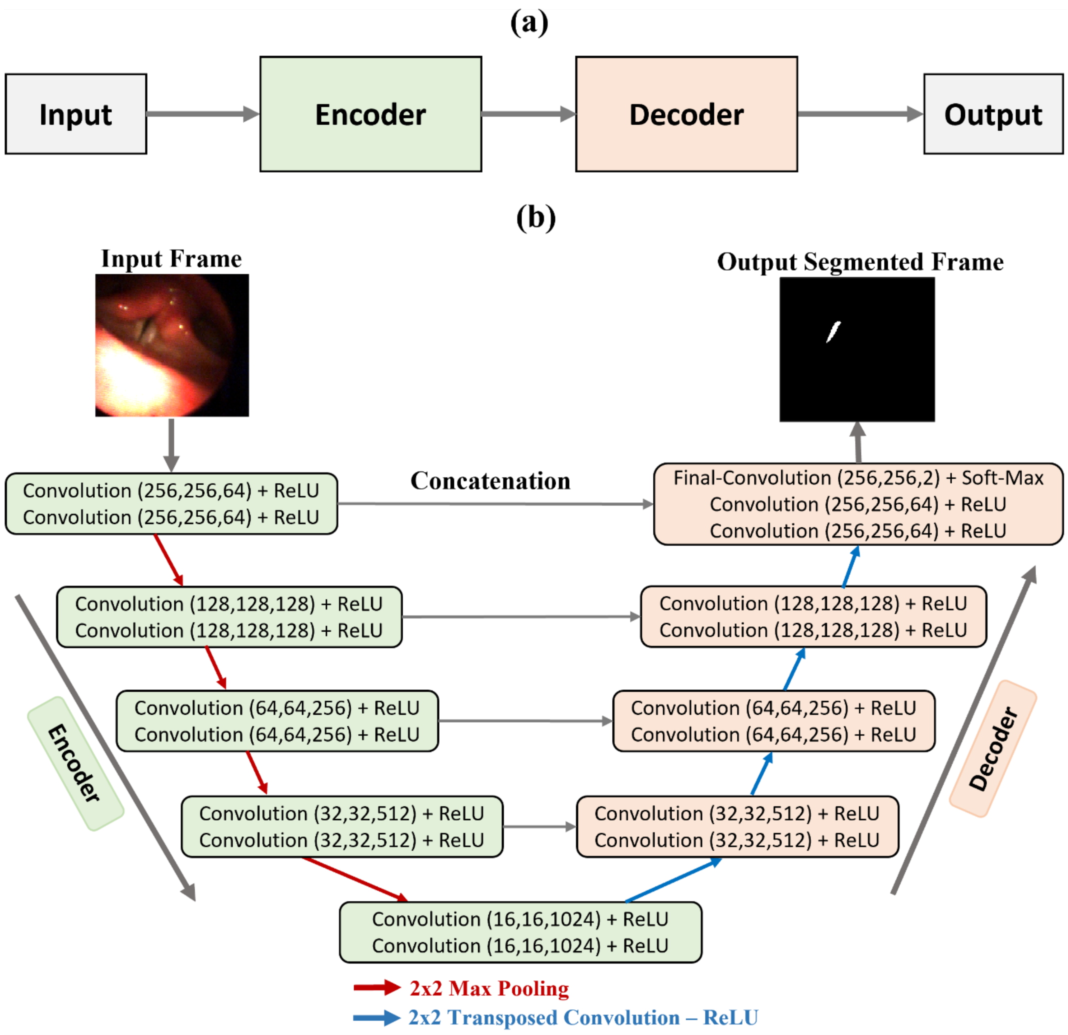

(a)
Input
Encoder
Decoder
Output
(b)
Input Frame
Output Segmented Frame
Concatenation
Convolution (256,256,64) + ReLU
Convolution (256,256,64) + ReLU
Convolution (128,128,128) + ReLU
Convolution (128,128,128) + ReLU
Convolution (64,64,256) + ReLU
Convolution (64,64,256) + ReLU
Convolution (32,32,512) + ReLU
Convolution (32,32,512) + ReLU
Convolution (16,16,1024) + ReLU
Convolution (16,16,1024) + ReLU
Convolution (32,32,512) + ReLU
Convolution (32,32,512) + ReLU
Convolution (64,64,256) + ReLU
Convolution (64,64,256) + ReLU
Convolution (128,128,128) + ReLU
Convolution (128,128,128) + ReLU
Final-Convolution (256,256,2) + Soft-Max
Convolution (256,256,64) + ReLU
Convolution (256,256,64) + ReLU
Encoder
Decoder
2x2 Max Pooling
2x2 Transposed Convolution – ReLU

Figure 2.3. Schematic diagram for the proposed deep neural network. Panel (a) shows the general encoder-decoder architecture of the U-Net. Panel (b) illustrates the detailed structure of the network. An example of a network input (HSV frame) and output (segmentation mask) is shown in the figure in which the captured glottal area is shown in white. The input image is first processed in the encoder (highlighted in green) through several layers of 3x3 convolutions-ReLU units. After each two layers of convolution-ReLU, a max pooling 2x2 layer was included (represented as red arrows). The outcome features of the encoder are then upsampled and processed within the decoder part (highlighted in orange) by several layers of convolution-ReLU, followed by transposed convolutions 2x2-ReLU (the blue arrows) for the downsampling. The last layer includes a final convolution, followed by a soft-max layer. The residuals (shown in light gray arrows) of each downsampled stage are concatenated from the encoder to decoder. The feature maps dimensions are shown in the figure.

The input images is first processed in the encoder (in green) which encodes the images into feature representations by extracting their main spatial features and context. The images in this contracting path is downsampled through multiple layers of 3x3 convolutions along with ReLU layers. The feature maps are then kept in the memory for latter concatenation with the decoder before performing a down sampling step (light gray arrows). After each two layers of convolution-ReLU, a max pooling 2x2 layer is included (red arrows) for downsampling to reduce the input image size while preserving the most predominant image features. This is done by a sliding window with a size of 2x2 pixels, where only the pixels with the maximum value in this window are considered. The extracted features from the encoder are then propagated and upsampled during an expansive path (the decoder in orange) by several layers of convolution-ReLU, followed by transposed convolutions 2x2-ReLU (the blue arrows) for upsampling. The encoder layers reconstruct/recover the dimension of the feature maps to ultimately match the input image resolution. After four decoder stages of upsampling, the last layer includes a 1x1 final convolution followed by a soft-max layer to classify each image pixel into two classes based on the resulted image features: a glottal area pixel (a value of 1) or a background pixel (a value of 0).

2.6.3. Segmentation CNN Training

To train a DNN, an optimization technique is used to tune the network parameters that yields the minimum difference between the predicted outcome (by the developed network) and the expected outcome (by the ground-truth data). Adam optimizer is considered to train the network. The deep learning network (U-Net) is first trained on a training dataset, which is created using the automated labelling tool (the hybrid method). The training dataset includes 2,050 automatically labeled/segmented frames and is selected from the color HSV recording. These segmented frames are evaluated through visual inspection to validate the accuracy of the hybrid labeling tool prior to training of the neural network. 20% of the training data are used as a validation dataset to evaluate the performance of the network during the training process and, accordingly, tune the network parameters to enhance its performance.

Because running speech imposes excessive laryngeal movements and frequently alters the spatial location of the VFs across the HSV frames, data augmentation is applied to the training images before training. This allows for a more generalized accurate model that can adapt to the variability in connected speech. Accordingly, the training images are translated with random shifts in both the vertical and horizontal directions, downscaled and upscaled, and randomly rotated.

These modified training images are used to train the constructed network, with a batch size of 10 for a maximum of 20 epochs. The trained network output is a binary image (segmentation mask) with the same dimensions as the input frames where only the pixels located in the glottal area are in white color while the other pixels are in black.

2.6.4. Segmentation CNN Evaluation

Several networks are trained with different architectures and training parameters. Networks with different number of encoder-decoder stages are tested (ranging from 3 to 6 stages which refers to the number of times the input frames are downsampled/upsampled). Different batch sizes of 4, 10, 16, and 32 are also considered during the training of these networks. The segmentation performance of each of the trained networks are evaluated against a testing dataset, where the best-performing network is determined based on the highest segmentation accuracy scores. The testing dataset is comprised of manually labelled HSV frames. This dataset is created using 600 frames from different phonation events including sustained VF vibration, onsets/offsets of phonation, and when VFs are not vibrating. These frames are different from the training images. The glottal edges in these frames are manually segmented to serve as ground truth by an expert. Four measures are used for assessing the performance of the trained networks on the test set: accuracy, Intersection over Union (IoU), Dice Coefficient (DC), and Boundary-F_1 (F_1) score. The first three metrics (accuracy, IoU, and DC) are used to assess the segmentation quality in segmenting the glottal area while the F_1 score is computed to assess the accuracy in detecting VF edges (glottal boundaries). The segmentation accuracy is calculated using the following equation:

$$Accuracy = \frac{\mathrm{TP}}{\mathrm{TP + FN}}, \qquad (2.17)$$

where TP refers to the true positive pixels which is the number of correctly predicted pixels inside the glottal region; in contrast, FN is the number of pixels incorrectly predicted as background (not glottal area pixels). The IoU metric is determined along with the accuracy to provide a more robust performance evaluation including the assessment of those pixels that are incorrectly classified. The IoU ranges between 0-1; a value of 0 refers to no overlap between the predicted glottal area and the ground-truth while a value of 1 refers to a perfect similarity between the estimated and ground-truth glottal area. IoU is calculated according to the following equation:

$$IoU = \frac{\mathrm{TP}}{\mathrm{TP + FP + FN}}\,. \qquad (2.18)$$

DC is additionally computed to compare the segmentation results of this model against other models in the literature that used DC as an assessment measure [142]. The equation used to compute the DC score is as follow:

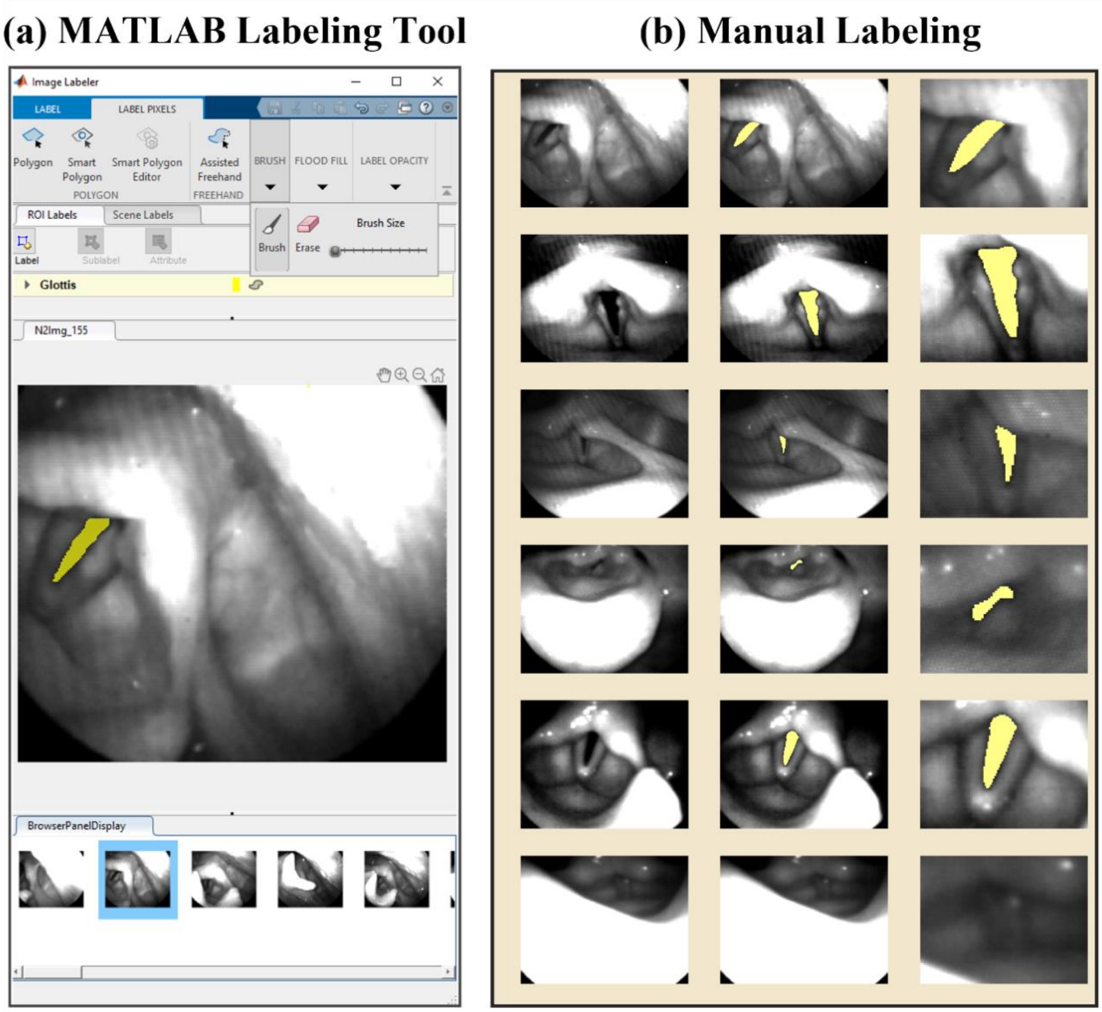
$$DC = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}\,. \qquad (2.19)$$

In addition to the above evaluation metrics, the F_1 score is calculated to assess the accuracy of the boundaries and edges of the glottal area and VF. The F_1 score allows us to measure how the predicted glottal/VF edges accurately match the ground-truth one. The F_1 score is calculated using the following equation:

$$F_1 = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP}+\mathrm{FN})}\,. \qquad (2.20)$$

2.6.5. Segmentation CNN Application to Monochrome HSV Data

The best-performing trained network with the highest segmentation accuracy scores on the vocally normal color HSV video is applied to the monochrome HSV recordings from both vocally normal and AdLD individuals (subject pool two). To do so, the network with the same architecture should be first retrained on these new monochrome images in order to take into account and adapt to the changed environment of the new frames. Hence, the network is retrained on 4,500 HSV monochrome images selected from HSV recordings belonging to vocally normal and AdLD speakers. These frames are selected during VF vibrations, phonation onsets/offsets, spasms of VFs in AdLD patients, no voicing (no VF vibration), and even obstructed views of VF (due to, e.g., epiglottis or arytenoid cartilages motion). The glottal area in these training frames is manually segmented using a MATLAB labelling tool (“Image Labeler”) as can be seen in Figure 2.4. Panel (a) includes a screenshot of the tool showing several options for annotation and an example of a labeled image in which the glottal area pixels are highlighted in yellow. Panel (b) illustrates 6 manually segmented HSV frames (before and after segmentation besides a zoomed-in view) as a result of using the labeling tool.


(a) MATLAB Labeling Tool

(b) Manual Labeling

Figure 2.4. MATLAB labeling tool used for manual segmentation of HSV frames (panel(a)); results of applying the labeling tool to manually segment the glottal area (highlighted in yellow) are shown for six different frames (panel (b)). The first and second column in panel (b) show the frames prior to and after the manual labeling; the third column shows the zoomed-in segmented frames.

The network is evaluated again to make sure that it adapts to the new monochrome HSV frames with adequate segmentation quality; hence, the retraining parameters (particularly batch size and number of epochs) are fine-tuned. This is done by testing the network on 1000 manually segmented frames that are not a part of the retraining process and using the four assessment measures discussed before for evaluation (accuracy, IoU, DC, and F_1 score). After fine-tuning and evaluating the network performance, it is implemented on the entire HSV videos of the participants (both vocally normal speakers and AdLD patients) to extract the change in the glottal area (glottal area waveform) during phonation. The glottal area and its edges are estimated spatially in each frame of the HSV recordings based on the spatial segmentation resulted from applying the DNN. The integral of the areas enclosed by the VF edges is considered as an estimate for the glottal area. The GAW is then determined as the sequence of the estimated areas across the image frames in each recording.

To automatically detect the left and right VF edges in HSV frames based on the extracted glottal area, the first image moment of inertia is computed for each row of the detected glottal area pixels (glottis center) on each HSV frame. Finding the first image moment localizes the center of brightness in the segmentation mask. After obtaining the center of brightness (center of glottal area) on each row of the image, a set of scattered points are attained and depicted on the original HSV frame. Figure 2.5. illustrates the sequential steps to obtain the glottal midline. The figure shows an original HSV frame along with its automatically detected segmentation mask (in black and white) which is obtained using the DNN tool. The segmentation mask shows a zoomed-in view for a better visualization of the segmented glottal area shape. The figure includes another image showing the original HSV frame overlaid with the detected glottal area besides another copy of the image showing the detected points that are related to the center of the glottal area.

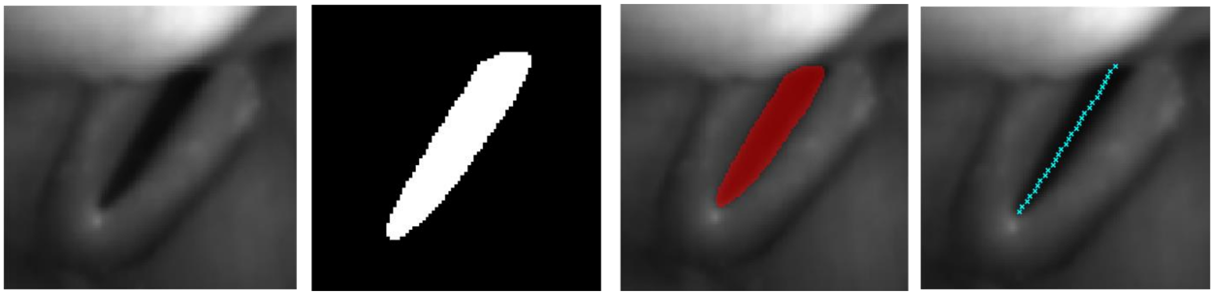


Figure 2.5. A schematic diagram of the required steps in detecting the glottal midline including sequence of images showing (from left to right) an original HSV frame, the segmentation mask (a zoomed-in view in black and white), automatically segmented glottal area (in red), and glottal midline (in cyan color).

After detecting the midpoints of the glottal area, the corresponding midline is predicted. The midline is estimated as a fitted second-order curve in order to accurately represent the glottal center. Based on the detected glottal midline (the fitted curve), the spatial locations of the left and right VF edges are automatically determined with respect to the location of the midline.

2.7. Study IV: Automated Measurements of Glottal Attack and Offset Time

The purpose of this study is to fulfill Aim 4 by addressing the following:

Q4: Are the glottal attack and offset times different between AdLD and normal controls?

H4.1: An automated algorithm can be developed to measure GAT and GOT with comparable accuracy to visual measurements.

H4.2: GAT and GOT will be significantly higher in AdLD versus normal controls.

H4.3: GAT and GOT will show more variability in AdLD subjects.

The successful implementation of the previous studies for spatial segmentation and HSV analysis allows to automatically capture the glottal area and VF edges in the HSV recordings during running speech. This facilitates the extraction of useful HSV-based measures associated with VF dynamics. The proposed HSV-based measures are chosen so that they can portray the VF dynamics during phonation onset and offset. The extraction of these measures helps fulfill Aim 4 by determining GAT and GOT. Noting that these analyses of generating HSV-based measurements are implemented using the monochrome data (subject pool two). By using this dataset, the measures are extracted from the HSV recordings during running speech for the sake of addressing the research question related to Aim 4. The measures include GAT and GOT. These two measures are automatically generated during the onset and offset of phonation through the extracted GAW and the segmented VF edges.

Measuring GAT and GOT is important to be studied as they are critical factors in investigating the pathophysiology of voice disorders [153]. GAT and GOT are physiological measurements that correspond to VF vibrations. GAT is defined as the time difference between the first oscillation and the first contact of the VFs. GOT represents the time difference between the last contact and the last oscillation of the VFs. Those two measures are mainly computed based on the change of the spatial location of the VF edges during phonation onsets and offsets. Hence, they are directly calculated from the HSV recordings after the successful implementation of the spatial segmentation tools, discussed before. It is also aimed to generate these two measures automatically from the HSV videos and validate them against visual analysis.

GAT is measured as the time delay between the rise of the energy of VF oscillation and VF contact during the onset of phonation. To do so, the normalized GAW and the average medial glottal contact waveforms are calculated from the GAW and the detected edges of the VFs. The energy of the two waveforms are then computed using a sliding window with a size of 30 ms. The GAW is defined as the area, measured in pixels, between the left and right VF edges. The average medial glottal contact waveform is identified as the average number of points (pixels) located along the VF length that are in contact. The energy of the GAW increases at the beginning of VF oscillation, and the energy of the average medial glottal contact waveform rises sharply with VF contact. Therefore, the delay between the two energy waveforms can determine the GAT. The GAT is computed as the unbiased delay between the first derivative (with a time step of 25ms) of the two energy contours, using cross-correlation technique. The main reason to select the cross-correlation method is that this technique provides an unbiased measurement of the time delay without any predefined thresholds/conditions – free of any operator intervention or bias (fully automated) [153].

Similar to the computation of the GAT, GOT is determined using a similar procedure. GOT represents the time delay between the drop of the energy of VF oscillation and VF contact during the offset of phonation. So, the GAW and the average medial glottal contact waveform are computed similar to the way they are computed for the GAT as discussed above. In contrast, during phonation offset, the energy of the GAW drops and dissipates after the last VF contact since the oscillation of the VFs damps, and the energy of the average medial glottal contact waveform drops sharply since the VFs start to be separated after the last VF contact. The GOT is also calculated as the unbiased delay between the first derivative of the two energy waveforms (with the same time step as in GAT) using again the cross-correlation method.

In order to fulfill Aim 4 and address its hypotheses, GAT and GOT are automatically computed using the above approach for the monochrome HSV recordings (subject pool two) from both the vocally normal speakers as well as the AdLD patients. The measurements are obtained using both the proposed automated algorithm and the visual analysis. In the visual analysis, three raters analyzed the HSV data from each participant using a camera playback software (Phantom Camera Control, PCC) where they can adjust the playback speed, gaussian filter setting, gain, brightness, and contrast of the video frames for a better visualization and better image quality [162]. The raters visually determined the timestamps corresponding to the first oscillation and contact frames as well as the last oscillation and contact frames for each vocalization [162]. Based on the difference between these timestamps, they computed the GAT and GOT measured in number of frames and in ms (dividing the number of frames by 000 fps). The raters compared and reviewed the vocalizations that showed large error between the raters (the GAT and GOT measures that resulted in more than 2.5ms for the contacts and 5ms for the oscillations). This was done to allow the raters to come to a consensus about their measurements. The visual and automated analysis of GAT and

GOT of the HSV recordings are compared to each other in order to validate the accuracy of the automated analysis in detecting the GAT and GOT.

After validating the automated approach, the mean and standard of deviation values of GAT and GOT are computed in each HSV recording during the reading of the six CAPE-V sentences as well as the Rainbow Passage. The goal of the analysis of these measurements to demonstrate how the GAT and GOT measures are affected by the impaired vocal function of AdLD disorder during the onset and offset of phonation in running speech. Hence, a statistical analysis is used to investigate whether there is any significant difference between the GAT/GOT measures of the AdLD group in comparison with the normal controls. The GAT and GOT measurements are considered as continuous dependent variables in the statistical model. The group type (vocally normal subjects vs AdLD patients) is considered as a categorical independent variable. To test the proposed hypotheses of Aim 4, the t-test is conducted to compare the GAT/GOT between the non-pathological group and AdLD groups.

2.8. Study V: Lumped-Element Modeling

The purpose of this study is to fulfill Aim 5 by addressing the following:

Q5: Can a simplified one-mass model be optimized to accurately match the vibratory behavior of VF extracted from HSV?

H5.1: A simplified one-mass model can successfully simulate both the vibratory and closure phases of VF motion.

H5.2: The particle swarm optimization technique will enable accurate optimization of the model to predict the experimental glottal area waveform.

H5.3: The optimized model parameters, obtained through inverse analysis of HSV data, can estimate the VF mass, elasticity, and viscosity indices.

This section introduces the last study in this dissertation in order to address Aim 5 and its related hypotheses. The study investigates the feasibility of generating biomechanical characteristics of VF vibrations. The dynamic vibration of the VF (the experimental data) is obtained based on the monochrome HSV data analysis. This is done by combining the extracted VF vibration with a model-based approach. Hence, biomechanical measures are generated based on a vocalized segment during VF vibration in running speech as a proof of concept. In order to develop these measures, HSV data are first processed using the spatial segmentation techniques proposed in the previous sections. The main goal of these video analysis techniques is to spatially capture the change in the glottal area during phonation across the video frames. Then, a one-mass lumped-element model is built [96, 97, 98, 99]. This model is designed such that each VF is described by one rigid mass coupled by a spring and a damper. The main model parameters associated with the VF properties are the masses, stiffness of the springs, and the damping parameters [95]. The model is combined with the resulted HSV analysis such that the extracted glottal area is utilized to optimize the model parameters. The model is optimized so that it can generate an oscillation behavior similar to the extracted glottal area during VF vibration from the HSV data.

In the present work, a one-mass model, in comparison with multi-mass models – was considered as being sufficient to simulate the glottal area variations during the VF vibrations observed in the HSV data. That is, the extracted VF motion from the HSV images only represents a two-dimensional oscillation (the opening and closing along VF length) and does not display the contact area between the VFs along its thickness during the closure phase. Accordingly, increasing the number of masses to mimic the contact area of the VF oscillation was not necessary in the present work, and using a one-mass model to simulate/optimize the VF movement is a reasonable assumption. In other words, simulating the first mode of vibration using a one-mass model is enough to capture the observed VF vibrations in HSV. This mode of vibration of the VF represents a specific vibratory pattern and shape in which every tiny part of the VF tissue oscillates sinusoidally at the same frequency, which refers to the fundamental frequency of VF vibration [163].

For optimization of the model with the experimental HSV data, different techniques can be used, e.g., particle swarm optimization and genetic algorithm [93, 103]. In this work, particle swarm approach is used to optimize and approximate the model parameters. This optimization algorithm has high efficacy in solving different applications, particularly in the field of biomechanics [93]. It is simple and can determine the optimal model solution using a few parameters. After optimizing the model, the main model parameters are quantified in order to estimate the corresponding biomechanical measures of the vibrating VF.

Each aforementioned step is discussed in detail in the following subsections. The first subsection (2.8.1) includes the model description, governing equations, mathematical representation of the parameters, and the time-integration method used for the numerical solution. The second subsection (2.8.2) describes the input parameter values and initialization of the model parameters. The last subsection (2.8.3) discusses a detailed description of the optimization procedure including a description of the utilized experimental data, optimized parameters, the optimization procedure, and the optimization output.

2.8.1. Biomechanical Modeling

In this study, a one-mass lumped model is implemented as the first step to simulate the VF vibrations. Figure 2.7 illustrates a schematic diagram of the lumped-element model that is utilized in the present work. As can be seen, the lumped model consists of a one mass (representing one VF), a spring (referring to VF elasticity), and a damping element (to simulate VF viscosity). In the illustrated diagram, the mass of the VF is denoted by *m*, VF elasticity is represented by *k*, and the damping coefficient of the VF is given by *c*. The diagram includes several other parameters that are used to describe the mathematical model. The subglottal pressure (*Ps*), the inlet glottis pressure (P_1), and the outlet glottis pressure (P_2) are shown in the figure with their relative locations. The glottal air flowrate, directed from the lungs, passing through the glottal constriction, toward the vocal tract is represented by Q_g. This mathematical model allows us to simulate the movement of the VF mass in one dimension (*x(t)*). In this model, VF thickness is indicated by *d*, the length of the VFs is denoted by *l*, and the width between the two VFs is identified by *w*.

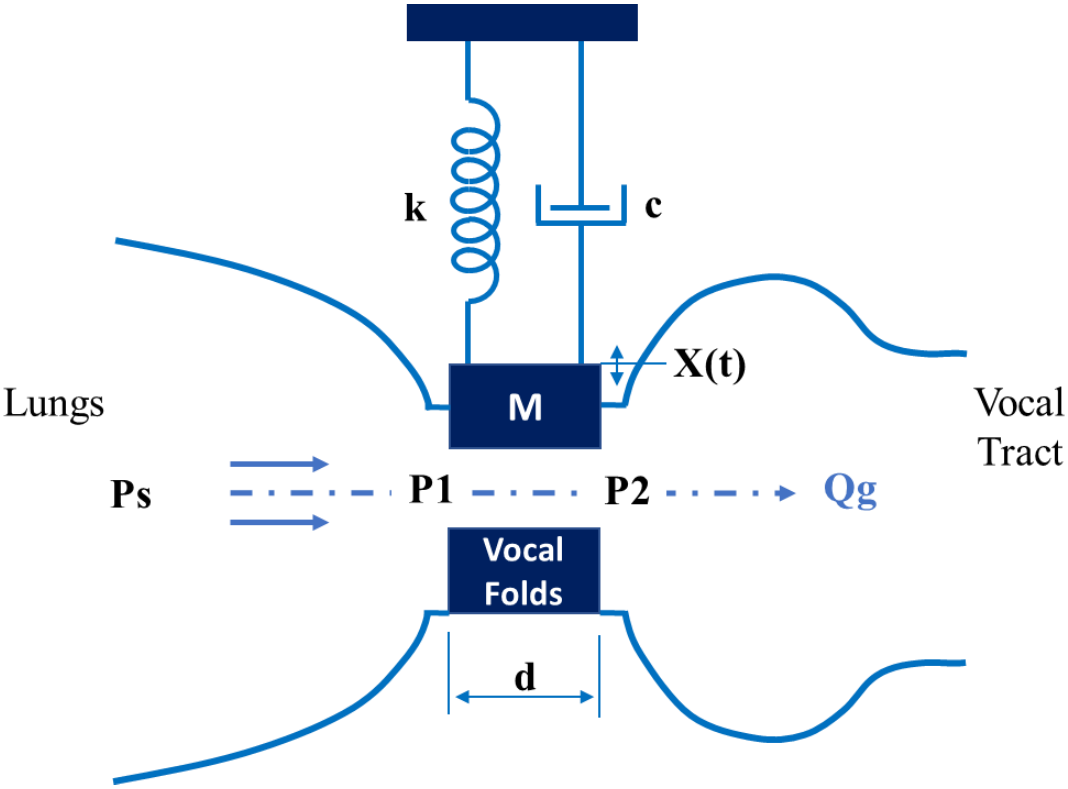


Figure 2.7. A schematic diagram of a one-mass lumped model to simulate VF vibration.

Below is a detailed explanation of each step to drive the mathematical equations used in this work. The mathematical representation and the differential equation of motion of the above oscillating system are mainly derived from the Newton's second law of motion:

$$\mathcal{F} = m\ddot{x}. \qquad (2.21)$$

The acceleration ($\ddot{x}$) of the VF depends on two variables: the net force $\mathcal{F}$ acting upon the VF and the mass *m* of the VF. From equation (2.21) and the schematic diagram showed above, the governing equation of the system can be derived as follows:

$$\sum \mathcal{F} = -c\big(t, x(t)\big)\dot{x} - kx + F\big(t, x(t)\big) = m\ddot{x}, \qquad (2.22)$$

where *F(t, x(t))* indicates the external forces that act upon the VF during vibration as a function of time *t* and displacement *x(t)*. Equation 2.22 can be written as:

$$m\ddot{x} + c\big(t, x(t)\big)\dot{x} + kx = F(t, x(t)). \qquad (2.23)$$

To derive the above equation (2.23) per unit mass of the VF, both sides of the equation can be divided by the VF mass. As a result, the elasticity index and the viscosity index associated with the vibrating mass can be obtained. These two indices are computed from the following formulas:

$$Elasticity\ Index = \frac{k}{m} \text{ and} \qquad (2.24)$$

$$Viscosity\ Index = \frac{c}{m}. \qquad (2.25)$$

These two parameters are considered as biomechanical output measures of the model in the present study. The Elasticity Index is related to the displacement trajectory of the VFs in the HSV recordings. The Viscosity Index is another important measure to study the biomechanics of VF oscillation. The viscosity of VF tissue is a biomechanical property that measures resistance to the velocity of VF tissue deformation [164]. With higher VF viscosity, the VF oscillations would be expected to be more damped, and a greater subglottal pressure with larger air force would be expected to maintain the same vibratory behavior of VF [164]. After defining the main model equation and its parameters, the only term that needs to be defined in the above equation is the external forces that act upon the VF. The external force *F(t, x(t))* is obtained based on the inlet glottal pressure (P_1), which is right before the glottal airflow enters the glottis, and the outlet pressure (P_2), which is right after the glottal airflow exists the glottis (see Figure 2.7). The following formula (2.26) is used to obtain the external force as a function of the inlet and the outlet glottal pressures P_1 and P_2 [165]:

$$F(t, x(t)) = \frac{1}{2} ld(P_1(t, x(t)) + P_2(x(t))). \qquad (2.26)$$

Experimental measurements have shown that P_1 and P_2 can be obtained from equation 2.27 using the P_S and Bernoulli pressure P_B [166]:

$$P_1(t,x(t)) = (P_S(t,x(t)) + 1.37P_B(t,x(t))), \qquad (2.27)$$

$$P_2(x(t)) = -0.50P_B(t,x(t)).$$

P_B in the above equation can be derived from the Bernoulli equation. That is, if the glottal airflow through an orifice (referred to the VFs) is ideal, lossless, and steady, the glottal inlet and outlet pressures will be identical and equal to P_B [167]. P_B can then be defined as the kinetic energy per unit volume attributed to the glottal airflow (Q_g) and can be computed from the following formula:

$$P_B(t,x(t)) = \frac{\rho {Q_g}^2(x(t))}{2A_g^2(x(t))}, \qquad (2.28)$$

in which A_g(t) represents the glottal area (the space between the two VFs). Q_g is computed using an empirical formula. This experimental formula was obtained by van den Berg and others [168], where they empirically estimated the resistance of the glottal airflow (P_s/Q_g) in an orifice. The following formula represents the derived empirical equation for determining Q_g as a function of $x(t)$ and P_s.

$$\frac{0.875\rho}{2d^2w(x(t))^2}Q_g(x(t))^2 + \frac{12\mu l}{dw(x(t))^3}Q_g(x(t)) - P_S(t,x(t)) = 0. \qquad (2.29)$$

Two air properties are included in the above equation, which are the air density ρ and the coefficient of viscosity μ. In this equation, w contributes to the nonlinearity of the model and depends on the trajectory and the displacement of the VF masses. Plugging in equations 2.27, 2.28, and 2.29 into equation 2.26, it can be seen the external force $F(t, x(t))$ is a function of the flowrate, which itself depends on the VF displacement $x(t)$. Moreover, as can be seen from equation (2.29), the formula is a quadratic equation, where the variable Q_g is the unknown variable. The positive root resulted from solving this quadratic equation is only considered to obtain the flowrate value. In order to compute $A_g(x(t))$, a resting position of the VF is defined such that when the displacement $x(t)$ of the mass is zero, the glottal area corresponds to the initial area A_{g0} between the two VFs, which is a constant. Accordingly, the change of the glottal area during the VFs vibration can be calculated using the following formula:

$$Ag(x(t)) = Ag_0 + lx(t). \qquad (2.30)$$

Another parameter that we define is the critical displacement X_c, which is the displacement at the initiation of the closure phase:

$$X_c = -\frac{A_{g0}}{l}. \qquad (2.31)$$

The closure phase in this study refers to when the glottal area is equal to zero in the experimental data (when the VFs are in contact). When the mass displacement reaches and exceeds a predefined critical displacement value X_c, a complete glottal closure happens, where the two VFs come into contact with each other. During the closure phase, when the mass exceeds the critical value X_c, the glottal area A_g(x(t)) becomes zero and, theoretically, it turns into a negative value. It should be noted that the glottal volume flowrate Q_g also becomes zero during the closure.

The derived differential equation and the defined variables and constants corresponding to the introduced lumped model can be summarized and rephrased into the set of the following equations:

$$
\begin{gathered}
m\ddot{x} + \mathrm{c}(\mathrm{t}, x(t))\dot{x} + kx = \mathrm{e}_1 P_S(t, x(t)) - \mathrm{e}_2 P_B\big(t, x(t)\big), \\
\mathrm{e}_1 = \frac{ld}{2}, \qquad (2.32) \\
e_2 = 1.87\frac{ld}{2},
\end{gathered}
$$

where e_1 and e_2 are constants. The parameterized variables $c(t, x(t))$ and $P_s(t, x(t))$ as well as the nonlinear function $P_B(t, x(t))$ depend on the status of the glottis (either an open glottis or a closed glottis). These three variables are computed as follows:

If $X(t) \geq X_c$ (the open glottis, referring to the opening phase):

$$c(t, x(t)) = 0,\ P_s(t, x(t)) = \overline{P}_s,\ P_B\big(t,\ x(t)\big) = \frac{\rho {Q_g}^2(x(t))}{2(Ag_0 + lx(t))^2} \qquad (2.33)$$

If $X(t) \leq X_c$ (the closed glottis, referring to the closure phase) happened at a specific t, then $\mathrm{t} = \mathrm{t}_{0c}$, taken as the initial moment of closure, then

$$c(t, x(t)) = c',\ P_s(t, x(t)) = P_{Smax},\ P_B\big(t, x(t)\big) = 0 \text{ when } \mathrm{t} - \mathrm{t}_{0c} \leq \mathrm{t}_c \qquad (2.34)$$

$\overline{P}_s$ refers to the typical value of the subglottal pressure. P_{Smax} indicates the pressure level to which the subglottal pressure is increased, acting as a built-up pressure during the closure time. The time at which the beginning of the closure phase happens is indicated by t_{0c}, which depends on both $x(t)$ and t. The subglottal pressure is parameterized as $P_s = f(t, P_{Smax}, t_c, x(t))$. This parameterization of the subglottal pressure allows to represent the change of the subglottal pressure during vocalization. The parametrization was performed as a step function, which is a simplified representation of the actual variations that occur in the subglottal pressure during VF vibration. Also, according to the above equation (2.34), the damping coefficient c' is parameterized as a function of t and $x(t)$ to simulate the additional viscous damping occurring during the closure between the two VFs. Therefore, the above equations provide a mathematical representation of the model parameters including five input parameters m, k, c', P_{Smax}, and t_c; and the model output would be the predicted glottal area ($A_g(x(t))$). These input parameters are being optimized using the optimization procedure as explained in section 2.8.3.

Based on the derived mathematical representation during the closure phase, the external forcing function becomes .5($P_s ld$), computed from equation (2.26 and 2.27), such that the forces (mainly coming from the P_s) act on the mass to open the VFs. During the VF contact, there is an increase in P_s, which is built up to help in pushing the VFs apart by a value of P_{Smax} during the course of the opening phase. Another increase is considered during this time in the damping coefficient, where the damping coefficient c is increased by an additional viscous damping c' to simulate the overdamping contact between the VFs. When the VFs start to open ($x(t) > X_c$), the damping coefficient returns back to the value c and the forcing function changes back to being derived from P_s and P_B.

