ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

A Deep Learning Approach for Quantifying Vocal Fold Dynamics During Connected Speech Using Laryngeal High-Speed Videoendoscopy

Ahmed M. Yousef,[a] Dimitar D. Deliyski,[a] Stephanie R. C. Zacharias,[b,c] Alessandro de Alarcon,[d,e] Robert F. Orlikoff,[f] and Maryam Naghibolhosseini[a]

[a] Department of Communicative Sciences and Disorders, Michigan State University, East Lansing [b] Head and Neck Regenerative Medicine Program, Mayo Clinic, Scottsdale, AZ [c] Department of Otolaryngology—Head and Neck Surgery, Mayo Clinic, Phoenix, AZ [d] Division of Pediatric Otolaryngology, Cincinnati Children's Hospital Medical Center, OH [e] Department of Otolaryngology—Head and Neck Surgery, University of Cincinnati, OH [f] College of Allied Health Sciences, East Carolina University, Greenville, NC

ARTICLE INFO

Article History:
Received October 7, 2021
Revision received January 30, 2022
Accepted February 28, 2022

Editor-in-Chief: Bharath Chandrasekaran
Editor: Jack J. Jiang

https://doi.org/10.1044/2022_JSLHR-21-00540

ABSTRACT

Purpose: Voice disorders are best assessed by examining vocal fold dynamics in connected speech. This can be achieved using flexible laryngeal high-speed videoendoscopy (HSV), which enables us to study vocal fold mechanics with high temporal details. Analysis of vocal fold vibration using HSV requires accurate segmentation of the vocal fold edges. This article presents an automated deep-learning scheme to segment the glottal area in HSV from which the glottal edges are derived during connected speech.
Method: Using a custom-built HSV system, data were obtained from a vocally healthy participant reciting the "Rainbow Passage." A deep neural network was designed for glottal area segmentation in the HSV data. A recently introduced hybrid approach by the authors was utilized as an automated labeling tool to train the network on a set of HSV frames, where the glottis region was automatically annotated during vocal fold vibrations. The network was then tested against manually segmented frames using different metrics, intersection over union (IoU), and Boundary F1 (BF) score, and its performance was assessed on various phonatory events on the HSV sequence.
Results: The designed network was successfully trained using the hybrid approach, without the need for manual labeling, and tested on the manually labeled data. The performance metrics showed a mean IoU of 0.82 and a mean BF score of 0.96. In addition, the evaluation assessment of the network's performance demonstrated an accurate segmentation of the glottal edges/area even during complex nonstationary phonatory events and when vocal folds were not vibrating, thus overcoming the limitations of the previous hybrid approach that could only be applied to the vibrating vocal folds.
Conclusions: The introduced automated scheme guarantees accurate glottis representation in challenging color HSV data with lower image quality and excessive laryngeal maneuvers during all instances of connected speech. This facilitates the future development of HSV-based measures to assess the running vibratory characteristics of the vocal folds in speakers with and without voice disorder.
Supplemental Material: https://doi.org/10.23641/asha.19798864

Voice disorders are typically observed in connected speech (Halberstam, 2004; Lowell, 2012; Maryn et al., 2010; Morrison & Rammage, 1993; Roy et al., 2005; Yiu et al., 2000). Videostroboscopy, which consists of an endoscope coupled with a stroboscopic light and a video camera, is the primary tool used in clinical settings to visualize

Correspondence to Maryam Naghibolhosseini: naghib@msu.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

and assess vocal fold vibration (Bless et al., 1987; Kitzing, 1985; Woo et al., 1994). Despite the widespread clinical use of videostroboscopy (Mafee et al., 2005; Slonimsky, 2019; Uloza et al., 2005; Verikas et al., 2009), it fails to capture the details of the intracycle vibratory characteristics of the vocal folds in running speech and during irregular vocal fold oscillations (Aronson & Bless, 2011; Mehta & Hillman, 2008; Stemple et al., 2000; Stojadinovic et al., 2002). This is because the strobe light is unable to precisely be synchronized to the acoustic signal. This is crucial when evaluating voice disorders, where it is essential to observe the detailed cycle-to-cycle vibrations of vocal folds due to aperiodic vocal fold vibrations (Patel et al., 2008; Zacharias et al., 2016).

Laryngeal high-speed videoendoscopy (HSV) is an advanced endoscopy technique that overcomes videostroboscopy's restrictions (Deliyski, 2010; Deliyski & Petrushev, 2003; Echternach et al., 2013; Patel et al., 2008; Zacharias et al., 2016). The high HSV frame rates allow visualization of the detailed motion of the vocal folds providing the opportunity to develop new tools to objectively analyze the entire vibratory cycles during phonation (Mehta et al., 2015; Naghibolhosseini et al., 2017, 2018a; Yousef et al., 2020, 2022; Zañartu et al., 2011). HSV offers the capability of studying connected speech including nonstationary phonation events during normal voice production (Naghibolhosseini et al., 2018b, 2018c; Popolo, 2018; Yousef et al., 2020, 2021a, 2021b; Yousef, Deliyski, Zacharias, & Naghibolhosseini, 2021), as well as aperiodic vibration (Brown et al., 2019; Deliyski et al., 2015; Mehta et al., 2011; Naghibolhosseini et al., 2021; Zenas et al., 2021) and singing (Echternach et al., 2013). A myriad of studies showed the usefulness of HSV as a powerful tool, particularly for the objective analysis of vocal fold vibrations—which can contribute to our understanding of complex voice production mechanisms (Deliyski, 2007; Deliyski & Hillman, 2010; Deliyski et al., 2008; Woo, 2020; Yousef et al., 2021b, Yousef, Deliyski, Zacharias, & Naghibolhosseini, 2021). However, using HSV remains a daunting task for clinicians since they must visually navigate through thousands of HSV frames. Clinical assessment of vocal fold vibration using videoendoscopic images is performed subjectively with visual inspection of the data. Several approaches have been proposed to overcome this evaluation challenge through providing more compact representations of the data; approaches such as kymograms (Švec & Schutte, 2012), phonovibrograms (Lohscheller & Eysholdt, 2008), glottovibrograms, and phasegrams (Döllinger et al., 2011; Herbst et al., 2013). Employing efficient quantitative methods for voice analysis using HSV would be valuable for clinical voice examination (Olthoff et al., 2007). Hence, extracting useful, quantitative measurements of the dynamic motion of the vocal folds in HSV recordings could allow the clinicians to obtain clinically relevant characteristics of the vocal fold oscillation during connected speech. Therefore, it is crucial to develop techniques to automatically analyze vocal fold vibration through segmenting vocal fold edges and glottal area.

Spatial segmentation methods of glottal edges/area were proposed for analysis of vocal fold vibrations mainly during isolated sustained vowels (Karakozoglou et al., 2012; Koç & Çiloğlu, 2014; Lohscheller et al., 2007; Mehta et al., 2011; Moukalled et al., 2009); methods such as threshold-based region growing technique (Lohscheller et al., 2007; Yan et al., 2006, 2007), histogram thresholding (Larsson et al., 2000; Mehta et al., 2010, 2011), active-contour modeling (ACM; Karakozoglou et al., 2012; Manfredi et al., 2006; Moukalled et al., 2009; Schenk et al., 2015), watershed transform (Osma-Ruiz et al., 2008), and level-set methods (Demeyer et al., 2009; Shi et al., 2015). More recently, we have developed an ACM-based glottal edge representation for HSV analysis in connected speech (Yousef et al., 2020). This method was applied to detect the glottal edges on kymograms, which were automatically extracted at different intersections of the vocal folds in the HSV data. This approach was based on deformation of an active contour to capture the edges of interest in the image through an iterative energy minimization procedure (Kass et al., 1988). This method not only has been able to address the sensitivity of prior techniques to image noise and intensity inhomogeneity, but also could tackle more challenging video quality in HSV data in connected speech (Yousef et al., 2020). However, the ACM method was still vulnerable to the excessive laryngeal maneuvers and inferior image quality (dim lighting) in some HSV frames. 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.

We enhanced the ACM method for connected speech by coupling the ACM method with an unsupervised machine learning method and introduced a hybrid approach (Yousef et al., 2021a). In the hybrid technique, a k-means clustering method was used to accurately localize the initialized active contours of the ACM method in the HSV kymograms – facilitating the deformation of these contours to efficiently capture the glottal edges. The unsupervised machine learning was specifically selected in order to have a fully automated hybrid method (Yousef et al., 2021a). This combination provided an advantage over most image processing techniques which, in contrast, showed difficulty in automatically segmenting the glottal area, requiring different degrees of manual interaction and visual inspection (Fehling et al., 2020; Kist & Döllinger, 2020).

