Automatic quantification of the glottal area in the stroboscopic videos using deep neural network

A project funded by AIISH Research Fund (2018-2019)

Sanction No.: SH/CDN/ARF-SLP-GK/2018-19 dated 30.10.18

Total Fund: Rs. 9,86,000/-

Project Report

Principal Investigator

Dr. Gopikishore P

Assistant Professor

Department of SLP

All India Institute of Speech and Hearing

Mysuru-06

Principal Investigator

Dr. Prasanta Kumar Ghosh

Associate Professor

Department of Electrical Engineering

Indian Institute of Science (IISc)-

Bangalore

Principal Investigator

Rahul Krishnamurthy

Assistant Professor

Department of ASLP

Kasturba Medical College,

Mangalore

Co - Investigator

Dr. Prakash T K

Associate Professor

Department of ENT

All India Institute of Speech and Hearing

Mysuru-06

Co - Investigator

Dr. Suja Sreedharan

Professor

Department of ENT

Kasturba Medical College,

Mangalore

Research Officer
Ms. Rashmi Singh
All India Institute of Speech and Hearing
Manasagangothri, Mysuru 570006

Research Officer
Ms. Divya Degala
Indian Institute of Science (IISc)
Bangalore

Acknowledgements

Our sincere gratitude to Director, Prof. M. Pushpavathi, All India Institute of Speech and Hearing, Mysuru, for sanctioning and providing all the support in completion of the project. We extend our gratefulness to all the participants, staff of Departments of SLP and DCS who helped in one way or the other in completion of this project report.

Dr. Gopikishore P
Principal Investigator

Dr. Prasanta Kumar Ghosh
Principal Investigator

Rahul Krishnamurthy
Principal Investigator

Dr. Prakash T K
Co Investigator

Dr. Suja Sreedharan
Co Investigator

Ms. Rashmi Singh
Research Officer

Ms. Divya Degala
Research Officer

Table of Contents

LIST OF TABLES

List of Figures

Chapter I – Introduction

Laryngeal videostroboscopy (LVS) is widely used in clinical practice for its ability to efficiently capture many salient vocal fold vibratory characteristics (Mehta, & Hillman, 2012). Even with emerging technologies such as high speed videoendoscopy (HSV), videokymygraphy (VKG); LVS maintains its predominant clinical role in laryngeal imaging due to its cost effectiveness, and ease of use. Lately, LVS is being widely used to monitor the changes in vocal fold status as response to treatment, which in turn results in generation of large amounts of uninterpreted imaging data. The current clinical practice involves the use of subjective (visual perceptual) observation methods, and standardized rating systems (Bless, Hirano, & Feder, 1987; Poburka, 1999; Stemple, Gerdemann, & Kelchner, 1998) to judge/assess LVS recordings. The use of subjective methods to judge LVS recordings becomes challenging and time consuming, while handling large amounts of imaging data. Further, the validity of such perceptual rating systems has been questioned due to reports of poor reliability (Nawka, & Konerding, 2012).

The advent of deep learning methods, and robust image processing strategies has made it possible to automate several image analyses tasks, and one such application is the quantification and automation of LVS image analysis. By using robust and automatic methods it is possible to reduce the workload on clinicians and provide them with more objective information than currently available in the clinical routine. However, such a task is challenging, as quantification and automation methods require accurate identification (localization) and segmentation of glottis from LVS recordings. Also, the localization and segmentation of glottis serve as essential first step towards developing methods to quantify several vocal fold vibration parameters, such as, periodicity and amplitude, mucosal wave, glottal closure, and symmetry of vibration, etc.

In literature, different image processing strategies for glottis segmentation have been reported. These strategies (techniques) are mainly categorized into two, namely, the semi-automatic, and automatic approaches. In semi-automatic methods, there is a certain level of clinician (user) involvement to achieve the desired segmentation. It usually involves a clinician (user) selecting a set of images (usually the ones with maximum and/or minimum glottal opening) from a video sequence, which act as representative frames. Following the frame selection, multiple seed points (vocal fold landmarks such as, the anterior and posterior commissure) are identified and selected to which certain techniques are applied to compute the glottis segmentation. On the other hand, fully automatic methods function without any clinician (user) involvement. Algorithms designed for fully automatic glottal segmentation are very few, and only some are specifically designed for LVS image analysis. A fully automatic glottis segmentation followed by quantification of glottal area would reduce the clinician workload, and provide objective information, which can be used for accurate diagnosis and prognostic monitoring.