The rest of this section discusses the numerical integration of these equations in order to simulate the theoretical glottal area waveform $A_g(x(t))$ as the main output of the system. The above differential equation is solved using the classical Runge-Kutta approach (4th order) as an effective time-integration method to determine the theoretical trajectory and displacement of the vibrating mass, as well as the theoretical glottal area. To do so, the above differential equation is rephrased in a form of two 1st order differential equations, which can be formulated as follows:

$$\dot{x} = V, \qquad (2.35)$$

$$\dot{V} = \frac{1}{m}[-\mathrm{c}(\mathrm{t}, x(t))V - kx + \mathrm{e}_1 P_S(t, x(t)) - \mathrm{e}_2 P_B\big(t, x(t)\big)]. \qquad (2.36)$$

Overall, the Runge-Kutta method computes the solution and performs the integration by iteratively updating the approximation at each iteration (time step). This update is done using a weighted average function evaluation. The target of computation is to numerically integrate the above differential equation and obtain $X(t)$. For a clearer representation of the execution of the method, the two differential equations mentioned above are combined into a single function *dxdt* as follow:

$$dxdt(\mathrm{t, x}) = [x_2, \frac{1}{m}[-\mathrm{c}(\mathrm{t}, x_1)x_2 - kx_1 + \mathrm{e}_1 P_S(t, x_1) - \mathrm{e}_2 P_B(t, x_1)], \qquad (2.37)$$

where x refers to a solution vector including two values: x_1, indicating the displacement x(t); and x_2, indicating the velocity V. Considering these definitions, the method is implemented through the following steps:

I. Initial values of the model are first determined for the solution vector x for both the initial displacement and velocity at $t = 0$ and $x(0) = \{x_o, V_o\}$, where $x_1 = x_o$ cm and $x_2 = V_o$. In addition, the time step Δt of the numerical solution is kept at 0.25 ms (1/4000 s) in order to match the frame rate of the high-speed camera for recording the experimental data.

II. At each time step, intermediate slopes K_1, K_2, K_3, and K_4 are approximated. These slopes provide an approximation of the derivative of the solution at three different stages within the time step itself: at the beginning, middle, and the end of each time step Δt. They are evaluated using the derivative function *dxdt* at these three timestamps. The purpose of these slopes is to assist in capturing the variation in the derivative function that happens within the time step Δt; this results in a better accuracy in approximating the solution across the time steps. K_1 and K_4 refer to the slope at the beginning and end of the time step Δt, while both K_2 and K_3 indicate the slop at the midpoint (*Δt/x*), evaluated using k_1 and k_2. The variable *i* here is considered as a loop counter to iterate across the time steps. Below is the mathematical representation of the computations implemented for each derivative:

$$\begin{aligned} k1 &= \Delta\mathrm{t} * dxdt(\mathrm{t(i)}, \mathrm{x(i)}) \\ k2 &= \Delta\mathrm{t} * dxdt(\mathrm{t(i)} + \Delta\mathrm{t}/2, \mathrm{x(i)} + k1/2) \\ k3 &= \Delta\mathrm{t} * dxdt(\mathrm{t(i)} + \Delta\mathrm{t}/2, \mathrm{x(i)} + k2/2) \\ k4 &= \Delta\mathrm{t} * dxdt(\mathrm{t(i)} + \Delta\mathrm{t}, \mathrm{x(i)} + k3) \end{aligned} \tag{2.38}$$

III. The approximated solution at the next time step is updated according to the computed slopes using the weighted average. The weighted average of these slopes enables better estimation of the solution at the next time step according to the following formula:

$$x(i+1) = x(i) + (1/6) * (k1 + 2k2 + 2k3 + k4) \tag{2.39}$$

IV. The steps I and IV are repeated iteratively until reaching the specified time duration of the numerical integration and computing the solution vector at each time step– returning the approximated value of the displacement across time.

The above numerical integration method is implemented using 64-bit MATLAB R2020b (MathWorks Inc., Natick, MA) as a powerful platform for building such models. After computing the theoretical displacement of the mass through the numerical integration technique, the simulated glottal area waveform is determined as the model output using equation 2.30. In the next section, the simulation as well as the values corresponding to the model parameters are discussed.

2.8.2. Model Parameters Initialization

In this subsection, the model parameters initialization and assumptions are discussed in order to simulate the VF oscillatory behavior. The results of this simulation are discussed in Chapter 3.

The lumped-element model that is discussed above is simulated to produce theoretical displacements of the VFs. The main output of the simulation is the theoretical glottal area waveform A_g. This theoretical area is optimized afterwards with the experimental one, which is discussed later in the following subsection. In order to simulate the model, the following model constants are used [169]: $\overline{P_s}$ = 8000 *dyn/cm*^{2}, A_{g0} = 0.05 *cm*^{2}, μ = 1.86×10−5 *g/(cm*^{2}*.s)*, ρ = 1.2 × 10−3 g/*cm*^{3}, *l* = 1.4 *cm*, *d* = 0.3 *cm*. The CGS system of units is used in this study. The damping coefficient *c* is considered zero during the VFs vibration when the VFs are open, yet when the contact occurs, the damping coefficient is considered to be *c*'.

P_s is not considered constant during the simulation; instead, it is parameterized as a function of the following: the simulation time measured in *s*, the contact duration between the two VFs (t_c in *s*), and P_{Smax} referring to the brief built up of pressure during closure as discussed before. During the opening phase of the VF oscillations, P_s is fixed at 8000 *dyn/cm*^{2} (~800*Pa*), but, during the closure phase, it is increased to a maximum value at P_{Smax} and returned back to 8000 *dyn/cm*^{2} when the VFs start to open again. In order to numerically solve the model, the initial displacement and velocity are defined as follows: x_o = .01*cm* and V_o = 0. These values are considered based on previous simulation studies [92, 169].

For the model simulation, the values for the mass *m*, the damping coefficient *c'* during closure, the spring stiffness *k*, maximum subglottal pressure P_{Smax}, and the closure time t_c are set to the following values: 0.24 *g*, 500 *g/s*, 5000 *g/s*^{2}, 10000 *dyn/cm*^{2}, and 2.75 *ms*. These values are chosen based on previous studies [169]. The simulation is carried out using these values to generate results that reflected the model's behavior prior to the optimization.

2.8.3. Model Optimization

After the mathematical representation and simulation of the proposed on-mass model, the model is optimized with experimental HSV data. Hence, from the above description of the proposed biomechanical model, the oscillatory pattern of the VF vibration can be represented using a parameter vector or a set, denoted by *q* with six optimizing parameters: $q = (\alpha, m, k, c', tc, P_{Smax})$. Since the units of the simulated and the experimental glottal area waveforms do not match, the scaling factor α is used in the optimization process to minimize the difference between the two waveforms in terms of the amplitudes. In summary, the input model parameters used in the optimization process are as follows: the mass *m*, spring stiffness *k*, damping coefficient during the closure phase *c'*, closure time *tc*, and the maximum subglottal pressure P_{Smax}. The last two optimizing parameters t_c and P_{Smax} are used to compute the subglottal pressure. The optimizing parameters are utilized as inputs to the mathematical model in order to obtain the simulated glottal area waveform A_g as an output from the simulation.

The optimization procedure aimed to optimize the theoretical/simulated glottal area waveform generated using the model with the experimental glottal area waveform. The experimental glottal area waveform is extracted using HSV from a vocally normal participant during a vocalized segment, where the glottal area is computed at each frame. The experimental glottal area is automatically detected using the developed DNN tool in this dissertation for the glottal area segmentation [170].

To optimize the theoretical/simulated glottal area waveform and obtain a good match with the experimentally extracted area waveform, an objective function is computed by calculating the sum of squared error between the simulated and experimental glottal area waveforms at every time step (or frame). The objective function is normalized by the experimental glottal area. Equation (2.40) shows the objective function (Obj) formula that is considered in the current study:

$$Obj(q_i) := \frac{\sum_{n=1}^{\#frames}[A_{Model}(\mathrm{n} \cdot \Delta \mathrm{t}) - A_{HSV}(\mathrm{n})]^2}{\sum_{n=1}^{\#frames} A_{HSV}^2(n)}, \qquad (2.40)$$

where q_i denotes the different sets of the optimizing parameters: q_i, i = 1,2,…, N. As such, each set q refers to a parameter vector that includes potential values of the six optimizing parameters, mentioned above. A_{Model} is the simulated glottal area while A_{HSV} represents the glottal area extracted from the HSV recording. The Δt indicates the simulation time step (considered as 0.25 ms), used to obtain the time corresponding to a specific frame number n. In order to minimize the above objective function, the particle swarm optimization (PSO) technique [171] is employed to determine the optimum values associated with each optimized parameter (the optimum parameter vector q*). Below is a brief description of the optimization method and its implementation in the present study.

PSO is a population-based stochastic method, which was first inspired by the behavior of birds flocking [171]. This technique works based on a population of individuals (called particles). These particles navigate a designated search space in steps with adjustable positions and velocities through an iterative procedure. As such, each particle has a position, which represents a candidate solution for the optimization problem in the search space; candidate solution refers to a possible set of values for the optimizing parameters (q_i). Moreover, each particle has a velocity value (particle's momentum/movement), which determines the direction of the particle that it can be guided towards a better position (potential solution). This potential solution corresponding to each particle is evaluated with each iteration step using the above objective function to identify how accurate this particle's solution/position is. After this evaluation at a specific step, an updated velocity value is computed and assigned to each particle, on account of which they adjust their positions in the next iteration step. After updating the velocity and position of each particle, the algorithm reevaluates the adjusted particle's position and iterates. In this iteration process, the algorithm iteratively updates the velocity value of each particle based on several factors: the particle's current velocity value, the best position/solution found by the particle throughout the completed iterations (called local/personal best), the best-observed position/solution among all the particles of the entire swarm (called global best). Throughout the iterations, all particles can coalesce towards a location, referring to a possible solution, in the search space where it may reflect the optimum solution (optimum particles' positions). Another way of convergence can happen in PSO where the local or the global best approaches a predetermined local optimum (acceptable level of error/threshold).

In the present study, the PSO algorithm is implemented using some annotations and assumptions. The iteration number of the algorithm is denoted as j, $j = 1,2,\dots, J$ with total iteration number of J = 400. The swarm included a total of N = 200 particles, and each particle is identified by unique i-value (same notations as in 2.40). A potential position of a specific particle i at certain iteration step is represented by $q_i(j)$ including a vector of six potential values of the considered optimizing parameter. Similarly, the particle's velocity is denoted by $v_i(j)$. Personal best position of each particle and the global best position of the entire swarm are identified by p_b and g_b. The algorithm is implemented through of the following steps:

I. Initializing the swarm particles with specified initial positions and random velocities. Same initial positions are given to all particles at the first iteration step acting as initial guesses of the target optimizing parameter. In the present work, the initial values for the six optimizing parameters α, m, k, $\underline{c'}$, t_c, and P_{Smax} are set as 120, 0.1 g, 40000 g/s^2, 400 g/s, 5×10−3 s, and 15000 dyn/cm^2, respectively. In other words, the particles' positions are chosen as $q_i(1)$ = (120, 0.1, 40000, 400, .005, 15000) in the first iteration. Also, a specific range was assigned to each optimization parameter value to constrain the search space as follows (listed as the same order as the initial values): α = 110 – 130, m = 0.04 – 0.30 g, k

= 10000 – 60000 g/s^2, c' = 300 – 800 g/s, t_c = 3 – 8×10−3 s, and P_{Smax} = 8000 – 20000 dyn/cm^2.

II. Evaluating a fitness value from computing the objective function outcome (Equation 2.40) using each particle's position ($q_i(j)$) in order to assess the performance (the error) of the potential solutions. The fitness value is a scalar value (the inverse of the error), which represents the quality of the candidate solution – the larger the fitness value, the lower the error, and the better the solution.

III. Updating the local best position and the global best position: (1) If the evaluated current $q_i(j)$ is better than p_b, p_b is updated; (2) if current p_b is better than g_b, g_b is updated. These updated best values are then used for the subsequent iteration of the algorithm.

IV. Updating the current velocity and, accordingly, the position of each particle. The velocity is adjusted first based on the current velocity, distance from the best position, and distance from the global best position. This adjusted velocity value is then added to the current particle's position according to the following formula [171]:

$$v_i(j+1) = \mathrm{W}[v_i(j)] + Z_1 r_1[p_{bi} - q_i(j)] + Z_2 r_2[g_b - q_i(j)], \qquad (2.41)$$

where r_1 and r_2 indicate random numbers from 0 to 1. Three constants are included in the equation: W, Z_1, and Z_2 whose values are chosen according to the standard values in MATLAB. The first constant W represents the inertia weight is given by 0.1-1.1 as an adaptive value which is reduced gradually over iterations. The following formula is utilized to update the weight at each iteration in order to improve the optimization procedure for better convergency [172]:

$$W(i) = w_{max} - \frac{w_{max} - w_{min}}{J} i \qquad (2.41)$$

W_{max} and W_{min} are given by 1.1 and 0.1 representing the maximum and minimum weights considered in the optimization process. Also, J refers to the maximum number of iterations that is considered by 400, as mentioned before. Adjusting W influences the relative weight given to each of the two other constants (Z_1 and Z_2). Z_1 refers to a cognitive component to control the particle's tendency to move to its best position (given by 1.49), and Z_2 is the social component, which influences the tendency of the particle to move towards the global best position obtained among the entire swarm (given by 1.49).

V. Moving and updating the particles' positions based on the updated velocities. The new position of the particles is determined using the following formula [171]:

$$q_i(j+1) = q_i(j) + v_i(j+1) \qquad (2.42)$$

VI. Repeating steps II-V until the best solution found, the optimization procedure converges to a specified acceptable error level, or the termination criterion is met. The termination criterion in this work is getting to the maximum number of iterations (400) after which the algorithm stops and returns the best solution found including the optimum values of the optimized parameters.

After carrying out the optimization procedure, the six optimized parameters are obtained (donated by q^*). These optimum parameters including $q_{(2,6)}$, except the scaling factor, (m^*, k^*, c'^*, tc^*, P_{Smax}^*) are used as input model parameters to obtain the theoretical glottal area A_g^* from the simulation. A_g^* refers to the glottal area simulated using the optimized model parameters. The remaining optimizing parameter, the scaling factor, is used to obtain the optimized glottal area such that $A_{model} = (\alpha^*)A_g^*$, which then can be compared with the experimental one A_{HSV}. The optimized parameters are also used to estimate the Elasticity Index k^*/m^* and the Viscosity Index c^*/m^* as the biomechanical measures corresponding to the investigated VF vibration. Additionally, these optimized parameters were validated with the typical values found in literature.

CHAPTER 3: RESULTS

3.1. Study I: Automated Detection of Vocal Fold Image Obstructions

A sample of manually labeled frames from the training dataset is shown in Figure 3.1. The figure depicts the two different classes of frames considered during the training: "Unobstructed and Obstructed Vocal Fold". Under the "Unobstructed Vocal Fold" class, frames from different phonatory events at various gestures in connected speech are depicted such as during VF sustained vibration, phonation onsets/offsets, and no vibration. In contrast, the "Obstructed Vocal Fold" group displays different configurations of VF obstructions that were observed during running speech. The figure presents partial/full obstructions due to epiglottis, arytenoid cartilages, laryngeal constriction, false VFs, or when VFs fall outside the view of the endoscope.

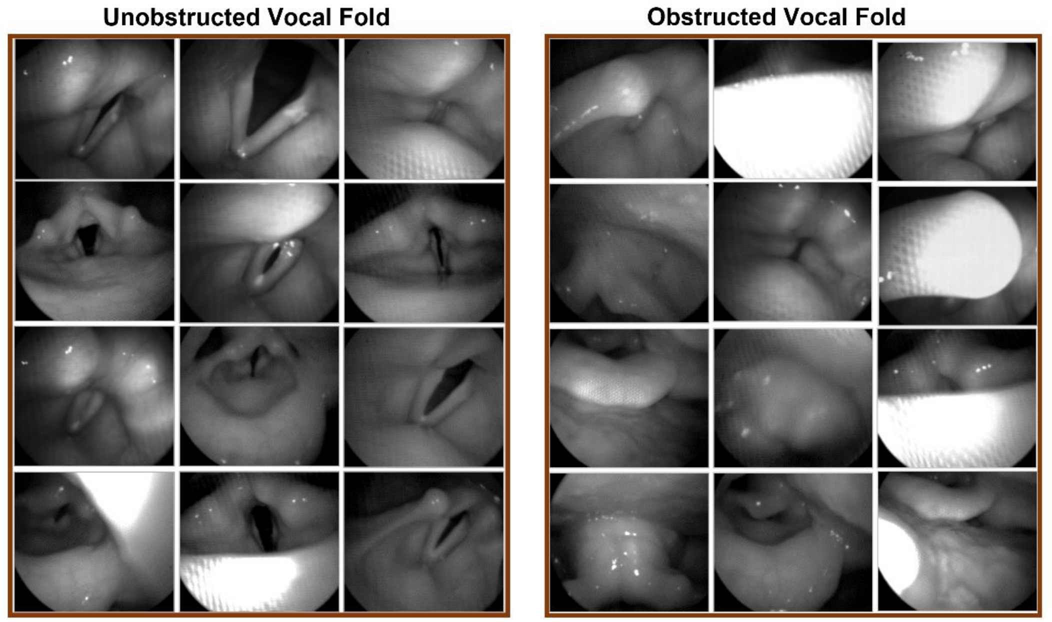


Figure 3.1. A sample of classified HSV images during connected using the manual analysis (visual classification). The two sets of three columns display the two different groups of frames: "Unobstructed Vocal Fold" showing the presence of the true vocal folds and "Obstructed Vocal Fold" demonstrating an obstructed view of vocal folds.

The results of applying the automated deep learning approach on the testing dataset is presented in Figure 3.2. A sample of random frames from the testing dataset is presented in the figure, which are classified and labeled by the developed automated approach. The figure depicts the classification outcome of the trained network for both "Unobstructed Vocal Fold" class (left side panels) and "Obstructed Vocal Fold" class (right side panels). For each class, different testing frames are shown displaying various unobstructed views of the VFs and different configurations of obstructed views of the VFs. Almost all the frames, included in the figure, show a correct classification of the developed tool. The figure includes only one misclassified frame in the "Obstructed Vocal Fold" class (the right-side frame in the second row), which was supposed to be classified in the "Unobstructed Vocal Fold" class.

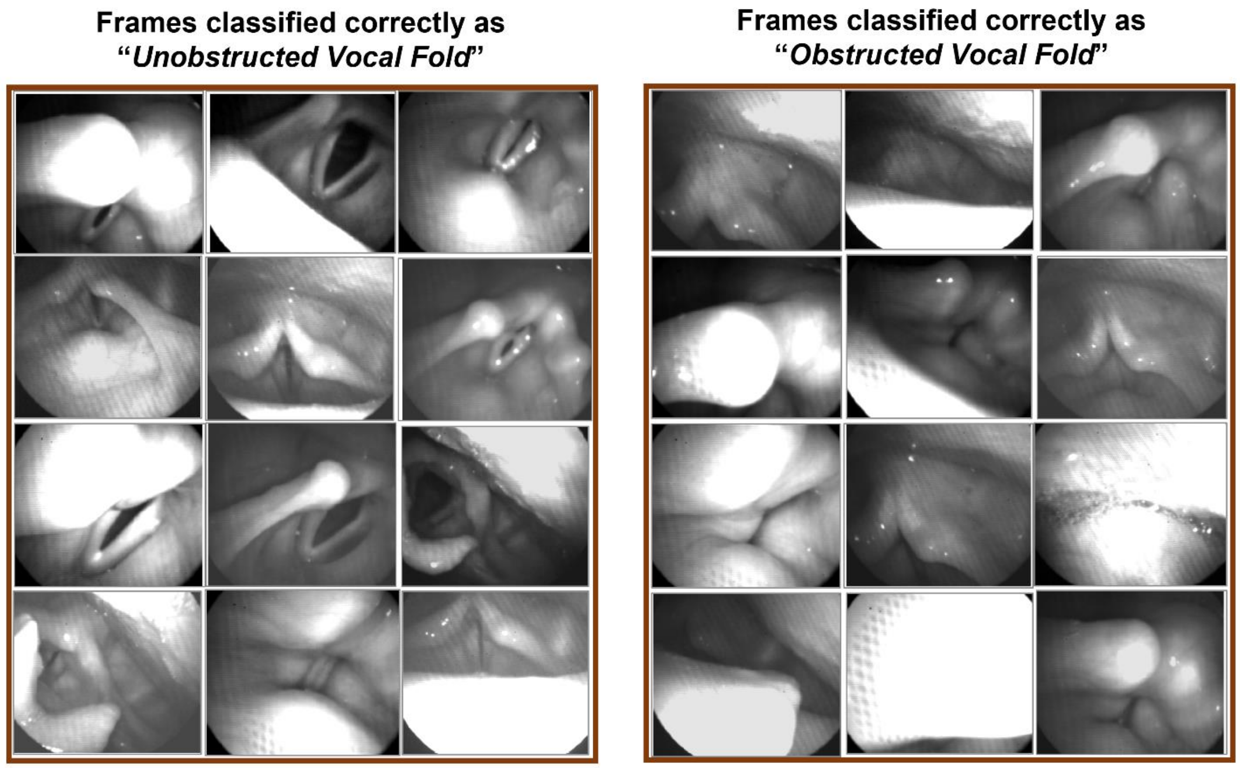
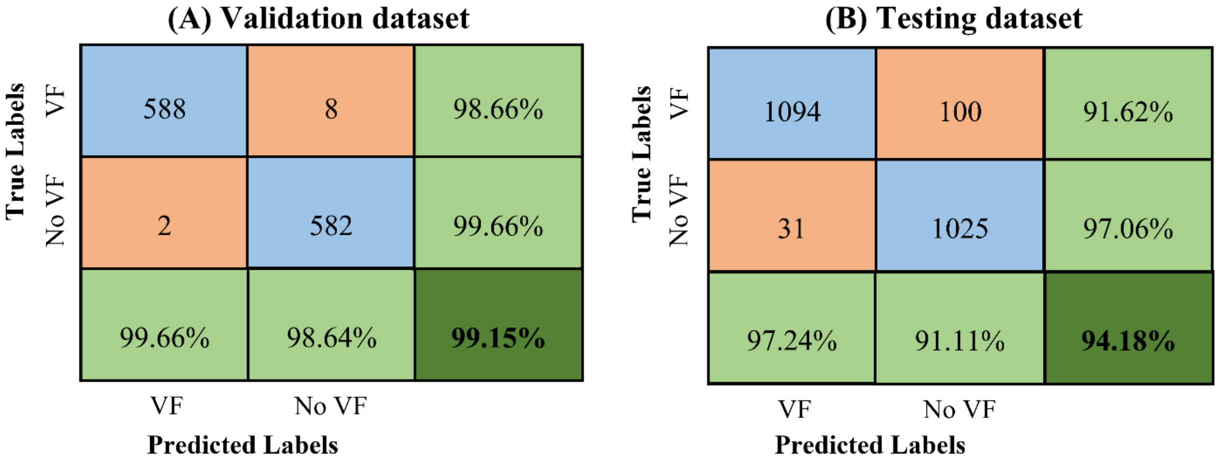

Frames classified correctly as
"Unobstructed Vocal Fold"
Frames classified correctly as
"Obstructed Vocal Fold"

Figure 3.2. The classification results using the automated deep learning approach on the testing dataset. The two sets of three columns display the correctly classified frames of the testing dataset as "Unobstructed Vocal Fold" (left side panels) and "Obstructed Vocal Fold" (right side panels).

The performance of the developed CNN is demonstrated using confusion matrices when the trained network is applied to the validation and testing datasets as shown in Figure 3.3. The horizontal and vertical labels in the two matrices represent the predicted outcome of the classifier and the true visual observation classes, respectively. The cells show the number of correctly classified (in blue) and misclassified (in orange) frames in the "Unobstructed Vocal Fold" class and "Obstructed Vocal Fold" class, which are represented in the figure as "VF" and "No VF", respectively. The associated accuracies of each classification are also represented in the green cells. As can be seen, the overall accuracies of detecting VFs in the validation and testing frames are 99.15% and 94.18% (shown inside the dark green cells). In both datasets, the network has a slightly higher accuracy in terms of recognizing the VFs in the frames than detecting an obstructed view of the VFs. This slight difference can be seen from the precision values of each class in the figure (in the two light green cells in the bottom row of the matrices): 99.66% versus 98.64% in the validating frames and 97.24% versus 91.11% in the testing frames, respectively. The two light green cells in the right columns of the matrices show the sensitivity and specificity of detecting VF obstruction in the frames in percent, which are 99.66% and 98.66% for the validation dataset and 97.06% and 91.62% for the testing dataset, respectively. Furthermore, the F1-score in the validation dataset was 0.99 for both the Unobstructed and Obstructed Vocal Fold classes, while these scores marginally dropped to 0.94 in the testing dataset.



(A) Validation dataset

True Labels \ Predicted Labels	VF	No VF	
VF	588	8	98.66%
No VF	2	582	99.66%
	99.66%	98.64%	**99.15%**

(B) Testing dataset

True Labels \ Predicted Labels	VF	No VF	
VF	1094	100	91.62%
No VF	31	1025	97.06%
	97.24%	91.11%	**94.18%**

Figure 3.3. Confusion matrices of the deep learning network, showing its performance on classification of the validation dataset (panel A) and the testing dataset (panel B). Blue and orange cells refer to the number of frames/images in each category, and the green cells represent the associated accuracy of each row and column – noting that the overall classifier's accuracy is highlighted in the dark green cells. The horizontal labels represent the predicted outcome of the classifier on the "Unobstructed Vocal Fold" class (VF) and "Obstructed Vocal Fold" class (No VF). The vertical labels refer to the ground-truth labels observed by the rater for each class.

The receiver operating characteristics curve of the developed CNN is illustrated in Figure 3.4. The figure depicts the curve (in blue) when the network was implemented for the validation dataset (panel A) and testing dataset (panel B). The closer the curve is to the upper left corner, the higher the overall accuracy of the CNN. The diagonal red line represents points where Sensitivity=1-Specificity. Along with the curve, the value for the area under the curve (AUC) is included in the figure. As can be seen, both validation and testing curves show almost the same behavior. AUC of the validation and testing dataset curves are almost 1.00.

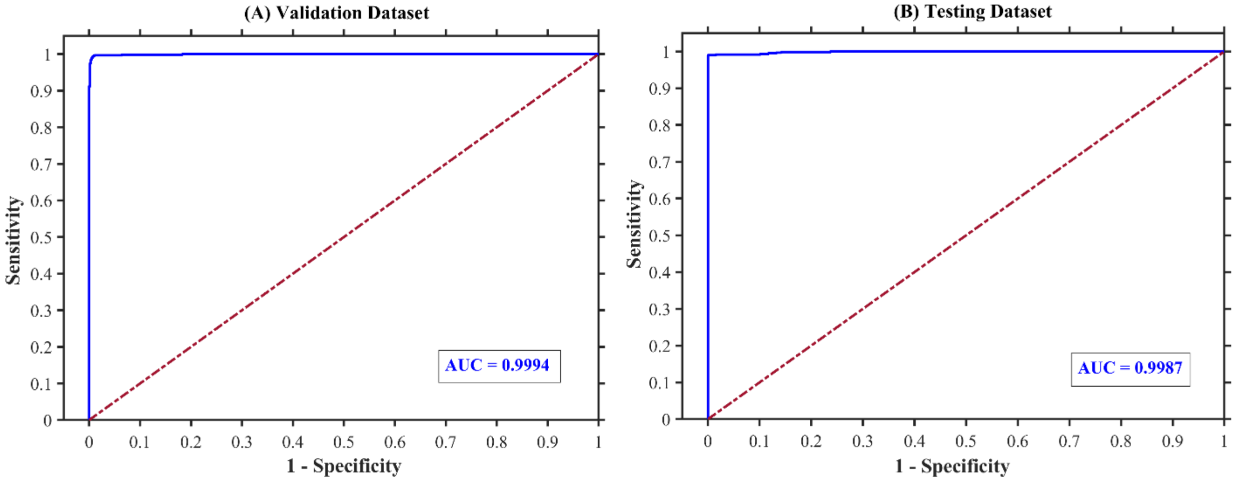


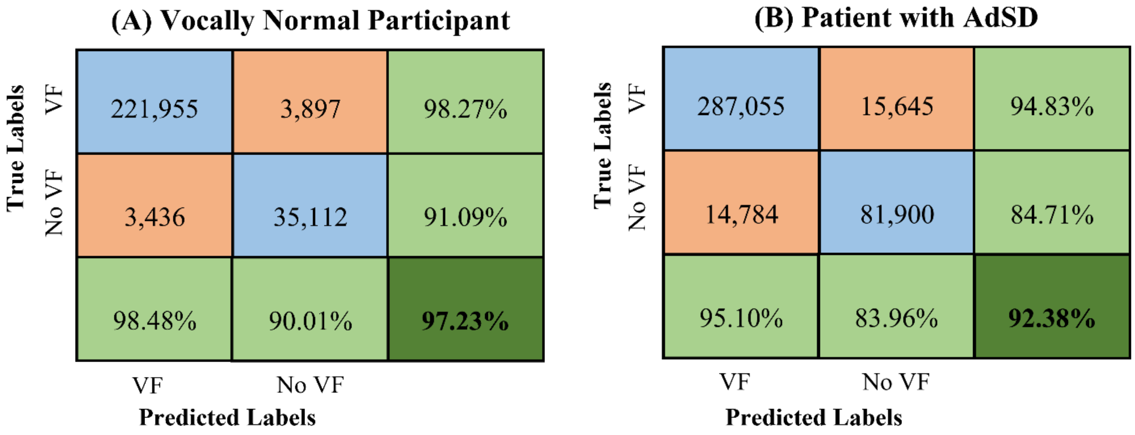
Figure 3.4. The sensitivity-specificity curve (receiver operating characteristics curve), in blue, for the validation dataset (panel A) and the testing dataset (panel B). AUC refers to the area under the sensitivity-specificity curve. The diagonal red line represents points where Sensitivity=1-Specificity.

The robustness evaluation of the proposed automated classifier was done through comparison between the CNN performance and manual analysis classification using two complete HSV recordings for a vocally normal participant and a patient with AdLD. The results of the comparison are listed in Table 3.1 for 264,400 HSV images from the vocally normal participant and 399,384 images from the patient. In the vocally normal participant, the manual analysis reveals that 38,497 out of 264,400 frames (14.56%) with an obstructed view of VFs, and the automated analysis shows almost a similar number of frames, 39,009 (14.75%). Likewise, in the patient, the manual and automated analysis demonstrate close number of frames for the obstructed VFs: 96,571 versus 97,545 out of 399,384 frames (24.18% versus 24.42%), respectively. The difference in the detected number of frames with an obstructed VFs between the automated technique and the manual observation is 512 (0.19%) and 974 (0.25%) for the vocally normal and disordered participant, respectively.

Table 3.1. Robustness evaluation: A comparison between the visual observation versus the automated technique in terms of number/percent of frames with obstructed view of vocal folds in the entire HSV recordings for a vocally normal participant and a patient with AdLD

	# HSV Frames	**# Obstructed Frames**	**% Obstructed Frames**	**Difference (# Frames)**	**Difference (%)**
Normal participant					
Visual observation	264,400	38,497	14.56	512	0.19
Automated analysis		39,009	14.75		
Patient with AdLD					
Visual observation	399,384	96,571	24.18	974	0.25
Automated analysis		97,545	24.42		

The two confusion matrices for the robustness evaluation are presented in Figure 3.5 for a detailed comparison of the proposed CNN against the manual analysis – using the same two HSV recordings in Table 3.1. The two matrices in Figure 3.5 have the same formatting and color code as in Figure 3.3. For the vocally normal participant in Figure 3.5 (panel A), the manual analysis shows 225,852 frames with an unobstructed view of the VFs and 38,548 frames with an obstructed VFs view; based on the manual analysis, the proposed CNN demonstrates correct classification of 221,955 (98.27%) and 35,112 frames (91.09%), respectively. In the HSV video for the patient (panel B), the automated method shows a successful identification of the unobstructed and obstructed VF view in 287,055 out of 302,700 frames (94.83%) and 81,900 out of 96,684 frames (84.71%). Furthermore, the developed automated approach shows an overall accuracy of 97.23% and 92.38% for the entire HSV video of the vocally normal participant and the patient, respectively. The overall accuracy is measured as the network ability in recognizing both the presence and absence of the VFs in the HSV frames correctly.



(A) Vocally Normal Participant

True Labels \ Predicted Labels	VF	No VF	
VF	221,955	3,897	98.27%
No VF	3,436	35,112	91.09%
	98.48%	90.01%	**97.23%**

(B) Patient with AdSD

True Labels \ Predicted Labels	VF	No VF	
VF	287,055	15,645	94.83%
No VF	14,784	81,900	84.71%
	95.10%	83.96%	**92.38%**

Figure 3.5. Confusion matrix of the developed deep learning network for classification of HSV recordings of a vocally normal participant (panel A) and a patient with AdLD (panel B). The blue and orange cells refer to the number of frames/images in each category, and the green cells represent the associated accuracy of each row and column – noting that the overall classifier accuracy is highlighted in the dark green cell. The horizontal labels represent the predicted outcome of the classifier on the “Unobstructed Vocal Fold” class (VF) and “Obstructed Vocal Fold” class (No VF). The vertical labels refer to the ground-truth labels, which are visually/manually observed for each class.

Similar to Figure 3.5, the evaluation metrics resulted from applying the proposed automated classifier on the two manually analyzed HSV videos for each class are represented as light green cells in Figure 3.5. In the normal participant recording, the automated technique has a sensitivity and specificity of 98.27% and 91.09% with respect to recognizing VFs obstruction in HSV frames whereas these values fall to 94.83% and 84.71% in the patient recording. The CNN precision scores are higher for the “Unobstructed Vocal Fold” class with 98.48% and 95.10% for the norm and disorder, respectively, compared to the “Obstructed Vocal Fold” class with 90.01% and 83.96%. A similar behavior was found with respect to F1-scores for both HSV videos – 0.98 and 0.95 for the “Unobstructed Vocal Fold” class and 0.91 and 0.84 for the “Obstructed Vocal Fold” class.

Figure 3.6 shows the resulted receiver operating characteristics curve of the introduced classifier (in blue) for the two HSV videos (for robustness evaluation) of the vocally normal participant (panel A) and the patient with AdLD (panel B). The figure shows the change in the network threshold of the binary classification with respect to sensitivity and specificity of the developed classifier in recognizing the VF obstruction over the entire video frames. As can be seen, the classifier shows a better performance with larger area under the curve when analyzing the normal participant's HSV sequence than that of the patient's. This is clear when comparing the two corresponding AUC of videos – shown in the bottom right corner of the panels in Figure 3.6. The AUC for the vocally normal participant is 0.99 while it marginally drops to 0.96 for the AdLD patient.

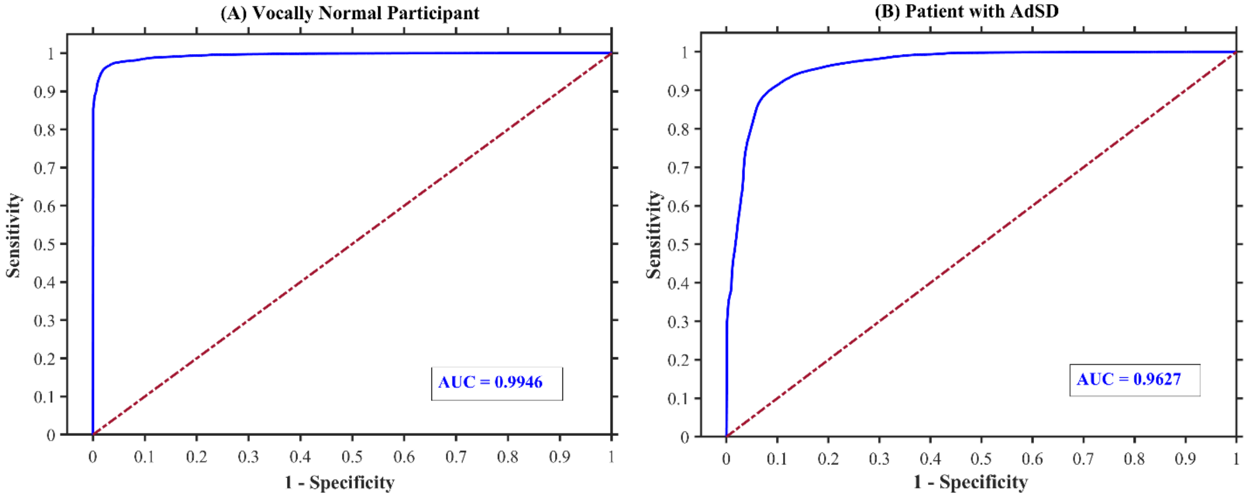


Figure 3.6. The sensitivity-specificity curve (receiver operating characteristics curve), in blue, of the developed deep learning network performance on binary classification of the entire two HSV videos of a vocally normal participant (panel A) and a patient with AdLD (panel B). AUC refers to the area under the sensitivity-specificity curve.

A detailed comparison between the automated technique against the visual/manual analysis is illustrated in Figure 3.7. The comparison is shown for each frame of the entire two HSV videos of the vocally normal participant (panel A) and the patient with AdLD (panel B). For each video sequence, the red and blue color represent the automated and manual method, respectively, for the instances during which VFs were visually obstructed. As can be seen in the figure, the results of the automated and visual detection display a similar pattern. Besides visual assessment, the accumulated overall accuracy (in solid black line), precision of detecting obstructed view of VFs (in dotted dark red line), and precision of detecting unobstructed view of VFs (in dashed green line) are also illustrated in the figure. The accuracies represent the performance of the developed classifier as a function of time for each HSV video. The accuracies were computed at accumulated time step of 1,000 frames over the entire length of the videos. That is, for each time step, the confusion matrix was generated to evaluate the performance of the automated technique versus the manual analysis on classifying the accumulated number of frames; these generated matrices were then used to compute the accumulated accuracies over the video sequence. As such, the values of the three accuracies at the end of each video (at the last frame) refer to the accuracies over the entire HSV video frames, which was shown in Figure 3.5. As can be seen in Figure 3.7, both the overall accuracy and the precision of recognizing unobstructed view of VFs have nearly similar behavior with high values across each video's frames. In line with the previous results, the precision of detecting obstructed view of VFs demonstrates slightly lower values than that of the unobstructed class over the entire video frames; the two curves also show different trends as well.