Despite the high accuracy and robustness of our hybrid scheme over the ACM approach, it required a

relatively high computational cost. Apart from the computational cost, the hybrid technique only worked during vocal fold vibrations but not during all nonstationary phonatory events nor when the vocal folds were not vibrating. Analyzing these nonstationary events and the vocal folds motion during all instances of connected speech would allow for studying different phenomena such as prephonatory adjustments, glottal offset, and attack times for norm and disorder (Naghibolhosseini et al., 2018b, 2018c, 2021).

To address the drawbacks of the hybrid method, in this study, we propose a deep learning-based approach as a more general and flexible tool to capture glottis area/edges during any phonatory events of connected speech. Deep learning has been shown to be a promising technique to detect the glottal edges/area in HSV data during sustained vocalization in several studies (Fehling et al., 2020; Gómez et al., 2020; Kist & Döllinger, 2020; Kist et al., 2020, 2021). These approaches used deep neural networks in the segmentation task and required manual labeling/annotation of the glottal edges/area in HSV frames to train the neural networks. Keeping in view that these previous studies used sustained phonation as their data set, expanding this to connected speech is an important next step. This study aims to build upon the hybrid method, previously proposed by the authors, and introduces a cost-effective and robust scheme based on deep neural networks. The scheme is developed as the first deep learning-based technique to automatically segment the glottal edge/area in the entire HSV data during connected speech, which also does not require manual labeling for the model training. This method is applicable to the various events that exist during running speech: stationary events as in sustained vocal fold vibrations and nonstationary events as in onsets/offsets of phonation and voice breaks, and even during no vibrations of the vocal folds. The proposed scheme is a combination of the hybrid method and a deep neural network. That is, since the hybrid method was accurate when applied to the HSV data during sustained portions of the connected speech, in this study, the hybrid method is being utilized as an automated labeling tool. The hybrid method is used accordingly to segment the vocal fold edges of a set of HSV frames. These segmented images serve as automated, labeled data for the purpose of training a deep neural network instead of manual labeling that can be a cumbersome and subjective task. The network is trained so that it can segment HSV frames in different complex phonatory events during connected speech even with inferior image quality.

The objectives of this work are to (a) develop an automated labeling technique based on our recently developed hybrid method, (b) design and train a deep neural network using the automated labeling tool, and (c) show the capability of the trained network in glottal edge/area representation in challenging color HSV data recorded during all instances of running speech.

Materials and Method

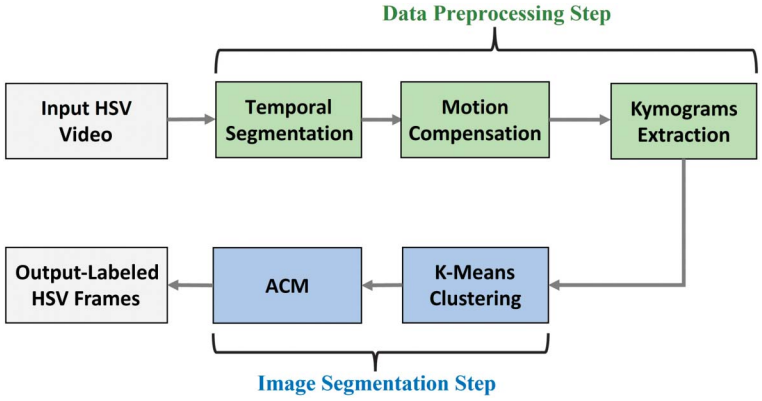
Data Collection

A vocally normal 38-year-old female participated in this study. The participant was examined at the Center for Pediatric Voice Disorders, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, United States. The examination was approved by the institutional review board. The experimental setup was designed using a custom-built color HSV system that captured video recordings at 4,000 frames per second (with 249-μs integration time) and spatial resolution of 256 × 256 pixels. The system was utilized to record the subject during reading of the "Rainbow Passage," which took 29.14 s. The recorded HSV sequence comprised of 116,543 frames in total, which was saved as an uncompressed 24-bit RGB in AVI format. The designed system utilized a FASTCAM SA-Z color high-speed camera (Photron Inc.) with a cache memory of 64 GB and 12-bit color image sensor. The camera was coupled with a 300-W xenon light source (model 7152A: PENTAX Medical Company) along with a 3.6-mm Olympus ENF-GP Fiber Rhinolaryngoscope (Olympus Corporation).

Automated Labeling Tool

This section provides information about developing the automated labeling tool. Here is an overview of why and how the automated labeling tool was implemented. An image segmentation tool was implemented to provide an adequate estimate of the glottal area in a set of HSV frames during vocal fold vibration (Yousef et al., 2021a). This set of segmented frames formed a training data set on which a deep neural network was trained to accurately segment the glottal area during connected speech in different phonatory events. That is, instead of using manual labeling to create the training data, an automated labeling tool was utilized. Regarding how the labeling tool was developed, the automated hybrid technique we previously introduced (Yousef et al., 2021a) served as the labeling tool. The hybrid image segmentation method consisted of several integrated algorithms, which can be divided into two main stages as shown in Figure 1: A data preprocessing step (including temporal segmentation, motion compensation, and kymograms extraction) and an image segmentation step (including k-means clustering and ACM). Each of the aforementioned stages will be discussed in detail in the following subsections. All the algorithms were developed and implemented using 64-bit MATLAB R2019b (MathWorks Inc.).

Figure 1. Workflow chart of the automated labeling tool. The gray boxes indicate the input (high-speed videoendoscopy [HSV] video data) and the output (labeled HSV frames with segmented glottal area), the green boxes show the data preprocessing steps, and the blue boxes represent the image segmentation steps. ACM = active-contour modeling.



Data Preprocessing

The first step before proceeding to segmenting the glottal area was to preprocess the video data automatically as shown in Figure 1. The temporal segmentation algorithm (Naghibolhosseini et al., 2018a) was first utilized to automatically extract the timestamps of the vocalized segments (phonation onsets and offsets) in the entire HSV recording with an unobstructed view of the vocal folds. Therefore, HSV frames with a visually obstructed view of the vocal folds were excluded from further processing in the automated labeling tool. Next, a motion compensation was applied to the vocalized segments. This was done using a gradient-based approach (Deliyski, 2005; Naghibolhosseini et al., 2017) to track the location of vibrating vocal folds. The location of the vocal folds was captured in a bounding box across frames. The frames were cropped to only enclose the vocal folds to eliminate any irrelevant tissues or image noise. For different vocalized segments in the video data, HSV kymograms were then extracted at various cross sections of the vocal folds, along the anteroposterior length. Therefore, onset, sustained phonation, and offset were included in the kymograms, ensuring that the full phonatory phases were obtained. For more details of each preprocessing step, please refer to Yousef et al. (2020).

Hybrid Method for Image Segmentation and Automated Labeling

After the automated preprocessing of the HSV recording and extracting the kymograms (see Figure 1), an automated labeling method was implemented as a hybrid technique (Yousef et al., 2021a). The hybrid approach was a combination of an unsupervised machine learning (k-means clustering) and ACM. The k-means method was initially used to segment the glottal area and classify each pixel as a glottal or a nonglottal pixel, and afterwards the ACM was used to locate the glottal edges. The hybrid technique (also known as k-means ACM) allowed for analytic representation of the glottal edges in the extracted kymograms. A set of features (i.e., pixel intensities of the red and green channels and the image gradient) were extracted from the HSV kymogram images. The blue channel was excluded due to high levels of image noise. The unsupervised clustering technique was then utilized based on the well-known k-means method (Jain et al., 1999) to cluster all pixels of the kymogram images into glottal and nonglottal pixels. All the glottal pixels were highlighted by a spline and formed the glottal area. This spline acted as an initialized contour for the ACM method (see Yousef et al., 2020, for complete description of the ACM approach). The ACM method was implemented on the kymograms to accurately locate the glottal edges at different cross sections along the vocal fold length. That is, instead of directly segmenting the vocal fold edges from the HSV frames, the glottal edges were first segmented in the kymograms and then the detected edges were registered back to the HSV frames. These segmented HSV frames during vocal fold vibration were utilized as automated, labeled data for the purpose of training a deep neural network. The network was trained such that it could also segment frames in complex nonstationary phonatory events, also frames with more challenging image quality.