Aim of the study

In the current project, we aimed to investigate the performance of two neural networks (NN), the deep neural network (DNN), and the U – Net based architecture in automatic localization and segmentation of glottis from LVS images.

Objectives of the study

1. To identify the best performing NN architecture by comparing the performance of DDN and the U – Net based architecture in automatic localization and segmentation of glottis from LVS recordings.
2. To investigate the efficacy of best performing NN architecture identified in objective 1 in automatic glottal localization and segmentation in some pathological (voice) conditions.

Chapter II – Method

All the stroboscopic video recordings used in this study were recorded as a part of the voice evaluations at two centers. It consists of stroboscopic recordings from patients visiting the All India Institute of Speech and Hearing, Mysore, and the Department of Audiology and SLP, Kasturba Medical College, Mangalore, India. A standard recording protocol was used at both the centers.

Data set

LVS recordings of 55 patients (one video per patient having multiple phonations) with vocal nodules *(N = 34),* Cysts *(N = 10),* and Polyps *(N = 11)* were used in this study. Each video recording had a 3 – 15 phonations, and the duration of a video varied from 11 – 84 seconds with an average duration of 44 (± 20) seconds. Each video was chosen to ensure that it consisted audible recording events with adequate view of the laryngeal inlet and glottis. These videos (resolution = 720× 576; frame rate = 25 frames per second) were spliced into images using a custom software. The data set used in this study consisted of 24970 individual images from 55 LVS recordings.

Ground truth segmentation

Supervised training and evaluation of different neural architectures require corresponding reference segmentations serving as Ground Truth (GT). For this purpose, a graphical user interface was developed using MATLAB to manually mark the boundaries of the glottis region. Three SLPs (annotators*: a1, a2, a3*) experienced in analysing stroboscopic recordings, manually annotated 2723 glottal images spliced from LVS recordings. These manual annotations served as reference (ground truth) to train, and evaluate neural architectures reported in the current study.

Neural network architecture

In the present study, we carried out experiments to investigate the performance of two neural network architectures, namely, 1) the Deep Neural Network (DNN), and 2) the U – Net based architecture in automatic localization and segmentation of glottis from LVS images. The technical details of the two neural network architectures have been purposefully omitted to declutter the technical intricacies for the readers. However, a brief overview of both the architectures in non-technical terms has been provided below.

Deep Neural Network (DNN) based architecture

The DNN based architecture (Figure 1) used in the present study was developed by Rao et al (2018) and has two main steps.

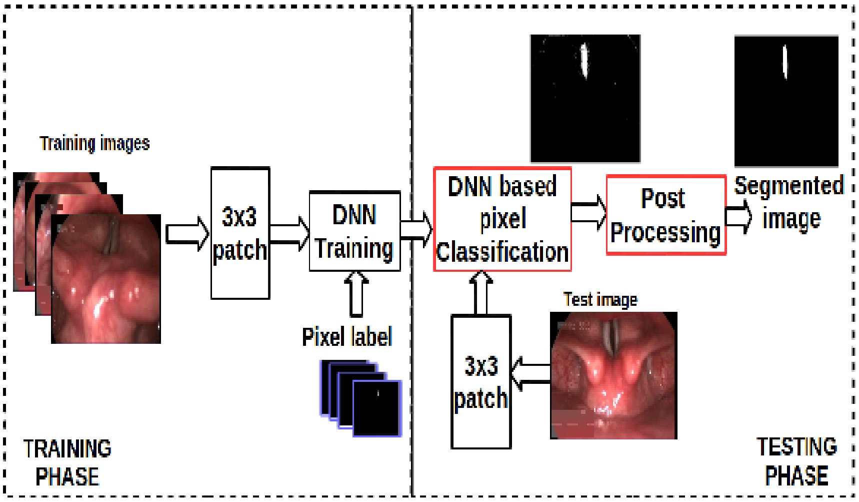

Figure 1: Block diagram of the proposed DNN approach

In the first step, each pixel in the image was classified to predict whether it belongs inside or outside the glottis region. In the second step, these pixels were clustered, classified as inside glottis regions, and filtered based on eccentricity and its orientation to find the final glottis segment. A much-detailed technical explanation on the development of this DNN based neural network architecture could be found elsewhere (See: Rao et al, 2018).