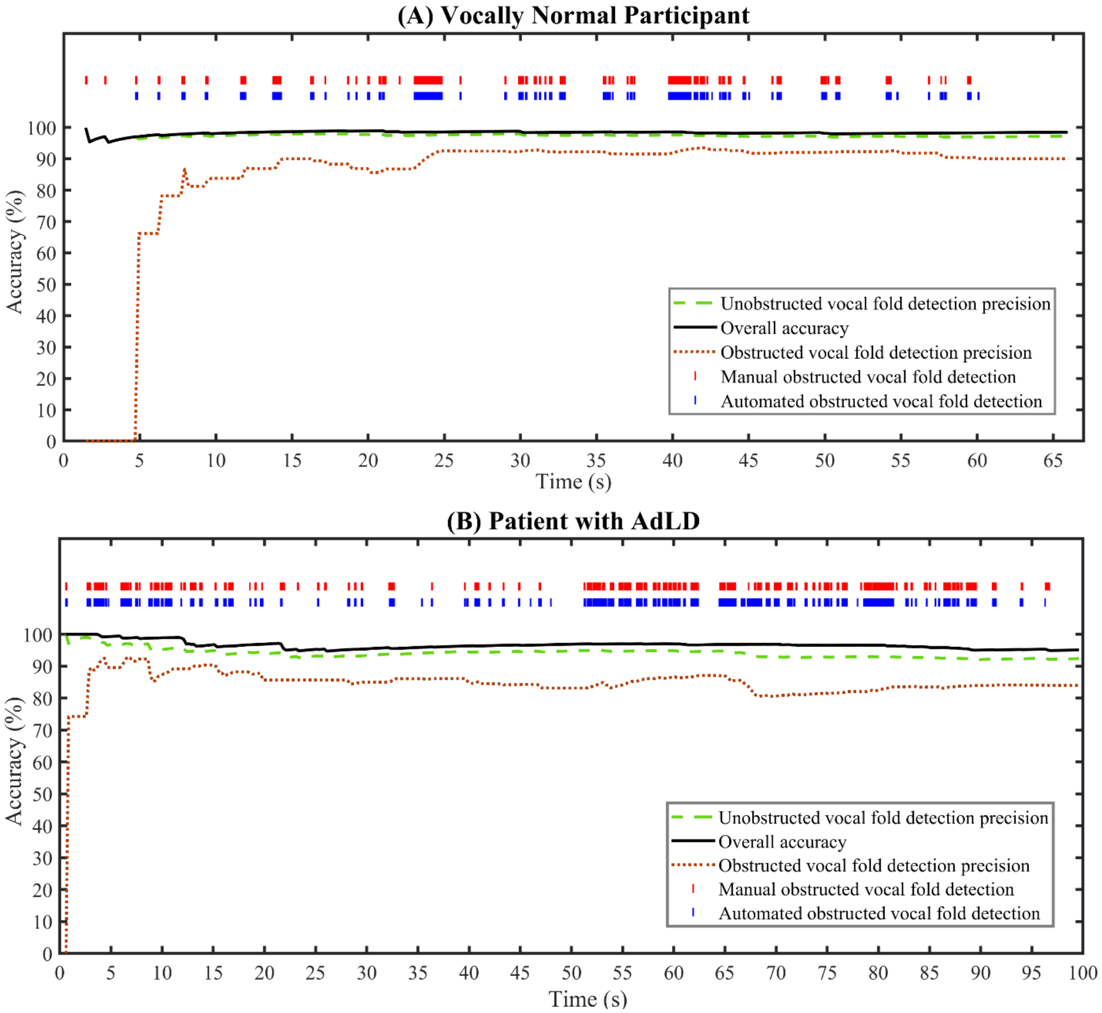


Figure 3.7. Comparison between automated (in blue) and manual (in red) analysis of the instances during which vocal folds are obstructed. The comparison shown for the entire two HSV videos of a vocally normal participant (panel A) and a patient with AdLD (panel B). The accumulated overall accuracy (in solid black line), precision of detecting obstructed view of vocal folds (in dotted brown line), and precision of detecting unobstructed view of vocal folds (in dashed green line) are also illustrated.

After validating the proposed method, the difference between the AdLD group and the normal control group was investigated. The AdLD patients showed a considerable difference in the average percentage of VF obstruction in comparison with the vocally normal controls. That is, the AdLD group exhibited an average obstruction percentage of 26.1% of the recorded HSV running speech sample, whereas the vocally normal speakers demonstrated a noticeable shorter average duration of obstruction (19.75%).

3.2. Study II: Image Segmentation of Vocal Fold Edges

3.2.1. Image Segmentation Approach: Active Contour Modeling (ACM)

Using the classical image processing technique for temporal segmentation for HSV color data preprocessing, the timestamps for all the vocalized segments of the HSV connected speech recording were extracted (except for the segments with epiglottic obstruction). Subsequently, the motion compensation was applied to each vocalized segment of the "Rainbow Passage" to capture the location of the vibrating VFs across the frames. The result of applying the motion window to three individual frames during five different vocalizations are depicted in Figure 3.8a-e. Each row in Figure 3.8 shows three frames for a different vocalization between the following frame numbers: 40,505-41,255 (Figure 3.8-a), 42,975-43,815 (Figure 3.8-b), 84,281-84,891 (Figure 3.8-c), 103,942-104,577 (Figure 3.8-d), and 109,548-110,363 (Figure 3.8-e). The individual frame numbers are shown on top of the HSV frames for each figure panel. The figure shows that the implemented motion window captures both the location and size of the vibrating VFs in different frames.

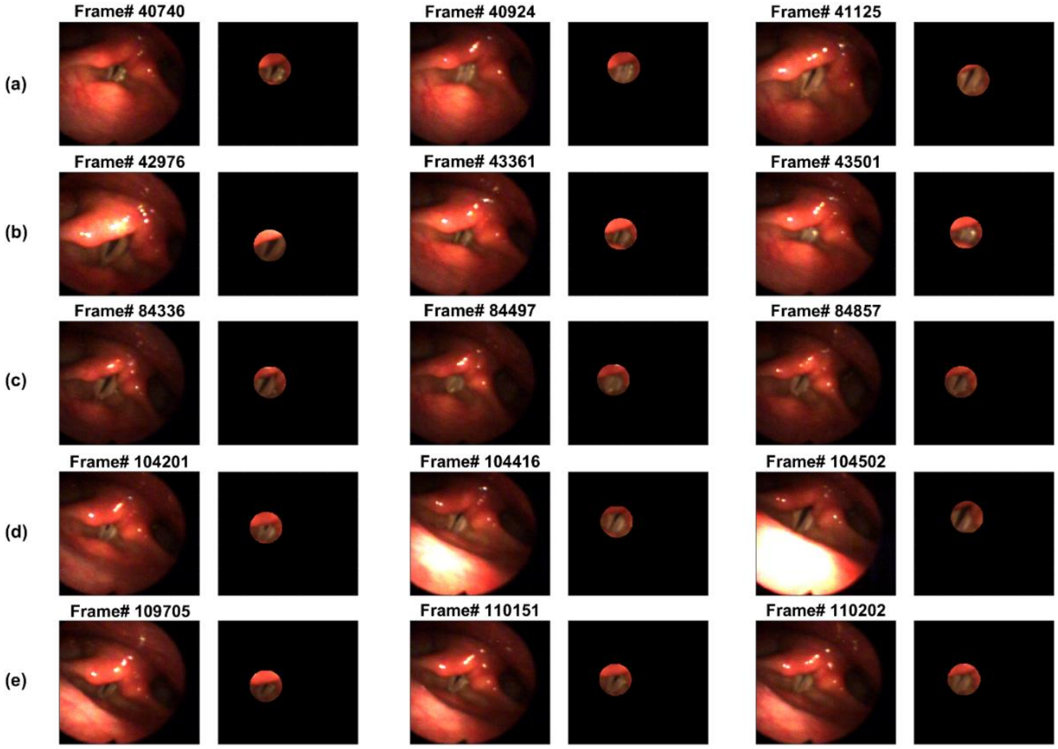


Figure 3.8. HSV frames along with the applied motion windows for three different frames at five different vocalized segments (panels (a-e)).

The kymograms were extracted for each vocalization of the "Rainbow Passage" after aligning the VFs across the frames using the motion compensation method. Examples of the extracted kymograms at the medial section of the VFs are shown in Figure 3.9. The kymograms for five different vocalizations are shown in panel a-e between the same frame numbers as in Figure 3.8. You can see the onset and offset of phonation in each kymogram, and that the darker glottal area is almost on a straight line across the frames for each kymogram.

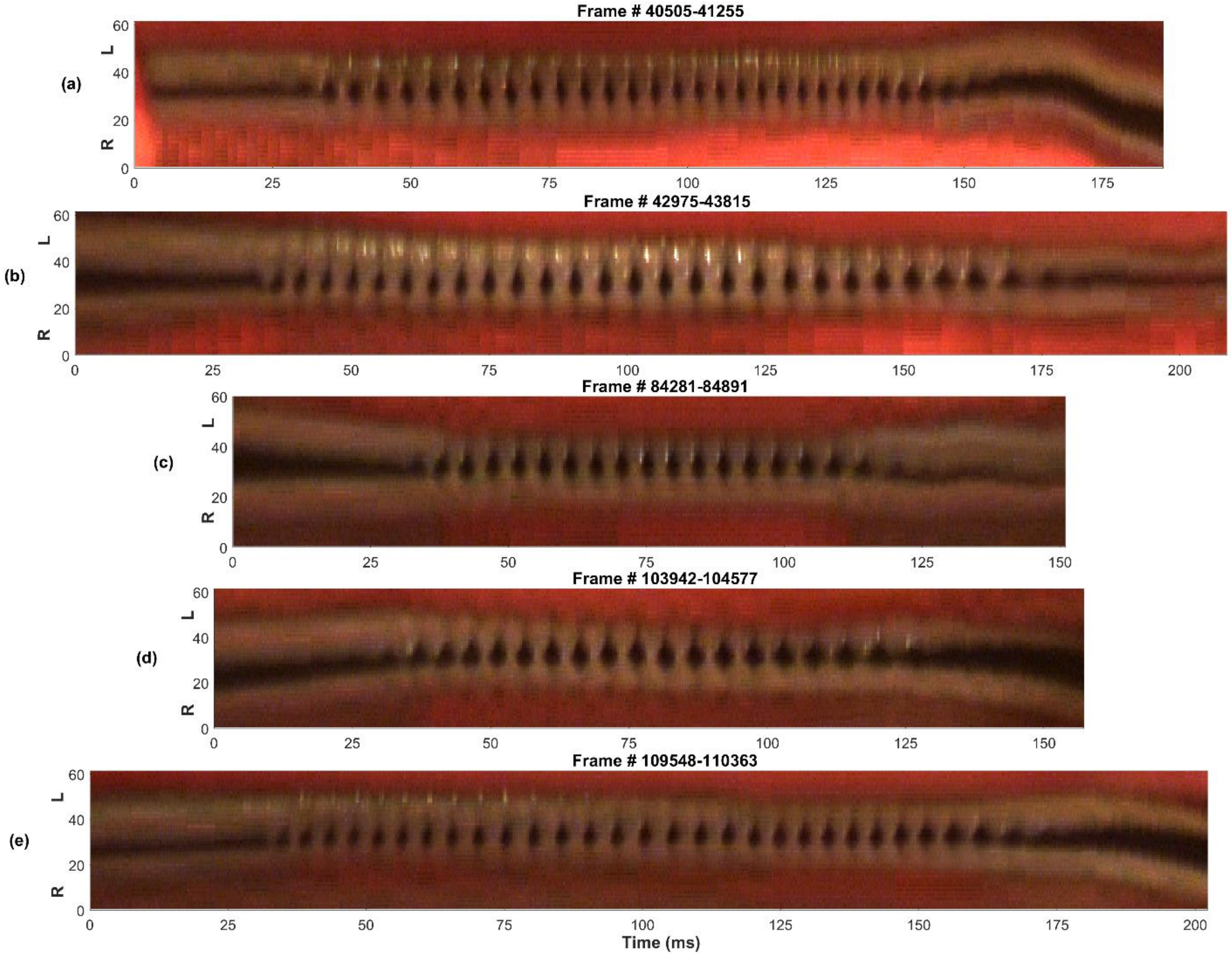


Figure 3.9. HSV kymograms at the medial section of the vocal folds for five different vocalized segments (panels (a-e)). The L and R on the y-axis refer to the left and right VFs, respectively.

The results of applying the ACM method to four kymograms, extracted at different cross sections of VF, are illustrated in Figure 3.10. That is, after the snake initialization using the first moment of inertia line (a horizontal line spanning through the center of the glottal areas in the kymograms), the active contour algorithm was applied to the kymograms. The upper and lower snakes (active contours) corresponding to the left and right VFs for four kymograms are shown in Figure 3.10. The number of frames shown in the figure is 546, between frame 40,585 and 41,165, including voicing onset, VFs vibration, and voicing offset. Two zoomed-in image segments are included in Figure 3.10 to better visualize the performance of the algorithm. As seen, the ACM approach detects both the left and right VF edges (solid green line and dotted yellow line, respectively) at different cross sections – providing an analytical representation of the glottal edges. Moreover, the algorithm is able to capture the edges before the phonation starts and after the phonation ends.

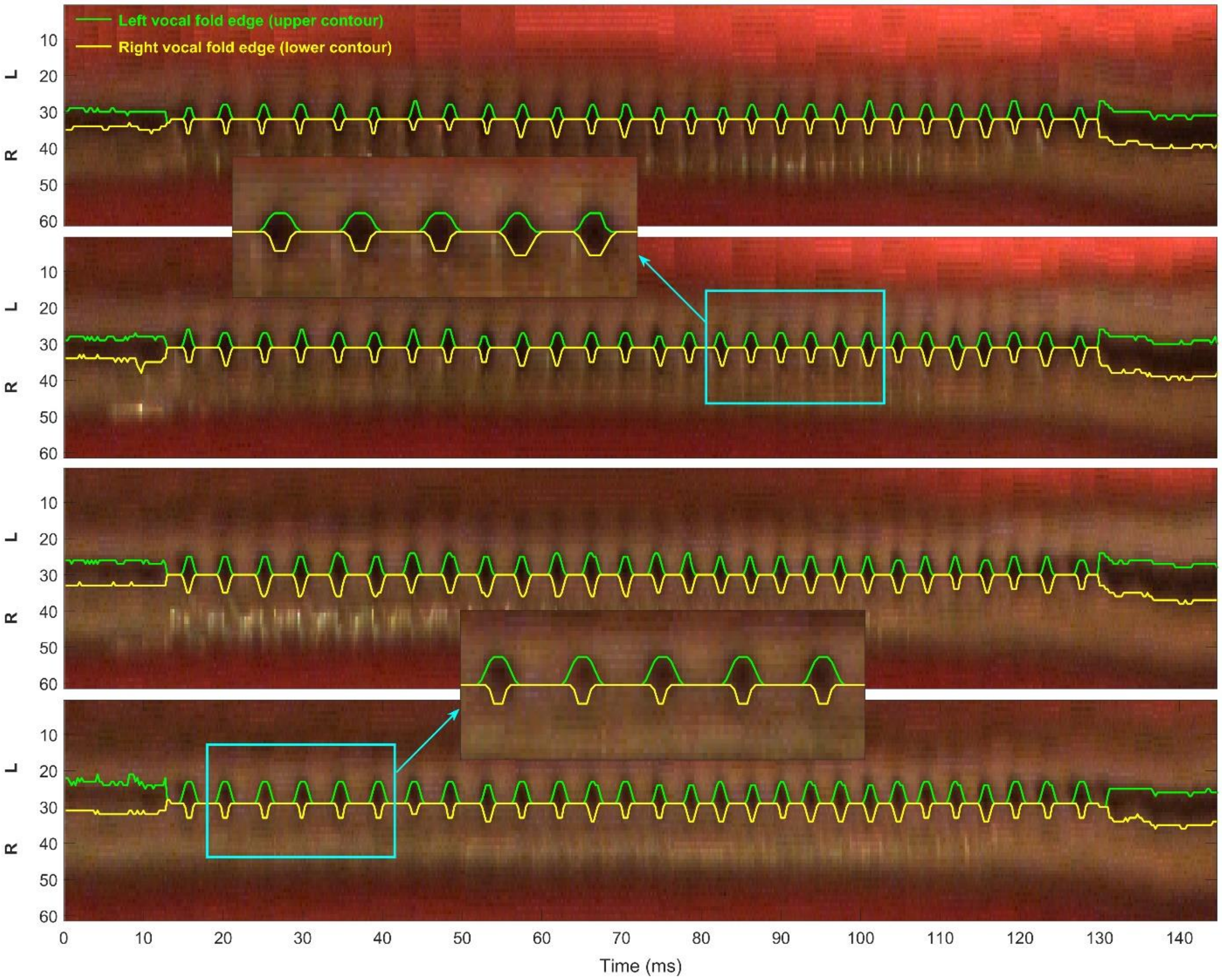

Left vocal fold edge (upper contour)
Right vocal fold edge (lower contour)
L
R
10
20
30
40
50
60
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
Time (ms)

Figure 3.10. Kymograms between frames 40,585 and 41,165 at four different cross sections of the vocal folds (panel a-d) along with the upper and lower active contours (solid green line and solid yellow line, respectively) corresponding to the left and right vocal folds. Two zoomed-in image segments are included to better visualize the performance of the algorithm. The L and R on the y-axis refer to the left and right VFs, respectively.

3.2.2. Image Segmentation Approach: The Hybrid Method

The following results demonstrate the implementation of the proposed hybrid method (ACM + k-means clustering) for the color HSV data. An example of 5 cropped HSV frames extracted from a vocalization after applying the temporal segmentation and motion compensation techniques, previously discussed, are illustrated at the top of Figure 3.11. This vocalization was extracted between frames 32,659 and 35,111. The frame numbers are shown at the top of panel (b)-(f). As seen, the motion window captures the size and the spatial location of the VFs during different phases of the vibratory cycle. After applying the motion window, the HSV kymograms were extracted at various cross sections of the VFs during each vocalized segment. Four kymograms, extracted at four different cross sections of the VFs during the same vocalization, are shown in Figure 3.11-g-j. The y-axis of the kymograms represents the left-right dimension of the HSV frame while the x-axis refers to the time (number of frames). Each kymogram in the figure displays the voicing onset and offset along with the vibration of the VFs.

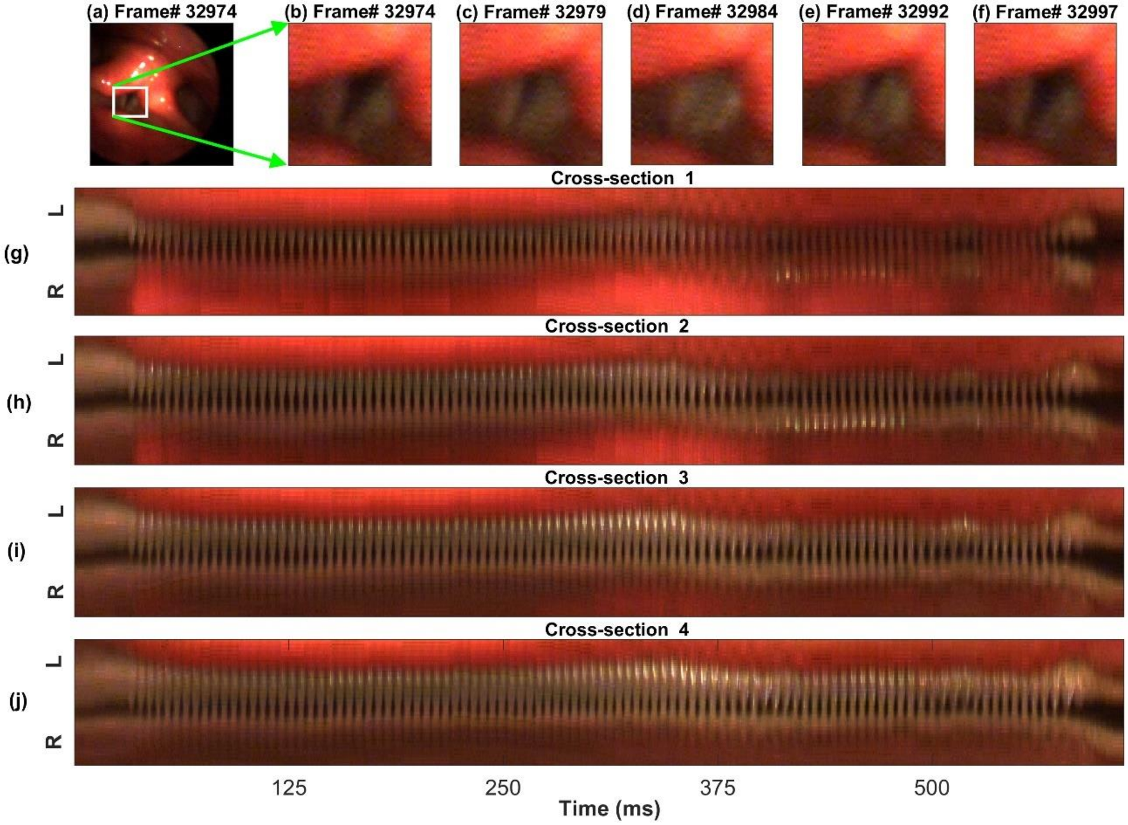


Figure 3.11. Panels (b)-(f): 5 HSV cropped frames for frame #32,974, 32,979, 32,984, 32,992, and 32,997 after applying the motion window to five different HSV frames (one HSV frame is shown in panel (a)). Panels (g)-(j): Four extracted kymograms at different cross sections of the vocal folds. The R and L on the y-axis indicate the right and left VFs in the HSV frames, respectively.

The k-means clustering technique was implemented for each kymogram. Different subsets of features were fed into the machine learning (ML) algorithm to determine the proper number and combination of features leading to an accurate VF edge representation. Figures 3.12-14 illustrate a comparison between the results of applying two different combinations of the features for glottal area/edge detection: i) red and green channel intensities as two features (panel (a) in the figures) versus ii) the image gradient along with the red and green channel intensities as three features (panel (b) in the figures). The results of utilizing the other subsets of the aforementioned features to perform the clustering showed poorer performance of the method in comparison with using the selected feature combinations in Figures 3.12-14. Figure 3.12 shows the result of applying the clustering technique to the kymogram shown in Figure 3.11-h between frame 32,709 and 35,061 (for a total of 143,167 data points). The scatter plot in Figure 3.12-a is generated by feeding the clustering algorithm the two intensity features: the green channel intensity and red channel intensity. The scatter plot in Figure 3.12-b is generated using the gradient feature along with both red and green channel intensities features. The glottal area cluster in the kymogram is shown by red diamonds and the non-glottal cluster is shown by blue circles. As seen, after adding the gradient feature to the intensity features in Figure 3.12-b, the two clusters can be distinguished in the scatter plot; in contrast, depending only on the intensity as a feature, it is relatively hard to divide the data points into two different clusters. The better performance of the ML method using the three features is more prevalent in Figure 3.13.

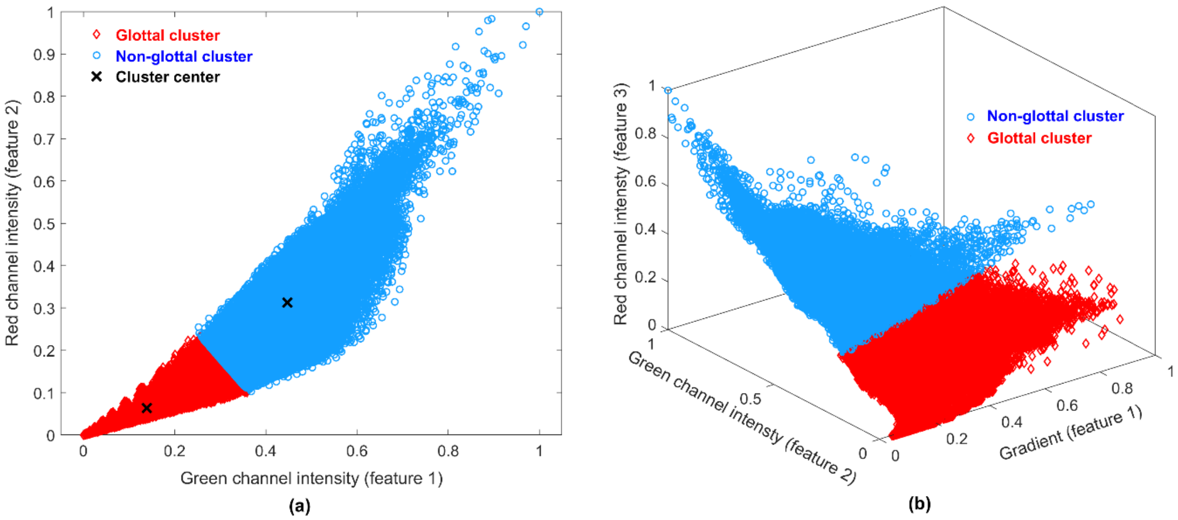


Figure 3.12. Scatter plots of the two clusters when applying the clustering method to the kymogram in Figure 3.11-h between frame 32,659 and 35,111: (a) using the green and red channel intensities as the features; and (b) using both green and red channel intensities along with the gradient as features.

Figure 3.13 shows the two clusters after applying the k-means clustering technique to the kymogram in Figure 3.11-h. The top figure illustrates the clustered regions using the two intensities as features and the bottom figure shows the result when using both the gradient and the intensities (green and red channel intensities) as three features. Figure 3.13 illustrates the clustered areas on the binary labeled kymogram so that only two distinct colors are shown representing the two clusters obtained. As seen, using the gradient in addition to the intensity allows us to capture more information about the glottal area, which well aligns with the results obtained from the former figure.

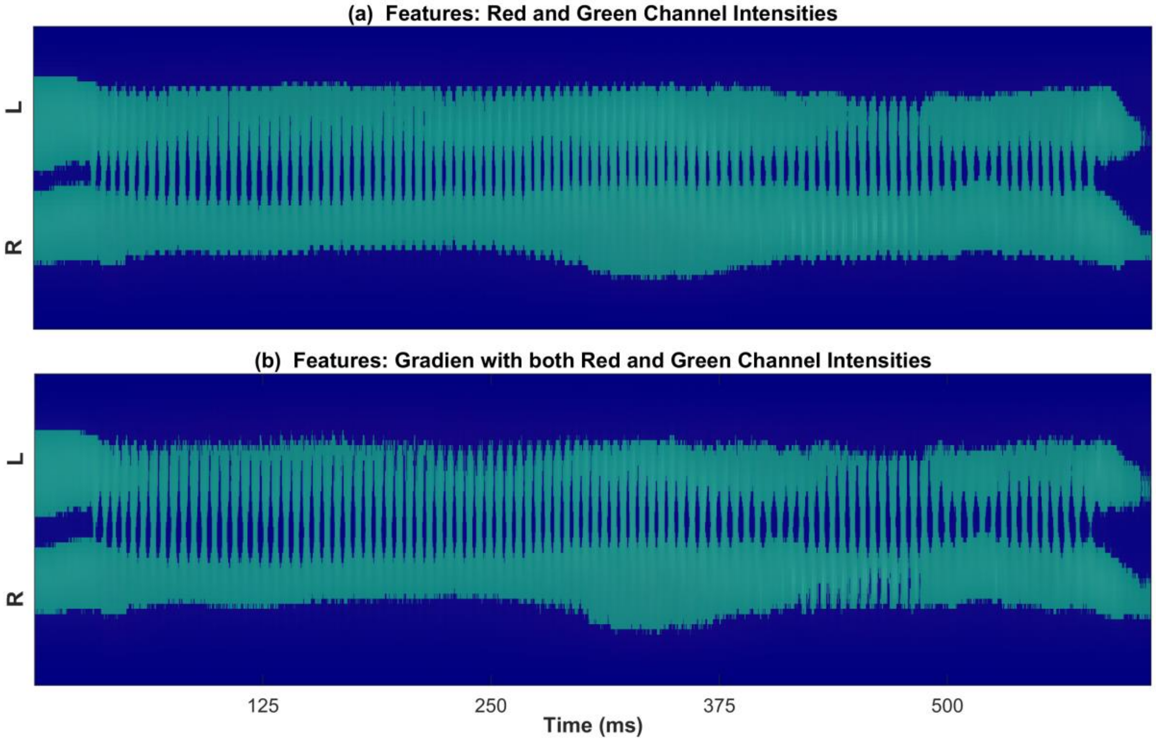

(a) Features: Red and Green Channel Intensities
L
R
(b) Features: Gradien with both Red and Green Channel Intensities
L
R
125
250
375
500
Time (ms)

Figure 3.13. The clustered kymogram (from Figure 3.11-h) by employing the k-means clustering algorithm using (a) the red and green channel intensities as features and (b) the gradient along with the red and green channel intensities as three features.

Figure 3.14 shows the detected edges of the glottal area based on the results of clustering. In this figure, only the glottal cluster region is shown with a white line in the original kymogram to have a better visual representation of the performance of the clustering method using the intensity features (panel (a)) and the gradient and intensity features (panel (b)). The comparison of panel (a) and (b) shows the improvement in clustering after adding the gradient feature to the intensity features. As can be seen in this figure, using only intensity features results in missing spatial information about the glottal area, particularly during the sustained vibration of the VFs. On the other hand, the glottal edges were detected more accurately when the gradient feature was used along with both the red and green channels intensities. This improvement is clear during the sustained oscillation of the VFs while it is not considerable during voicing onsets and offsets.

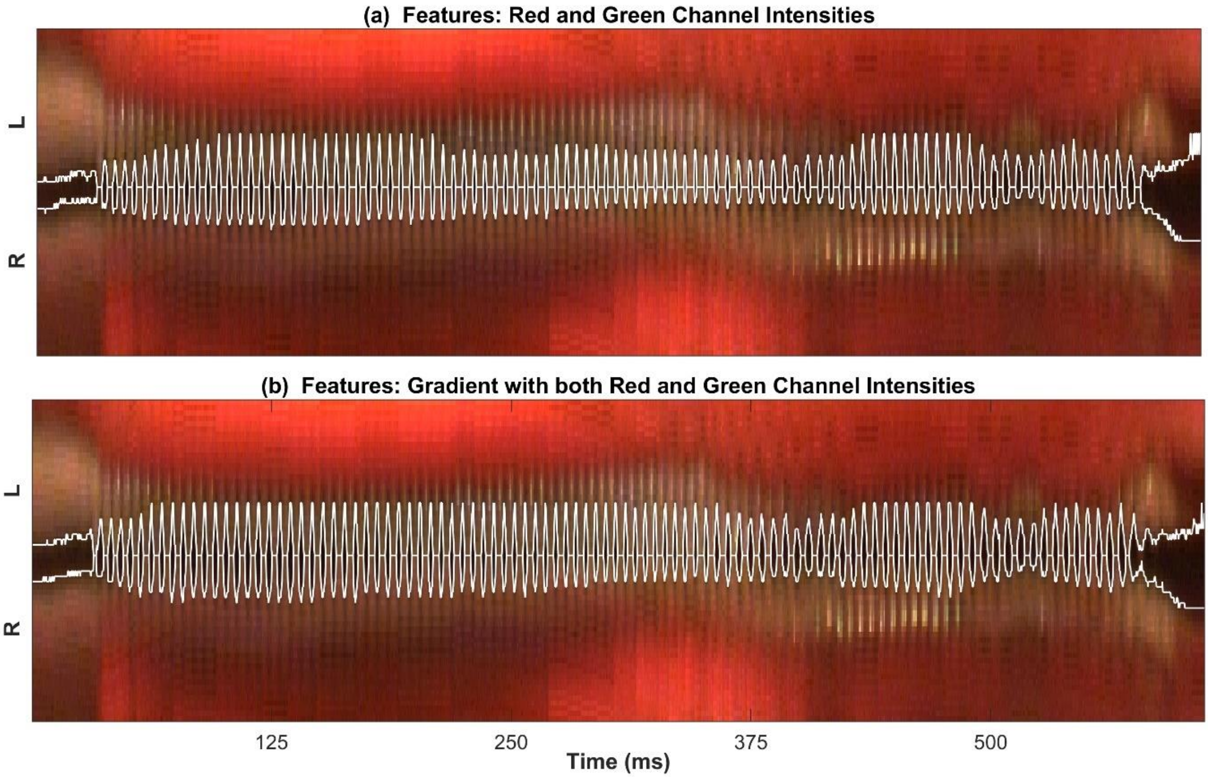

(a) Features: Red and Green Channel Intensities
L
R
(b) Features: Gradient with both Red and Green Channel Intensities
L
R
125
250
375
500
Time (ms)

Figure 3.14. The detected glottal edges based on the results of k-means clustering algorithm using (a) the green and red channel intensities as features and (b) using the gradient along with both red and green channel intensities as three features.

The preliminary segmented glottal edges as a result of applying the clustering technique were used as inputs to the ACM method. Figure 3.15 shows how using k-means clustering as an initialization step for the ACM impacts the accuracy of the method. The results are presented in four kymograms extracted at four different vocalizations. The detected glottal edges using the ACM alone and the developed machine-learning-based hybrid method are shown for two decent quality kymograms (between frames 40,505 to 41,255, panel (b), and 103,992 to 104,522, panel (d)) and for two challenging kymograms (between frames 18,975 to 19,803, panel (a) and 98,105 to 98,651, panel (c)). The figure depicts the result of applying the ACM method alone along with the performance of the hybrid method at each vocalization. Although the ACM performed better for the top kymograms in panel (b) and (d) in comparison with the (more challenging) kymograms at the top of panel (a) and (c), this method missed the glottal edges for several cycles as seen in the top figures in panel (b) and (d). The ACM was not able to capture the glottal edges for many glottal cycles in the dim kymograms as seen in the top figures in panel (a) and (c). In contrast, the hybrid method showed a considerable enhancement in the performance and high accuracy as it detected the glottal edges precisely for all the kymograms, as seen in the bottom kymograms in panel (b) and (d), also in panel (a) and (c) with inferior quality and challenging kymograms.

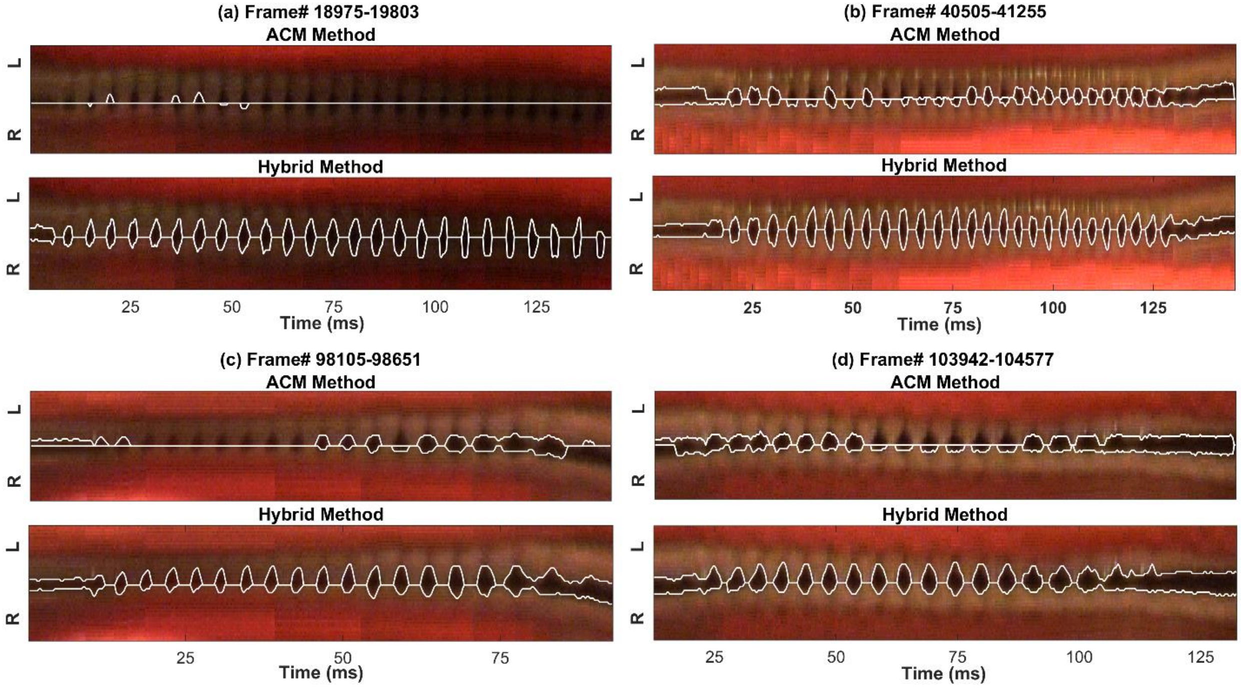


Figure 3.15. The detected glottal edges using the ACM method (top kymograms in panel (a)-(d)) versus the hybrid method (bottom kymograms in panel (a)-(d)) for the kymograms extracted at four different vocalizations.

In Figure 3.16, five HSV frames are presented from each of the four different vocalizations in Figure 3.15 along with the detected glottal edges by the hybrid method. This figure shows the captured edges after registering the glottal edges from the kymograms back to the HSV frames. For each vocalization, the five frames are chosen to show several frames from a different phase of a vibrating cycle of the VFs. As can be seen in Figure 3.16, the hybrid method was able to track the left and right VF edges accurately during the VF vibration in different frames and vocalizations.

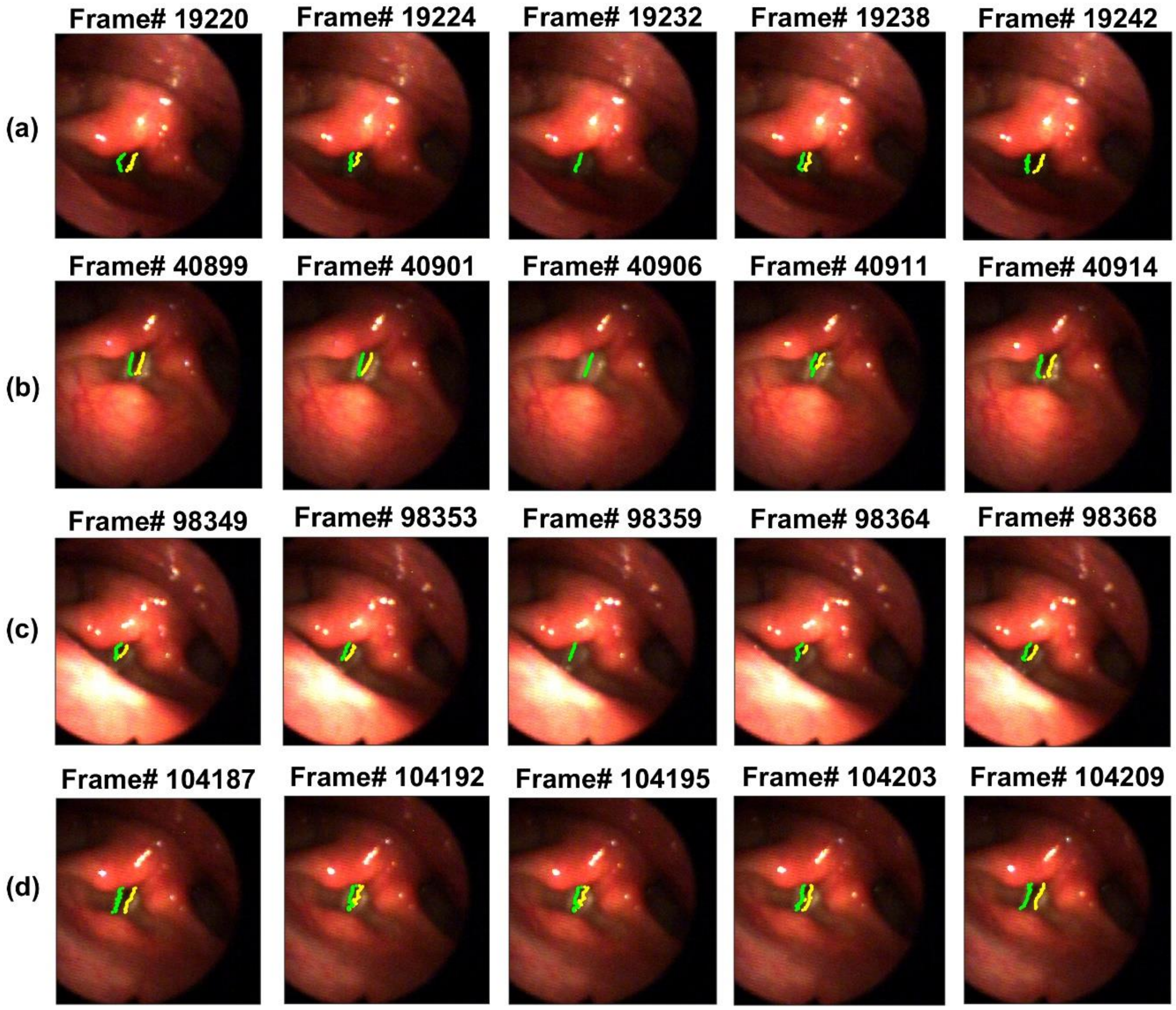


Figure 3.16. Five HSV frames from four different vocalizations (panel (a)-(d): between frame 18,975-19803, 40,505-41,255, 98,105-98,651, and 103,942-104,577 after implementing the hybrid method to spatially register the edges of the vibrating vocal folds.