Deep Neural Network

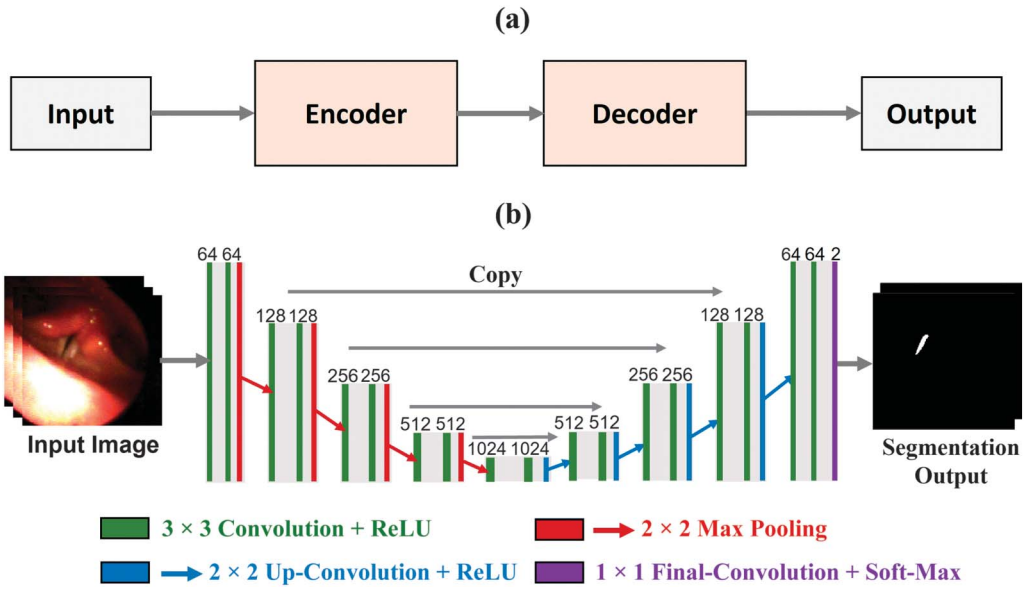
Network Architecture

In the present work, the deep learning network (U-Net) architecture was used, which is a fully convolutional neural network architecture. U-Net was introduced by Ronneberger et al. (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 illustrates a schematic

diagram of the deep neural network used in this work, which shows the proposed U-Net architecture based on the work of Ronneberger et al. The network was implemented using 64-bit MATLAB R2019b (MathWorks Inc.) on a CPU. As seen in Figure 2, Panel a, the input HSV frame is provided to the encoder as a 256 × 256 RGB color image. The encoder then encodes the input HSV frame into feature representations by extracting the main spatial features and the context of the input image. The encoder (i.e., contracting path) was designed as a typical fully convoluted neural network (Long et al., 2015) encompassing repeated 3 × 3 convolutions. Each convolution was followed by a rectified linear unit (ReLU), which converts numerical values of less than zero in the convolution input to a value of zero while keeping the values above zero the same (Petersen et al., 2018). Hence, the ReLU accelerates the computations and enhances the model performance. The feature maps were then kept in the memory for latter concatenation before performing a down sampling step. The down sampling was done to reduce the size of the input frames such that most of the unique features in the input image were preserved; therefore, this step decreased the number of redundant features in the image and lowered the number of the parameters for training the network. For the down sampling, a 2 × 2 max pooling with a stride of 2 was used such that the number of feature maps was doubled at each down sampling stage. Max pooling is a method to reduce the dimensions of the input images while retaining the most predominant image features (Lee et al., 2009). This was done by a sliding window with a size of 2 × 2 pixels, where only the pixels with the maximum value in this sliding window were considered. After four stages of down sampling at the end of the contracting path, dropout was applied, where it discarded contiguous regions in the feature map (rather than dropping random areas) in order to avoid overfitting during the training (Ronneberger et al., 2015).

As shown in Figure 2, the decoder (i.e., the expansive path) had a similar path as of the encoder but inverted. The decoder semantically projects the features extracted by the encoder onto the pixel space—allowing for reconstructing the output image. Instead of pooling, each decoding stage encompassed an upsampling technique of the feature map using a transposed 2 × 2 convolution (up-convolution) that halved the number of feature maps. The upsampling was done to recover the size of the feature maps and to make the output image have the same dimensions as of the input through compensating for the reduction of the resolution caused by the pooling (Wu et al., 2009). Each up-convolution was followed by a ReLU. The upsampling convolution was concatenated with the corresponding feature map in the encoder path, which was previously stored. In the original architecture of Ronneberger

Figure 2. Schematic diagram for the deep neural network in this work. Panel a shows the general encoder–decoder architecture of the U-Net. Panel b illustrates the detailed structure of the network. The HSV frames serve as the input. The input images are downsampled during the contracting path (the encoder) through multiple layers of 3 × 3 convolutions along with rectified linear unit (ReLU) layers (in green), followed by several 2 × 2 max pooling layers (in red). The extracted features from the encoder are then propagated and upsampled during an expansive path (the decoder) using multiple layers of both 3 × 3 convolution besides ReLU in green and 2 × 2 up-convolution besides ReLU in blue. The last layer involves a 1 × 1 final convolution followed by a soft-max layer (in purple). The dimensions of the feature maps are also included in the figure. Residuals are propagated from encoder to decoder via concatenation (shown in gray arrows). The segmentation results return an output of binary images, where the bright area represents the segmented glottal area.



et al. (2015), the feature map was cropped to match the corresponding up-sampled convolution; however, in this work, a padding of 1 for the convolutions was utilized to have a matched input and output dimensions. After concatenation, repeated 3×3 convolution along with ReLU were applied. This encoder–decoder structure made the network architecture symmetric. The final layer consisted of a 1×1 convolution with two filters (corresponding to the number of classes in the present work) followed by a pixel-wise soft-max layer (Kouretas & Paliouras, 2019). The soft-max layer is an activation function that assigned decimal probabilities (0–1) to each pixel in the image representing the probability of each pixel to be either a glottal or nonglottal pixel (Kouretas & Paliouras, 2019). The soft-max layer was then followed by a pixel classification layer, classifying each pixel in the input image as either glottal or nonglottal/background, which was the final outcome of the network.

Network Training

To train a deep neural network, an optimization technique is used to tune the network parameters that yield the minimum difference between the predicted outcome (by the developed network) and the expected outcome (by the ground-truth data). Adam optimizer was considered in the present work to train the developed network as an iterative stochastic gradient descent optimizer (see Kingma & Ba, 2014, for the complete description and details of implementing Adam optimizer). The initial learning rate was chosen to be 0.001 during the training process; the learning rate is the amount by which the network parameters are updated or tuned during the training. Three parameters were considered as the hyper-parameters of the optimizer ($\beta1$, $\beta2$, and Epsilon); noting that the hyper-parameters are parameters whose values are utilized to control the learning/training process. The decay rates of the first- and second-order moments were set to $\beta1 = 0.9$ and $\beta2 = 0.999$. The Epsilon was set to 10^{-8}, which refers to an extremely small number to prevent any division by zero during the implementation. The U-Net was trained on the training data set, which was created using the automated labeling tool (the hybrid method). The training data set was composed of 2,050 automatically segmented frames. These segmented frames were evaluated through visual inspection to validate the accuracy of the automated glottal area segmentation prior to training of the neural network. Twenty percent of the training data were used as a validation data set to evaluate the performance of the network during the training process and, accordingly, tune the network parameters to enhance its performance. Noting that these validation frames were randomly chosen from the training data set. A testing data set was also created using manually segmented frames to assess the network performance, which is discussed in detail in the following subsection.

In connected speech, there are excessive laryngeal maneuvers across frames that could shift and alter the location of the vocal folds in the three-dimensional space. Therefore, different data augmentation techniques were applied to the training data set to enhance the generalization ability of the trained neural network. This augmented data was used for training the model. To do so, the training frames were randomly rotated between −30° and +30°, translated along the horizontal and vertical directions with a range between −64 and +64 pixels, and upscaled and downscaled by a factor ranging from 0.5 up to 2. In addition, other augmentation methods were performed via altering the contrast/brightness (by adding and scaling random amount of brightness between −0.2–−0.08 and 0.5–1.5, respectively), adding random Gaussian noise with zero mean and variance of 0–0.01, and applying random Gaussian blurring with the standard deviation of 0–1.5. The final ratio of the augmented frames to the original frames in the modified training data set was 1:3.

The constructed network was trained using the modified training data set, including the augmented frames, with a batch size of 10 for a maximum of 20 epochs. During the training process, the training data set was shuffled before each epoch and similarly, the validation data was shuffled before each network validation. Early stopping of the training process was used when noticing a plateau in the validation accuracy. The predicted output of the trained network returned an image, where each pixel was classified into either a "glottal pixel," located inside the glottal area, or a "background pixel," which was outside the glottal region. Those pixels labeled "glottal pixels" were assigned a value of one while the remaining pixels ("background pixels") were assigned values of zero. That is, a binary image was constructed as a segmentation mask.

Network Testing and Evaluation

Sixteen different networks were trained. The U-Net architecture of these networks were altered with respect to the number of the encoder–decoder depth levels ranging from 3 to 6 levels, the level refers to the number of times the input frames were downsampled or upsampled during processing. Different batch sizes of 4, 10, 16, and 32 were considered during the training of these networks. The segmentation performance of each of the trained networks was evaluated against a testing data set, where the best-performing network (the proposed one in this work) was determined based on the highest segmentation accuracy scores. In this comparison, the intersection over union (IoU) score was mainly used as a segmentation accuracy measure to compare between the different networks, which will be discussed later in this section along with the other evaluation metrics used in this study.

The testing data set was comprised of manually labeled HSV frames. This data set was created using 600

HSV frames from different phonation events including sustained vocal fold vibration, onsets/offsets of phonation, and when vocal folds were not vibrating. These frames were selected randomly and were different from the training data set. The glottal edges in these frames were manually segmented to serve as ground truth by an expert. After creating the testing data set, a quantitative evaluation of the developed segmentation method was carried out—in addition to the visual inspection of the method's segmentation performance. This assessment step was done as the final performance evaluation of the proposed approach.

To assess the performance of the best-performing network using the test data set, an area-based metric (class accuracy score and IoU) and a boundary-based metric (Boundary F1) were used. The area-based metrics were considered to evaluate the accuracy of the developed network in segmenting the glottal area whereas the boundary-based metric was utilized to evaluate the accuracy in detecting the glottal edges. Evaluating both the glottal area and its edges was useful to assess the overall performance, particularly, in cases where the glottal boundary was accurately detected but some pixels in the glottal area were misclassified.

The class accuracy score determined the percentage of the correctly predicted pixels for a specific class (e.g., the "glottal class"). The accuracy score is the ratio between the correctly classified pixels and the total number of pixels in a specific class as identified by the ground truth data (Estrada & Jepson, 2009). The accuracy was calculated as:

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{1}$$

where TP (true positive pixels) was the number of correctly classified pixels as glottal area by the network and FN (false negative pixels) referred to the incorrectly classified pixels as nonglottal pixels. Since this accuracy metric cannot give a complete evaluation of the proposed method performance, IoU metric was used as another area-based metric for a more rigorous performance assessment. This is because, accuracy can be a misleading measure since it can excessively present false positive cases (referred to the incorrectly classified pixels as glottal pixels), whereas IoU metric penalizes false positive pixels. IoU metric also provides a statistical measurement of the segmentation accuracy and is commonly used in related literature for evaluating segmentation performance (Gómez et al., 2020; Kist & Döllinger, 2020; Kist et al., 2021). It can take a value from zero (no similarity/overlap) to one (perfect similarity/overlap) so, the larger the IoU, the better the network performance. IoU was computed using the following (Csurka et al., 2013):

$$IoU = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{2}$$

where FP is the false positive pixels. In addition, the dice coefficient (DC; Dice, 1945) was computed as an extra metric to evaluate the overall match between the ground truth and the automatically segmented glottal area. DC is similar to IoU and both are positively correlated. The main reason to calculate DC is for comparison of this study with the literature that also used DC as a verification metric. DC was calculated using the following equation:

$$\text{DC} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{3}$$

Although the area-based metrics allowed for evaluation of the segmented glottal area, these metrics did not evaluate the precision of the glottal area boundaries (i.e., vocal fold edges). Therefore, the Boundary F1 (BF) score was included in the analysis as a contour-based metric to evaluate the accuracy of the segmented glottal area boundaries (glottal edges). BF score refers to the weighted average of the accuracy and precision. This boundary score allowed for the measurement of F1 accuracy between the predicted glottal boundary using the proposed segmentation method against the ground truth boundary. The BF score was considered as a measure of how the estimated glottal boundary using the proposed approach was close to the spatial location of the ground-truth boundary. BF (F1) score was computed according to the following equation (Csurka et al., 2013):

$$\text{BF} = \frac{2 \times \text{Precision} \times \text{Accuracy}}{\text{Precision} + \text{Accuracy}} \tag{4}$$

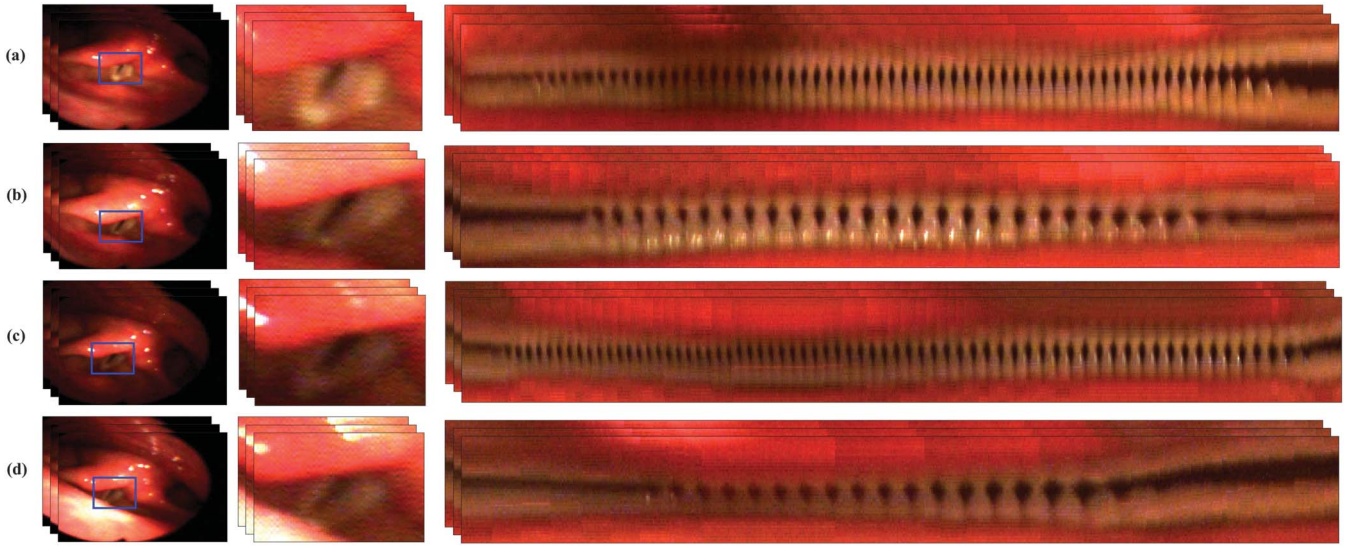
where

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Results

Figure 3 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 vocal folds 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

Figure 3. 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 the 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.



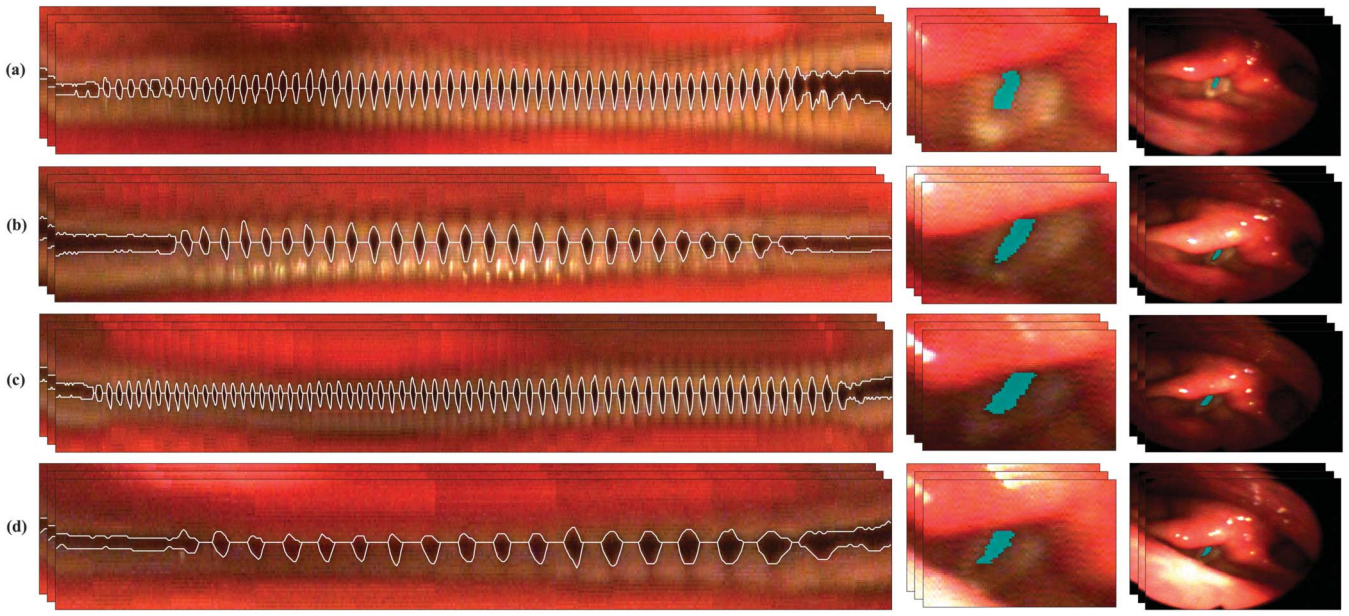
at different cross sections of the vibrating vocal folds (represented by a stacked kymograms in the figure). Examples of the extracted kymograms at the medial intersection of the vocal folds showing the variation in the glottal region across the frames can be seen in the right side of the figure. The kymograms span through the entire vocalization—clearly representing the vibratory patterns and behavior, namely, phonation onset, the sustained vibration of vocal folds, and phonation offset.