3.3. Study III: Deep-Learning-based Representation of Vocal Fold Dynamics

3.3.1. Deep Learning Approach: Segmenting Network on Color HSV Data

In this section, results of implementing the DNN to the color HSV video is illustrated. Overall, the automated labelling tool (the hybrid method that was discussed in the previous section) was able to segment the glottal area in the training HSV frames on which the neural network was trained. The DNN was successfully trained on the automatically segmented frames in the training dataset. The trained network was then tested on manually labeled HSV frames yielding a promising performance. Below is the summary of the results obtained from this work: starting with results of implementing the temporal segmentation method, the labeling tool (the hybrid method), the neural network, and the generation of the glottal area waveforms.

Figure 3.17 shows the results of each preprocessing step at four different vocalized segments between frame numbers 4,261-5,551 (panel (a)), 42,999-43,774 (panel (b)), 84,900-86,118 (panel (c)), and 98,162-98,542 (panel (d)). In each panel, the outcome of applying the temporal segmentation, motion compensation, and kymogram extraction for a vocalization is illustrated. As shown, the utilized motion compensation specifies the true location of the VFs in the cropped frames. The stacked frames/cropped images refer to the sequence of image sections during the vocalized segments of the connected speech. These frames were used to generate multiple kymograms at different cross sections of the vibrating VFs (represented by a stacked kymograms in the figure). Examples of the extracted kymograms at the medial intersection of the VFs showing the variation in the glottal region across the frames can be seen in the right side of the figure. The kymograms span through the entire vocalization – clearly representing the vibratory patterns and behavior, namely, phonation onset, the sustained vibration of VFs, and phonation offset.

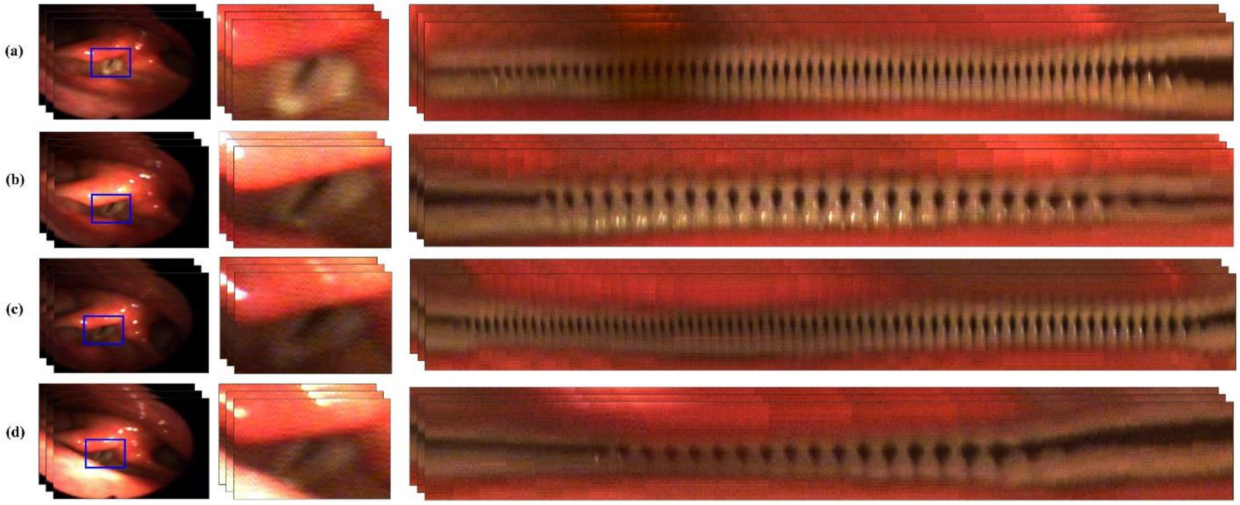


Figure 3.17. Results of applying temporal segmentation, motion compensation, and kymograms extraction at four different vocalized segments between frames: 4,261-5,551 (panel (a)), 42,999-43,774 (panel (b)), 84,900-86,118 (panel (c)), and 98,162-98,542 (panel (d)). The stacked frames/image sections refer to the sequence of frames and the cropped images during each vocalized segment. The stacked kymograms, at each vocalized segment, represent the multiple kymograms extracted at different cross sections of the vibrating vocal folds.

For each kymogram, the hybrid method (aka k-means-ACM) was applied to segment and detect the glottal edges during vocalizations. Figure 3.18 illustrates the results of implementing the k-means-ACM algorithm at various kymograms of Figure 3.17 that were extracted from different vocalizations. As illustrated in Figure 3.18, the k-means-ACM technique was able to accurately segment the edges of right and left VFs, shown in solid white lines in the kymograms (left panels of the figure). The glottal edges were then registered back to each HSV frame in the cropped images (see the mid panels in Figure 3.18) and the original HSV frames (shown in the right-side panels). This was done to segment the glottal area in each image; the glottal areas are shown in cyan in mid and right-side panels.

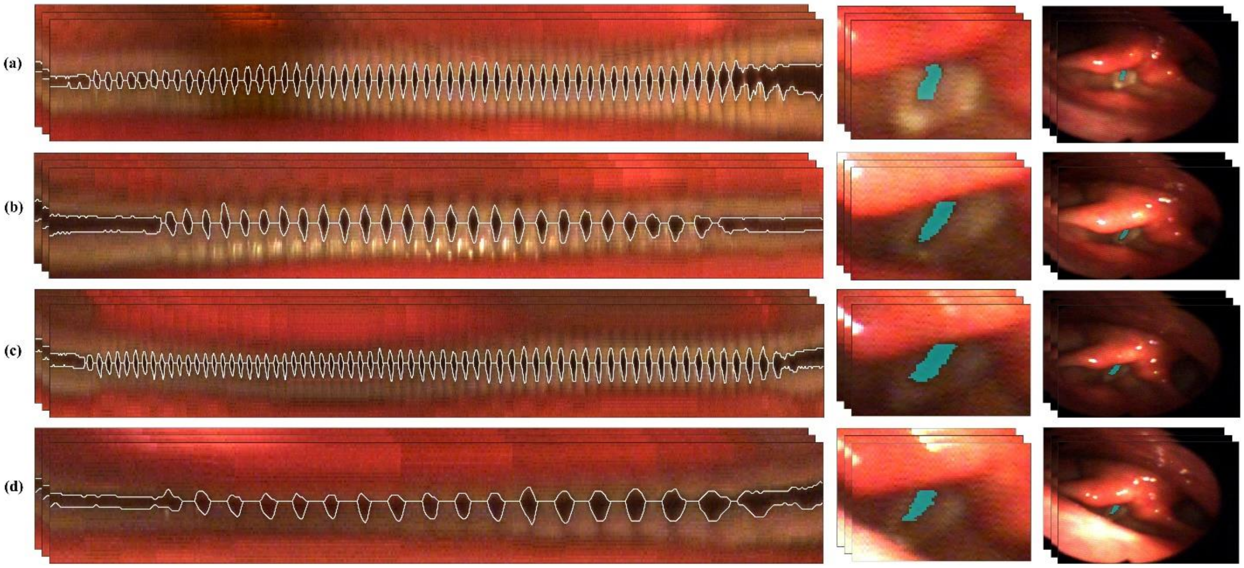


Figure 3.18. Results of applying k-means-ACM at four different vocalized segments between frames: 4,261-5,551 (panel (a)), 42,999-43,774 (panel (b)), 84,900-86,118 (panel (c)), and 98,162-98,542 (panel (d)).

Figure 3.19 shows the results of training the proposed DNN for two different vocalizations (panel (a) and (b)). Each panel shows the result for four frames, extracted from a different vocalization. The results in Figure 3.19 are displayed for the following frame numbers: #41,658, #41,738, #41,880, and #41,986 in panel (a) and frame #104,061, #104,162, #104,311, and #104,460 in panel (b). For each frame, the original HSV frame along with the associated binary segmentation masks are depicted for both k-means-ACM (the automated labelling tool) and the proposed DNN. The segmented glottal areas using the k-means-ACM (in cyan color) and DNN (in yellow color) are overlaid on top of each other (in the right-side panels of Figure 3.19) to demonstrate their differences. The DC and the BF (aka F1) scores that are associated with evaluating the similarities between the two segmented areas are included in the figure as well. As shown in the segmented frames and by the scores, the DNN demonstrates a relatively similar performance to the k-means-ACM on most of the presented frames in accurately segmenting the glottal regions. Most of the frames in the figure show that DC > 0.80 and BF > 0.9. In addition, it can be seen that the introduced network can even outperform the k-means-ACM in some frames (e.g., frame #41,658, #104,311, and #104,460) providing smoother glottal edges.

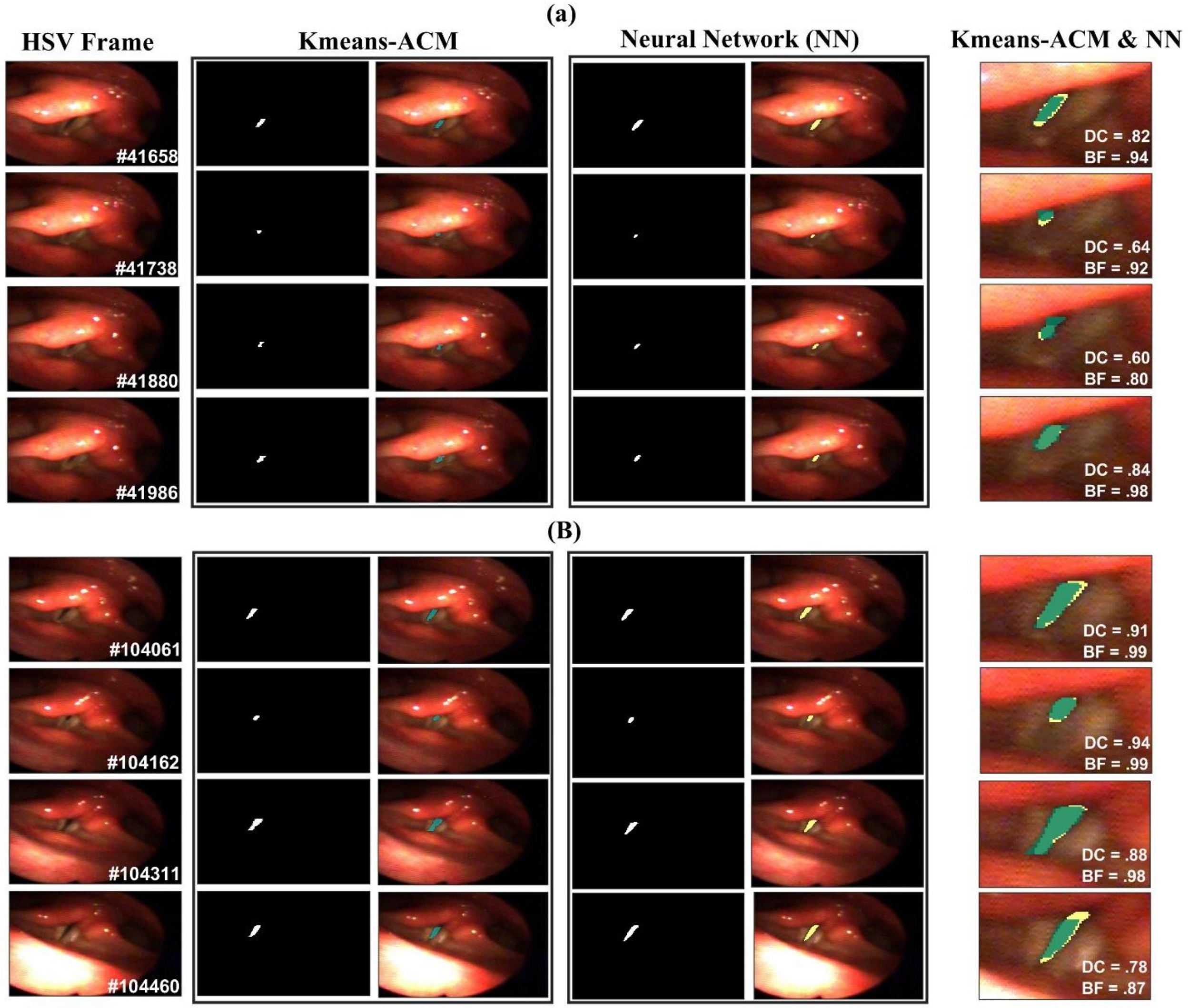


Figure 3.19. Results of implementing the k-means-ACM and the trained DNN; the segmented HSV frames along with the associated binary segmentation masks are shown for eight different frames extracted from two different vocalizations (a and b). (a) for Frame #41,658, #41,738, #41,880, and #41,986. (b) for frame #104,061, #104,162, #104,311, and #104,460. The segmented glottal areas using the k-means-ACM and DNN are shown in cyan and yellow color, respectively. The DC and the BF scores associated with the two segmented areas, overlaid on each other, are included at the lower right corner of the images.

Figure 3.20 illustrates the performance of the proposed DNN on HSV frames extracted from three different vocalized segments (panel (a)-(c)): frame numbers 40,505-41,204 (panel (a)), 98,732-99,451 (panel(b)), and 106,118-108,084 (panel (c)). These frames were selected among those that were not used for training or testing the network, showing the performance of the network for new frames. The network was implemented on the entire frame sequence of each vocalization – segmenting the glottal regions across frames. The glottal area of each frame in the sequence was computed and plotted in the figure during each vocalized segment to see how the algorithm can capture the glottal area variations at the onsets and offsets. The HSV frames in Figure 3.20 (indicated by red dots in the glottal area waveforms) were selected during different behaviors of the VFs. As such, for the two vocalizations in panel (a) and (b), the segmented frames are extracted near the voicing offset and onset at 138-172.5ms and 14-33.5ms, respectively. The segmented frames shown in panel (c) were extracted during the sustained oscillation of VFs between 222.5 and 229.5ms – representing sudden larger degree of VFs abduction during the sustained vibration.

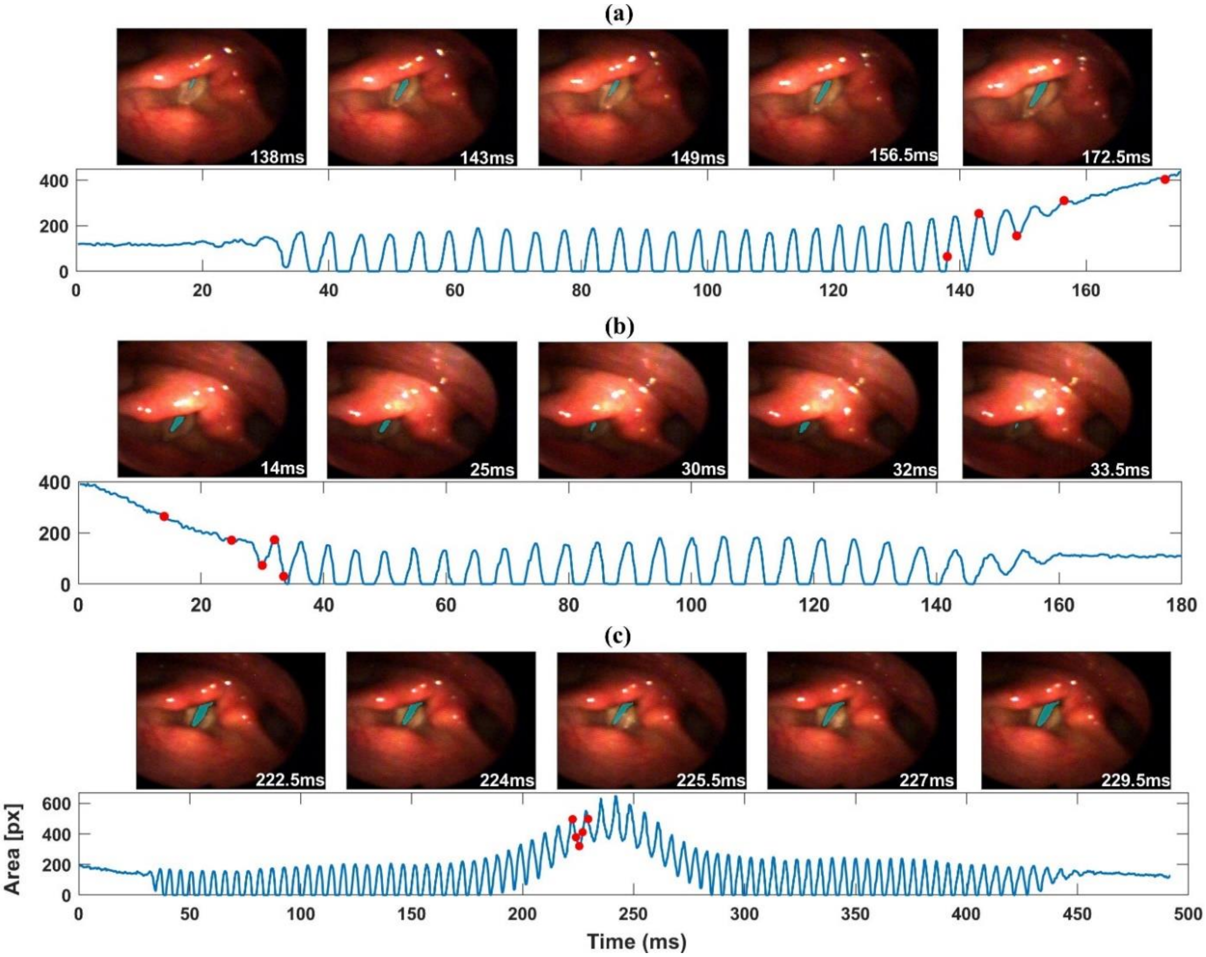


Figure 3.20. The glottal area waveform as well as five segmented HSV frames after applying the trained neural network at three different vocalized segments between frames: 40,505-41,204 (panel (a)), 98,732-99,451 (panel(b)), and 106,118-108,084 (panel (c)). The selected frames are marked by red dots on the glottal area waveforms.

Besides the visual inspection, the network was also tested against manually labelled frames (testing dataset) in order to provide a quantitative evaluation of the segmentation performance. When the proposed network was applied to the testing dataset, the results revealed promising accuracy scores and a good match between the predicted glottal area in comparison with the manually segmented glottal area in the testing frames. As such, the results demonstrated that the mean IoU and DC of the segmented glottal region were 0.82 (STD: 0.26) and 0.88 (STD: 0.25), respectively; STD refers to the standard deviation. In addition, the contour-based evaluation metric (BF score) showed a mean value of 0.96 (STD: 0.12) in terms of detecting the glottal area boundary.

3.3.2. Deep Learning Approach: Neural Network on Monochrome HSV Data

The network which was trained and evaluated before on the color HSV dataset and whose results were discussed in the previous subsection was retrained on and applied to the monochrome HSV dataset. The results of this application are presented here. A sample of the images from the training dataset are shown in Figure 3.21. The top panel shows HSV frames from the vocally normal subjects and the bottom panel depicts HSV images from the AdLD subjects. Each panel illustrates a variety of image qualities and different gestures of VFs during several phonation tasks in running speech. The images display the VFs in different spatial locations, orientations, scales, and brightness conditions. Also, various states of VF behavior are shown, e.g., during vibrations, adduction/abduction with various degrees, and when the VFs are not vibrating. Additionally, other images show partial and full obstructed view of VFs due to the movement of the epiglottis, the constrictions of the arytenoid cartilages, and when the VFs fall outside the endoscope view.

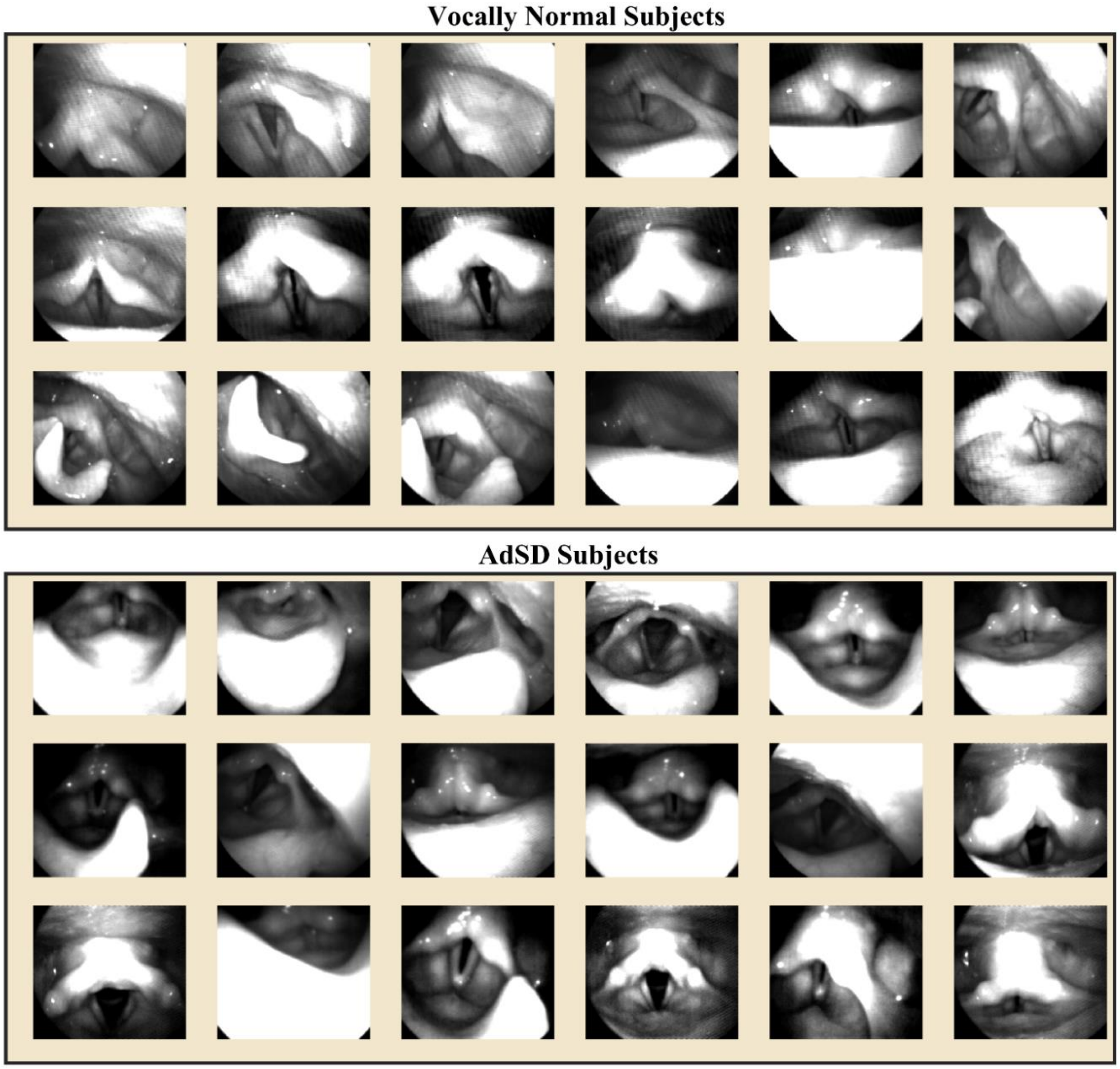


Figure 3.21. A sample of images considered in the training of the developed deep neural network. Top and bottom panels show HSV frames from the vocally normal subjects and AdLD (referred as AdSD subjects in the figure) patients, respectively.

After building the training dataset and manually segmenting the images using the labeling tool, an independent test set was created from new images, different than the training ones, from the HSV recording of an AdLD subject. A sample of the testing HSV frames is depicted in Figure 3.22. The figure includes two different sets of frames: individual testing images (top panel) and a set of consecutive frames. The top panel displays random HSV testing frames showing both a clear view of VFs with different gestures and a partial/full covered view of VFs with different obstructions. Additionally, in the bottom panel, a sequence of HSV frames is illustrated that shows several cycles of VF vibration. The short sequence includes 27 images, and these images were selected with a step size of 12 frames from the HSV recording. The sequence displays a partial obstruction of VF by the right arytenoid cartilage. Due to the laryngeal maneuvers, the right arytenoid cartilage came very close to the endoscope in these frames displaying a bright area. It can be noticed that during the oscillation of VFs, the arytenoids are also moving, which is clear when comparing the first and last images in the sequence. That is, in the first frame, most of the VF tissues are covered by the right arytenoid cartilage while it is almost unobstructed in the last frame.

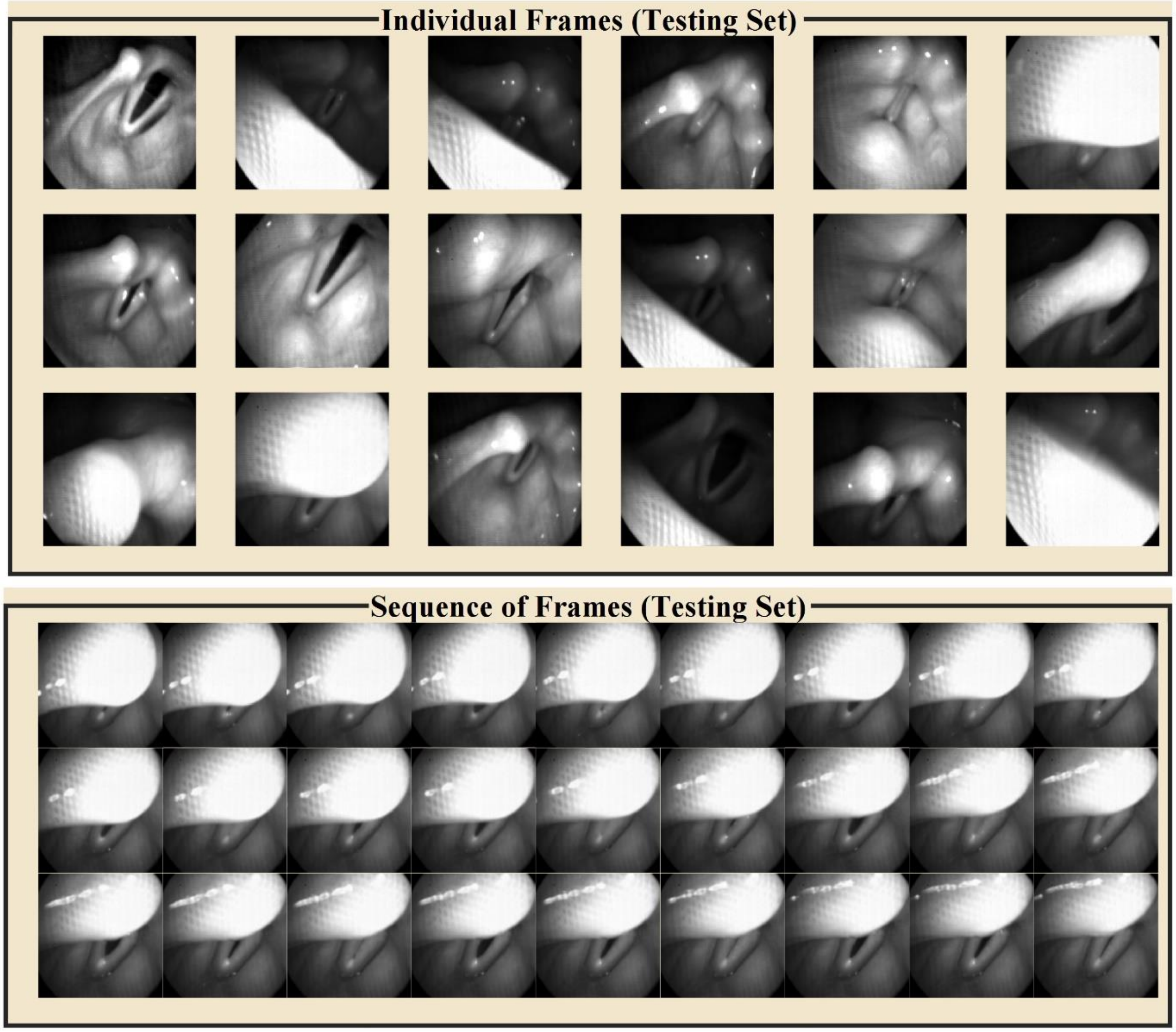

Individual Frames (Testing Set)
Sequence of Frames (Testing Set)

Figure 3.22. A sample of random individual frames (top panel) and a short sequence of HSV frames (bottom panel), selected from the test dataset to evaluate the developed neural network.

The performance of the developed DNN on the testing dataset is shown in Figure 3.23. The figure illustrates the results of applying the developed DNN on 14 testing frames. Some of these frames are selected from Figure 3.22 (top panel) that display different VF configurations with relatively open glottal area and various image qualities in addition to a few testing frames (not in Figure 3.22) for a better segmentation evaluation. The testing frames are divided into two panels for clarity. For each frame, the original HSV image, before segmentation, is displayed besides two binary segmentation masks: one resulted from manual segmentation of the glottal area (ground truth labeling) and another one resulted from the automated segmentation by the DNN. Additionally, a zoomed-in view of the segmented glottal area is shown in the last column of each test image, where the manually detected area (in yellow) and the automatically segmented one (in blue) are overlaid on top of each other. Moreover, the DC values, which allow for a quantitative measurement of the similarity between the two detected glottal regions, are included inside each segmented frame. Hence, both subjective and quantitative evaluation of each frame is provided in the figure. As can be seen from the visual information in the figure, the automated approach demonstrates a large degree of match with the manual segmentations. Furthermore, this high match is reflected in the DC scores since most of the frames achieved DC values above 0.85. It can be noticed that, among the 14 images in Figure 3.23, two images show an obstructed view of the VF and glottal area. In these two images, the automated DNN results in a complete black mask, similar to the manual mask, with a DC of 1 (perfect match) – indicating the high ability of the network to recognize the absence of the glottis in these challenging images.

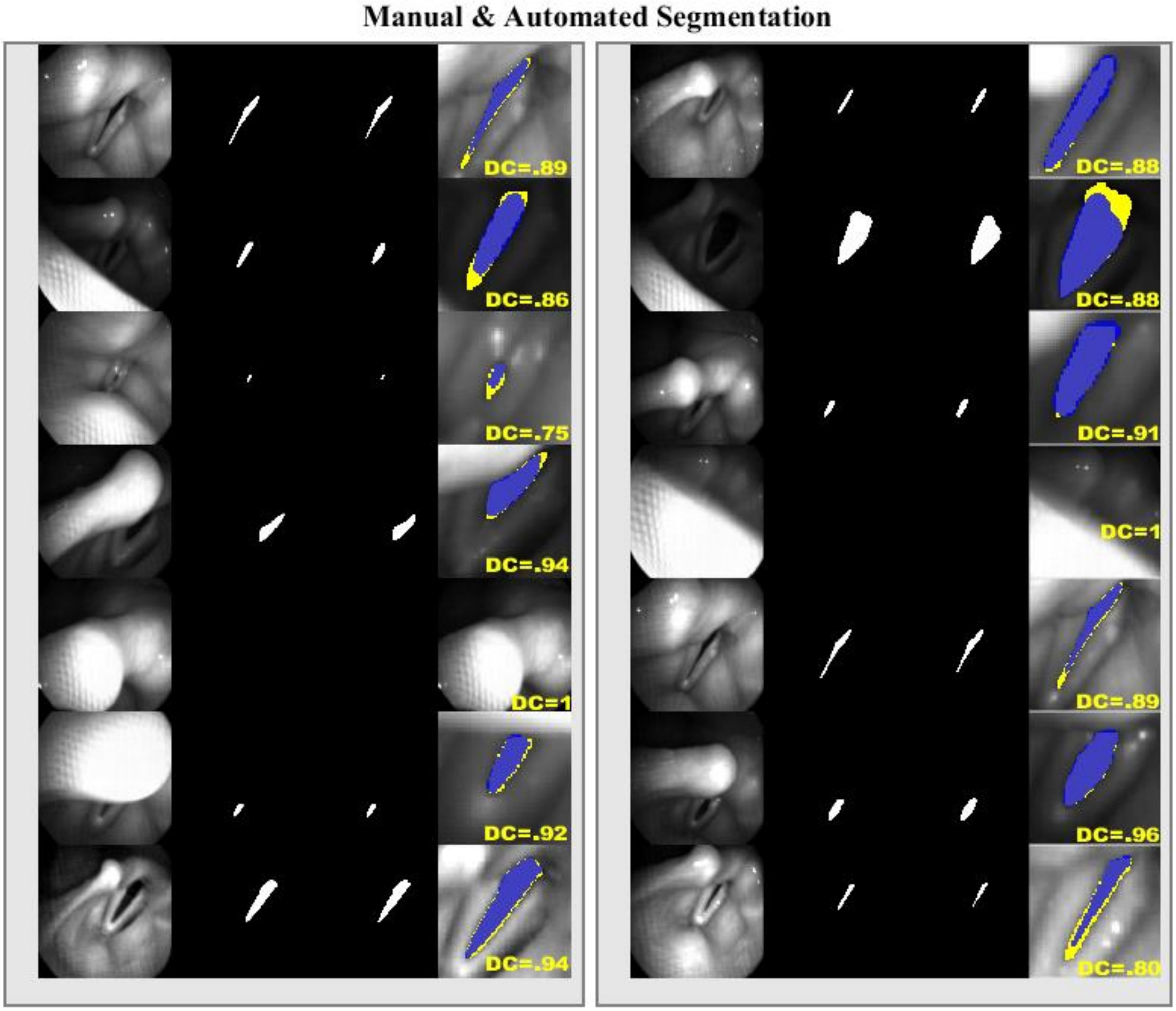


Figure 3.23. Performance of applying the developed neural network to segment the glottal area in a sample of 14 testing frames. The results of the different testing frames are displayed in two panels (7 per each panel). The original HSV images are illustrated besides the corresponding segmentation masks of the manual (second column in each panel) and automated analysis (third column in each panel). The segmented areas of each testing image are shown in the last column of each frame, where the manual (in yellow) and the automated (in blue) segmented areas are overlaid on top of each other.

In addition to the evaluation of the DNN shown in the previous figure (Figure 3.23) on individual frames, the DNN was also assessed by applying it to the sequence of testing frames presented in Figure 3.22 (bottom panel). This sequence was a subset of the original testing set (1,000 frames) on which the performance of the neural network was demonstrated in Figure 3.24.

This subset included 166 consecutive HSV frames, which contained about 9 abduction-adduction cycles of VF vibrations. Among the 166 images, the results of implementation to 32 frames are displayed in the top panel of Figure 3.24. These sample frames, displayed in the top panel, were selected such that the glottal area is relatively large to facilitate the visual evaluation between the manual and the automated segmentation performances for the reader. As such, on each sample image in the top panel, the manually segmented glottal region and the automatically segmented one are overlaid on top of each other to demonstrate the discrepancy/match between them. Manual annotation is highlighted in yellow while the DNN segmentation is depicted in blue color. Also, the corresponding time (in ms) of the frame is included on each consecutive frame in yellow font. As shown, there is a considerable agreement between the results of the developed DNN and the manual labeling in the displayed frames.

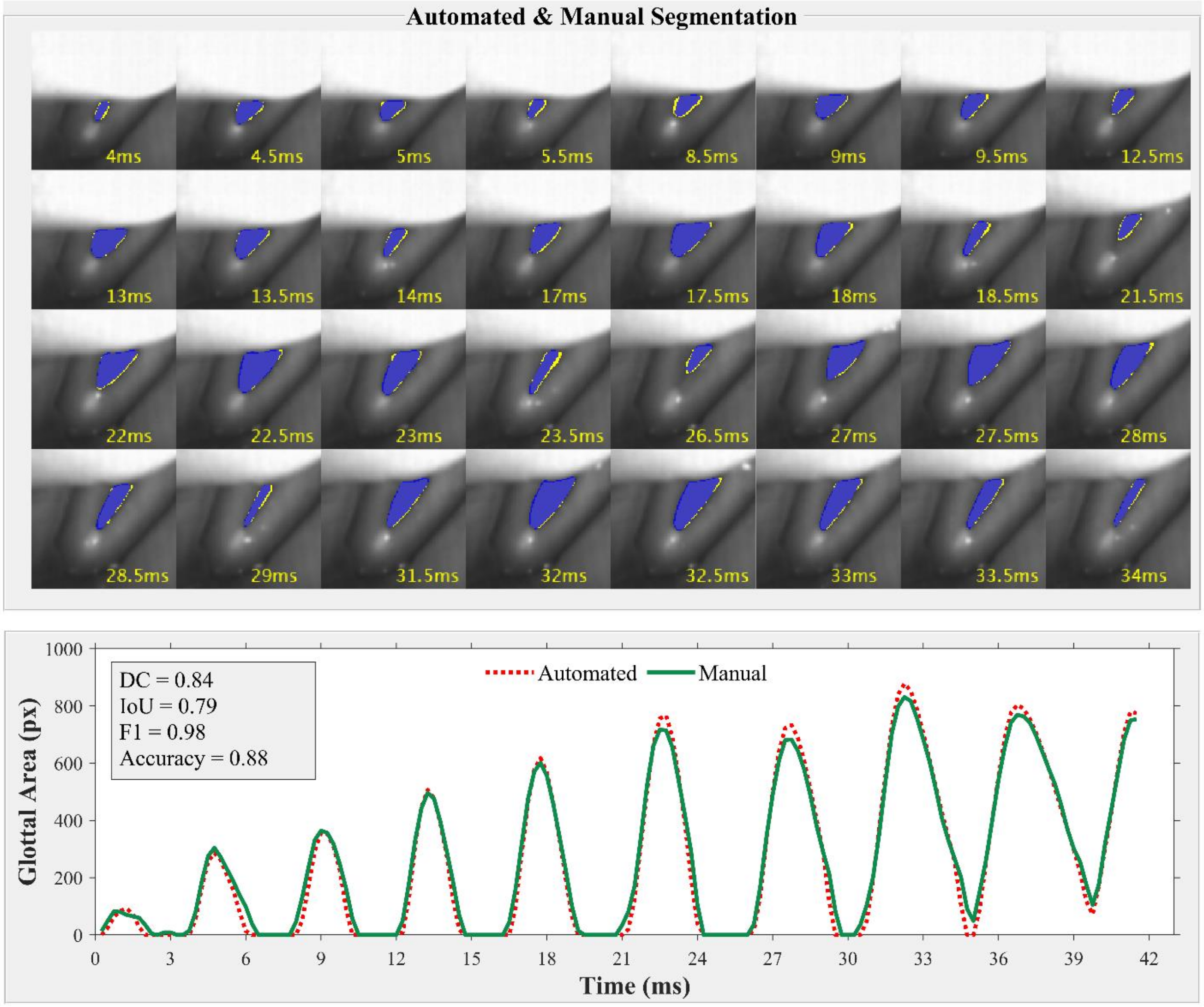


Figure 3.24. Results of applying the developed automated approach to a testing sequence of HSV frames (166 consecutive images). The top panel shows a comparison between the manual (in yellow) versus automated (in blue) glottal area segmentation on 32 frames, selected from the sequence; the corresponding time of each frame is shown in yellow font. The bottom panel illustrates the segmented glottal area variation across the 166-frame sequence (in ms) when computed via the manual (in solid green line) and the automated method (in dotted red line). DC, IoU, F1, and accuracy scores are also included in the bottom panel.

As can be seen in Figure 3.24, the glottal area waveform of the testing sequence is plotted in the bottom panel. The plot demonstrates the change in the segmented glottal area (measured in pixels) across the 166-frame sequence (measured in ms) as a result of using the manual analysis (green line) and the automated DNN (red dotted line). Both glottal area waveforms match well and exhibit a similar behavior across most of the sequence frames – demonstrating the promising performance of the developed model. Also, there is a slight discrepancy observed between the manually and automatically computed areas during few instances when the VFs are fully abducted. During these instances, the DNN slightly overestimates the glottal area. In addition to the subjective evaluation, a quantitative assessment of the DNN was also considered by providing three metrics. The DC, F_1, and accuracy scores were computed to quantitatively assess the segmentation quality of the DNN on the testing sequence. As shown, the mean DC, IoU, F_1, and accuracy values are 0.84, 0.79, 0.98, and 0.88, respectively – demonstrating the high segmentation quality in terms of both the detected glottal region and its edges.

The developed automated DNN was also tested on the entire testing dataset, which consisted of 1000 HSV frames in order to provide a quantitative evaluation of its segmentation performance. When the DNN was applied to the entire testing images, including both segregated frames as well as short sequences of VFs vibration in consecutive frames, the results revealed high segmentation accuracy and a good match between the estimated glottal region against the manually segmented one. The outcome mean scores of IoU, DC, and accuracy were 0.81, 0.86, and 0.89, respectively, revealing a high similarity between the manual and automated analysis as well as the promising performance of the developed DNN in detecting the glottal region in the testing frames. In addition, the automated model demonstrated high precision in estimating the glottal boundaries and VF edges with a mean F_1 score of 0.93.

In addition to the results presented above in terms of the performance and the validation of the proposed tool in detecting the glottal area (using, e.g., IoU) and its boundary/edges (using, e.g., F_1), another method was developed for left and right VF detection. The following results illustrate the performance of the developed tool in detecting the left and right VF edges. As can be seen, Figure 3.25 shows three HSV frames. The first frame demonstrates the precise detection of the midline points in cyan color along with the fitted second-order curve in yellow color. The second image exhibits the HSV frame including only the midline for clarity. The third image represents the detected left and right VF edges plotted with the midline showing the successful description of the computed midline in capturing and matching the shape of the glottal area.

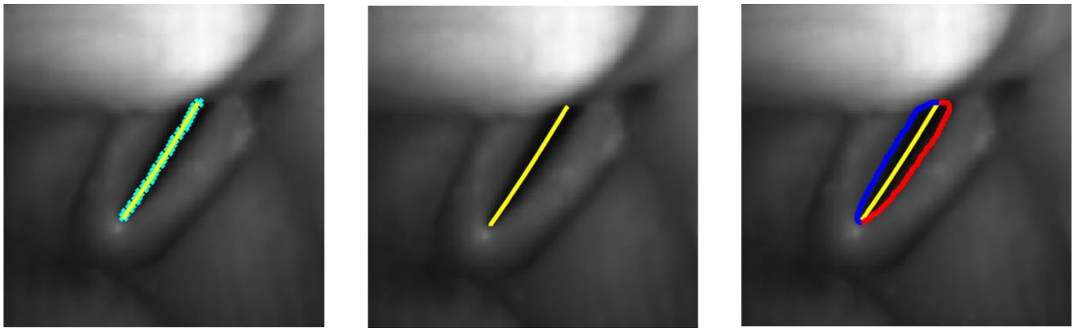


Figure 3.25. A schematic diagram of the detected glottal line on a HSV frame showing the midline in yellow color along with the detected left and right vocal fold edges (in red and blue color , respectively).

Figure 3.26 shows the results of the DNN when applied to 12 individual HSV frames showing different configurations and gestures of the VF in terms of its location, size, image quality, and being partially obstructed. Seven frames were presented in the figure organized in five rows. Each row shows different segmentation results except the first one which depicts the original HSV frames. The second and third row exhibit the glottal area segmentation results in a form of segmentation mask and original frames with highlighted glottal area in red. The last two rows show the detecting of the midline in yellow and the left (in red) and right (in red) VF edges. As can be seen, the developed tool to detect the left and right VF edges was able to accurately capture the various shapes of the glottal area.

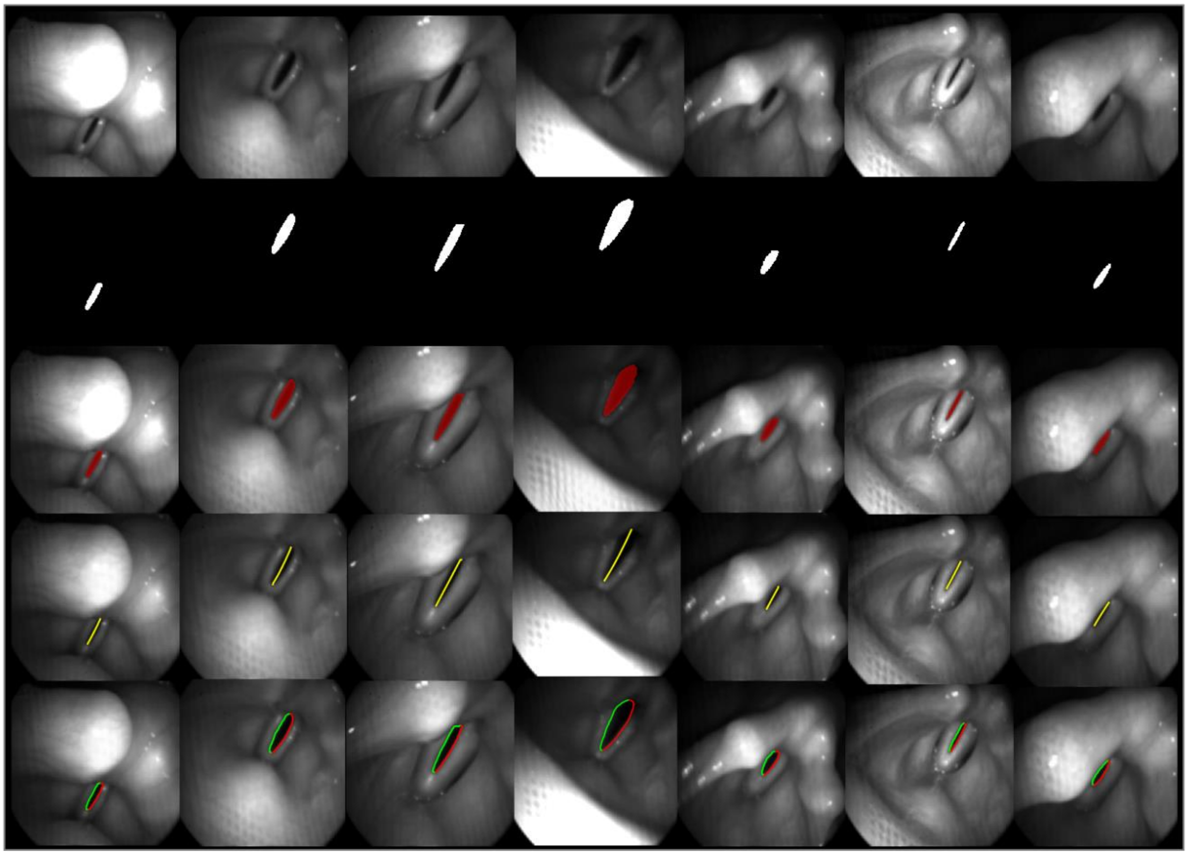


Figure 3.26. Results of the automated segmentation on seven different HSV frames with different vocal fold gestures. Five rows shown in the figure represent the original HSV frames, segmentation masks (glottal area in white), segmented glottal area (in red), the glottal midline (in yellow), and the left and right vocal fold edges (highlighted in red and green color).

Another figure is included in order to demonstrate the performance of the developed tool for midline and VF edge detection. Figure 3.27 includes results related to the automated segmentation in terms of detecting the glottal area and the VF edges in a sequence of around 760 frames. The top panel illustrates the segmented frames displaying the segmented glottal area in red and VF edges in blue and red for the right and left VFs, respectively (along with the detected glottal midline in yellow). Bottom panel depicts the extracted glottal area waveform (measured in pixels) including the timestamps where the frames were extracted shown in red dots. The timestamps are also included in the frame images in the figure in white color. As shown in the figure, the developed tool was able to capture the glottal area as well as the left and right VF edges during different temporal locations within the presented HSV sequence. The sequence represents the high-quality performance of the DNN tool in capturing the detailed shape and sizes of the glottal area –

including when the VFs are widely open as well as when the VFs are vibrating. Also, the detected glottal midline was able to follow the different gestures of the glottal area.

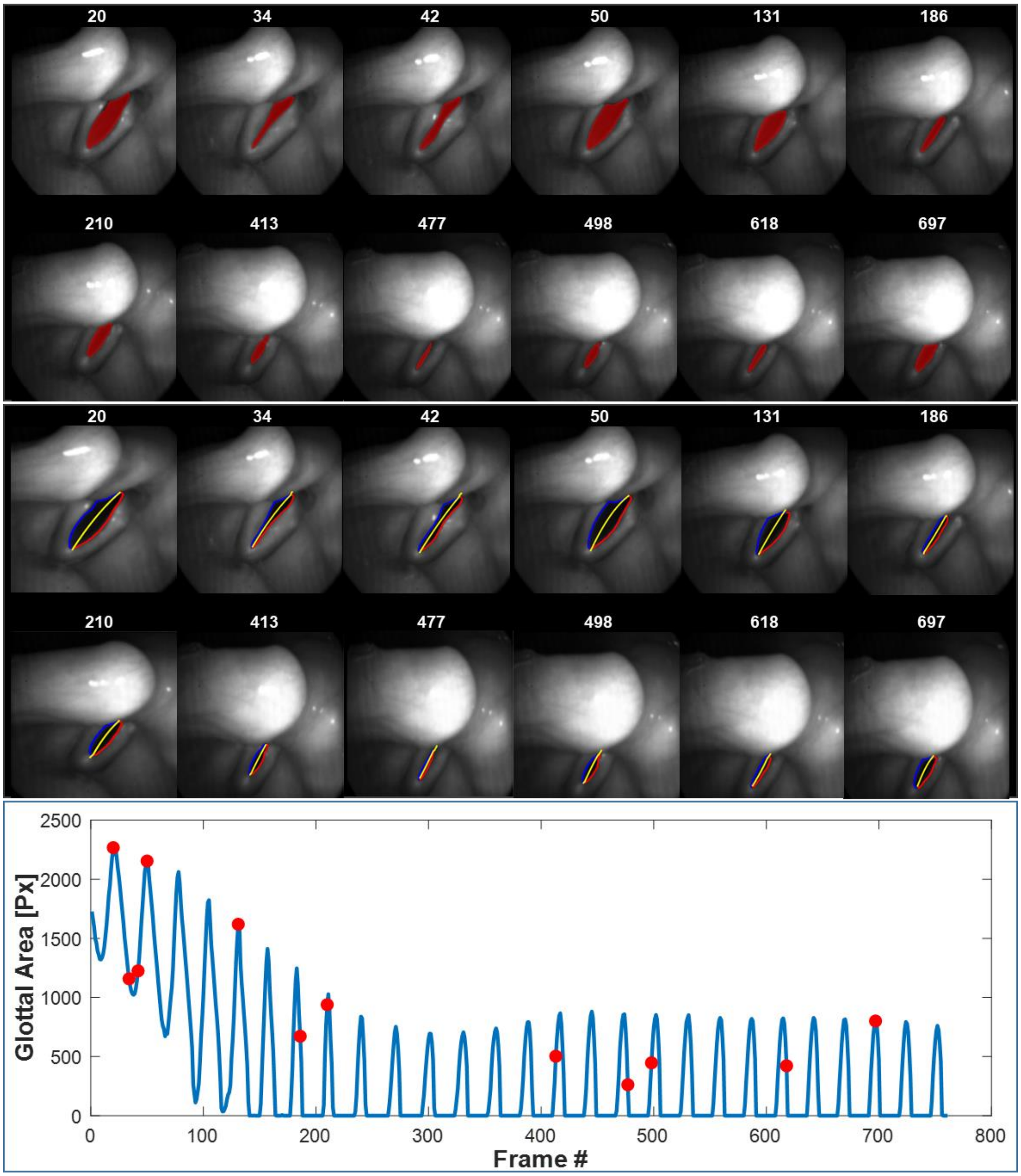


Figure 3.27. Automatically segmented glottal area and vocal fold edges in a sequence of 12 HSV frames depicted in the top panel. The segmented area, midline, left and right vocal fold edges are highlighted in red, yellow, red, and blue. Bottom panel illustrates the extracted glottal area waveform (measured in pixels). The timestamps where the frames were extracted are shown in red dots and included on frame images.

3.4. Study IV: Automated Measurements of Glottal Attack and Offset Time

In this section, the deep learning tool, introduced earlier, was successfully implemented and was able to accurately determine the glottal area across the HSV frames during each the different vocalized segments in each monochrome video of each subject. Accordingly, the edges of the VF were precisely determined based on the segmented glottal area/boundary. This accurate segmentation enabled the successful computation of the GAT and GOT measurements corresponding to the onset and offset of each vocalization, respectively. This section demonstrates the results associated with developing the GAT and GOT measures.

Results from applying the developed deep learning tool to a set of sequential frames that were captured during a phonation onset are presented in Figure 3.28. The figure illustrates the outcomes of this implementation. The top panel displays a subset of 12 HSV frames selected from various timestamps throughout the sequence and segmented using the automated tool. The segmented glottal area is shown in green, and the corresponding timestamps were indicated on each segmented image in yellow font. As can be seen, the segmented frames demonstrate the accurate detection of the glottal area via capturing the complex details of the glottis. The timestamps, at which the segmented frames were selected, were marked by red dots on the generated GAW, which is illustrated in the bottom panel to help with temporal referencing. The automatically developed GAW in the bottom panel provides an accurate visual representation of the change in the glottal area and VF behavior during the initiation of phonation in terms of pixel measurements.

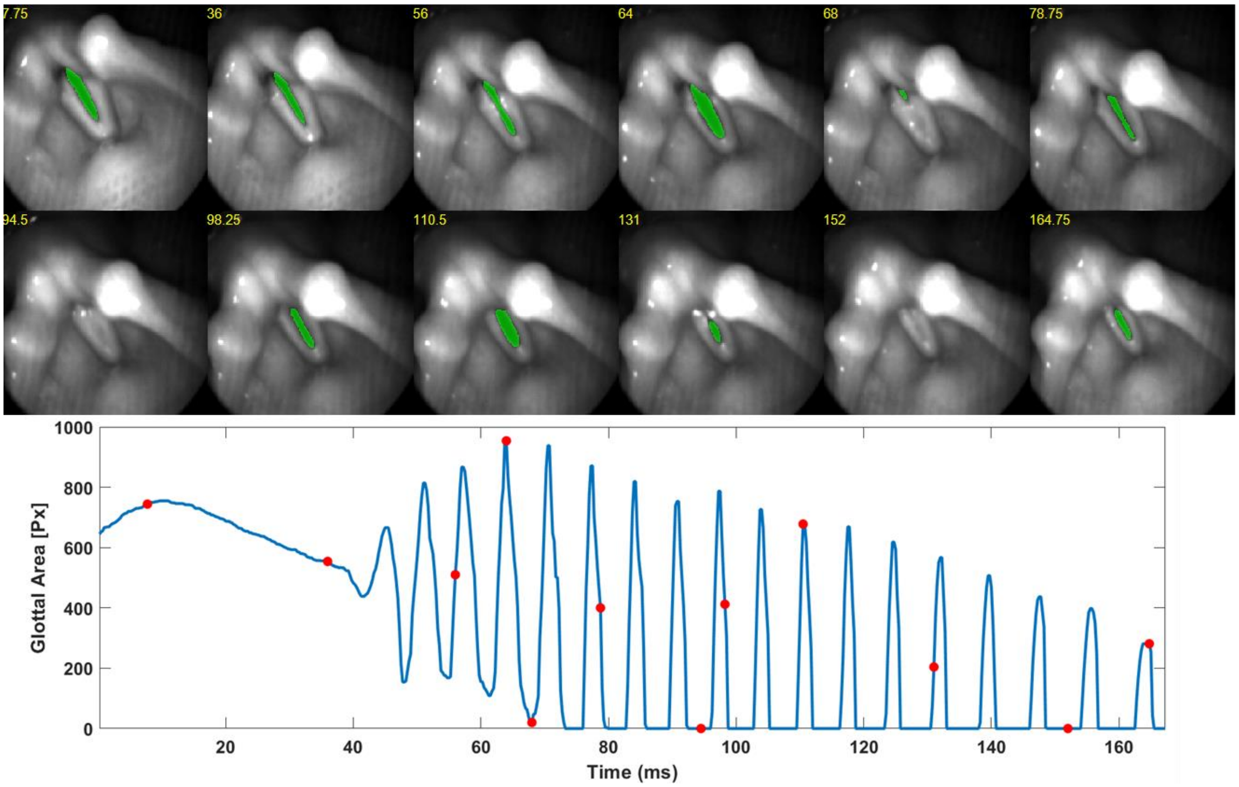


Figure 3.28. Results of applying the developed deep learning tool to a sequence of frames during phonation onset. The top panel shows automated glottal area segmentation (highlighted in green) on 12 HSV frames, selected from different timestamps within the sequence. The corresponding time is indicated on each frame in yellow font. The bottom panel illustrates the segmented glottal area variation (measured in pixels) across the sequence during the onset of phonation.

Results of the automated measurement of the GAT during phonation onset of a vocalization during running speech are shown in Figure 3.29. The automatically generated GAW (normalized) and the average medial glottal contact waveform are displayed in the top two panels, in red and blue, respectively. The GAW and contact waveform energy contours are also illustrated in the figure bellow the associate waveforms with the same related colors. At the bottom panel, the cross-correlation's results are shown in a green curve. On the cross-correlation graph, the automatically computed GAT is indicated by the time difference between the two horizontal dashed lines. As shown in the figure, the automated algorithm was able to detect the energy rise of both the GAW and the contact waveform and, accordingly, compute the delay between the two energy lines. In this example, the automated algorithm reveals a delay time of 14.75 ms – referring to the GAT value during this particular phonation onset.

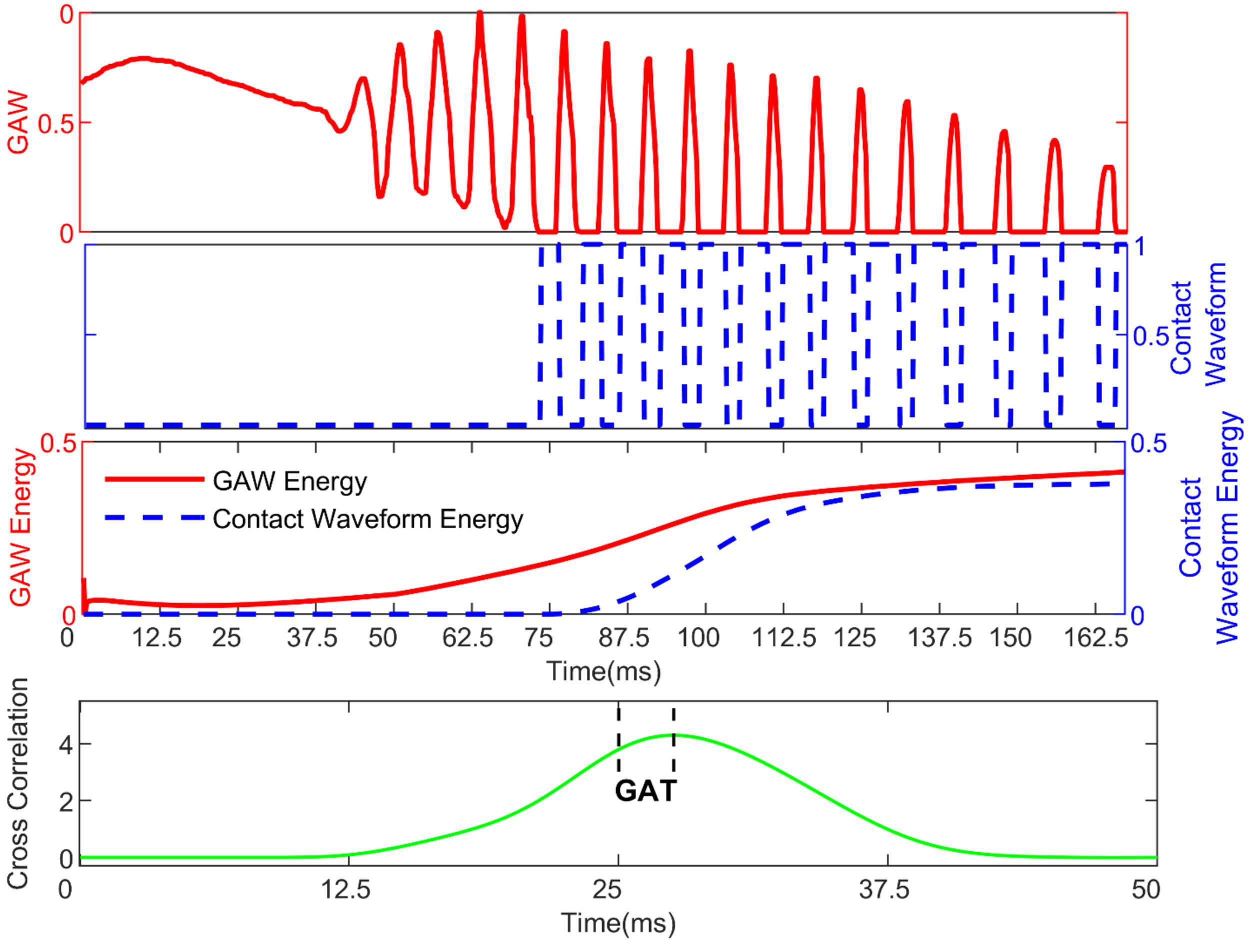


Figure 3.29. Results of the automated measurement of the glottal attack time (GAT) during a phonation onset. The top two panels show the magnitudes of the normalized glottal area waveform (GAW) in red color and the average medial glottal contact waveform in blue. The bottom two panels illustrate the energy contours corresponding to each waveform along with the outcome of the cross-correlation in green color. The measured GAT is marked on the cross-correlation graph as the time delay between the two horizontal dashed lines (14.75 ms).

The outcome from implementing the DNN tool to a sequence of HSV frames during the offset of phonation is introduced in Figure 3.30. The figure has the same formatting as Figure 3.28 – including two panels (top panel showing a sample of 12 segmented frames within the sequence and bottom panel showing the generated GAW during phonation offset). As shown in the top panel, although the segmented frames were displayed during a short period of time (short sequence) during running speech, different image quality and altered configurations/views/sizes of the VFs can be observed across the different frames. Despite that, as can be seen, the automated segmentation tool was able to precisely detect the glottal area regardless of these variations.

Moreover, the segmented GAW, shown in the figure, exhibits accurate illustration of the dynamic characteristics in the glottal area and VF behavior during the offset of phonation. The plotted GAW accurately represents and captures not only the oscillation during the steady-state vibration potion but also the small-amplitude oscillations existed toward the end of the phonation offset.

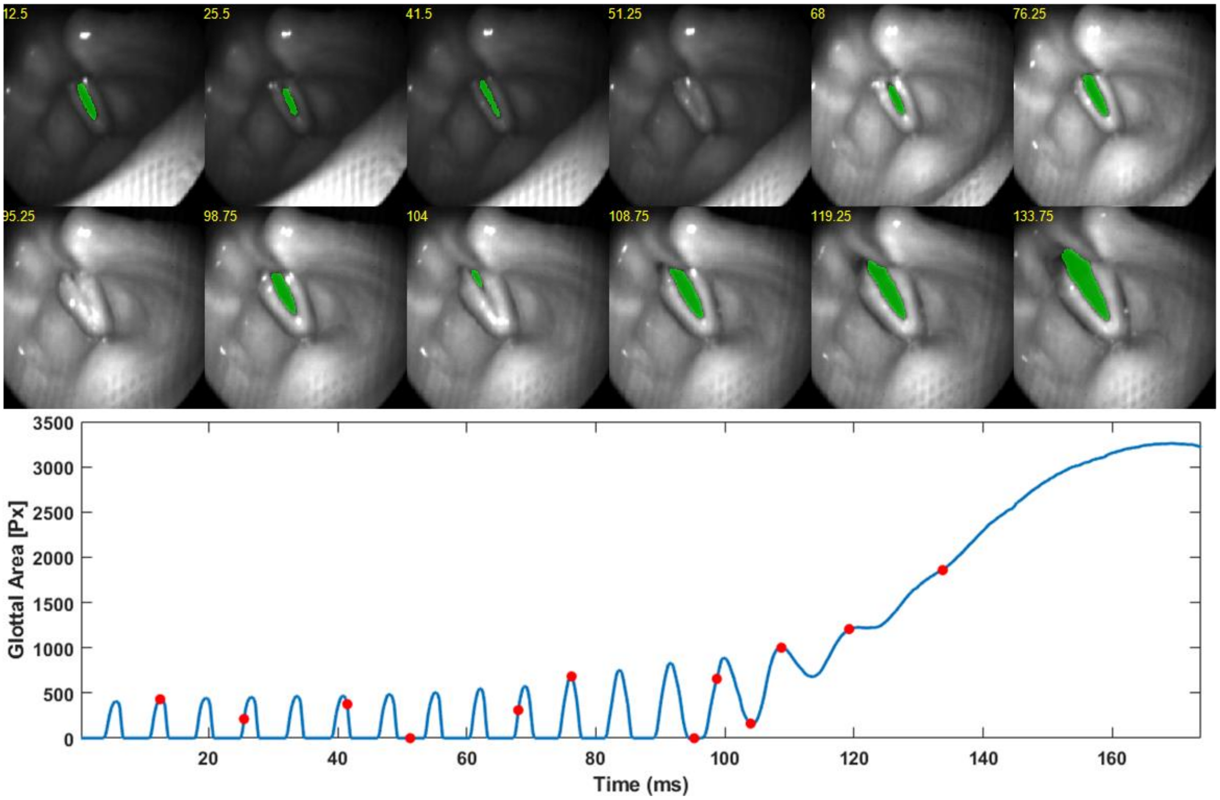


Figure 3.30. Results of applying the developed deep learning tool to a sequence of frames during phonation offset. The top panel shows automated glottal area segmentation (highlighted in green) on 12 HSV frames, selected from different timestamps within the sequence. The corresponding time is indicated on each frame in yellow font. The bottom panel illustrates the segmented glottal area variation (measured in pixels) across the sequence during the offset of phonation.

Figure 3.31 depicts the results of the automated GOT measurement during a phonation offset selected from a running speech sample. Similar to Figure 3.29, the top two panels display the normalized GAW and average medial contact waveform, automatically generated using the segmentation tool and visually represented in red and blue, respectively. The derived energy contours of each waveform are illustrated. The two energy waveforms show the accurate representation of the drop of the oscillation energy corresponding to the damping motion of the VF at the end of phonation. As can be seen in the figure, there is a time lag between the two energy lines which can be precisely observed from the cross-correlation graph. This lag is shown between the dashed lines indicating the offset time between the two waveforms which is 28 ms in this phonation offset sample.

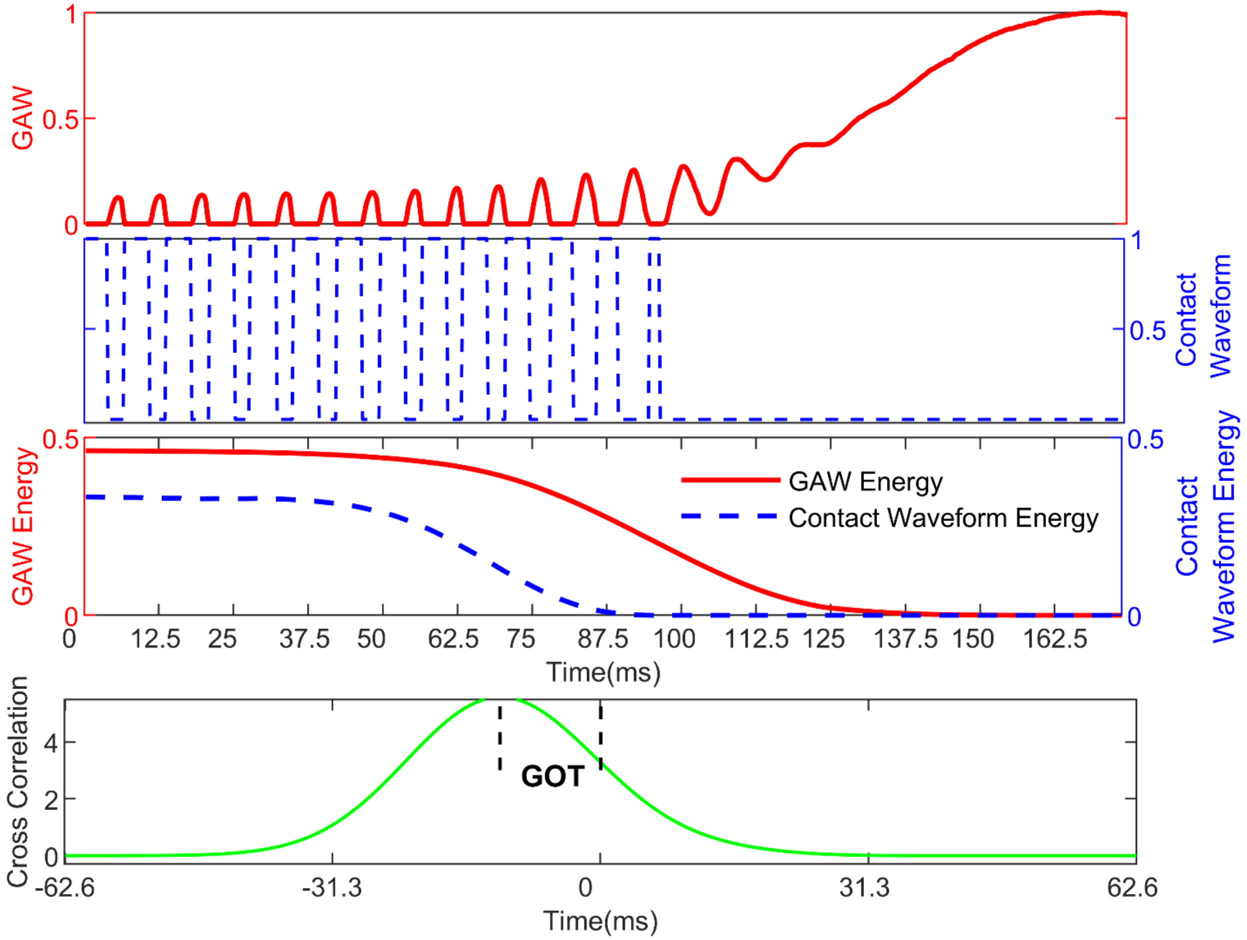


Figure 3.31. Results of the automated measurement of the GOT during a phonation offset. The top two panels show the magnitudes of the normalized glottal area waveform (GAW) in red color and the average medial glottal contact waveform in blue. The bottom two panels illustrate the energy contours corresponding to each waveform along with the outcome of the cross-correlation in green color. The measured GOT is marked on the cross-correlation graph as the time delay between the two horizontal dashed lines (28 ms).

The developed automated method for computing GAT and GOT was first validated against the visual analysis. This was done by applying the automated method to the HSV recordings of all the subjects and generating the GAT and GOT values during the different phonation onsets and offsets in each recording. The mean values of GAT and GOT during each recording was obtained using the automated algorithm in addition to the corresponding visual measurements. The values computed from the two methods were compared. The results of this comparative analysis is shown in Figure 3.32 for both the vocally normal subjects (N) and the AdLD patients. As can be seen in the figure, the solid blue and green lines indicate the automated measurements of GAT and GOT, respectively, whereas the dashed lines indicate the visual measurements. Overall, the automated measurements precisely detect the GAT and GOT values – showing a close alignment and agreement with the visual analysis in most of the subjects with minimal differences. It can also be observed that the automated measurements demonstrate more accurate values of GAT and GOT in comparison with the AdLD patients – showing a marginally better agreement with the manual measures. In addition, the overall automated computation of the GAT showed slightly more accurate values compared to GOT.

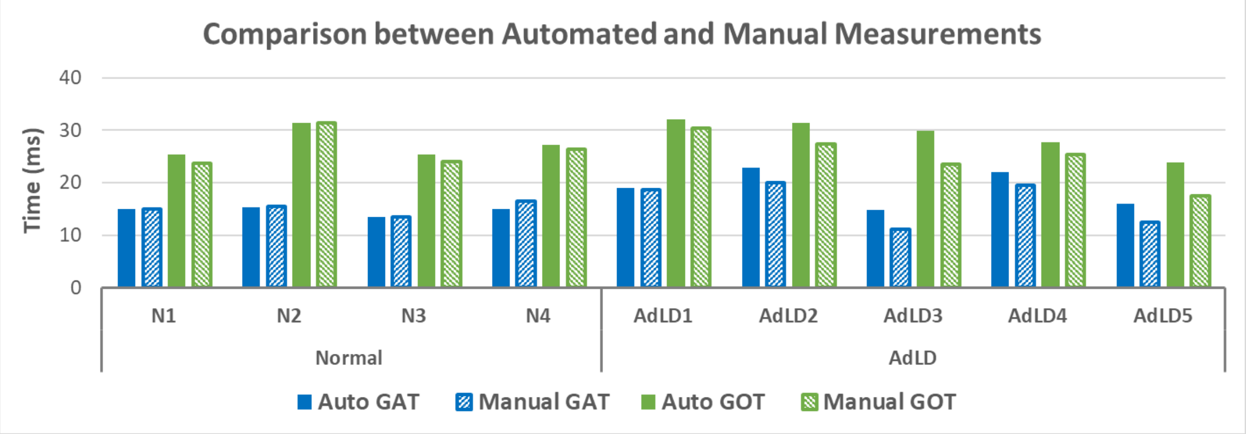


Figure 3.32. Results of the comparison between the automated measurements and the visual measurements of the GAT and GOT, both measured in ms, for the vocally normal participants (N) and the AdLD. The automated measurements of GAT and GOT are shown in solid blue and green bars, respectively. The manual measurements of GAT and GOT are illustrated in dashed blue and green bars.

Overall, the analytical comparison revealed a minimal average discrepancy between the automated and the manual measurements. The average difference for the mean GAT between the automated and manual analysis was 1.6 ms across all the recordings. The mean GOT showed a slightly higher average difference of 2.7ms between the automated and visual measurements. Moreover, an additional quantitative analysis was carried out between the automated and the visual analysis to compare the magnitudes of the GAT and GOT within the levels of the vocalized segments across various subjects. The comprehensive statical analysis demonstrated a strong and significant correlation between the automated and manual measurements in both GAT and GOT.

A high correlation coefficient of Pearson r = 0.93 was found for GAT, suggesting a significant level of agreement between the two ways of measurements. Likewise, GOT measurements demonstrated a strong correlation coefficient of r = 0.91. An independent t-test was performed using the various vocalized segments in order to investigate to what degree the measurements of the automated and manual analysis differ for GAT and GOT. It was observed that there is no statistical difference between the automated and visual measurements – indicating a high level of similarity between the two methods of measurement. That is, the resulted p-value from the t-test conducted between automated and manual measurements was 0.86 in the GATs and 0.77 in the GOTs, refereeing to a minimal statistical discrepancy between the manual versus the automated approach.

After validating the automated algorithm with the visual measurements, the automated method was used to compute the GAT and GOT values for all the subjects where a comparison can be made between the vocally normal subjects against the AdLD patients. Figure 3.33 and Figure 3.34 provide the mean values (shown in blue) along with STD (shown in light orange) of the GAT and GOT measurements, respectively, for each normal control and AdLD participant. As shown in Figure 3.33, the mean GAT values of almost all the AdLD patients were higher compared to the vocally normal individuals with a noticeable difference. In addition, the figure shows a discrepancy in the mean GAT measurements in the AdLD individuals ranging from 14.8 to 22.9 ms than in the normal controls with minimal fluctuation in the values (14.9 – 15.2 ms). In addition, as can be seen from the figure, the STD shows high values in AdLD in comparison with the vocally normal group – referring to the high variability observed in AdLD. Overall, AdLD group had higher average GAT values with (18.95 ms) than the vocally normal group (14.65 ms). The statistical analysis revealed that this difference was statistically significant between the two groups (p-value < 0.001). Conversely, the mean GOT values demonstrated a slight increase in the AdLD group (varying between 23.8 and 32.1 ms) versus the normal controls (ranging from 25.3 to 31.4 ms). Hence, the mean GOT value across all the AdLD patients, marked by 28.9 ms, didn’t demonstrate a statistical significant difference (p-value = 0.2) compared to the vocally normal controls, having an average value of 27.3 ms. Similar to the variability found in the GAT results, the GOT demonstrated larger values of STD in the AdLD group compared to the normal controls.

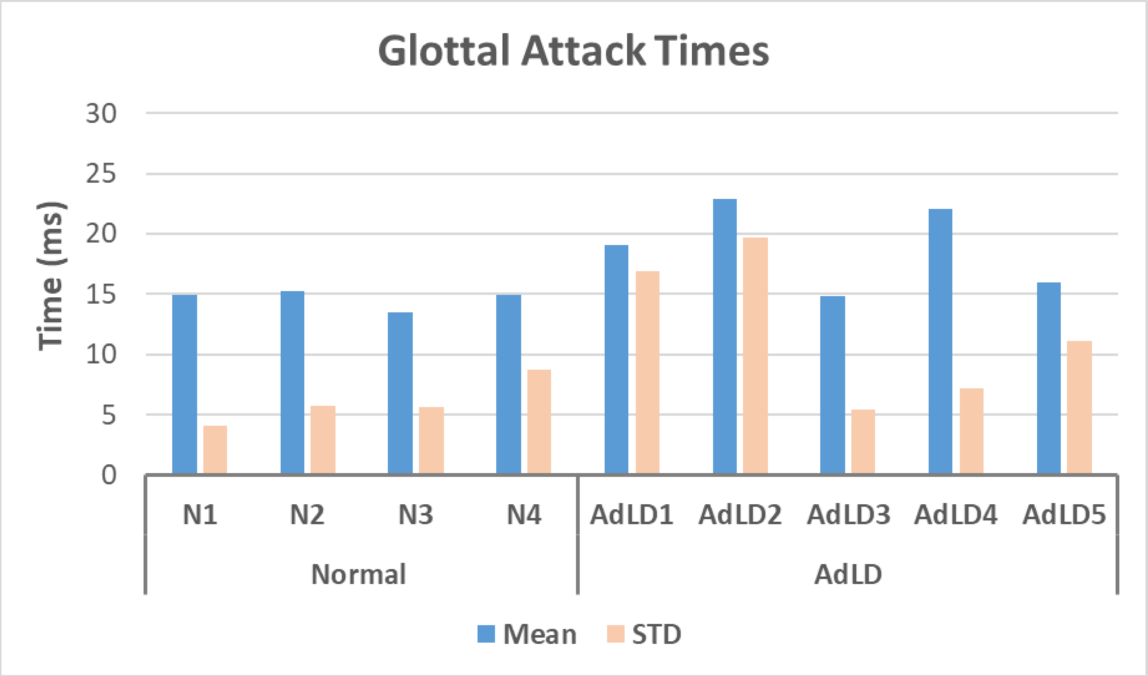


Figure 3.33. Results of the automated measurements of GAT (measured in ms) between the vocally normal participants (N) and the AdLD. The mean and STD values are shown in blue and light orange bars, respectively.