For each kymogram, the hybrid method (also known as k-means ACM) was applied to segment and detect the glottal edges during vocalizations. Figure 4 illustrates the results of implementing the k-means ACM algorithm at various kymograms of Figure 3 that were extracted from different vocalizations. As illustrated in Figure 4, the k-means ACM technique was able to accurately segment the edges of right and left vocal folds, 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 4) 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 5 shows the results of training the proposed deep neural network for two different vocalizations (Panels a and b). Each panel shows the result for four frames, extracted from a different vocalization. The results in Figure 5 are displayed for the following frames: #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 labeling tool) and the proposed deep neural network. The segmented glottal areas using the k-means ACM (in cyan color) and NN (in yellow color) are overlaid on top of each other (in the right-side panels of Figure 5) to demonstrate their differences. The DC and the BF 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 NN 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 6 illustrates the performance of the proposed deep neural network on HSV frames extracted from three different vocalized segments (Panels 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

Figure 4. Results of applying k-means active-contour modeling 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).



capture the glottal area variations at the onsets and offsets. The HSV frames in Figure 6 (indicated by red dots in the glottal area waveforms) were selected during different behaviors of the vocal folds. As such, for the two vocalizations in Panels a and b, the segmented frames are extracted near the voicing offset and onset at 138–172.5 ms and 14–33.5 ms, respectively. The segmented frames shown in Panel c were extracted during the sustained oscillation of vocal folds between 222.5 and 229.5 ms—representing sudden larger degree of vocal folds abduction during the sustained vibration.

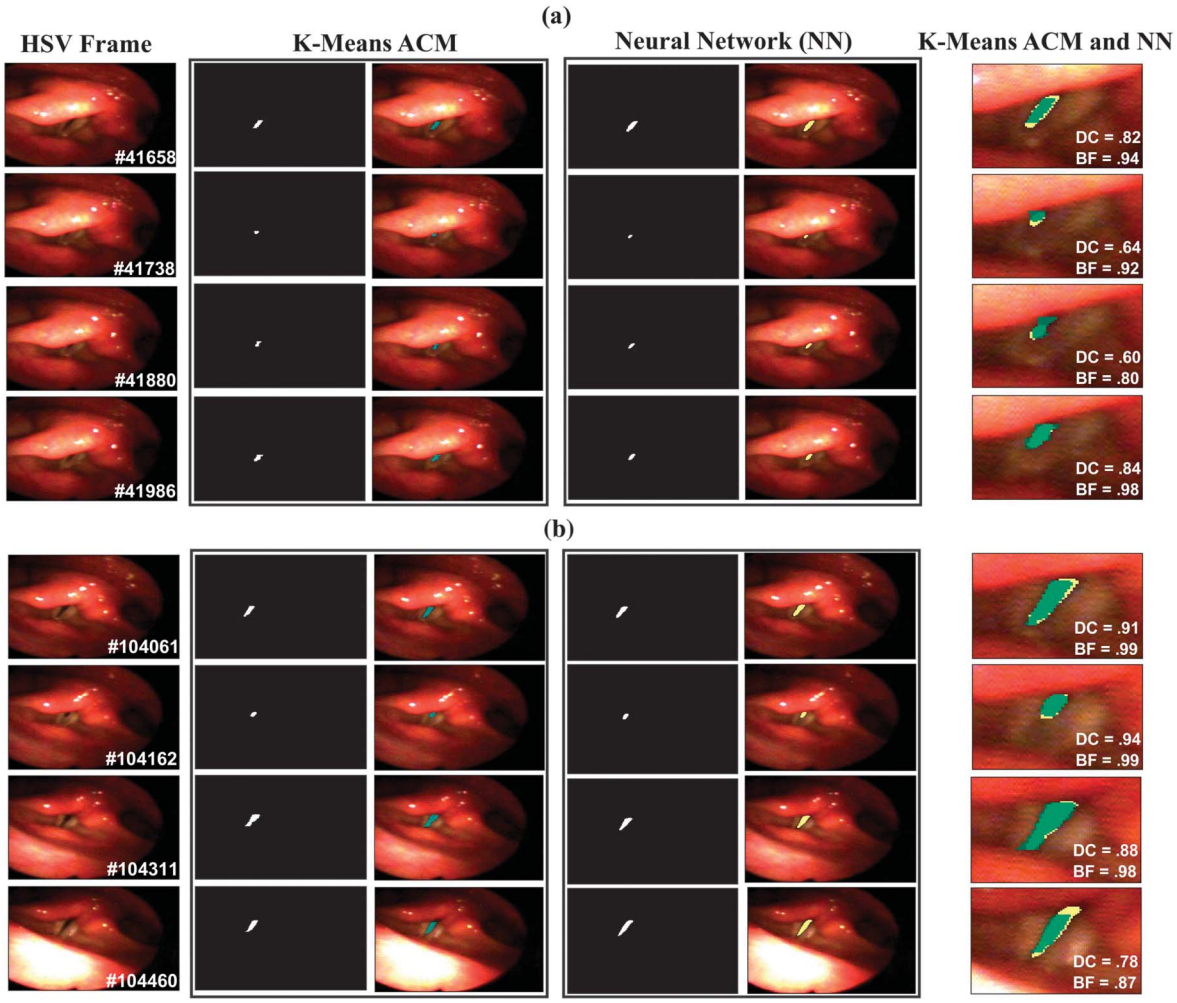
See Supplemental Material S1 displays the performance of the introduced network during multiple, consecutive vocalized segments in the HSV data during running speech. The top figure in the video shows a sequence of 6,115 HSV frames, where the glottal area is segmented (in cyan color) using the deep neural network. The associated glottal area waveform is depicted in the bottom two figures of the video; the moving red star refers to the computed glottal area value, synchronized with the displayed frame in the top figure. The middle panel shows the variation in the glottal area (computed in pixels) of the last 50 frames in the running video. The bottom panel illustrates the glottal area calculated during the entire video sequence. This video demonstrates the successful performance of our approach on the 6,115 subsequent frames, selected arbitrarily. The video also shows the result of segmentation during different phases of glottal closure and opening. In addition, the video displays the accurate segmentation of the glottal area in different phonation events: voicing onsets, voicing offsets, irregular vocal fold vibrations, voice breaks, and instances where the vocal folds are obstructed by the epiglottis.

Besides the visual inspection, the network was also tested against manually labeled frames (testing data set) in order to provide a quantitative evaluation of the segmentation performance. When the proposed network was applied to the testing data set, 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 (SD = 0.26) and 0.88 (SD = 0.25), respectively; SD refers to the standard of deviation. In addition, the contour-based evaluation metric (BF score) showed a mean value of 0.96 (SD = 0.12) in terms of detecting the glottal area boundary.

Discussion

We have recently developed two spatial segmentation approaches to represent the vocal fold edges/areas during connected speech in HSV data. The first approach was implemented using ACM and showed promising performance, but it was vulnerable to excessive image noise and very dim lighting conditions in connected speech data (Yousef et al., 2020). This technique can be best used for HSV data collected using rigid videoendoscopy due to higher image quality. The second technique was designed