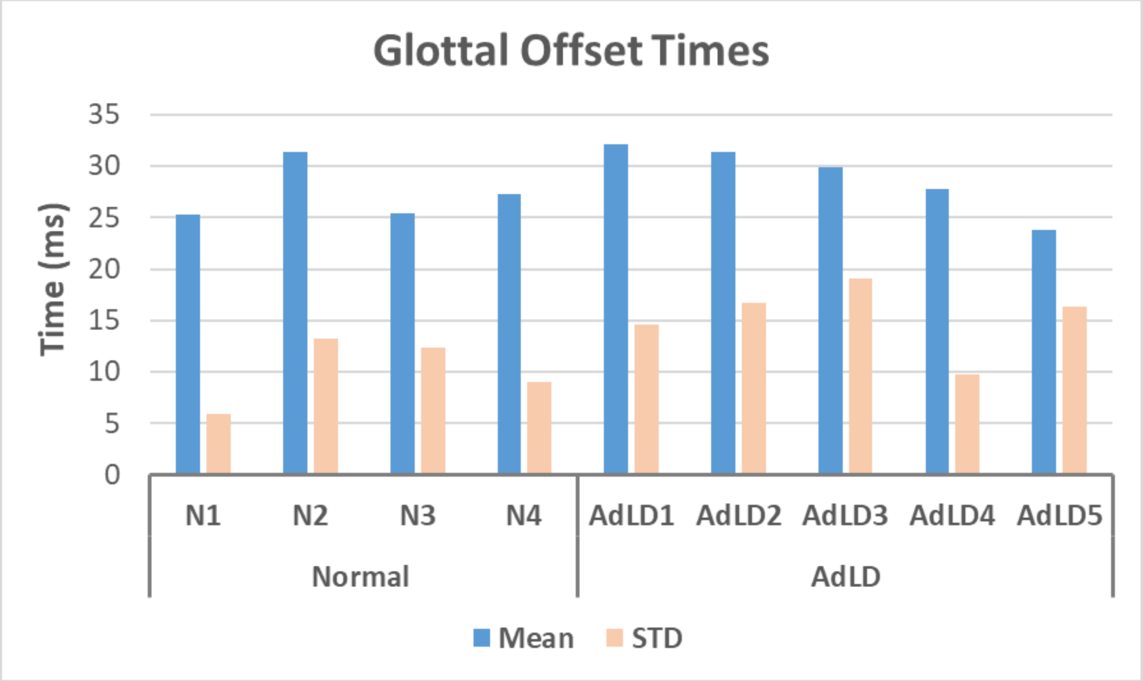


Figure 3.34. Results of the automated measurements of GOT (measured in ms) between the vocally normal participants (N) and AdLD. The mean and STD values are shown in blue and light orange bars, respectively.

3.5. Study V: Lumped Modeling

This section introduces both the simulation and the optimization results obtained using the developed single lumped model of the VFs. The model was built using the one-mass approach such that it can generate oscillatory behavior comparable with the one observed from the HSV data. Hence, after building the model, the experimental data were extracted using the segmentation tool developed, introduced earlier. This experimental data were represented by the automatically generated glottal area waveform. By matching the experimentally extracted glottal area waveform with the simulated one, the optimization was conducted in order to infer biomechanical measurements of the dynamics of the VFs. For that purpose, a vocalized segment extracted from the HSV of a vocally normal participant during running speech – showing VF vibrations – was considered for the analysis. The results corresponding to the modeling simulation and optimization are described in this section.

The results of the simulation are illustrated in the following two figures. The simulation results of the model, presented in these two figures, were based on the input parameters that were chosen earlier in the second Chapter (Methodological Approach). The model constants related to A_{g0}, μ, ρ, l, and d were selected by 0.05 cm^2, 1.86×10−5 $g/(cm^2.s)$, 1.2 × 10−3 g/cm^3, 1.4 cm, and 0.3 cm, respectively. The damping coefficient was considered here as zero and was given a value of 500 g/s *(c')* only during the closure phase of the VFs. The values of m and k were set to 0.24 g and 5000 g/s^2. Also, the subglottal pressure, which was considered another variable here, was parameterized using the following values during the closure phase (P_{Smax} and t_c): 10000 dyn/cm^2 and 2.75 *ms*. Noting that during the opening phase of the VFs, the subglottal pressure was given a value of 8000 dyn/cm^2.

Based on the listed model input parameters and constants, the simulation results are generated in the following two figures. Figure 3.35 shows the resulting glottal area waveform measured in cm^2 as a function of time. The figure illustrates the behavior of the VF vibrations and how the area between the VFs change along with time. Also, the closure phase is represented in this figure as a dashed line below zero showing how the high value of the viscous damping during the closure phase act toward greatly attenuating the momentum energy and damping the motion of the VFs due to their contact.

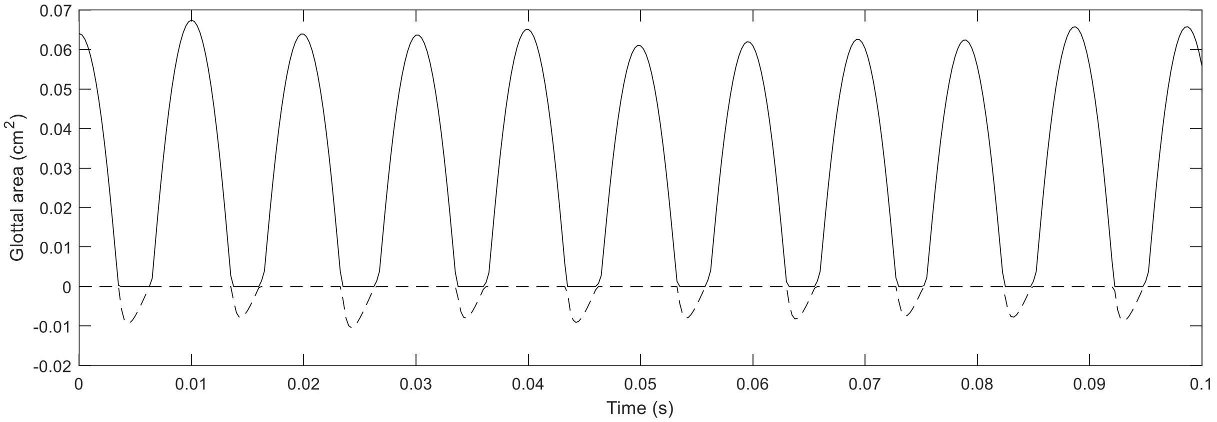


Figure 3.35. Simulation results of the theoretical glottal area waveform as a function of time. The dashed line shows the movement of the vocal folds during the closure phase.

Figure 3.36 illustrates several simulation results corresponding to different model parameters. The figure depicts the change of the mass displacement, glottal air volume flowrate, the subglottal pressure, the elastic restoring force induced by the spring, total external force acting on the vibrating mass, and the damping coefficient during the closure phase. Each subplot shows the behavior and the variation of each variable along with time during the simulation. The displacement subplot refers to the spatial movement of the VF mass, measured in cm. The glottal air flowrate illustrates a similar behavior to the displacement – demonstrating the change in the air flow corresponding to the abduction (maximum flowrate) and adduction (no flowrate) of the VFs. Also, the subglottal pressure and the damping coefficient plots reflect the fluctuation of their values during the closure (maximum build-up pressure and a large value of viscous damping) and the opening of the VFs (typical value of the subglottal pressure and zero damping). In addition, the behavior of the total external force acting upon the VF mass during vibration is illustrated in the figure as well as the characteristics of the force exerted due to the elastic spring forces. Overall, as can be seen from Figure 3.35 and 3.36, the model can capture the oscillatory behavior of VFs and even represent some degree of nonlinearity in the numerical solution. In addition to that, the contact between the VFs was also incorporated into the model as a form of closure time and shown in the simulation plots.

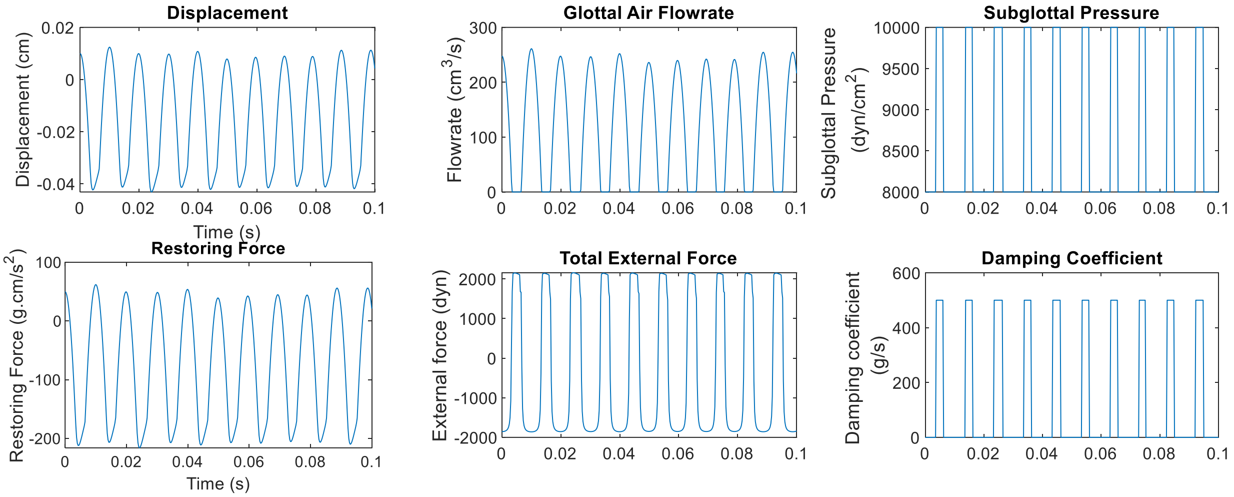


Figure 3.36. Simulation results of the mass displacement, glottal air volume flowrate, subglottal pressure, elastic restoring force, total external force acting on the vibrating mass, and the damping coefficient during the closure phase.

After simulating the proposed lumped model, it was optimized against the experimental data. The optimization was conducted using the simulated glottal area waveform generated from the model along with the glottal area waveform automatically extracted from the HSV data using a vocalized segment of the VF vibration. The optimization procedure was carried out using initial values and constrained ranges associated with the optimization vector *q (α, m (g), k (g/s^2), c' (g/s), tc (s),* P_{Smax} *(dyn/cm^2))* based on the optimization parameters mentioned previously in the second Chapter. The initial values were chosen such that *q* = {120, 0.1 *g*, 40000 *g/s^2*, 400 *g/s*, 0.005 s, 15000 *dyn/cm^2*}, and the constrained ranges associated with each optimization parameter were selected by {110 – 130, 0.04 – 0.30 g, 10000 – 60000 *g/s^2*, 300 – 800 *g/s*, 0.003 – 0.008 s, 8000–20000 *dyn/cm^2*}, respectively.

After running the optimization technique (particle swarm method), the objective function was computed at each iteration of the optimization. Figure 3.37 depicts the convergence plot associated with the optimization approach showing the objective function value at each iteration during the optimization process. The figure shows the normalized error (related to the objective function) along with the iteration number. As can be seen, the objective function value decreases and converges to a minimum value of 0.04155, exhibiting a plateau-like behavior after a specific number of iterations (around 60 iterations). Hence, the eventual objective function returned a normalized optimization error of 0.04155 between the theoretical/simulated glottal area waveform in comparison with the experimental glottal area waveform.

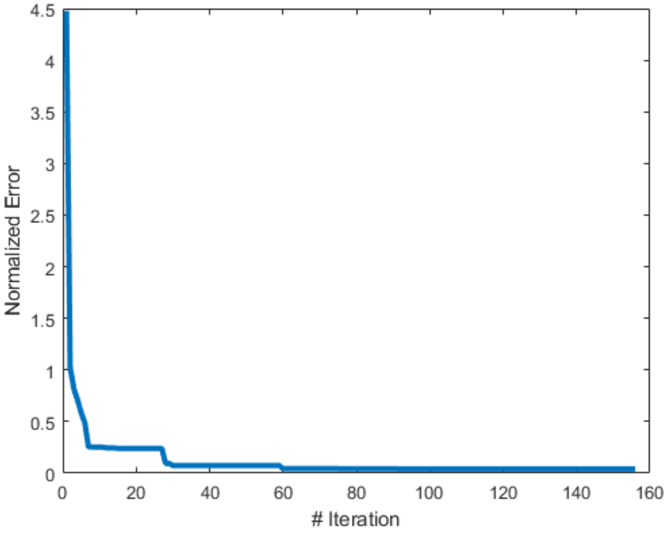


Figure 3.37. Results of the objective function value (normalized error) at each iteration during the optimization process.

Upon the successful completion of the optimization process, the optimization method returned the optimized parameters associated with the model. These generated optimized parameters indicate that their optimal combination was achieved after the iterative optimization procedure and refining the solution. The outcome of the optimization procedure results in a set of optimized values of 124.67, 0.0501 g, 11,787 *g/s*^{2}, 414.69 *g/s*, 0.0030005 *s*, and 8,571.3 *dyn/cm*^{2} corresponding to the optimizing parameters: scaling factor, the mass, the spring stiffness, the damping coefficient during closure, the closure time, and the maximum subglottal pressure, respectively. Based on the optimized parameters, the biomechanical measure of the elasticity index was 235,269 *1/s*^{2} and the viscosity index was 8,277 *1/s*.

The obtained optimized parameters were utilized as inputs to the developed lumped model in order to visualize the resulted simulation. After incorporating the optimized parameters, the simulation of the VF movement was generated. Figure 3.38 illustrates the simulation results of the optimized model. The figure plots the computed glottal area waveform as a function of frame numbers (converted from time). The displayed theoretical glottal area waveform was multiplied by the optimized scaling factor in order to match the experimental glottal area (which was measured in pixels). The figure shows the oscillatory behavior of the mass (VF) including the simulated closure time.

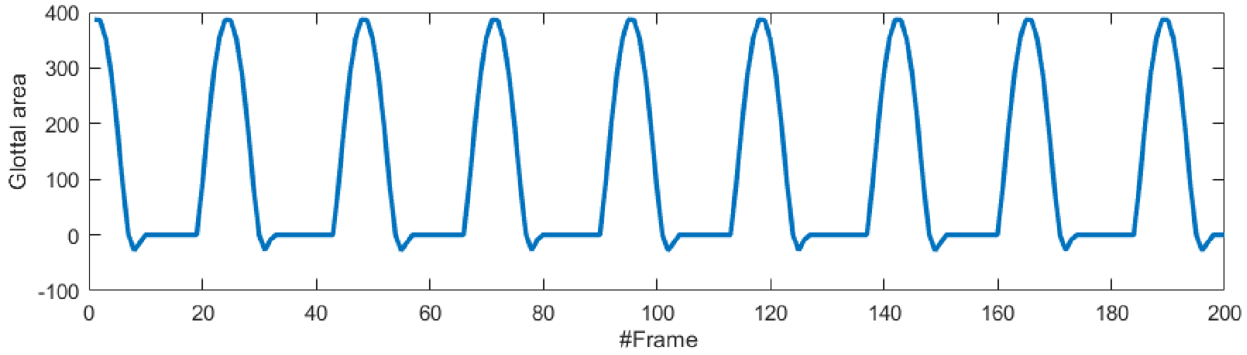


Figure 3.38. Simulation results of the theoretical glottal area waveform as function of time using the optimized parameters.

In addition, the behavior of the optimized subglottal pressure during the simulated VF vibration is illustrated in Figure 3.39. The figure demonstrates the corresponding changes in the subglottal pressure as a function of time. As can be seen in the figure, the pressure oscillates between the typical subglottal pressure value (8000 *dyn/cm*^{2}) and the maximum pressure, which was the outcome of the optimization procedure (8571 *dyn/cm*^{2}). In addition, the plot exhibits the adduction time (the optimized closure time which is t_c) during which the subglottal pressure value was maintained at the maximum optimized pressure (referring to the build-up pressure during the closure of the VFs.

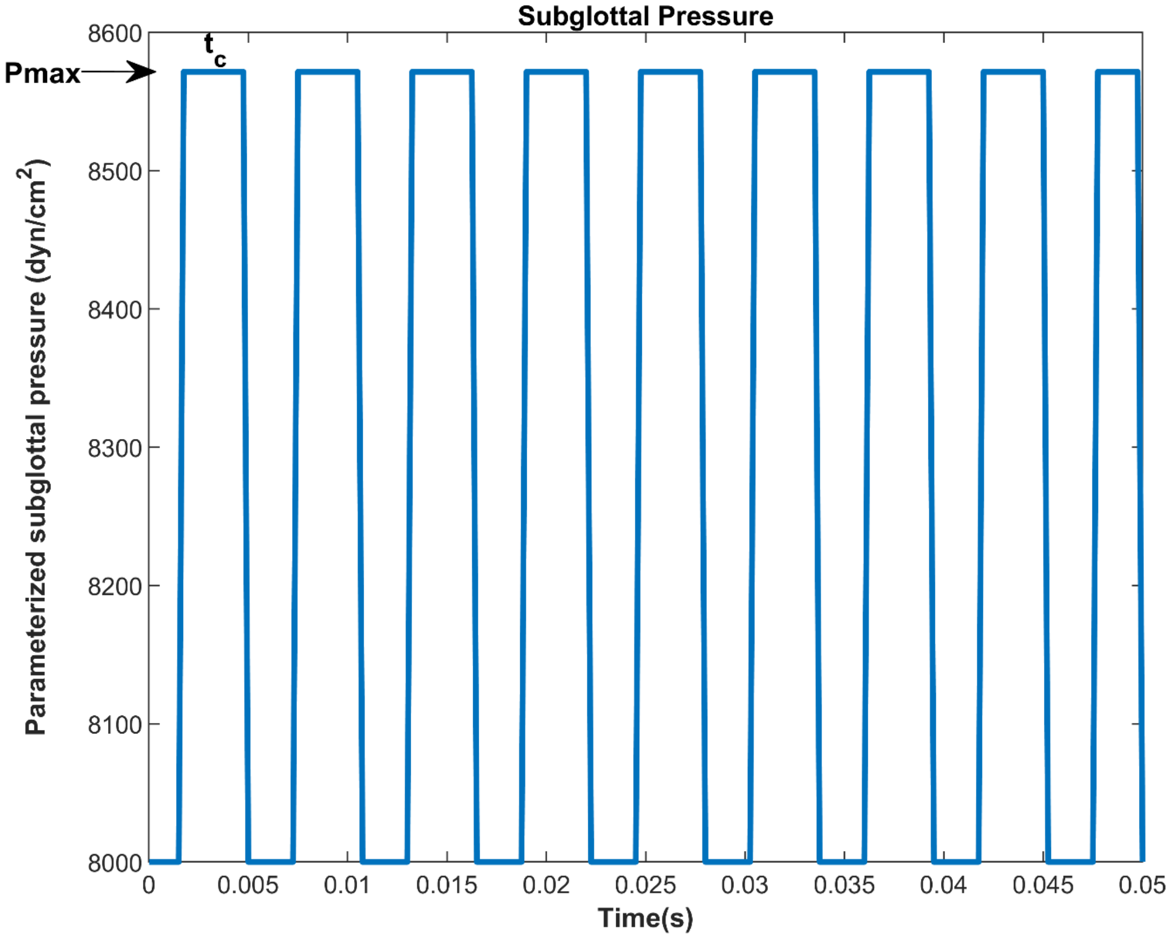


Figure 3.39. Results of the change in the parameterized subglottal pressure during the vibration of the vocal folds as a function of time using the optimized parameters. P_{max} refers to the maximum optimized build-up pressure, and t_c indicates the optimized closure time.

In order to compare between the optimized theoretical glottal area waveform and the experimentally generated one, both waveforms were plotted in the same figure for better visualization as can be seen in Figure 3.40. the simulated glottal area (resulted from using the optimized parameters) is plotted in blue along with the experimental glottal area, shown in red dotted line. The simulated movement of the VFs during closure was depicted in the dotted blue line – referring to the negative area. As shown, the figure reveals a relatively similar behavior between the simulation and the experimental data where the optimized simulation model captures the main vibratory characteristics of the experimentally extracted glottal area.

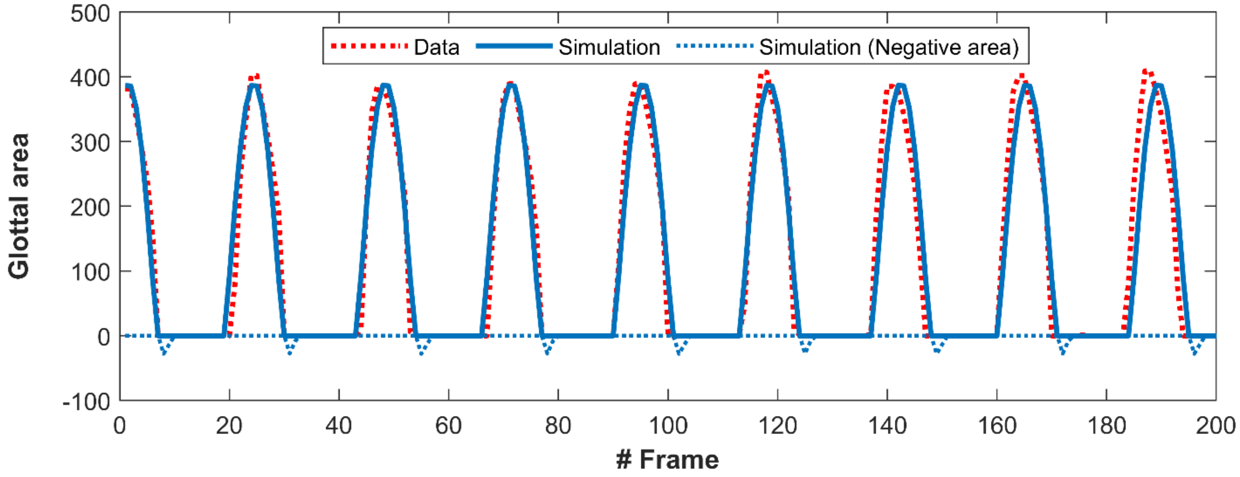


Figure 3.40. Optimization results between the simulated (in blue) and the experimental glottal area (in red) waveforms – overlaid on top of each other. The dotted blue line shows the movement of the simulated vocal folds during the closure time.

CHAPTER 4: DISCUSSION

4.1. Study I: Automated Detection of Vocal Fold Image Obstructions

The purpose of this study was intended to fulfill Aim 1 by addressing the following:

Q1.1: Can DNN accurately classify HSV frames in AdLD during connected speech regardless of the excessive laryngeal maneuvers?

H1.1: DNN can accurately classify HSV frames based on whether these frames display an obstructed view of the VFs.

Q1.2: Does the presence of AdLD affect the durations over which VFs are visually obstructed in HSV during running speech?

H1.2: The duration of the visual obstruction of the VFs will be longer in AdLD versus normal controls during connected speech.

A deep learning technique was successfully developed as a classifier to automatically detect the VF obstruction in HSV data, recorded during connected speech. The introduced automated framework was developed and implemented based on HSV recordings of vocally normal individuals and patients with AdLD. A robust training dataset was created through a sample of visually labeled HSV frames, displaying various obstructed and unobstructed VF views. The deep neural network was built using CNN and was successfully trained and validated on the dataset to classify HSV frames into two classes: frames with or without VFs obstruction. The overall visual evaluation of the performance of the trained network showed high capability in recognizing the VF obstruction in HSV video frames.

The results of implementing the trained CNN on a testing dataset, which was created from an HSV dataset on which the network was not trained, demonstrated high classification capability of the network in detecting different obstructions of the VFs with overall accuracy of 94.18%. This indicated how the presented automated approach was flexible and general toward classifying the VF obstruction of new HSV data from different participants with a high sensitivity and specificity of 97.24% and 91.11%. The developed network also returned high F1-score of 0.94 when applied to the testing dataset. This high F1-score revealed the high precision of the developed framework toward classifying the different obstructed views of the VFs in the HSV frames.

A robustness evaluation was done to assess the performance of the trained CNN-based classifier by a thorough comparison between the results of the automated method against the manual analysis of two HSV recordings. The two videos were selected from two different participants: one from a vocally normal person and another from a patient with AdLD. The comparison was conducted on the entire of the two HSV videos (consisting of 264,400 and 399,384 frames; over half a million HSV images in total). This massive number of images (663,784 frames) were manually classified by a rater to compare the developed CNN against visual observation of the rater. The results of the comparison revealed a promising performance of the automated classifier against the visual analysis. The percentage of the total number of frames in the two HSV videos that showed an obstructed view of the VFs was almost the same between the manual observation and the automated analysis – 14.56% versus 14.75% in the vocally normal individual's HSV video and 24.18% versus 24.42% in the patient's HSV video, respectively. As found in this study, the patient showed a higher number of frames with an obstructed view of the VFs, which can be explained by excessive laryngeal spasms in the AdLD patient.

The high robustness of the developed technique in classifying laryngeal HSV data against the enormous number of manually labeled frames is an apparent advantage over the previously introduced classifiers. This is because the previous deep learning models were tested against considerably limited sizes of images at around 720 [136, 137, 138], 1,176 [140], and 5,234 laryngoscopic images [139], which were extremely smaller than the number of images used for assessing the introduced automated technique. The comparison between the manual and automated classification on each individual frame was further extended over the two HSV recordings by generating confusion matrices. Different metrics were used, based on the resulted confusion matrices, to provide a detailed evaluation of the developed CNN performance in this comparison: sensitivity, specificity, precision, F1-Score, and accuracy for detecting the obstructed/unobstructed view of the VFs in the HSV frames. Overall, the proposed deep learning approach demonstrated high robustness when applied to the two HSV videos against the visual observation with overall accuracies above 92%. The automated technique showed a better performance with higher accuracy when identifying the VFs in the frames than when recognizing an obstructed view of the VFs. The reason for this was that the VFs can be obstructed in different ways and configurations in connected speech – imposing a more challenging view for the automated approach. Furthermore, the developed network showed higher overall accuracy in the vocally normal participant's recording (97.23%) than in the patient's recording (92.38%). This is because, in running speech, patients with AdLD demonstrate an increase in laryngeal maneuvers and complex VF obstructions than vocally normal persons, imposing more challenging conditions for the developed technique to maintain a high classification accuracy. These complex obstructions could be due to, for instance, epiglottis, left/right arytenoid cartilages, laryngeal constriction, false VFs, or any combination of these. In addition, the developed network had few challenges in detecting the frames with partial VF obstructions. This is because, in the manual labeled data used for training, it was challenging for the rater to exactly determine partial VF obstruction – such that if more than 50% of the VFs was obstructed, the frame would be classified as a frame with VF obstruction.

This study is the first work that developed and applied a fully automated deep learning approach in order to detect VF obstructions in HSV data during connected speech. To the best of our knowledge, there are no other studies in literature that used a state-of-the-art deep learning technique as a classifier for frame selection on HSV recordings; instead, several studies implemented deep learning schemes to laryngoscopic images [139, 140]. The HSV data used in the present study in running speech exhibit lower image quality along with excessive laryngeal movements and significant changes in glottal posture, which impose considerable challenges upon applying the deep learning approaches compared to high-quality laryngoscopic images. In spite of these challenges, the developed approach was highly successful as a classifier in automatically selecting HSV frames based on the presence and absence of the VFs. The introduced technique achieved overall classification accuracy of 94.18% on the testing dataset, which is even a comparative accuracy against the accuracies found in literature using the better-quality/less challenging laryngoscopic images at 86-96% [139, 140]. This, therefore, demonstrates that the present deep leaning-based approach not only proved its high robustness in classifying HSV data against a huge number of manually labeled frames, but also revealed a promising performance in HSV data with challenging image quality in connected speech. Accordingly, hypothesis H1.1 was accepted.

After validating the accuracy of the proposed classification network which addresses H1.1, it was implemented to investigate the differences between the AdLD patients and the vocally normal individuals in terms of the durations within the HSV recordings when the VF was visually obstructed. This investigation was performed to address H1.2. The comparative results demonstrated that there was a noticeable difference in the durations of the visual obstruction of the VFs in connected speech between AdLD (with an average obstruction of 26.1% across all the participants) versus normal controls (with average obstruction of 19.7% across the different individuals). The outcome of the comparison supported the acceptance of H1.2. The reason for this difference stems from the impaired laryngeal control and excessive movements of the laryngeal tissues in AdLD – leading to frequent obstructions of the VF view – in comparison with the vocally normal subjects. Overall, the results demonstrated the applicability and the potential of this measurement in studying the differences between AdLD and normal controls. Therefore, the durations of visual obstruction might be a good measurement that could be used in future for determining the severity of the AdLD; however, a larger sample size would be useful to emphasize the findings and investigate the clinical relevance of this measure.

4.2. Study II: Image Segmentation of Vocal Fold Edges

4.2.1. Image Segmentation Approach: ACM

The purpose of this study was intended to fulfill Aim 2 by addressing the following:

Q2: Can VF edges be accurately and robustly segmented in HSV data during running speech in the presence of image noise?

H2.1: The dark glottal area can be successfully silhouetted against the brighter surrounding VF tissue.

H2.2: ACM can accurately segment VF edges in HSV data with excessive image noise during VF vibrations.

The temporal segmentation technique, developed in a previous study [119], was successfully utilized to determine the vocalized segments of the "Rainbow Passage." Subsequently, the motion compensation precisely located the vibrating VFs across the frames during the extracted vocalizations. After applying the motion compensation and determining the location of the vibrating VFs, digital kymograms were successfully extracted at various intersections of the VFs. The vibrating VFs always appeared on an almost straight line in the extracted kymograms, which was necessary for a better performance of the spatial segmentation algorithm.

The automated snake initialization tool was successfully developed and accurately located a line that spanned through the glottis center in the extracted kymograms. The adjusted moment line was introduced because the results revealed how vulnerable the first moment of inertia line was toward the noise in the kymogram image. Based on the results, the proposed adjusted moment line demonstrated a better estimation than using only the first moment of inertia line for finding the center of the glottis. Obtaining an accurate snake initialization line was a necessary step toward a better performance of the ACM modeling algorithm and its convergence.

The ACM was successfully implemented for the kymograms of the vocalized segments of the "Rainbow Passage." The application of ACM allowed the analytic representation of the VF edges at different cross sections of VFs from the anterior to the posterior commissure. The performance of the algorithm exceeded the challenging quality of the HSV images. From 76 vocalizations of the "Rainbow Passage", the visual observation of the detected edges and the HSV kymograms showed that the algorithm's error was not more than one pixel for 67 vocalizations (88%) deeming it successful for precise detection of the glottis boundaries. Due to dim lighting in some of the frames in the kymograms of the other 9 vocalizations, the active contour modeling (ACM) was not able to find the glottal edges. The visual observation also could not determine the glottal edges due to the lighting issue for these kymograms.

This study showed the feasibility of automatic VF edge detection using the proposed ACM method in challenging data obtained using a color high-speed camera – leading to the acceptance of hypothesis H2.1 and H2.2. Color images are preferred over monochrome images by clinical specialists since color images allow them to evaluate the health of the tissues while observing and evaluating the vibrations of the VFs. Therefore, this study used a color high-speed camera to demonstrate that the proposed algorithm can be applied to color images. Moreover, the goal of this study was to develop an algorithm that works for the most challenging conditions given color images. Color images are challenging to analyze compared to the monochrome images due to the inherently higher dynamic range (image quality) of monochrome images and to the significantly more accurate representation of the gradients of the edges in the monochrome images. Despite the edge uncertainties on the color images, the paired active contour was not attracted to erroneous edges, and it maintained optimal rigidity. Since this work shows the robustness of the spatial segmentation method in the most challenging conditions due to color images, this method can be a promising image processing technique to detect VF edges in HSV data regardless its image quality.

After registering back the segmented edges in the kymograms to the HSV frames, based on the visual inspection of the results, the implemented active contour modeling successfully detected the edges of the vibrating VFs across the frames during each vocalized segment. This method not only has been able to address the sensitivity of prior image segmentation techniques to image noise and intensity inhomogeneity, but also could tackle more challenging video quality in HSV data in connected speech. Despite the promising performance of ACM, it was vulnerable to very dim lighting conditions in connected speech data, where the kymograms had inferior lighting conditions. This issue occurred due to the high sensitivity of the active contours toward their initialization, creating a challenge to accurately localize the contours near the glottal edges. Moreover, since ACM is an iterative method, it required a relatively long time for convergence because the analysis is done at all cross-sections of the VFs for each vocalization which could include thousands of frames. Therefore, this technique can be best used for HSV data collected using rigid videoendoscopy due to higher image quality. The following section discusses an advanced method (the hybrid approach) that can provide an enhanced performance and overcome the limitation of ACM on its dependency to the contour initialization and the high computational cost to be implemented even with inferior dim lighting image quality during connected speech.

4.2.2. Image Segmentation Approach: The Hybrid Method

This study was focused to completely address Aim 2 in combination with the previous one by fulfilling the following:

Q2: Can VF edges be accurately and robustly segmented in HSV data during running speech in the presence of image noise?

H2.3: A clustering technique can be combined with ACM to build a hybrid method improving the edge segmentation accuracy of ACM during vocalization and when VFs are not vibrating.

The temporal segmentation and motion compensation algorithms were successful in capturing the location of the vibrating VFs in a cropped motion window, which prepared the HSV frames for kymogram extraction. The HSV kymograms were generated at different cross sections of the VFs during each vocalization. The moment of inertia was used to successfully determine a horizontal line spanning through the center of the VFs in each kymogram, which was an important step before applying the hybrid spatial segmentation method to the kymograms.

The selection and extraction of the appropriate features were done in order to implement the unsupervised ML technique (i.e., k-means clustering). A different number and combination of features were fed into the ML algorithm to determine the salient subset of features for the development of the method. These features included the intensities of red and green channels and the image gradient. It was found that using these three features was the most appropriate combination of features in terms of obtaining an adequate clustering performance. Given the three considered features, the implemented clustering algorithm was able to precisely cluster the kymograms' pixels into two clusters (glottal area pixels and non-glottal area pixels). Subsequently, the edges of the clustered glottal area pixels were spatially segmented, returning the top and bottom initialization contour lines corresponding to the left and right VFs, respectively.

After obtaining the initial contours from the clustering technique, they were used as inputs to the ACM method to enhance its performance in segmenting the VF edges. The ACM method was successfully applied to the kymograms utilizing the initialized contours. The main weakness of the ACM method is the sensitivity to the contour initialization, which should be selected to be close to the glottal edges. In this study, using the clustering technique to initialize the active contours significantly improved the accuracy of the hybrid ACM in comparison with using the ACM alone. This hybrid method allowed for the accurate representation of the edges of the vibrating VFs in the kymograms at different intersections of the VFs. A comparison between the new machine-learning-based hybrid method against the ACM alone was conducted in order to show to what extent the new hybrid technique enhanced the performance of the VF edge representation in comparison with using only the ACM approach. The performance of the hybrid method was compared with that of the ACM by applying the two methods on two decent quality kymograms and two kymograms with dim lighting and degraded qualities. The results of the comparison revealed a significant improvement in edge detection by the hybrid method over using the ACM alone. This enhancement was more noticeable in the lower quality kymograms. This indicated how the proposed hybrid method was less vulnerable to the noise in the image compared to the ACM, which failed to detect the edges in the presence of significant noise in the kymograms. In addition, the computational cost of the hybrid method was half of the ACM technique.

After applying the hybrid method, the segmented edges in the kymograms, which were extracted at different VF cross sections, were registered back to the HSV spatial frames to detect the VF edges in each individual HSV frame. The performance of the proposed hybrid method was tested through visual inspection of the detected VF edges in the HSV kymograms of different vocalization segments of the “Rainbow Passage.” Out of 76 vocalizations, the visual inspection of the detected VF edges in the extracted kymograms demonstrated that the developed hybrid technique successfully captured the glottal edges for 74 vocalizations with an error less than ±1 pixel. This yielded a high accuracy of 97.4% in VF edge representation using the hybrid method for HSV data during connected speech. The only other study performing the same task that we can compare our work with was our previously developed ACM method [34], which detected the glottal edges accurately in 88% of the vocalizations in the same HSV sample. There are no other known studies of automated VF segmentation of HSV recordings during connected speech. The current study presented several of the vocalizations, where the ACM method failed. The higher accuracy and performance of the hybrid method, as were shown in this study, reveals its superiority over the ACM method; hence, hypothesis H2.3 was accepted. The extracted kymograms of the two vocalizations in which the hybrid method did not perform accurately had extremely dim lighting across most of the frames, which also made the visual detection of the glottal boundaries impossible, making it challenging to create an accurate reference manually.

The hybrid method in this study is the first ML-based approach developed for VF segmentation during connected speech. The recently developed deep learning approaches for VF segmentation were all employed for HSV analysis during sustained vocalization with higher image quality [142, 143, 144, 145]. The developed hybrid method is fully automated, while the deep learning techniques, previously developed, required manual labelling of a part of the dataset in order to train the deep neural networks. Moreover, the prior deep learning methods are all spatial segmentation techniques; however, the hybrid method in this study is a spatiotemporal method that would lead to a higher robustness in case of irregular VF closure. The hybrid method in this study relies on the accurate performance of the developed motion compensation method; however, this is not an issue with the HSV analysis during sustained vocalization due to the little change in the VF location across frames. Since there is no known gold-standard accurate method to fully capture the VF edges from HSV data during connected speech, visual inspection was performed to serve as reference for validating the performance of the developed technique. It should be noted that this study showed the feasibility of the hybrid method for VF edge representation (in HSV data) during connected speech in one participant with no history of voice disorder.

The proposed hybrid approach showed a promising performance for HSV data with the most challenging images, obtained by a color HSV system. This facilitates the future implementation of the proposed method on less challenging monochromatic images since a monochrome camera provides a higher sensitivity and dynamic range with better pixel representation. This will potentially lead to a higher accuracy and faster performance of the hybrid method for monochromatic HSV data. This study aimed to show the feasibility of this approach for color HSV images, which is preferred over monochromatic images by many voice specialists since color images allow them to better evaluate the health of the tissues. Although the promising performance of the hybrid method was shown during VF oscillation, the algorithm did not perform accurately before and after the onset and offset of VF vibration. This was due to the deviation of the motion window from the VF location before and after the oscillation. However, this did not contradict the purpose of this study, which was to track the edges of the VFs during vocalization. Therefore, the application of the hybrid technique would be efficient during the more sustained portions of phonation (vibratory portions) during connected speech as it provides accurate detection of the edges of the VFs during vocalized segments in running speech regardless the inferior image quality. In the next section, the development of an algorithm to automatically detect the edges of the VFs when adducted and not vibrating will be discussed which would be valuable in studying laryngeal maneuvers as well as phonation onsets and offsets during connected speech.

4.3. Study III: Deep-Learning-based Representation of Vocal Fold Dynamics

4.3.1. Deep Learning Approach: Segmenting Network on Color HSV Data

This study intended to partially address Aim 3 by fulfilling the following:

Q3: Can the GAW be automatically extracted given the inferior image quality in the fiberoptic HSV and the excessive laryngeal movements in AdLD during running speech?

H3.1: The hybrid method can be used as an automated labeling tool to train a robust DNN on detecting the glottal area in HSV during running speech.

The present technique in this study used the power of the developed hybrid method, discussed in the previous section, and deep learning to overcome the challenges of the previous two image segmentation methods (ACM and the hybrid technique) in terms of detecting the glottal area during all phonatory tasks (including nonstationary portions) and when VFs are not vibrating. Hence, the proposed deep learning approach can be used as a robust and cost-effective tool for segmenting the glottal edges—regardless of the image quality and the phonatory tasks during running speech.

This study showed the successful utilization of the previously developed hybrid segmentation technique as an automated labeling tool to form a training data set. In the hybrid method, k-means clustering technique was successfully applied to cluster the kymogram's pixels into two clusters (glottal area and nonglottal area). The edges of the glottal area cluster were roughly segmented as initialized contours for the ACM method, which was then implemented to accurately segment the edges of the vibrating VFs in the kymograms. The combination of k-means and ACM yielded a precise detection of the VF edges, which were registered back to the original HSV frames to segment the glottal area. The hybrid method showed an accurate performance but mainly during

VFs sustained oscillation as mentioned before. Hence, the hybrid method was applied to segment a set of frames during those instances of sustained vocalizations in the HSV data. This allowed for automatic labeling of a huge number of HSV frames. A subset of these segmented/labeled frames were sufficient to create a training data set to train a deep neural network as a more robust segmentation technique that can work during different phonatory events other than the sustained vibrations. Using the hybrid method as an automated labeling tool offered a huge advantage over the manual labeling, which is commonly used in the literature [142, 143, 144, 145]. That is, the proposed deep neural network was trained using only automatically segmented frames (utilizing the hybrid approach) without the need for any manual labeling. So, one advantage of this method is that larger training data sets can be formed using the developed automated labeling tool in a cost-effective and objective manner, which is favorable for training deep-learning techniques.

The deep neural network was built based on the U-Net architecture. Several networks with different configurations were successfully trained on the automatically labeled data set. Since the quality and performance of the automated labeling tool was evaluated in the previous study (the hybrid method), the automatically labeled data set was sufficient to successfully train the networks. In addition, to ensure the training process using the automatic labeling was appropriate, we have evaluated the automatically segmented frames via visual assessment before the training; furthermore, the trained networks were assessed against manually labeled frames (ground truth data). Among the trained networks, we found that the network, which was trained using a batch size of 10 and built with encoder–decoder depth of four had the best performance on the testing data set (the ground truth data) with the highest mean IoU (0.82). The other networks with different encoder–decoder depths and batch sizes showed poorer performance and lower IoUs.

The visual evaluation of the HSV data of the female subject showed that the best trained network (the proposed one) outperformed the automated labeling tool (the hybrid method)—demonstrating better accuracy in segmenting the glottal edges and area, and higher robustness toward image noise based on what we found in our visual assessment. This promising performance of the trained network indicated the acceptance of hypothesis H3.1. In addition, the developed network showed a considerably lower complexity because it did not depend on several image processing steps to achieve the segmentation task as in the hybrid approach. Overall, the visual inspection of the performance of the introduced network showed a successful segmentation when implemented on the video frames. The accurate representation of the glottal area using the developed network enabled the precise measurement of the variation of the glottal area over time (glottal area waveform). While the glottal area might be influenced due to relative motion of the endoscope and tissues during phonation in connected speech, it is still an important measure, which allows to evaluate the oscillation of VFs in the HSV data [173].

Although the network was trained on frames segmented during sustained VF vibration, it was generalizable and was able to correctly segment frames during more complex nonstationary events such as in onsets/offsets of phonation, voice breaks, irregular VF vibrations, and when VFs were not vibrating—overcoming our previous method's limitation. Also, we found that the performance of the proposed approach was relatively stable and did not vary between the different phonatory tasks. Moreover, since the proposed network was trained on HSV frames that were segmented using the developed automated labeling tool, it was important to also validate the network by comparing it against manually segmented HSV frames. Hence, a separate manually labeled data set (testing data set) was created, where the glottal area in a set of new HSV frames were manually segmented, to test and quantify the performance of the proposed network. Different metrics were utilized to evaluate the network's performance against the manually segmented frames: a contour-based metric (BF score) to evaluate the detected boundary of the segmented glottal area (glottal edges) and an area-based metric (IoU) to assess the segmented glottal area itself. The introduced network showed a high mean BF score of 0.96 (LD = 0.12) indicating high accuracy of the network in localizing the edges of the glottal area (i.e., VF edges). Furthermore, the developed network achieved a mean IoU of 0.82 (LD = 0.26) and a mean DC of 0.88 (LD = 0.25), signifying high precision in detecting the glottal area.

This study introduced the first deep learning-based scheme for automatically segmenting glottal area in connected speech. So, there are no other studies that utilized the state-of-the-art deep neural network for glottal area segmentation in running speech to compare with. The recently introduced/utilized deep learning models applied deep neural networks to segment glottal area in grayscale [145, 144] and RGB [142] HSV data during sustained phonation using rigid endoscopes, but not during running speech using flexible endoscopy as in this study. HSV data in running speech, however, exhibit even lower image quality and excessive laryngeal maneuvers leading to considerable changes in the spatial location of the VFs. These constraints impose more challenges for the deep neural networks to successfully segment the glottal area in HSV in connected speech. Despite these challenges, the introduced approach showed a mean IoU of 0.82 and DC of 0.88, which are even above the baseline scores mentioned in literature that utilized a less challenging and higher quality HSV data with IoU of 0.799 [143, 145] and DC of 0.85 [142]. This comparison though was on a different data set but showing how the proposed method achieved a promising performance on a more challenging data demonstrates the high competitiveness of our approach against the other related methods. Furthermore, the previous deep learning approaches for HSV analysis [142, 143, 144, 145] were entirely utilized for only spatial segmentation. Among these studies, Fehling et al. [142] was the only research group that designed deep neural networks that could keep the HSVs temporal information, and they evaluated the segmentation conformity over the course of time. However, the sequences they utilized were quite short. In contrast, the introduced deep learning model is a spatiotemporal technique, where the HSV data are first preprocessed using a temporal segmentation algorithm to extract the vocalized segments on which the proposed deep neural network was applied on long HSV sequences. This spatiotemporal feature enhances the robustness of the proposed model toward, for example, irregular VF closure.

The present work was conducted to demonstrate the high capability and robustness of a new deep learning-based technique for automatically segmenting connected speech in challenging images using a color HSV data from one subject. It should be noted that the current work applied the developed method on color HSV data, which have smaller dynamic ranges in comparison with monochrome images. This will guarantee a higher accuracy of this method when applied to monochrome data with a higher dynamic range. In the next section, this approach will be applied to a larger sample size from individuals with and without voice disorders and on HSV data recorded using a monochrome camera with less challenging image quality.

4.3.2. Deep Learning Approach: Neural Network on Monochrome HSV Data

This study was conducted to pursue Aim 3 through addressing the following:

Q3: Can the GAW be automatically extracted given the inferior image quality in the fiberoptic HSV and the excessive laryngeal movements in AdLD during running speech?

H3.1: The hybrid method can be used as an automated labeling tool to train a robust DNN on detecting the glottal area in HSV during running speech.

H3.2: This trained DNN will be successfully implemented for the automated extraction of the GAW in AdLD and normal controls even with its challenging image conditions.

H3.3: The glottal midline along with the left and right VF edges can be successfully captured based on the segmented glottal area.

In the current study, the aim was to provide a DNN model using a larger sample size, compared to the previous study, to achieve a segmentation task of the glottis. To do so, the DNN which was developed and discussed in the previous section is retrained using a sample of both vocally normal participants and AdLD patients to validate its robustness/efficacy and provide a reliable automated generalizable tool for HSV analysis in running speech. On top of that, the literature on analyzing AdLD via HSV has been almost non-existent during connected speech. Hence, the main objective of this work is to develop and validate an accurate methodology to identify VF edges and glottal area in AdLD and vocally normal participants. Apart from the poor image quality of HSV in connected speech (found in the present video data), AdLD disorder imposes excessive laryngeal tissue movements, which makes analyzing HSV recording of AdLD patients more difficult. So, successful implementation of an automated DNN method to segment glottal edges in a challenging HSV dataset as for AdLD would allow for analysis and measurements of VF dynamics in AdLD during running speech. This can provide information on the characteristics of the impaired voice production in AdLD, which can potentially reduce AdLD misdiagnosis.

The present DNN approach was built based on the efficient architecture of the developed deep learning network discussed in the previous section. The built DNN was then successfully trained using a robust training dataset, which was created from HSV recordings in running speech of vocally normal subjects and AdLD patients. A sample of HSV frames were randomly selected from the recordings – including VF gestures in several phonatory events and different obstructed views of VFs – in which the glottal area was manually segmented by two raters after coming into a consensus for the detection of the area. This procedure guaranteed a fair representation of the various running speech events in the training/validation dataset and pixel-accurate segmentation of the manually annotated frames.

Different architecture modifications and training strategies were considered to obtain a generalized high segmentation performance of our previously designed U-Net DNN. The quantitative evaluation of these different trained DNNs was done on a test set, including manually segmented frames from an independent HSV recording of an AdLD patient on which the DNNs were not trained. The test set was important to allow for an unbiased assessment of the developed DNN and a realistic estimate of the model performance when testing on a set of new images, different from the training frames. Among the various trained DNNs, we found that the DNN with an architecture of four encoder-decoder stages and retrained with a batch size of 16 showed the best performance. That is, this optimum network showed a parallel improvement in all the four different assessment metrics (the mean DC, IoU, F1, and accuracy scores) on the test set. In contrast, the other trained networks (built with different architectures and trained using different batch sizes) demonstrated poorer performance and lower assessment scores on testing frames. The implementation of the best performing DNN on the testing dataset showed high quality of glottal area segmentation with a mean IoU score of 0.81, DC of 0.86, and accuracy of 0.89 when compared to the manual annotation. In addition to those area-based metrics (like IoU and DC), the developed DNN was also evaluated with respect to the precision in predicting the glottal boundary (the VF edges) using the F1 score. This dual evaluation was particularly important in two cases: (1) when the glottal edges were accurately segmented but some pixels inside the glottal area were missed or incorrectly predicted by the network and (2) when the glottal area was precisely detected whereas the pixels located on VF edges were incorrectly estimated by the network. The tested DNN demonstrated a high mean F1 score of 0.93, signifying high accuracy of the DNN in detecting the edges of VFs and a good match between the estimated glottal edges and the manually represented edges by the raters. Despite that, some discrepancies between the manual and the automated segmentations were mainly found near glottal edges. One reason could be that due to the poor image quality in the whole dataset including the testing frames, the edges were blurry, which made their manual segmentation challenging and sometimes inaccurate. So, since the model did not face that challenge, there is a high possibility that the model even outperformed the manual segmentation by providing accurate VF edges. This observation was clear when the VFs are fully open and right before and after the closure phase. Another discrepancy, though slight, was detected during the full abduction posture, particularly when there was no vibration including during the inhalation instances. This is due to that when the glottal area became large, VF edges became blurrier, and the glottis became brighter yielding a more challenging condition for the model to perform an accurate segmentation.

In addition to the quantitative evaluation, the overall visual assessment of the best performing DNN demonstrated the high quality in detecting glottal area/edges in various phonatory tasks that frequently occurred during the running speech. As such, this accurate performance was shown in the imaging data that included sustained and irregular VF vibrations, offsets/onsets of phonations, voice breaks, and when VFs were not moving. The trained DNN was not only able to detect the presence of VFs in the frames when they were clearly displayed in the image, but also was able to recognize the absence of VFs in the frames when they were visually obstructed by, for example, the epiglottis, arytenoid cartilages, and other laryngeal constrictions. This capability of our developed DNN was crucial in view of the fact that excessive laryngeal activities occur often in running speech. Furthermore, the results of developing an automated method to detect the glottal midline as well as the left and right VF edges based on the detected glottal area showed a promising performance. The introduced method was able to capture the edges of the VFs even in complex glottal area shapes and various sizes (including wide opening and during vibrations). The high performance was in line with what was found regarding the high quantitative score of detecting the glottal boundary which allowed the feasibility of the introduced glottal midline detection tool. Also, even with poor HSV quality and excessive laryngeal AdLD spasms, the tool was able to detect the glottal midline as a fitted second-order curve and, hence, capture the edges of the VF. Therefore, based on both the quantitative and visual evaluation of the developed DNN, hypothesis H3.2 and H3.3 were accepted.

It was found that the model was slightly more accurate in normal speakers than in AdLD patients. That is because the AdLD disorder frequently showed uncontrolled laryngeal tissue movements that lead to a partial or full coverage of the VFs during phonation, which might have been more challenging for the model to identify. The ability of the developed technique to recognize these different postures and obstructions of VFs and avoid incorrect glottis segmentations makes it a robust and efficient method to analyze VF dynamics in connected speech. Furthermore, the precise segmentation of the glottal edges/area using the introduced method allowed for the accurate measurement of the glottal area waveform in running speech. In comparison with the manual segmentation of the glottal area in a sequence of HSV frames, the automated method provided even smoother glottal area waveform such that there were no considerable changes in the area across frames – indicating the accurate detection of the glottal area.

The promising glottis segmentation quality demonstrated by the proposed approach in running speech addresses an existing literature gap in terms of the automated analysis of VF dynamics in HSV data. That is, previous image segmentation methods, including deep learning models were developed and evaluated using only rigid HSV in sustained phonation, [142, 143, 144, 145] not flexible HSV in connected speech as in the present work. Bridging this gap is essential to analyze VF vibratory characteristics and function in voice disorders, which are mostly present during the running speech. In addition, HSV data recorded in connected speech impose poorer image quality and higher variability of the spatial location of VFs due to the excessive movements of laryngeal tissues than in sustained phonation. This creates more difficulties for the DNNs to achieve successful glottis segmentation in running speech HSV data. Despite that, the proposed method showed a mean IoU and DC scores, even larger than the baseline accuracies found in the previous deep-learning methods that were tested on high-quality HSV data with IoU of 0.799 [143, 145] and DC of 0.85 [142]. This comparison shows that our proposed technique, though tested on a more challenging dataset, has a considerable competitiveness against the related deep learning approaches in the literature.

This study fills another research gap, which is the limited number of studies that analyzed AdLD disorder using HSV. To the best of our knowledge, this work is one of the earlies attempts in literature to provide accurate methodologies for quantifying VF vibrations during running speech, which enable us to investigate the VF dynamics in AdLD further using HSV in connected speech. Additionally, unlike the other deep-learning methods in the literature, the automated approach developed in the present study is the first deep learning-based approach for segmenting glottal area in HSV data obtained from AdLD subjects. The promising performance of the introduced method in detecting the glottal area change in running speech on challenging HSV data of AdLD subjects with excessive laryngeal movements can facilitate the development of HSV-based measures. Such measures allow for quantifying the vibratory behavior of VFs and the prephonatory adjustments such as the measurement of glottal attack and offset times in vocally normal speakers versus AdLD patients as an effective approach to evaluate the severity of this disorder, which will be discussed in the upcoming section.

4.4. Study IV: Automated Measurements of Glottal Attack and Offset Time

The purpose of this study was intended to fulfill Aim 4 by addressing the following:

Q4: Are the glottal attack and offset times different between AdLD and normal controls?

H4.1: An automated algorithm can be developed to measure GAT and GOT with comparable accuracy to visual measurements.

H4.2: GAT and GOT will be significantly higher in AdLD versus normal controls.

H4.3: GAT and GOT will show more variability in AdLD subjects.

The goal of this study was to develop an automated algorithm for measurements of GAT and GOT from HSV in connected speech as objective measures that could potentially facilitate the diagnosis of AdLD in future. In order to achieve that goal, the automated segmentation tool that was developed and discussed earlier in the previous section was implemented on the monochrome recordings. The segmentation technique showed successful detection of the glottal area during the various onset and offset of phonation in running speech. The DNN tool demonstrated high-performance capabilities in detecting the varied sizes, geometries, locations, and configurations of the glottal area and VF that are commonly existed during the different phonation onsets and offset in running speech. Being able to capture this variability and transitional states extending from various degrees of VF opening and small-amplitude oscillations to steady-state vibrations, even with the presence of inferior variable image quality, demonstrated the high reliability and consistent accuracy of the developed tool in glottal area segmentation during the onset and offset of phonations.

The successful segmentation outcome in capturing and quantifying the dynamic change in the glottal area facilitated the automated measurements of the GAT and GOT. Based on the segmented glottis, the contact between VF was successfully determined. This allowed the precise computation of the energy contours associated with both the dynamic vibration of VF (represented in GAW) and the VF contact (represented in GCW). By computing these two contours, the delay in time was successfully calculated using the cross-correlation technique between the rise in the energy contours, in case of the onset phonation (GAT), and the drop in the energy contours, in case of the offset of phonation (GOT). The automated algorithm showed efficient measurements of the GAT and GOT for the vocally normal group as well as the AdLD group.

In order to validate the developed automated approach, visual analyses were carried out by three raters to obtain manual measurements of the GAT and GOT through visually determining the timestamps between the first oscillation and first contact (referring to the GAT) and between the last contact and last oscillation (referring to the GOT). The manual and the automated measurements were obtained from each recording. The comparative analysis showed a close agreement between the automated and visual measurements of GAT and GOT in most of the recordings with minimal differences. As a measure of the developed approach accuracy, the average difference between the automated and manual measurements – computed based on the mean of each recording – exhibited a small value of 1.6 ms for GAT and 2.7 ms for GOT. This minor difference between the automated and the visual analysis was even lower than the error found among the three raters (up to around 4.5 ms,) which was considered as an acceptable deviation in the measurements of the GAT and GOT [162]. Additionally, the statistical analysis performed between the automated and the manual measurements within different vocalizations demonstrated a strong correlation between the two measurements in computing both GAT (r =0.93) and GOT (r= 0.91), indicating the high similarity degree between the two measures. These findings were also supported by the conducted t-test where no significant difference found between the automated versus the visual analysis in GAT, also GOT with p-values of 0.86 and 0.77, respectively. These comparative results reflect the reliability and accuracy of the automated analysis technique compared to the visual analysis in estimating GAT and GOT – leading to the acceptance of H4.1

Furthermore, results demonstrated that the automated algorithm was marginally more precise in determining the GAT compared to GOT across the different subjects. This small deviation in accuracy was primarily derived from the longer durations of the GOT that exhibited minimal amplitudes of VF oscillations toward the end of the offset which was difficult to define for the raters – causing small discrepancies with the automated method. Similarly, the automated measures obtained for the vocally normal group showed a marginally elevated precision for both GAT and GOT in comparison with the AdLD patients. This was mainly due to the irregularity found in the dynamic vibration of the VF in AdLD patients as well as the excessive phonatory breaks – making the analysis more challenging for both the automated and the visual analysis. Another likely cause of the difference between the automated and the manual analysis, though minimal, arose from the inferior image quality in the recordings and the blurriness found just prior to the adduction of the VFs, creating difficulty in determining the first and last contact frames.

After validating the introduced automated approach, it was utilized to compute the GAT and GOT values across all the recordings in order to draw conclusions in terms of the differences between the normal controls and the AdLD. The results revealed that, overall, the GAT measures were longer in the AdLD patients in comparison with the vocally normal participants. The statistical analysis demonstrated a significant difference (p-value < 0.001) between the average GAT of the AdLD (18.95 ms) versus normal controls (14.65 ms). This significant difference of GAT supported part of H4.2 that indicates significantly higher GAT in AdLD versus normal controls. This finding was supported by a previous study that demonstrated a delay between the onset of phonation and the activation of the laryngeal muscles [174]. Also, the results found in the present study were in agreement with the findings found in literature that showed longer GAT in

AdLD than normal controls [175, 153]. In contrast, although the present automated analysis demonstrated a slight increase in the mean GOT of the AdLD group (28.9 ms) versus the vocally normal group (27.3 ms), the difference was not statistically significant (p-value = 0.2). This insignificant difference of GOT rejected part of H4.2 that indicates significantly higher GOT in AdLD versus normal controls. Furthermore, the results demonstrated that there was a larger variability in the measurements of the GAT and GOT, within the AdLD group, with a particularly greater variability observed in GOT. In opposition, within the normal controls, this variability was less, especially the GAT which showed a consistent measure with a minimal range of variability across the vocally normal individuals. This finding showed the acceptance of H4.3. The explanation for this primarily lay in the irregular/inconsistent behavior of the VF vibrations in AdLD along with the impaired neurological dysfunction impacting laryngeal muscle control [66]. Hence, given the statistical significance between AdLD versus normal controls besides the more consistency and less variability found within the normal controls, GAT can be a valuable clinical measure compared to GOT. A larger sample size can substantiate the findings on the impact of AdLD on GAT and GOT in order to indicate the clinical significance of the introduced measures.

4.5. Study V: Lumped Modeling and Optimization of Vocal Fold Vibration

The purpose of this work was intended to fulfill Aim 5 by addressing the following:

Q5: Can a simplified one-mass model be optimized to accurately match the vibratory behavior of VF extracted from HSV?

H5.1: A simplified one-mass model can successfully simulate both the vibratory and closure phases of VF motion.

H5.2: The particle swarm optimization technique will enable accurate optimization of the model to predict the experimental glottal area waveform.

H5.3: The optimized model parameters, obtained through inverse analysis of HSV data, can estimate the VF mass, elasticity, and viscosity indices.

The goal of this study was to introduce a biomechanical model that can mimic the mechanical vibrations of the VFs. A lumped-element model was built using a one-mass model. The model was designed such that each VF tissue was described by a rigid mass coupled by springs and dampers. The model was combined with experimental data where the change in the glottal area was extracted from a vocalized segment in the monochrome HSV data. The aim was to build this model and optimize its parameters so that the model can generate an oscillation behavior similar to the extracted glottal area waveform from the HSV. For optimization, particle swarm technique was utilized to achieve the optimization task which was commonly utilized in literature [93, 103].

The one-mass lumped element model was successfully implemented in order to provide a relatively close behavior to the VF vibrations. To do so, the model was developed such that it can incorporate the dynamic behavior in VFs during both the vibration and the closure phases. Hence, several parameters were successfully considered in the model simulation. The model included parameters related to the VF mass and the spring stiffness (representing the elastic behavior of the VF). In addition to that, during VF closure, an extra damping coefficient was considered as a parameter in order to simulate the characteristics of the damped motion during VF adduction. It is expected to see a relative increase in the subglottal pressure during VF closure – referring to a build-up pressure – which helps in pushing the VF apart and complete the vibratory cycle. In order to incorporate this build-up pressure, a variable subglottal pressure was considered as a function of time where, during the closure time, an increase in the value of the subglottal pressure occurred to reach a maximum pressure. This maximum pressure during closure as well as the closure duration time were also incorporated into the model as an attempt to obtain a realistic simulation and capture the dynamic oscillatory behavior of the VFs.

After incorporating the proposed parameters into the one-mass model, the modified model was successfully implemented and simulated. Although the model had a limited degree of freedom (using only one mass to represent VF), the model was able to sufficiently mimic the oscillatory behavior of the VF during both vibration and the adduction phase – leading to the acceptance of H5.1. In addition to building the model, several simulations were carried out with different model parameters in order to make sure that the model offered acceptable performance suitable to be optimized with experimental data. By using the additional damping parameter during the closure phase of the VF, the model was able to provide a behavior close to the VFs when they are in contact. That is, during the adduction of the VFs, the simulated VFs showed an overdamped movement (due to the increased value of the viscous damping parameter, c') in order to emulate the impact of collision between the VFs. Moreover, the subglottal pressure was effectively incorporated into the model simulation as a variable parameter. The simulation results showed the accurate representation of amplifying the subglottal pressure, reflecting the buildup behavior during the closure phase of the VF. Also, the model was able to estimate the variation in the glottal airflow as well as the change in the glottal area waveform resulted from the oscillatory behavior of the simulated VFs.

After the successful simulation of the model, an optimization technique was utilized. Particle swarm method was successfully used and employed to optimize the theoretical glottal area waveform resulted from the model simulation with the experimental glottal area waveform extracted using HSV from a vocalized segment. Optimizing the model with a vocalized segment during VF vibration was considered because the VF properties such as the mass and elasticity were not expected to considerably vary, allowing for better optimization. Most of the previous models for optimizing lumped models with HSV-based VF oscillation conducted using samples during VF vibration, as considered in the present work as well [93, 103, 104, 105, 84, 106, 92].

Six model parameters were successfully optimized including a scaling factor, the mass, spring stiffness, damping coefficient during closure, the closure time (t_c), and the maximum subglottal pressure (buildup pressure). The scaling factor was used because the units of the simulated and the experimental glottal area waveforms did not match so using the scaling factor was able to minimize the difference between the two waveforms. The outcome of the optimization demonstrated a relatively good match between the simulated and the experimental behavior during both the vibratory portions when the VFs were open and the closure portions when the VFs are in contact. The successful optimization revealed the efficacy of the modeling parameters to produce close vibratory behavior compared to the experimental one; hence, H5.2 was accepted. Also, the optimized parameters showed an agreement with the typical ranges found in literature. The value of the optimized VF mass (0.05 g) lay within the expected range reported in previous studies, which was between 0.016 – 0.10 g, and the optimized elasticity (11,787 g/s^2) revealed close agreement as well with several studies that reported a wide range of VF elasticity values (6,000 – 32,000 g/s^2) [176, 177, 178, 179]. Moreover, the optimized subglottal pressure in the present study demonstrated a values between 8,000 – 8,571 dyn/cm^2, which fell within the typical pressure values found in literature that showed a great consensus regarding the value of 8,000 dyn/cm^2 with a range of 4,000 – 14,000 dyn/cm^2 [169, 165, 179]. The optimized VF closure time at 0.003 s in the introduced model exactly approximated the value that was observed during the HSV recording – revealing the success of the optimization process in capturing the experimental closure phase time. By obtaining the optimized parameters, it was able to refer and estimate biomechanical measurements of the VFs including the elasticity index and the viscosity index associated with the vibrating VFs. Therefore, hypothesis H5.3 was accepted.

The successful development and application of such a simple model like the one-mass model in this study using a vocalized segment during running speech is considered one of the earliest attempts in literature of achieving such optimization using an improved one-mass model and during a vocalized segment in running speech. Overall, the successful implementation of the introduced model in the present work can open up a new line for future research and act as a tool that can be further developed and used for estimating important non-invasive biomechanical features of the VFs during connected speech which cannot be extracted using the conventional assessment/experimental clinical tools. Also, the successful validation of the related hypotheses of this study demonstrated the potential of the simplified model introduced here in quantifying some biomechanical properties of VFs using HSV data. By leveraging the advantages of simple models, better understanding of the temporal dynamics in running speech and the impaired vocal function in voice disorders such as AdLD can be explored.

4.6. Limitations and Directions for Future Studies

This dissertation while providing pioneering HSV analysis methods and valuable insights into studying laryngeal mechanisms and vocal function in AdLD during running speech, is not without its limitations. The aim of this section is to examine the constraints associated with the present work and potential directions for future investigation.

In study I, a new approach was developed to detect laryngeal activity in AdLD as a potential measure of severity by classifying the durations of VF image obstructions in HSV. Although the approach demonstrated promising accuracy, it showed difficulty in detecting the partially obstructed VF images in HSV. This is mainly because, in the manual labeled images used for training the developed DNN, it was challenging for the rater to exactly determine partial vocal fold obstruction. This shortage can be avoided by modifying the criteria used for partial obstructions or by adding one more category of classification – instead of obstructed and unobstructed VF image, three classes of unobstructed, partially obstructed, and fully obstructed can be considered. Also, the developed measurement in this study was only based on whether the VF images were obstructed or not but did not identify the type of these obstructions. This limitation can be a potential direction for future research as well where different types of VF image obstructions can be detected such as obstructions due to epiglottis, arytenoid, laryngeal compressions, etc. Being able to classify these different types of tissue obstructions might lead to useful insights into the specific spasmodic behavior of AdLD or other neurological voice disorders. In addition, with a larger sample size, the clinical relevance of this measure can be further emphasized, and diagnostically relevant information could be extracted. For a future direction, the findings of the present study could assist in developing appropriate passages and speech texts with minimal obstructed view of the VFs for a more effective voice assessment in connected speech.

Study II developed two image segmentation techniques for capturing VF edges in HSV during running speech. The first technique ACM showed difficulties in detecting VF edges in excessive image noise and inferior image quality in the HSV data. The second method was the hybrid approach. The hybrid technique addressed the ACM challenges and was able to accurately detect VF edges in poor HSV image quality. However, the algorithm encountered some difficulties to find the VF edges before and after the steady-state vibrations. This might be due to the deviation of the motion window from the location of the VFs prior and after VF oscillations – yielding inaccurate extraction of the kymogram images during the onset and offset of phonation. Due to this deviation, the vibrating VFs were not exactly in the center of the motion window. However, this was not an issue in the present study since the main goal was to capture the VF edges during vocalization and the vibratory phase of the VF. Therefore, for future directions, using the introduced ACM technique can be much more effectively applied to future applications using video data recorded via rigid HSV with higher image quality and less image noise. For more challenging conditions in running speech with inferior image quality, the developed hybrid tool can provide accurate performance in detecting VF edges, particularly, when applied to the more stationary phonation tasks in running speech to analyze the vibratory behavior of VFs using kymograms.

Study III provided a quantitative representation of VF dynamics in AdLD during running speech using HSV. Although the method demonstrated accurate detection of the glottal area and its edges given the poor image quality and excessive laryngeal maneuvers in AdLD, there is a potential for further improvement. The main challenge found, though infrequent, was to mistakenly detect dark spots located toward the corners or the edges of the HSV frames as glottal area. Prospective research path can address this limitation by cropping the HSV frames before applying the DNN for glottal area segmentation functioning as a preprocessing step. This cannot only improve the segmentation accuracy by avoiding these misclassified pixels in the images, but also reduce the computational cost. This is because the input images to the DNN would be cropped (becoming smaller in size) which would require less time to be processed by the developed DNN. Moreover, using the developed tool opens new avenues of future research where it enables the extraction of objective HSV-based measures from both vocally normal individuals and patients with voice disorders during VF vibrations. Having access to such automated measures would benefit future clinicians to analyze the huge datasets of HSV in connected speech that could potentially facilitate voice disorder diagnosis.

Study IV developed and analyzed automated measurements of GAT and GOT using HSV in running speech. In spite of the substantial findings of this study demonstrating that GAT can be a potential candidate to assess AdLD in running speech, larger sample size is needed in order to emphasize these findings and differentiate between vocally normal adults and AdLD. Another related limitation of this study highlights the need to compare AdLD with other neurological voice disorders such as MTD and ET. This comparison is beneficial to reveal the main differences among the different voice disorders using the develop automated measures – leading to a better understanding of the difficulty associated with the misdiagnosis of these patients. Moreover, a potential research point exists in combining the different potential measures found in the present work in order to offer an assessment procedure of AdLD. That is, the measurement of the frequency and intensity of the laryngeal tissue activities can be combined with the GAT and GOT measurements as a more effective way for assessing the severity of AdLD. Moreover, the present study only investigated samples of running speech. This may lead to an important future research avenue that can study the impact of the different speech tasks and phonetic contexts on the severity of different voice disorders' symptoms compared with normal controls. For example, these future studies may benefit from comparing sustained phonation, the different CAPE-V sentences, and the various phonetic sounds included in the Rainbow Passage. These future investigations may uncover groundbreaking findings regarding the correlations between the phonetic context and the severity of various neurological voice disorders.

The last study, despite demonstrating the successful simulation/optimization of a simple lumped model to capture the vibratory characteristics of VF, has limitations. The limitation mainly arises from the simplifications/assumptions that were considered for model implementation. Describing VF as a one mass with a single degree of freedom compared to multi-mass models imposed few constrains. Including multiple masses allow for more precise simulation of the intricate VF vibrations, particularly in the closure phases. However, the present model was built to relatively overcome the prior limitation and simulate the closure phase by including extra damping coefficient to emulate the impact of VF collision. So, future investigations can be done in this direction to compare the simulation of the present model with multi-mass models. Moreover, although the subglottal pressure in this model was parameterized as a step function to simulate the built-up pressure impact during closure, it was an idealized representation of the actual variations in the subglottal pressure during VF oscillations. That is, the sudden increase in the pressure at the beginning of the closure phase should be simulated as a smooth rise; similarly, the sharp drop of the pressure right after the closure phase can also be replaced by a smooth drop. This can achieve a closer emulation of reality during VF vibration, enhancing the fidelity of the simulation. Another limitation of the present model was that the VF was simulated with zero damping coefficient during the opening phases. This limitation also represents another potential point to be further studied in future by incorporating a damping coefficient of the VF during oscillation in addition to the extra damping during the closure phases. Also, the limitation of relying solely on the Bernoulli equation to model the airflow underscores the need of incorporating advanced aerodynamic models to precisely simulate the turbulent effects occurring during VF vibrations. Lastly, the fact that the model was optimized during VF vibrations highlights the potential in advancing the present model to also optimize the onset and offset of phonations in future investigations.

CHAPTER 5: CONCLUSION

In order to gain a more comprehensive understanding of the impaired vocal function in AdLD and the persistent issue of misdiagnosis, this dissertation introduced one of the earliest endeavors in literature to utilize advanced HSV technology in studying this disorder during connected speech. Despite the challenges posed with HSV analysis in running speech and the lack of effective automated methods, various automated approaches and techniques were successfully developed in the present work to bridge this huge gap in literature and facilitate investigating the dynamic characteristics of VF in AdLD. Toward that endeavor, this dissertation presented five different studies aiming to tackle the previous challenges. The conclusion of each study is summarized as follows:

Study I introduced an automated tool to classify HSV frames by detecting the instances during which the image of the VFs is optically obstructed. This tool enabled exploring how the presence of AdLD impact the durations over which VFs were visually obstructed in HSV during running speech – indicating the degree of laryngeal activities in AdLD. The developed technique can accurately perform an automated frame selection task in HSV recordings, even with a challenging image quality, to recognize and classify HSV frames with clear view of VFs. Also, it can provide precise analysis to investigate laryngeal maneuvers in AdLD patients within running speech. By using this tool, the analysis showed that there are remarkable differences in the durations of the visual obstruction of VFs between AdLD and the vocally normal individuals during connected speech. These durations of visual obstruction might be a good measurement that could be used for determining the severity of AdLD. Overall, utilizing this tool can be useful to provide insights into the impaired laryngeal activities in AdLD or other voice disorders in terms of the spasmodic behavior of the laryngeal tissue movements.

Study II proposed a novel image segmentation tool for VF edge detection in HSV-based kymograms during the vibratory segments in running speech that can accurately perform automated analysis even with the poor image quality and noise existed in the transnasal HSV. Developing this automated tool addressed the lack of effective spatial segmentation methods amenable to HSV analysis in connected speech. The temporal segmentation and the motion compensation approaches used in this study successfully extracted the timestamps of the vocalized segments and localized the vibrating VFs of the HSV recording of the “Rainbow Passage”. The study showed the successful development of two image segmentation approaches for VF edges using ACM and an unsupervised ML technique (k-means clustering). The combination between these two approaches resulted in a powerful hybrid method for spatial segmentation, which revealed a promising performance in precisely capturing the VF edges in kymogram images across frames. This hybrid method helped significantly improve the performance of the ACM method in terms of addressing the dependency of ACM to contour initialization, enhancing the edge representation accuracy, mitigating the sensitivity towards image noise, and providing a lower computational cost.

Study III presented quantitative representation of VF dynamics in AdLD during connected with the groundbreaking application of HSV for disorder examination. This study extracted the GAW and glottal edges from HSV data in connected speech. Developing these analyses offered a considerable contribution to the existing literature in terms of offering a distinctive quantitative portrayal of the impaired behavior of VF vibrations. The study showed the successful implementation of an automated labeling tool for spatial segmentation of the glottal area in HSV frames, based on our previously developed hybrid method in study II, which formed a large training data set to effectively train the DNN. The developed DNN even outperformed the labeling tool by improving the segmentation accuracy in the glottal areas, enhancing the robustness toward poor image quality/noise, lowering the computational cost, and increasing the flexibility to accurately performing segmentation during complex events as in phonation onsets/offsets and voice breaks. The study also showed the successful implementation of the developed DNN and the accurate representation of VF dynamics in running speech HSV recordings of vocally normal individuals and patients with AdLD. The developed approach accurately segmented the glottal area, overcoming the inferior image quality and the excessive laryngeal spasms of AdLD recordings. Moreover, based on the detected glottal area, another proposed edge detection algorithm was successfully developed to capture the left and right VF edges along with the glottal midline. The high segmentation accuracy of the developed tools was demonstrated in onsets/offsets of vibration, voice breaks, prephonatory adjustments, and regular/irregular VF vibrations. These tools can aid clinicians in addressing the diagnostic challenges and early detection of this disorder.

Study IV investigated the pathological vocal function during phonation onset and offset of AdLD using HSV in connected speech – bridging a huge gap in literature. In this work, an automated analysis was successfully conducted to measure the GAT and GOT from vocally normal participants and AdLD to investigate the differences between the two groups. These analyses were carried out through developing an automated method for measurements. The automated measurements revealed minor, non-significant differences in comparison with the visual analysis – showing strong correlations between the two methods. The automated measurements demonstrated two main findings. That is, AdLD patients showed significantly longer GATs compared to the vocally normal group, and more variability was observed in both GATs and GOTs of the AdLD patients due to the considerable irregularity in their impaired VF vibrations. Since this study is considered one of the earliest attempts in literature to investigate these measurements in running speech for AdLD, these findings can serve as a foundational baseline for future research utilizing larger sample size and different voice disorders. The developed automated approach for glottal attack and offset time measurement can be valuable in the clinical practice. Obtaining such measures enables the exploration of clinically relevant information to address diagnostic challenges in AdLD.

Study V developed a lumped-element model that can determine the biomechanical characteristics of VF using an HSV running speech sample. A one-mass model was successfully implemented and simulated in this study. Also, an optimization procedure, based on the particle swarm optimization technique, was performed in order to optimize the model parameters with the experimental HSV data in terms of the glottal area waveforms. The optimization procedure with six parameters – representing the main characteristic of the oscillatory behavior of VFs – was successful in determining biomechanical measure of VF (including VF mass, elasticity and viscosity) and generating a waveform closely matched with the HSV-based one. Although the model was built using only one mass with limited degree of freedom, the model was able to sufficiently emulate VF vibrations observed in HSV. This work contributed to an existing gap in literature where the previous studies used this inverse analysis technique neither simulating connected speech samples studying the impaired VF vibrations in AdLD. The study showed the potential of a simplified model like a one-mass model that can still quantify biomechanical properties of VFs using the HSV running speech sample. Overall, the successful implementation of the developed model in the present study paves the path toward a new line for future research where it can be used as a tool that can be further developed and used for estimating clinically relevant non-invasive biomechanical features of VFs in connected speech.

BIBLIOGRAPHY

[1] R. L. Wegel, "Theory of vibration of the larynx," J. Acoust. Soc. Am., vol. 1, pp. 1-21, 1930.

[2] J. Van den Berg, "Myoelastic-aerodynamic theory of voice production," J. Speech Hear. Res., vol. 1, pp. 227-244, 1958.

[3] J. Van den Berg, "Physiology and physics of voice production," Acta Physiol. Pharmacol. Neerl., vol. 5, pp. 40-55, 1956.

[4] R. Husson, "Etude des phénomènes physiologiques et acoustiques fondamentaux de la voix chantée," Thesis, Paris, France, 1950.

[5] G. Portman, R. Humbert, J. Robin, P. Laget and R. Husson, "Etude electromyrographique des corde vocals chez l'homme," Compt. Rend. Soc. Biol. , vol. Paris 149, pp. 286-300, 1955.

[6] G. Fant, In: Acoustic Theory of Speech Production, with Calculations Based on X-Ray Studies of Russian Articulations, Mouton and Co. N.V., The Hague, 1960.