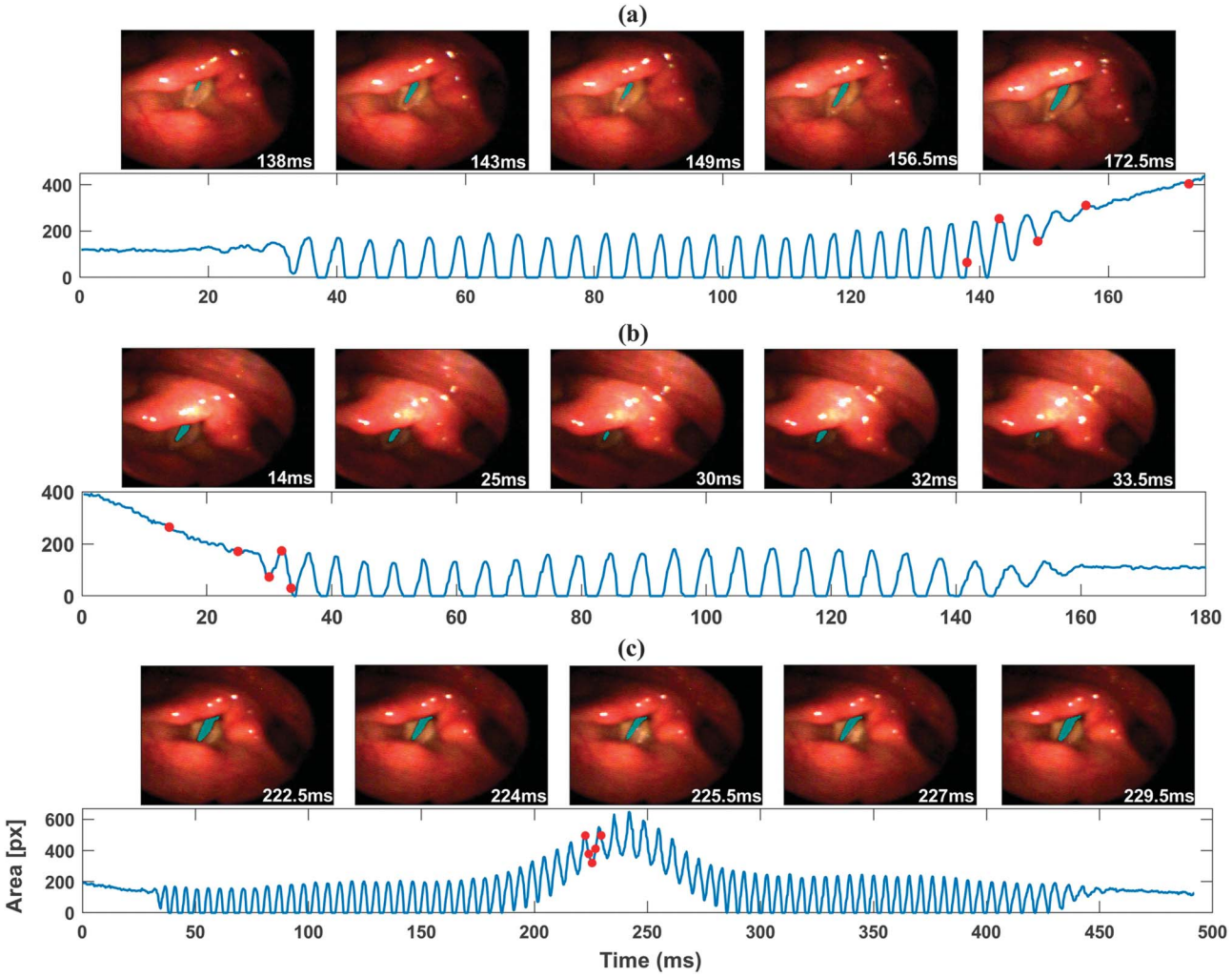
Figure 5. Results of implementing the k-means active-contour modeling (ACM) and the trained deep neural network (NN); the segmented HSV frames along with the associated binary segmentation mask are shown for eight different frames extracted from two different vocalizations (a and b). (a) for Frame #41,658, #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 NN 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. HSV = high-speed videoendoscopy.



as a spatiotemporal hybrid method employing the developed ACM method and combining it with an unsupervised machine learning approach, k-means clustering (Yousef et al., 2021a). The approach was robust against inferior image quality and showed effective capability, particularly in segmenting the glottal edges during vocal fold vibrations but encountered challenges in presence of excessive laryngeal movements during nonstationary phonation tasks like onsets/offsets of phonation. The hybrid method has the benefit of extracting the edges of vocal folds from the spatiotemporal information in HSV data. This method would provide accurate edges of the vocal folds mainly during the more sustained portions of phonation. The present technique uses the power of the developed hybrid method and deep learning to overcome the challenges of the hybrid method in terms of detecting the glottal area during all phonatory tasks (including nonstationary portions) and when vocal folds are not vibrating. Hence, the proposed 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.

The present work 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

Figure 6. The glottal area waveform as well as five segmented high-speed videoendoscopy 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.



implemented to accurately segment the edges of the vibrating vocal folds in the kymograms. The combination of k-means and ACM yielded a precise detection of the vocal fold edges (Yousef et al., 2021a), which were registered back to the original HSV frames to segment the glottal area. The hybrid method showed an accurate performance but mainly during vocal folds sustained oscillation. 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 (Fehling et al., 2020; Gómez et al., 2020; Kist & Döllinger, 2020; Kist et al., 2020, 2021). That is, our 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 (Yousef et al., 2021a).

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 our previous work (Yousef et al., 2021a), 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. 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 vocal folds in the HSV data (Deliyski et al., 2008).

Although the network was trained on frames segmented during sustained vocal fold 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 vocal fold vibrations, and when vocal folds 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 automated labeling tool we developed, 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 by 0.96 (SD = 0.12) indicating high accuracy of the network in localizing the edges of the glottal area, (i.e., vocal fold edges). Furthermore, the developed network achieved a mean IoU of 0.82 (SD = 0.26) and a mean DC of 0.88 (SD = 0.25), signifying high precision in detecting the glottal area.

This work is 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 (Kist & Döllinger, 2020; Kist et al., 2020) and RGB (Fehling et al., 2020) 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 vocal folds. 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 (Gómez et al., 2020; Kist & Döllinger, 2020) and DC of 0.85 (Fehling et al., 2020). 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 (Fehling et al., 2020; Gómez et al., 2020; Kist & Döllinger, 2020; Kist et al., 2020, 2021) were entirely utilized for only spatial segmentation. Among these studies, Fehling et al. (Fehling et al., 2020) was the only 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 vocal fold 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. 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. 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 in future when applied to monochrome data with a higher dynamic range. In future work, after applying this method to the HSV data of patients with voice disorders, objective HSV measures will be developed to characterize voice disorders. Having access to such automated measures would benefit future clinicians from HSV analysis in connected speech as it provides detailed vocal folds vibratory information that could potentially facilitate voice disorder diagnosis.

Conclusions

Developing approaches to automatically segment/detect the glottal area/edges in HSV is a critical step for vocal function analysis in connected speech, where voice disorders typically reveal themselves. This work introduces an efficient deep-learning model that can provide a quantitative representation of the glottal area from HSV during running speech. A successful implementation of an automated labeling tool for training deep-learning approaches was performed. The tool was based on our previously developed hybrid method incorporating multiple image processing steps: temporal segmentation, motion compensation, and spatial segmentation using k-means clustering with ACM. Since the hybrid method was accurate across the HSV frames with sustained vocal fold vibrations, it was used as an automated labeling tool to segment the glottal area to form a large training data set for a deep neural network. The developed network even outperformed the labeling tool (our prior hybrid method) by improving the segmentation accuracy, 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 present work showed the feasibility and the promising capability of the introduced deep-learning scheme to segment the glottal area in even challenging color HSV data during connected speech. This facilitates the future utilization of the developed model for HSV analysis in running speech from more vocally normal participants as well as patients with voice disorders.

Acknowledgments

The authors would like to acknowledge the support by the National Institute on Deafness and Other Communication Disorders Grant K01DC017751 (PI: Naghibolhosseini, Maryam), "Studying the Laryngeal Mechanisms Underlying Dysphonia in Connected Speech" and Grant R01DC007640 (PI: Deliyski, Dimitar), "Efficacy of Laryngeal High-Speed Videoendoscopy."

References

Aronson, A. E., & Bless, D. (2011). *Clinical voice disorders*. Thieme.

Bless, D. M., Hirano, M., & Feder, R. J. (1987). Videostroboscopic evaluation of the larynx. *Ear, Nose & Throat Journal, 66*(7), 289–296.

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

Csurka, G., Larlus, D., Perronnin, F., & Meylan, F. (2013). What is a good evaluation measure for semantic segmentation? In T. Burghardt, D. Damen, W. Mayol-Cuevas, & M. Mirmehdi (Eds.). *Proceedings of the British Machine Vision Conference* (Vol. 27, No. 2013, pp. 32.1–32.11). BMVA Press. https://doi.org/10.5244/C.27.32

Deliyski, D. D. (2005). Endoscope motion compensation for laryngeal high-speed videoendoscopy. *Journal of Voice, 19*(3), 485–496. https://doi.org/10.1016/j.jvoice.2004.07.006

Deliyski, D. D. (2007). Clinical feasibility of high-speed videoendoscopy. *Perspectives on Voice and Voice Disorders, 17*(1), 12–16. https://doi.org/10.1044/vvd17.1.12

Deliyski, D. D. (2010). Laryngeal high-speed videoendoscopy. In K. A. Kendall & R. J. Leonard (Eds.). *Laryngeal evaluation: Indirect laryngoscopy to high-speed digital imaging* (Chap. 28).

Deliyski, D. D., & Hillman, R. E. (2010). State of the art laryngeal imaging: Research and clinical implications. *Current Opinion in Otolaryngology & Head and Neck Surgery, 18*(3), 147–152. https://doi.org/10.1097/MOO.0b013e3283395dd4

Deliyski, D. D., & Petrushev, P. (2003). Methods for objective assessment of high-speed videoendoscopy. *Proceedings Advances in Quantitative Laryngology,* 1–16.

Deliyski, D. D., Petrushev, P. P., Bonilha, H. S., Gerlach, T. T., Martin-Harris, B., & Hillman, R. E. (2008). Clinical implementation of laryngeal high-speed videoendoscopy: Challenges and evolution. *Folia Phoniatrica et Logopaedica, 60*(1), 33–44. https://doi.org/10.1159/000111802

Deliyski, D. D., Powell, M. E., Zacharias, S. R., Gerlach, T. T., & de Alarcon, A. (2015). Experimental investigation on minimum frame rate requirements of high-speed videoendoscopy for clinical voice assessment. *Biomedical Signal Processing and Control, 17,* 21–28. https://doi.org/10.1016/j.bspc.2014.11.007

Demeyer, J., Dubuisson, T., Gosselin, B., & Remacle, M. (2009). Glottis segmentation with a high-speed glottography: A fully-automatic method. In *3rd Advanced Voice Function Assessment International Workshop*.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*(3), 297–302. https://doi.org/10.2307/1932409

Döllinger, M., Lohscheller, J., Svec, J., McWhorter, A., & Kunduk, M. (2011). Support vector machine classification of vocal fold vibrations based on phonovibrogram features. In F. Ebrahimi (Ed.), *Advances in vibration analysis research*. https://doi.org/10.5772/15200

Echternach, M., Döllinger, M., Sundberg, J., Traser, L., & Richter, B. (2013). Vocal fold vibrations at high soprano fundamental frequencies. *The Journal of the Acoustical Society of America, 133*(2), EL82–EL87. https://doi.org/10.1121/1.4773200

Estrada, F. J., & Jepson, A. D. (2009). Benchmarking image segmentation algorithms. *International Journal of Computer Vision, 85*(2), 167–181. https://doi.org/10.1007/s11263-009-0251-z

Fehling, M. K., Grosch, F., Schuster, M. E., Schick, B., & Lohscheller, J. (2020). Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. *PLOS ONE, 15*(2), Article e0227791. https://doi.org/10.1371/journal.pone.0227791

Gómez, P., Kist, A. M., Schlegel, P., Berry, D. A., Chhetri, D. K., Dürr, S., & Döllinger, M. (2020). Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. *Scientific Data, 15*(2), 186. https://doi.org/10.1371/journal.pone.0227791

Halberstam, B. (2004). Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. *Journal for Oto-Rhino-Laryngology, Head and Neck Surgery, 66*(2), 70–73. https://doi.org/10.1159/000077798

Herbst, C. T., Herzel, H., Švec, J. G., Wyman, M. T., & Fitch, W. T. (2013). Visualization of system dynamics using phasegrams. *Journal of the Royal Society Interface, 10*(85), 20130288. https://doi.org/10.1098/rsif.2013.0288

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *Association for Computer Machinery, 31*(3), 264–323. https://doi.org/10.1145/331499.331504

Karakozoglou, S.-Z., Henrich, N., D'Alessandro, C., & Stylianou, Y. (2012). Automatic glottal segmentation using local-based active contours and application to glottovibrography. *Speech Communication, 54*(5), 641–654. https://doi.org/10.1016/j.specom.2011.07.010

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision, 1*(4), 321–331. https://doi.org/10.1007/BF00133570

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv: 1412.6980.

Kist, A. M., & Döllinger, M. (2020). *Efficient biomedical image segmentation on EdgeTPUs at point of care*. IEEE. https://doi.org/10.1109/ACCESS.2020.3012722

Kist, A. M., Gómez, P., Dubrovskiy, D., Schlegel, P., Kunduk, M., Echternach, M., & Döllinger, M. (2021). A deep learning enhanced novel software tool for laryngeal dynamics analysis. *Journal of Speech, Language, and Hearing Research, 64*(6), 1889–1903. https://doi.org/10.1044/2021_JSLHR-20-00498

Kist, A. M., Zilker, J., Gómez, P., Schützenberger, A., & Döllinger, M. (2020). Rethinking glottal midline detection. *Scientific Reports, 10*(1), 20723. https://doi.org/10.1038/s41598-020-77216-6

Kitzing, P. (1985). Stroboscopy-a pertinent laryngological examination. *Journal of Otolaryngology, 14*(3), 151–157. https://doi.org/10.1016/S0095-4470(19)30693-X

Koç, T., & Çiloğlu, T. (2014). Automatic segmentation of high speed video images of vocal folds. *Journal of Applied Mathematics, 2014,* 1–16. https://doi.org/10.1155/2014/818415

Kouretas, I., & Paliouras, V. (2019). Simplified hardware implementation of the softmax activation function. In *2019 8th international conference on Modern Circuits and Systems Technologies (MOCAST)* (pp. 1–4). IEEE. https://doi.org/10.1109/MOCAST.2019.8741677

Larsson, H., Hertegard, S., Lindestad, P. A., & Hammarberg, B. (2000). Vocal fold vibrations: high-speed imaging, kymography, and acoustic analysis: A preliminary report. *The Laryngoscope, 110*(12), 2117–2122. https://doi.org/10.1097/00005537-200012000-00028

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning* (pp. 609–616). ACM Digital Library. https://doi.org/10.1145/1553374.1553453

Lohscheller, J., & Eysholdt, U. (2008). Phonovibrogram visualization of entire vocal fold dynamics. *The Laryngoscope, 118*(4), 753–758. https://doi.org/10.1097/MLG.0b013e318161f9e1

Lohscheller, J., Toy, H., Rosanowski, F., Eysholdt, U., & Döllinger, M. (2007). Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Medical Image Analysis, 11*(4), 400–413. https://doi.org/10.1016/j.media.2007.04.005

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 3431–3440). IEEE. https://doi.org/10.1109/CVPR.2015.7298965

Lowell, S. Y. (2012). The acoustic assessment of voice in continuous speech. *SIG 3 Perspectives on Voice and Voice Disorders, 22*(2), 57–63. https://doi.org/10.1044/vvd22.2.57

Mafee, M. F., Valvassori, G. E., & Becker, M. (2005). *Imaging of the neck and head* (2nd ed.). Thieme. https://doi.org/10.1055/b-006-160969

Manfredi, C., Bocchi, L., Bianchi, S., Migali, N., & Cantarella, G. (2006). Objective vocal fold vibration assessment from videokymographic images. *Biomedical Signal Processing and Control, 1*(2), 129–136. https://doi.org/10.1016/j.bspc.2006.06.001

Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N., & De Bodt, M. (2010). Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels. *Journal of Voice, 24*(5), 540–555. https://doi.org/10.1016/j.jvoice.2008.12.014

Mehta, D. D., Deliyski, D. D., Quatieri, T. F., & Hillman, R. E. (2011). Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings. *Journal of Speech, Language, and Hearing Research, 54*(1), 47–54. https://doi.org/10.1044/1092-4388(2010/10-0026)

Mehta, D. D., Deliyski, D. D., Zeitels, M. S., Zañartu, M., & Hillman, R. E. (2015). *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). Pacific Voice & Speech Foundation.

Mehta, D. D., Deliyski, D. D., Zeitels, S. M., Quatieri, T. F., & Hillman, R. E. (2010). Voice production mechanisms following phonosurgical treatment of early glottic cancer. *Annals of Otology, Rhinology and Laryngology, 119*(1), 1–9. https://doi.org/10.1177/000348941011900101

Mehta, D. D., & Hillman, R. E. (2008). Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. *Current Opinion in Otolaryngology & Head and Neck Surgery, 16*(3), 211–215. https://doi.org/10.1097/MOO.0b013e3282fe96ce

Morrison, M. D., & Rammage, L. A. (1993). Muscle misuse voice disorders: Description and classification. *Acta Oto-Laryngologica, 113*(3), 428–434. https://doi.org/10.3109/00016489309135839

Moukalled, H. J., Deliyski, D. D., Schwarz, R. R., & Wang, S. (2009). Segmentation of laryngeal high-speed videoendoscopy in temporal domain using paired active contours. In C. Manfredi (Ed.), *Proceedings of the 10th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications MAVEBA* (pp. 137–140). Firenze University Press.

Naghibolhosseini, M., Deliyski, D. D., Zacharias, S. R., de Alarcon, A., & Orlikoff, R. F. (2017). A method for analysis of the vocal fold vibrations in connected speech using laryngeal imaging. In C. Manfredi (Ed.). *Proceedings of the 10th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications MAVEBA, 10* (pp. 107–110). Firenze University Press.