[7] J. Tonndorf, "Die Mechanik bei Stimmlippenschwingungen und bei Schnarchen," Z. HalsNasen, u. Ohrenheilk., vol. 12, pp. 241-245, 1925.

[8] I. R. Titze, T. Riede and P. Popolo, "Nonlinear source-filter coupling in phonation: vocal exercises," J. Acoust. Soc. Am., vol. 123, no. 4, p. 1902–1915, 2008.

[9] M. Zañartu, "Acoustic coupling in phonation and its effect on inverse filtering of oral airflow neck surface acceleration," Ph.D. Thesis, Purdue University, West Lafayette, IN., 2010.

[10] B. H. Story, I. R. Titze and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am., vol. 100, no. 1, p. 537– 554, 1996.

[11] H. Takemoto, K. Honda, S. Masaki, Y. Shimada and I. Fujimoto, "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," J. Acoust. Soc. Am., vol. 119, p. 1037–1049, 2006.

[12] H. M. Hanson, "Glottal characteristics of female speakers: acoustic correlates," J. Acoust. Soc. Am., vol. 101, no. 1, p. 466–481, 1997.

[13] D. H. Klatt and L. C. Klatt, "Analysis synthesis and perception of voice quality variations among male and female talkers," J. Acoust. Soc. Am. , vol. 87, no. 2, p. 820–856, 1990.

[14] I. Titze, Principles of Voice Production, Englewood Cliffs, NJ: Prentice-Hall, 1994.

[15] M. Huffman, "Measures of phonation type in Hmong," J Acoust Soc Am, vol. 81 , no. 2, pp. 495-504, 1987.

[16] Y. Koike, H. Takahashi and T. Calcaterra, "Acoustic measures for detecting laryngeal pathology," Acta Otolaryngol, vol. 84, no. 1-6, pp. 105-117, 1977.

[17] E. Yumoto, W. Gould and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," J Acoust Soc Am., vol. 71, no. 6, pp. 1544-1550, 1982.

[18] J. Hillenbrand, R. Cleveland and R. Erickson, "Acoustic correlates of breathy vocal quality," J Speech, Lang Hear Res., vol. 37, no. 4, pp. 769-778, 1994.

[19] G. Kempster, B. Gerratt, K. Abbott, J. Barkmeier-Kraemer and R. Hillman, "Consensus Auditory-Perceptual Evaluation of Voice: development of a stan dardized clinical protocol," Am J Speech Lang Pathol, vol. 18, pp. 124-132, 2009.

[20] R. Zraick, G. Kempster, N. Connor, S. Thibeault, B. Klaben, Z. Bursac, C. Thrush and L. Glaze, "Establishing validity of the Con sensus Auditory-Perceptual Evaluation of Voice (CAPE-V)," Am J Speech Lang Pathol, vol. 20, p. 14–22, 2011.

[21] M. Rothenberg and J. Mashie, "Monitoring vocal fold abduction through vocal fold contact area," Journal of Speech Language and Hearing Research, vol. 31, p. 338–351, 1988.

[22] T. Luang-Thongkum, "Phonation types in Mon-Khmer languages," UCLA Working Papers in Phonetics, vol. 67, p. 29– 48, 1987.

[23] P. Kitzing, "Stroboscopy–a pertinent laryngological examination," J Otolaryngol., vol. 14, no. 3, pp. 151-157, 1985.

[24] D. M. Bless, M. Hirano and R. J. Feder, "Videostroboscopic evaluation of the larynx," Ear, Nose & Throat Journal, vol. 66, no. 7, pp. 289-296, 1987.

[25] P. Woo, J. Casper, R. Colton and D. Brewer, "Aerodynamic and stroboscopic findings before and after microlaryngeal phonosurgery," J Voice, vol. 8, no. 2, pp. 186-194, 1994.

[26] J. C. Stemple, L. E. Glaze and B. G. Klaben, Clinical voice pathology: Theory and management, Cengage Learning, 2000.

[27] A. Stojadinovic, A. R. Shaha, R. F. Orlikoff, A. Nissan, M.-F. Kornak, B. Singh, J. O. Boyle, J. P. Shah, M. F. Brennan and D. H. Kraus, "Prospective functional voice assessment in patients undergoing thyroid surgery," Ann Surg., vol. 236, no. 6, p. 823–832, 2002.

[28] D. D. Mehta and R. E. Hillman, "Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods," Curr. Opin. in Otol., Head. and Neck. Surg., vol. 16, no. 3, pp. 211-215, 2008.

[29] A. E. Aronson and D. Bless, Clinical Voice Disorders, Thieme, 2011.

[30] R. Patel, S. Dailey and D. Bless, "Comparison of high-speed digital imaging with stroboscopy for laryngeal imaging of glottal disorders," Ann. of Otol., Rhinol & Laryngol, vol. 117, no. 6, pp. 413-424, 2008.

[31] S. R. C. Zacharias, C. M. Myer, J. Meinzen-Derr, L. Kelchner, D. D. Deliyski and A. de Alarcón, "Comparison of videostroboscopy and high-speed videoendoscopy in evaluation of supraglottic phonation," Ann. of Otol., Rhinol & Laryngol., vol. 125, no. 10, pp. 829-837, 2016.

[32] D. D. Deliyski, Laryngeal high-speed videoendoscopy, in: Laryngeal evaluation: Indirect laryngoscopy to high-speed digital imaging, New York: Thieme Medical Publishers, 2010, pp. 243-270.

[33] M. Echternach, M. Döllinger, J. Sundberg, L. Traser and B. Richter, "Vocal fold vibrations at high soprano fundamental frequencies," The Journal of the Acoustical Society of America, vol. 133, no. 2, pp. EL82-EL87, 2013.

[34] A. M. Yousef, D. D. Deliyski, S. R. C. Zacharias, A. de Alarcon, R. F. Orlikoff and M. Naghibolhosseini, "Spatial segmentation for laryngeal high-speed videoendoscopy in connected speech," J Voice, vol. 37, no. 1, pp. 26-36, Nov 27;S0892-1997(20)30408-2, 2023.

[35] A. M. Yousef, D. D. Deliyski, S. R. Zacharias, A. de Alarcon, R. F. Orlikoff and M. Naghibolhosseini, "A Hybrid Machine-Learning-Based Method for Analytic Representation of the Vocal Fold Edges during Connected Speech," Applied Sciences, vol. 11, no. 3, p. 1179, 2021.

[36] A. M. Yousef, D. D. Deliyski, S. R. Zacharias, A. de Alarcon, R. F. Orlikoff and M. Naghibolhosseini, "Automated detection and segmentation of glottal area using deep-learning neural networks in high-speed videoendoscopy during connected speech," in In 14TH INTERNATIONAL CONFERENCE ADVANCES IN QUANTITATIVE LARYNGOLOGY, VOICE AND SPEECH RESEARCH (AQL), Bogota, Colombia, 2021.

[37] M. Naghibolhosseini, D. D. Deliyski, S. R. Zacharias, A. de Alarcon and R. F. Orlikoff, "A method for analysis of the vocal fold vibrations in connected speech using laryngeal imaging," in Manfredi C (Ed.) Proceedings of the 10th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications MAVEBA, Firenze University Press, Firenze, Italy, 2017.

[38] A. M. Yousef, D. D. Deliyski, M. Zayernouri, S. R. Zacharias and M. Naghibolhosseini, "Vocal Fold Detective Edge Analysis in High-speed Videoendoscopy during Running Speech in Adductor Spasmodic Dysphonia," in Proceedings of the 15th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research (AQL), Phoenix, AZ, 2023 March 30-April 1.

[39] A. M. Yousef, D. D. Deliyski, S. R. Zacharias and M. Naghibolhosseini, "Deep-Learning-Based Representation of Vocal Fold Dynamics in Adductor Spasmodic Dysphonia during Connected Speech in High-Speed Videoendoscopy," Journal of Voice, pp. S0892-1997(22)00263-6, 2022. Online ahead of print.

[40] M. Naghibolhosseini, A. M. Yousef, M. Zayernouri, S. R. Zacharias and D. D. Deliyski, "Deep Learning for High-Speed Laryngeal Imaging Analysis," in Proceedings of the 3rd International IEEE Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Amity University, Dubai, UAE, 2023.

[41] D. D. Mehta, D. D. Deliyski, T. F. Quatieri and R. E. Hillman, "Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings," Journal of Speech, Language, and Hearing Research, vol. 54, no. 1, pp. 47-54, 2011.

[42] D. D. Deliyski, M. E. Powell, S. R. Zacharias, T. T. Gerlach and A. de Alarcon, "Experimental investigation on minimum frame rate requirements of high-speed videoendoscopy for clinical voice assessment," Biomed. Signal. Process. and Control, vol. 17, pp. 51-59, 2015.

[43] M. Zañartu, D. D. Mehta, J. C. Ho, G. R. Wodicka and R. E. Hillman, "Observation and analysis of in vivo vocal fold tissue instabilities produced by nonlinear source-filter coupling: A case study," Journal of the Acoustical Society of America, vol. 129, no. 1, pp. 326-339, 2011.

[44] D. D. Mehta, D. D. Deliyski, Zeitels, S. M, M. Zañartu and R. E. Hillman, Integration of transnasal fiberoptic high-speed videoendoscopy with time-synchronized recordings of vocal function, innormal & abnormal vocal folds Kinematics: High speed digital phonoscopy (HSDP), optical coherence tomography (OCT) & narrow band imaging, vol. 12, San Fransisco, CA: Pacific Voice & Speech Foundation, 2015, pp. 105-114.

[45] M. Naghibolhosseini, D. D. Deliyski, S. R. C. Zacharias, A. de Alarcon and R. F. Orlikoff, "Studying vocal fold non-stationary behavior during connected speech using high-speed videoendoscopy," The Journal of the Acoustical Society of America, vol. 144, no. 3, pp. 1766-1766, 2018.

[46] M. Naghibolhosseini, N. Heinz, C. Brown, F. Levesque, S. R. C. Zacharias and D. D. Deliyski, "Glottal attack time and glottal offset time comparison between vocally normal speakers and patients with adductor spasmodic dysphonia during connected speech," in 50th Anniversary Symposium: Care of the Professional Voice, Philadelphia, 2021.

[47] M. Naghibolhosseini, D. D. Deliyski, S. R. C. Zacharias, A. de Alarcon and R. F. Orlikoff, "Glottal attack time in connected speech," in The 11th International Conference on Voice Physiology and Biomechanics ICVPB, East Lansing, MI, 2018.

[48] C. Brown, M. Naghibolhosseini, S. R. C. Zacharias and D. D. Deliyski, "Investigation of high-speed videoendoscopy during connected speech in norm and neurogenic voice disorder," in Michigan Speech-Language-Hearing Association (MSHA) Annual Conference, East Lansing, MI, 2019.

[49] D. D. Deliyski, "Clinical feasibility of high-speed videoendoscopy," in Perspectives on Voice and Voice Disorders, vol. 17, American Speech-Language-Hearing Association, 2007, pp. 12-16.

[50] D. D. Deliyski and R. E. Hillman, "State of the art laryngeal imaging: Research and clinical implications," Current Opinion in Otolaryngology & Head and Neck Surgery, vol. 18, no. 3, pp. 147-152, 2010.

[51] P. Woo, "Objective measures of stroboscopy and high speed video," Advances in Oto-Rhino-Laryngology, vol. 85, pp. 25-44, 2020.

[52] C. Watts, C. Nye and R. Whurr, "Botulinum toxin for treating spasmodic dysphonia (laryngeal dystonia): a systematic Cochrane review," Clin Rehabil, vol. 20, no. 2, pp. 112-122, 2006.

[53] K. E. Castelon, I. Trender-Gerhard, 1. C. Kamm and e. al., "Servicebased survey of dystonia in Munich," Neuroepidemiology, vol. 21, p. 202–206, 2002.

[54] J. M. Schweinfurth, M. Billante and M. S. Courey, "Risk factors and demographics in patients with spasmodic dysphonia," Laryngoscope, vol. 112, no. 2, pp. 220-223, 2002.

[55] J. Stemple, N. Roy and B. Klaben, Clinical Voice Pathology: Theory and Management, 6th ed., Plural Publishing, 2018.

[56] S. M. Cohen, K. J and R. N, "Prevalence and causes of dysphonia in a large treatment-seeking population," Laryngoscope, vol. 122, no. 2, pp. 343-348, 2012.

[57] E. A. Nash and C. L. Ludlow, "Laryngeal muscle activity during speech breaks in adductor spasmodic dysphonia," Laryngoscope, vol. 106, p. 484–489, 1996.

[58] V. Vanderaa and L. A. Vinney, "Laryngeal Sensory Symptoms in Spasmodic Dysphonia," J Voice, 2021.

[59] N. Mor, K. Simonyan and A. Blitzer, "Central voice production and pathophysiology of spasmodic dysphonia," Laryngoscope, vol. 128, p. 177–183, 2018.

[60] N. Roy, A. Mazin and S. N. Awan, "Automated acoustic analysis of task dependency in adductor spasmodic dysphonia versus muscle tension dysphonia," Laryngoscope, vol. 124, pp. 718-724, 2014.

[61] M. Cannito and P. Johson, "Spastic dysphonia: a continuum disorder," J Commun Disord., vol. 14, pp. 215-223, 1981.

[62] D. W. Chen and J. Ongkasuwan, "Spasmodic dysphonia," International ophthalmology clinics, vol. 58, no. 1, pp. 77-87, 2018.

[63] D. K. Chetri, A. L. Merati, J. H. Blumin and e. al., "Reliability of the perceptual evaluation of adductor spasmodic dysphonia," Ann Otol Rhinol Laryngol, vol. 117, p. 159–165, 2008.

[64] M. B. Higgins, C. D. H and L. Shulte, "Phonatory air flow characteristics of adductor spasmodic dysphonia and muscle tension dysphonia," J Speech Lang Hear Res, vol. 42, p. 101–111, 1999.

[65] N. Roy, "Functional dysphonia," Curr Opin Otolaryngol Head Neck Surg, vol. 11, p. 144–148, 2003.

[66] N. Roy, M. Gouse, S. C. Mauszycki, R. M. Merrill and M. E. Smith, "Task specificity in adductor spasmodic dysphonia versus muscle tension dysphonia," The Laryngoscope, vol. 115, no. 2, pp. 311-316, 2005.

[67] D. K. Chhetri, A. H. Mendelsohn, J. H. Blumin and G. S. Berke, "Long-term follow-up results of selective laryngeal adductor denervation–reinnervation surgery for adductor spasmodic dysphonia," Laryngoscope , vol. 116, p. 635–642, 2006.

[68] N. Roy, D. M. Bless, D. Heisey and C. N. Ford, "Manual circumlaryngeal therapy for functional dysphonia: an evaluation of short- and long-term treatment outcomes," J Voice, vol. 11, p. 321–331, 1997.

[69] M. P. Cannito, G. E. Woodson, T. Murry and e. al., "Perceptual analyses of spasmodic dysphonia before and after treatment," Arch Otolaryngol Head Neck Surg., vol. 130 , p. 1393–1399, 2004.

[70] D. T. Weed, B. S. Jewett, C. Rainey and e. al., "Long-term follow-up of recurrent laryngeal nerve avulsion for treatment of spastic dysphonia," Ann Otol Rhinol Laryngol., vol. 105, p. 592–601, 1996.

[71] E. P. Silverman, C. Garvan, R. Shrivastav and e. al., "Combined modality treatment of adductor spasmodic dysphonia," J Voice., vol. 26, p. 77–86, 2012.

[72] C. M. Sapienza, S. Walton and T. Murry, "Adductor spasmodic dysphonia and muscular tension dysphonia: acoustic analysis of sustained phonation and reading," J Voice, vol. 14, p. 502–520, 2000.

[73] C. J. Rees, P. D. Blalock, S. E. Kemp, S. L. Halum and J. A. Koufman, "Differentiation of adductor-type spasmodic dysphonia from muscle tension dysphonia by spectral analysis," Otolaryngol Head Neck Surg, vol. 137, p. 576–581, 2007.

[74] R. Leonard and K. Kendall, "Differentiation of spasmodic and psychogenic dysphonias with phonoscopic evaluation," Laryngoscope , vol. 109, p. 295–300, 1999.

[75] M. D. Morrison and L. A. Rammage, "Muscle misuse voice disorders: description and classification," Acta oto-laryngologica, vol. 113, no. 3, pp. 428-434, 1993.

[76] B. Halberstam, "Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels," ORL, vol. 66, no. 2, pp. 70-73, 2004.

[77] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy and M. De Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels," J Voice, vol. 24, no. 5, pp. 540-555, 2010.

[78] S. Y. Lowell, "The acoustic assessment of voice in continuous speech," SIG 3 Perspectives on Voice and Voice Disorders, vol. 22, no. 2, pp. 57-63, 2012.

[79] D. D. Deliyski and R. E. Hillman, "State of the art laryngeal imaging: Research and clinical implications," Current Opinion in Otolaryngology & Head and Neck Surgery, vol. 18, no. 3, p. 147–152, 2010.

[80] A. Olthoff, C. Woywod and E. Kruse, "Stroboscopy versus high-speed glottography: A comparative study," The Laryngo scope, vol. 117, no. 6, pp. 1123-1126, 2007.

[81] P. S. Popolo, "Investigation of flexible high-speed video nasolaryngoscopy," J Voice, vol. 32, no. 5, pp. 529-537, 2018.

[82] D. D. Deliyski and P. Petrushev, "Methods for objective assessment of high-speed videoendoscopy," Proc Adv in Quant Laryngol, p. 1–16, 2003.

[83] C. Brown, D. D. Deliyski, S. R. C. Zacharias and M. Naghibolhosseini, "Glottal attack and offset time during connected speech in adductor spasmodic dysphonia," in Virtual Voice Symposium: Care of the Professional Voice, Philadelphia, 2020.

[84] C. Tao, Y. Zhang and J. J. Jiang, "Extracting physiologically relevant parameters of vocal folds from high-speed video image series," IEEE Transactions on Biomedical Engineering, vol. 54, no. 5, pp. 794-801, 2007.

[85] J. J. Jiang, C. E. Diaz and D. G. Hanson, "Finite element modeling of vocal fold vibration in normal phonation and hyperfunctional dysphonia: implications for the pathogenesis of vocal nodules," Annals of Otology, Rhinology & Laryngology, vol. 107, no. 7, p. 603–610, 1998.

[86] I. Tokuda, M. Zemke and M. Kob, "Biomechanical modeling of register transitions and the role of vocal tract resonators," J. Acoust. Soc. Am., vol. 127, p. 1528–1536, 2010.

[87] A. Palaparthi, T. Riede and I. R. Titze, "Combining multiobjective optimization and cluster analysis to study vocal fold functional morphology," IEEE Transactions on Biomedical Engineering, vol. 61, no. 7, pp. 2199-2208, 2014.

[88] V. M. Espinoza, M. Zañartu, J. H. Van Stan, D. D. Mehta and R. E. Hillman, "Glottal aerodynamic measures in women with phonotraumatic and nonphonotraumatic vocal hyperfunction.," Journal of Speech Language, and Hearing Research,, vol. 60, no. 8, pp. 2159-2169, 2017.

[89] A. Giovanni, D. Demolin, C. Heim and J. M. Triglia, "Estimated subglottic pressure in normal and dysphonic subjects," Annals of Otology, Rhinology & Laryngology, vol. 109, no. 5, pp. 500-504, 2000.

[90] K. Ketelslagers, M. S. De Bodt, F. L. Wuyts and P. Van de Heyning, "Relevance of subglottic pressure in normal and dysphonic subjects," European archives of oto-rhino-laryngology, vol. 264, no. 5, pp. 519-523, 2007.

[91] P. Zhuang, J. T. Swinarska, C. F. Robieux, M. R. Hoffman, S. Lin and J. J. Jiang, "Measurement of phonation threshold power in normal and disordered voice production," Annals of Otology, Rhinology & Laryngology, vol. 122, no. 9, pp. 555-560, 2013.

[92] M. Döllinger, U. Hoppe, F. Hettlich, J. Lohscheller, S. Schuberth and U. Eysholdt, "Vibration parameter extraction from endoscopic image series of the vocal folds," IEEE Transactions on Biomedical Engineering, vol. 49, no. 8, pp. 773-781, 2002.

[93] P. Gómez, A. Schützenberger, S. Kniesburges, C. Bohr and M. Döllinger, "Physical parameter estimation from porcine ex vivo vocal fold dynamics in an inverse problem framework," Biomechanics and modeling in mechanobiology, vol. 17, no. 3, pp. 777-792, 2018.

[94] K. Ishizaka and J. L. Flanagan, "Acoustic Properties of Longitudinal Displacement in Vocal Cord Vibration," The Bell System Technical Journal, vol. 56, no. 6, p. 889–918, 1977.

[95] B. D. Erath, M. Zañartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer and S. D. Peterson, "A review of lumped-element models of voiced speech," Speech Communication, vol. 55, no. 5, pp. 667-690, 2013.

[96] J. Awrejcewicz, "Numerical Analysis Of The Oscillations Of Human Vocal Cords," Nonlinear Dynamics, vol. 2, no. 1, p. 35–52, 1991.

[97] J. Liljencrants, "A translating and rotating mass model of the vocal folds," J STL-QPSR, vol. 1, p. 1–18, 1991.

[98] P. Šidlof and J. Horáček, "Vocal fold motion and voice production: Mathematical modelling and experiment," in In Forum Acusticum 2005, Budapest, Hungary, 2005.

[99] J. L. Flanagan and L. L. Landgraf, "Self-Oscillating Source for Vocal-Tract Synthesizers," IEEE Transactions on Audio and Electroacoustics, vol. 16(1), p. 57–64, 1968.

[100] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," J. Acoust. Soc. Am., vol. 97, p. 1249–1260, 1995.

[101] K. Ishizaka, M. Matsuidara and T. Kaneko, "Input acoustic-impedance measurement of subglottal system," J. Acoust. Soc. Am., vol. 60, p. 190–197, 1976.

[102] I. T. Tokuda, J. Horáček, J. G. Svec and H. Herzel, "Comparison of biomechanical modeling of register transitions and voice instabilities with excised larynx experiments," Journal of the Acoustical Society of America, vol. 122, no. 1, p. 519–531, 2007.

[103] P. Gómez, K. S, S. A, B. C and D. M, "Degrees of freedom in a vocal fold inverse problem," in In: International conference on bioinformatics and biomedical engineering, Springer, Berlin, 2017.

[104] R. Schwarz, U. Hoppe, M. Schuster, T. Wurzbacher, U. Eysholdt and J. Lohscheller, "Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model," IEEE transactions on biomedical engineering, vol. 53, no. 6, pp. 1099-1108, 2006.

[105] T. Wurzbacher, R. Schwarz, M. Döllinger, U. Hoppe, U. Eysholdt and J. Lohscheller, "Model-based classification of nonstationary vocal fold vibrations," The Journal of the Acoustical Society of America, vol. 120, no. 2, pp. 1012-1027, 2006.

[106] X. Qin, S. Wang and M. Wan, "Improving reliability and accuracy of vibration parameters of vocalfolds based on high-speedvideoand electroglottography," IEEE T Bio-Med Eng, vol. 56, no. 6, p. 1744–1754 , 2009.

[107] R. Fraile, M. Kob, J. I. Godino-Llorente, N. Saenz-Lechon, V. J. Osma Ruiz and J. M. Gutierrez-Arriola, "Physical simulation of laryngeal disorders using a multiple-mass vocal fold model," Biomed. Signal Process. & Control, vol. 7, no. 1, p. 65–78, 2012.

[108] M. Kob, Physical modeling of the singing voice, Aachen, Berlin: Unversity of Technology.

[109] I. R. Titze, "The Myoelastic Aerodynamic Theory of Phonation," In: The National Center for Voice and Speech, 2006.

[110] M. E. Smith, G. S. Berke, B. R. Gerratt and J. Kreiman, "Laryngeal paralyses: theoretical considerations and effects on laryngeal vibration," J. Speech Hear. Res., vol. 35, p. 545–554, 1992.

[111] T. Koizumi, S. Taniguchi and S. Hiromitsu, "Two-mass models of the vocal cords for natural voice synthesis," J. Acoust. Soc. Am., vol. 82, pp. 1179-1192, 1987.

[112] B. Benjamin and G. Croxson, "Vocal nodules in children," Ann. Oto. Rhinol. Laryngol., vol. 99 , no. 5, p. 530–533, 1987.

[113] T. Koizumi, S. Taniguchi and F. Itakura, "An analysis-by-synthesis approach to estimation of vocal cord polyp features," Laryngoscope , vol. 103, p. 1035–1042, 1993.

[114] Y. Zhang and J. J. Jiang, "Chaotic vibrations of a vocal fold model with a unilateral polyp," J. Acoust. Soc. Am., vol. 115, p. 1266–1269, 2004.

[115] E. W. Massey and G. W. Paulson, "Essential vocal tremor: clinical characteristics and response to therapy," South. Med. J., vol. 78, p. 316–317, 1985.

[116] J. A. Logemann, H. B. Fisher, B. Boshes and E. R. Blonsky, "Frequency and coocurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," J. Speech Hear. Disord., vol. 43, p. 47–58, 1978.

[117] Y. Zhang and J. J. Jiang, "Nonlinear dynamic mechanism of vocal tremor from voice analysis and model simulations," J. Sound Vib., vol. 316, p. 248–262, 2008.

[118] Y. Zhang, J. J. Jiang and D. A. Rahn, "Studying vocal fold vibrations in Parkinson's disease with a nonlinear model," Chaos , vol. 15, p. 033903, 2005.

[119] M. Naghibolhosseini, D. D. Deliyski, S. R. Zacharias, A. de Alarcon and R. F. Orlikoff, "Temporal segmentation for laryngeal high-speed videoendoscopy in connected speech," J. of Voice, vol. 32, no. 2, pp. 256.e1-256.e12, 2018.

[120] T. Koç and T. Çiloğlu, "Automatic segmentation of high speed video images of vocal folds," Journal of Applied Mathematics, pp. 1-16, 2014.

[121] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt and M. Döllinger, "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos," Medical Image Analysis, vol. 11, no. 4, p. 400–413, 2007.

[122] S.-Z. Karakozoglou, N. Henrich, C. D'Alessandro and Y. Stylianou, " Automatic glottal segmentation using local-based active contours and application to glottovibrography," Speech Communication, vol. 54, no. 5, p. 641–654, 2012.

[123] H. J. Moukalled, D. D. Deliyski, R. R. Schwarz and S. Wang, "Segmentation of laryngeal high-speed videoendoscopy in temporal domain using paired active contours," in Manfredi C (Ed.) Proceedings of the 10th International Workshop on Models and Analysis of VocaL Emissions for Biomedical Applications MAVEBA, Firenze University Press, Firenze, Italy, 2009.

[124] Y. Yan, X. Chen and D. Bless, "Automatic tracing of vocal-fold motion from high-speed digital images," IEEE Transactions on Biomedical Engineering, vol. 53, no. 7, p. 1394–1400, 2006.

[125] Y. Yan, E. Damrose and D. Bless, "Functional analysis of voice using simultaneous high-speed imaging and acoustic recordings," Journal of Voice, vol. 21, p. 604–616, 2007.

[126] D. D. Mehta, D. D. Deliyski, S. M. Zeitels, T. F. Quatieri and R. E. Hillman, "Voice production mechanisms following phonosurgical treatment of early glottic cancer," Annals of Otology, Rhinology and Laryngology, vol. 119, no. 1, pp. 1-9, 2010.

[127] J. Demeyer, T. Dubuisson, B. Gosselin and M. Remacle, Glottis segmentation with a high-speed glottography: A fullyautomatic method, In: 3rd Adv. Voice Funct. Assess. Int. Workshop, 2009.

[128] Y. Yan, G. Du, C. Zhu and G. Marriott, "Snake based automatic tracing of vocal-fold motion from high-speed digital images," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'12), 2012.

[129] Y. Zhang, E. Bieging, H. Tsui and J. J. Jiang, "Efficient and effective extraction of vocal fold vibratory patterns from high-speed digital imaging," Journal of Voice, vol. 24, p. 21–29, 2010.

[130] S. Zhou, J. Wang, S. Zhang, Y. Liang and Y. Gong, "Active contour model based on local and global intensity information for medical image segmentation," Neurocomputing, vol. 186, pp. 107-118, 2016.

[131] G. Sulong, H. Abdulaali and S. Hassan, "Edge detection algorithms vs-active contour for sketch matching: Comparative study," Research Journal of Applied Sciences, Engineering and Technology, vol. 11, no. 7, pp. 759-764, 2015.

[132] M. Kass, A. Witkin and D. Terzopoulos, "Active contour models," International Journal of Computer Vision, vol. 1, no. 1, pp. 321-331, 1987.

[133] C. Manfredi, L. Bocchi, S. Bianchi, N. Migali and G. Cantarella, "Objective vocal fold vibration assessment from videokymographic images," Biomedical Signal Processing and Control, vol. 1, no. 2, pp. 129-136, 2006.

[134] G. Hinton, "Deep Learning — A Technology With the Potential to Transform Health Care," Journal of the American Medical Association, vol. 320, no. 11, p. 1101, 2018.

[135] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, T. S and J. Dean, "A guide to deep learning in healthcare," Nature Medicine, vol. 25, no. 1, pp. 24-29, 2019.

[136] M. S, V. G. O, M. E. D, A. Laborai, L. Guastini, G. Peretti and L. S. Mattos, "Learning-based classification of informative laryngoscopic frames," Comput Methods Programs Biomed, vol. 158, pp. 21-30, 2018.

[137] I. Patrini, M. Ruperti, S. Moccia, L. S. Mattos, E. Frontoni and E. De Momi, "Transfer learning for informative-frame selection in laryngoscopic videos through learned features," Med Biol Eng Comput, vol. 58, no. 6, pp. 1225-1238, 2020.

[138] A. Galdran, P. Costa and A. Campilho, "Real-Time Informative Laryngoscopic Frame Classification with Pre-Trained Convolutional Neural Networks," in In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 2019.

[139] J. Ren, X. Jing, J. Wang, X. Ren, Y. Xu, Q. Yang, L. Ma, Y. Sun, W. Xu, N. Yang and J. Zou, "Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique," The Laryngoscope, vol. 130, no. 11, pp. E686-E693, 2020.

[140] H. Xiong, P. Lin, J. G. Yu, J. Ye, L. Xiao, Y. Tao, Z. Jiang, W. Lin, M. Liu, J. Xu and W. Hu, "Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images," EBioMedicine, vol. 48, pp. 92-99, 2019.

[141] W. K. Cho, J. L. Yeong, H. A. Joo, I. S. Jeong, Y. Choi, S. Y. Nam, S. Y. Kim and S. Choi, "Diagnostic Accuracies of Laryngeal Diseases Using a Convolutional Neural Network-Based Image Classification System," The Laryngoscope, 2021.

[142] M. K. Fehling, F. Grosch, M. E. Schuster, B. Schick and J. Lohscheller, "Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network," PLoS ONE, vol. 15, no. 2: e0227791, 2020.

[143] P. Gómez, A. M. Kist, P. Schlegel, D. A. Berry, D. K. Chhetri, S. Dürr, M. Echternach, A. M. Johnson, S. Kniesburges, M. Kunduk, Y. Maryn, A. Schützenberger, M. Verguts and M. Döllinger, "BAGLS, a multihospital benchmark for automatic glottis segmentation," Scientific Data, vol. 7, no. 1, p. 186, 2020.

[144] A. M. Kist, J. Zilker, P. Gómez, A. Schützenberger and M. Döllinger, "Rethinking glottal midline detection," Scientific Reports, vol. 10:20723, 2020.

[145] A. M. Kist and M. Döllinger, "Efficient biomedical image segmentation on EdgeTPUs at point of care," IEEE Access, vol. 8, pp. 139356-139366, 2020.

[146] A. Kist, P. Gómez, D. Dubrovskiy, P. Schlegel, M. Kunduk, M. Echternach, R. Patel, M. Semmler, C. Bohr, S. Dürr, A. Schützenberger and M. Döllinger, "A Deep Learning Enhanced Novel Software Tool for Laryngeal Dynamics Analysis," Journal of Speech, Language, and Hearing Research, pp. 1-15, 2021.

[147] M. Döllinger, T. Schraut, L. A. Henrich, D. Chhetri, M. Echternach, A. M. Johnson, M. Kunduk, Y. Maryn, R. R. Patel, R. Samlan and M. Semmler, "Re-Training of Convolutional Neural Networks for Glottis Segmentation in Endoscopic High-Speed Videos," Applied Sciences, vol. 12, no. 19, p. 9791, 2022.

[148] C. L. Ludlow, "Treatment approaches for spasmodic dysphonia: limitations of current approaches," Curr Opin Otolaryngol Head Neck Surg, p. 160–165, 2009.

[149] G. S. Berke, K. E. Blackwell, B. R. Gerratt, A. Verneil, K. S. Jackson and J. A. Sercarz, "Selective laryngeal adductor denervation–reinnervation: a new surgical treatment for adductor spasmodic dysphonia," Ann Otol Rhinol Laryngol, vol. 100, p. 227–231, 1999.

[150] E. Yiu, L. Worrall, J. Longland and C. Mitchell, "Analysing vocal quality of connected speech using Kay's computerized speech lab: a preliminary finding," Clin Linguist & Phon, vol. 14, no. 4, p. 295–305, 2000.

[151] F. Boutsen, M. P. Cannito, M. Taylor and B. Bender, "Botox treatment in adductor spasmodic dysphonia: a meta-analysis," J Sp Lang Hear Res, vol. 45, p. 469– 481, 2002.

[152] A. Yousef, D. D. Deliyski, S. R. Zacharias and M. Naghibolhosseini, "Detection of Vocal Fold Image Obstructions in High-Speed Videoendoscopy During Connected Speech in Adductor Spasmodic Dysphonia: A Convolutional Neural Networks Approach," Journal of Voice, pp. S0892-1997(22)00027-3, Mar 15; S0892-1997(22)00027-3, 2022. Online ahead of print.

[153] R. Orlikoff, D. Deliyski, R. Baken and B. Watson, "Validation of a glottographic measure of vocal attack," J Voice, vol. 23, no. 2, pp. 164-168, 2009.

[154] M. Cannito and G. Kondraske, "Rapid manual abilities in spasmodic dysphonic and normal female subjects," J Speech Hear Res, vol. 33, p. 123–133, 1990.

[155] N. Roy, "Differential diagnosis of muscle tension dysphonia and spasmodic dysphonia," Current Opinion in Otolaryngology & Head and Neck Surgery, vol. 18, no. 3, pp. 165-170, 2010.

[156] O. Russakovsky, J. Deng, J. Krause, S. Satheesh, S. Ma, Z. Huang, K. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, "ImageNet large scale visual recognition challenge," Int J Comput Vis, vol. 115, pp. 211-252, 2015.

[157] T. Hirasawa, K. Aoyama, T. Tanimoto, S. Ishihara, S. Shichijo, T. Ozawa, T. Ohnishi, M. Fujishiro, K. Matsuo, J. Fujisaki, T. Tada and e. al., "Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images," Gastric Cancer, vol. 21, pp. 653-660, 2018.

[158] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv: 1412.6980., 2014.

[159] J. Hartigan and M. Wong, "A K-means Clustering Algorithm," Applied Statistics, vol. 28, pp. 100-108, 1979.

[160] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in SODA ‘07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Philadelphia, 2007.

[161] A. Amini, T. Weymouth and R. Jain, "Using dynamic programming for solving variational problems in vision," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 9, p. 855–867, 1990.

[162] M. Naghibolhosseini, S. R. Zacharias, S. Zenas, F. Levesque and D. D. Deliyski, "Laryngeal Imaging Study of Glottal Attack/Offset Time in Adductor Spasmodic Dysphonia during Connected Speech," Applied Sciences, vol. 13, no. 5, p. 2979, 2023.

[163] R. D. Blevins, Formulas for Natural Frequency and Mode Shape, Reprint Edition., 2001.

[164] I. R. Titze, "The concept of muscular isometrics for optimizing vocal intensity and efficiency," J Res Singing, vol. 14, pp. 15-25, 1979.

[165] E. Cataldo, F. R. Leta, J. Lucero and L. Nicolato, "Synthesis of voiced sounds using low-dimensional models of the vocal cords and time-varying subglottal pressure," Mechanics Research Communications, vol. 33, no. 2, pp. 250-260, 2006.

[166] I. R. Titze, "Comments on the myoelastic-aerodynamic theory of phonation," Journal of Speech, Language, and Hearing Research, vol. 23, no. 3, pp. 495-510, 1980.

[167] J. L. Flanagan, "Some properties of the glottal sound source," Journal of Speech and Hearing Research, vol. 1, no. 2, pp. 99-116, 1958.

[168] J. VAN DEN BERG, J. T. ZANTEMA and J. Doornenbal, "On the air re sistance and the Bernoulli effect of the human larynx," J. Acoust. Soc. Amer., vol. 29, pp. 626-631, 1957.

[169] E. Cataldo and C. Soize, "Voice signals produced with jitter through a stochastic one-mass mechanical model," Journal of voice, vol. 31, no. 1, pp. 111-e9, 2017.

[170] A. M. Yousef, D. D. Deliyski, S. R. Zacharias, A. de Alarcon, R. F. Orlikoff and M. Naghibolhosseini, "A Deep Learning Approach for Quantifying Vocal Fold Dynamics during Connected Speech using Laryngeal High-Speed Videoendoscopy," Journal of Speech, Language, and Hearing Research, vol. 65, no. 6, pp. 2098-2113, 2022.

[171] J. Kennedy and R. Eberhart, "Particle swarm optimization," in In Proceedings of ICNN'95-international conference on neural networks, 1995.

[172] J. Xin, G. Chen and Y. Hai, "A particle swarm optimizer with multi-stage linearly-decreasing inertia weight.," in In International joint conference on computational sciences and optimization IEEE, 2009 .

[173] D. D. Deliyski, P. P. Petrushev, H. S. Bonilha, T. T. Gerlach, B. Martin-Harris and R. E. Hillman, "Clinical imple mentation of laryngeal high-speed videoendoscopy: Challenges and evolution," Folia Phoniatrica et Logopaedica, vol. 60, no. 1, pp. 33-44, 2008.

[174] N. G. De Biase, G. P. Korn, P. Lorenzon, M. Padovani, M. Moraes, G. Madazio and L. C. P. Vilanova, "Dysphonia severity degree and phonation onset latency in laryngeal adductor dystonia," Journal of Voice, vol. 24, no. 4, pp. 406-409, 2010.

[175] W. Chen, P. Woo and T. Murry, "Vibratory Onset of Adductor Spasmodic Dysphonia and Muscle Tension Dysphonia: A High-Speed Video Study," J. Voice, vol. 34, p. 598–603, 2020.

[176] M. P. De Vries, H. K. Schutte and G. J. Verkerke, "Determination of parameters for lumped parameter models of the vocal folds using a finite-element method approach," The Journal of the Acoustical Society of America, vol. 106, no. 6, pp. 3620-3628, 1999.

[177] M. S. Howe and R. S. McGowan, "On the single-mass model of the vocal folds.," Fluid dynamics research, vol. 42, no. 1, p. 015001, 2010.

[178] F. Avanzini, P. Alku and M. Karjalainen, "One-delayed-mass model for efficient synthesis of glottal flow," in In Seventh European Conference on Speech Communication and Technology, 2001.

[179] F. Avanzini, "Simulation of vocal fold oscillation with a pseudo-one-mass physical model," Speech Communication, vol. 50, no. 2, pp. 95-108, 2008.