Naghibolhosseini, M., Deliyski, D. D., Zacharias, S. R., de Alarcon, A., & Orlikoff, R. F. (2018a). Temporal segmentation for laryngeal high-speed videoendoscopy in connected speech. *Journal of Voice, 32*(2), 256.e1. https://doi.org/10.1016/j.jvoice.2017.05.014

Naghibolhosseini, M., Deliyski, D. D., Zacharias, S. R., de Alarcon, A., & Orlikoff, R. F. (2018b). *Glottal attack time in connected speech*. The 11th International Conference on Voice Physiology and Biomechanics ICVPB. East Lansing, MI.

Naghibolhosseini, M., Deliyski, D. D., Zacharias, S. R., de Alarcon, A., & Orlikoff, R. F. (2018c). Studying vocal fold non-stationary behavior during connected speech using high-speed videoendoscopy. *The Journal of the Acoustical Society of America, 144*(3), 1766–1766. https://doi.org/10.1121/1.5067811

Naghibolhosseini, M., Heinz, N., Brown, C., Levesque, F., Zacharias, S. R. C., & Deliyski, D. D. (2021). *Glottal attack time and glottal offset time comparison between vocally normal speakers and patients with adductor spasmodic dysphonia during connected speech*. 50th Anniversary Symposium: Care of the Professional Voice, June 2–6, 2021.

Olthoff, A., Woywod, C., & Kruse, E. (2007). Stroboscopy versus high-speed glottography: A comparative study. *The Laryngoscope, 117*(6), 1123–1126. https://doi.org/10.1097/MLG.0b013e318041f70c

Osma-Ruiz, V., Godino-Llorente, J. I., Sáenz-Lechón, N., & Fraile, R. (2008). Segmentation of the glottal space from laryngeal images using the watershed transform. *Computerized Medical Imaging and Graphics, 32*(3), 193–201. https://doi.org/10.1016/j.compmedimag.2007.12.003

Patel, R., Dailey, S., & Bless, D. (2008). Comparison of high-speed digital imaging with stroboscopy for laryngeal imaging of glottal disorders. *Annals of Otology, Rhinology & Laryngology, 117*(6), 413–424. https://doi.org/10.1177/000348940811700603

Petersen, P., Raslan, M., & Voigtlaender, F. (2018). *Topological properties of the set of functions generated by neural networks of fixed size*. arXiv: 1806.08459.

Popolo, P. S. (2018). Investigation of flexible high-speed video nasolaryngoscopy. *Journal of Voice, 32*(5), 529–537. https://doi.org/10.1016/j.jvoice.2017.08.017

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Roy, N., Gouse, M., Mauszycki, S. C., Merrill, R. M., & Smith, M. E. (2005). Task specificity in adductor spasmodic dysphonia versus muscle tension dysphonia. *The Laryngoscope, 115*(2), 311–316. https://doi.org/10.1097/01.mlg.0000154739.48314.ee

Schenk, F., Aichinger, P., Roesner, I., & Urschler, M. (2015). Automatic high-speed video glottis segmentation using salient regions and 3D geodesic active contours. *Annals of the BMVA, 2015*(3), 1–15.

Shi, T., Kim, J. H., Murry, T., Woo, P., & Yan, Y. (2015). Tracing vocal fold vibrations using level set segmentation method. *International Journal for Numerical Methods in Biomedical Engineering, 31*(6), e02715. https://doi.org/10.1002/cnm.2715

Slonimsky, E. (2019). Laryngeal imaging. *Operative Techniques in Otolaryngology-Head and Neck Surgery, 30*(4), 237–242. https://doi.org/10.1016/j.otot.2019.09.003

Stemple, J. C., Glaze, L. E., & Klaben, B. G. (2000). *Clinical voice pathology: Theory and management*. Cengage Learning.

Stojadinovic, A., Shaha, A. R., Orlikoff, R. F., Nissan, A., Kornak, M.-F., Singh, B., Boyle, J. O., Shah, J. P., Brennan, M. F., & Kraus, D. H. (2002). Prospective functional voice assessment in patients undergoing thyroid surgery. *Annals of Surgery, 236*(6), 823–832. https://doi.org/10.1097/00000658-200212000-00015

Švec, J. G., & Schutte, H. K. (2012). Kymographic imaging of laryngeal vibrations. *Current Opinion in Otolaryngology & Head and Neck Surgery, 20*(6), 458–465. https://doi.org/10.1097/MOO.0b013e3283581feb

Uloza, V., Saferis, V., & Uloziene, I. (2005). Perceptual and acoustic assessment of voice pathology and the efficacy of endolaryngeal phonomicrosurgery. *Journal of Voice, 19*(1), 138–145. https://doi.org/10.1016/j.jvoice.2004.01.009

Verikas, A., Uloza, V., Bacauskiene, M., Gelzinis, A., & Kelertas, E. (2009). Advances in laryngeal imaging. *European Archives of Oto-Rhino-Laryngology, 266*(10), 1509–1520. https://doi.org/10.1007/s00405-009-1050-4

Woo, P. (2020). Objective measures of stroboscopy and high-speed video. *Advances in Oto-Rhino-Laryngology, 85,* 25–44. https://doi.org/10.1159/000456681

Woo, P., Casper, J., Colton, R., & Brewer, D. (1994). Aerodynamic and stroboscopic findings before and after microlaryngeal phonosurgery. *Journal of Voice, 8*(2), 186–194. https://doi.org/10.1016/S0892-1997(05)80311-X

Wu, X., Zhang, X., & Wang, X. (2009). Low bit-rate image compression via adaptive down-sampling and constrained least squares upconversion. *IEEE Transactions on Image Processing, 18*(3), 552–561. https://doi.org/10.1109/TIP.2008.2010638

Yan, Y., Chen, X., & Bless, D. (2006). Automatic tracing of vocal-fold motion from high-speed digital images. *IEEE Transactions on Biomedical Engineering, 53*(7), 1394–1400. https://doi.org/10.1109/TBME.2006.873751

Yan, Y., Damrose, E., & Bless, D. (2007). Functional analysis of voice using simultaneous high-speed imaging and acoustic recordings. *Journal of Voice, 21*(5), 604–616. https://doi.org/10.1016/j.jvoice.2006.05.011

Yiu, E., Worrall, L., Longland, J., & Mitchell, C. (2000). Analysing vocal quality of connected speech using Kay's computerized speech lab: A preliminary finding. *Clinical Linguistics & Phonetics, 14*(4), 295–305. https://doi.org/10.1080/02699200050023994

Yousef, A. M., Deliyski, D. D., Zacharias, S. R., de Alarcon, A., Orlikoff, R. F., & Naghibolhosseini, M. (2020). Spatial segmentation for laryngeal high-speed videoendoscopy in connected speech. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2020.10.017

Yousef, A. M., Deliyski, D. D., Zacharias, S. R. C., de Alarcon, A., Orlikoff, R. F., & Naghibolhosseini, M. (2021a). A hybrid machine-learning-based method for analytic representation of the vocal fold edges during connected speech. *Applied Sciences, 11*(3), 1179. https://doi.org/10.3390/app11031179

Yousef, A. M., Deliyski, D. D., Zacharias, S. R. C., de Alarcon, A., Orlikoff, R. F., & Naghibolhosseini, M. (2021b). Automated

detection and segmentation of glottal area using deep-learning neural networks in high-speed videoendoscopy during connected speech. In *14th International Conference Advances In Quantitative Laryngology, Voice And Speech Research (AQL)* (pp. 29–30). https://doi.org/10.1016/j.jvoice.2020.10.017

Yousef, A. M., Deliyski, D. D., Zacharias, S. R. C., & Naghibolhosseini, M. (2021). *Automated glottis obstruction detection in high-speed videoendoscopy during connected speech for patients with adductor spasmodic dysphonia: A deep-learning scheme*. 50th Anniversary Symposium: Care of the Professional Voice, June 2–6, 2021.

Yousef, A. M., Deliyski, D. D., Zacharias, S. R., & Naghibolhosseini, M. (2022). 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*. https://doi.org/10.1016/j.jvoice.2022.01.028

Zacharias, S. R., Myer, C. M., Meinzen-Derr, J., Kelchner, L., Deliyski, D. D., & de Alarcón, A. (2016). Comparison of videostroboscopy and high-speed videoendoscopy in evaluation of supraglottic phonation. *Annals of Otology, Rhinology & Laryngology, 125*(10), 829–837. https://doi.org/10.1177/0003489416656205

Zañartu, M., Mehta, D. D., Ho, J. C., Wodicka, G. R., & Hillman, R. E. (2011). Observation and analysis of in vivo vocal fold tissue instabilities produced by nonlinear source-filter coupling: A case study. *The Journal of the Acoustical Society of America, 129*(1), 326–339. https://doi.org/10.1121/1.3514536

Zenas, S., Heinz, N., Levesque, F., Deliyski, D. D., Zacharias, S. R. C., & Naghibolhosseini, M. (2021). *Vocal folds obstruction during high-speed videoendoscopy in connected speech*. National Conference on Undergraduate Research (NCUR), April 12–14, 2021.