U – Net with and without Segnet based architecture

The U – Net based architecture used in the present study was developed by Degala et al., (2020). Even in this architecture, the glottis segmentation was posed as a classification problem and was approached using the following two steps. The first step was the ***Glottis Detection Network (GDN)***, which was used to detect the frames with glottis, when a sequence of frames was given. The architecture of the glottis detection network (GDN) is shown below in Figure - 2.

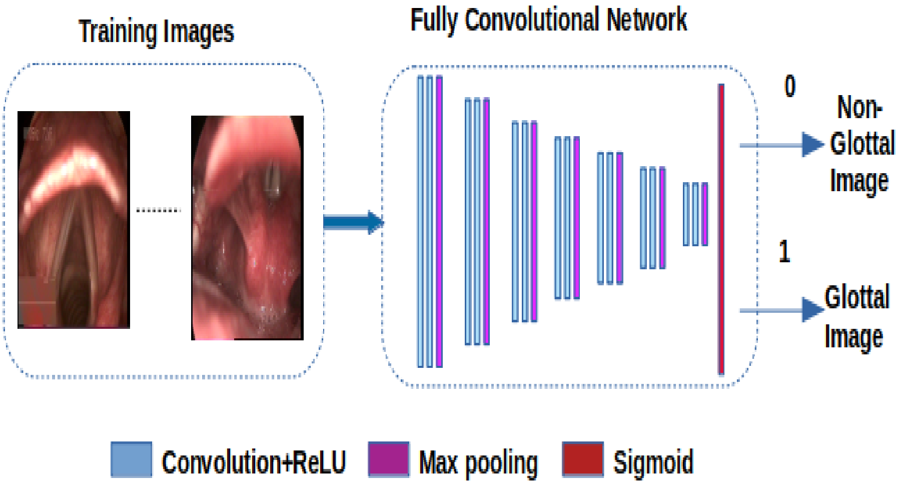

Figure 2: Block diagram summarizing the architecture of glottis detection network (GDN)

The second step involved the ***Glottis Segmentation Network (GSN),*** which was used to segment the glottis from those frames selected from step 1. The architecture of the segmentation network is shown in Figure 3. A much-detailed technical explanation on the development of this U – Net based neural network architecture could be found elsewhere (See: Degala et al., 2020).

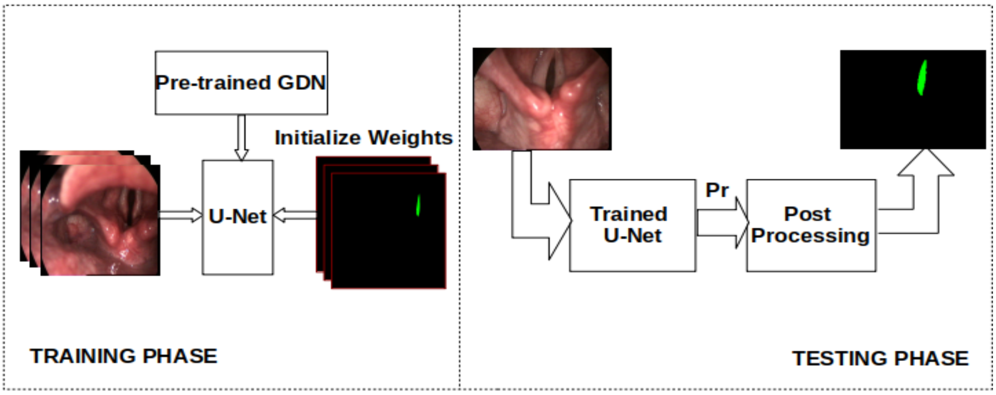

Figure 3: Block diagram of the proposed U – Net based glottis segmentation network (GSN)

Training, testing, and evaluation of neural network architectures

Training and testing

From 55 LVS recordings, a set of 24970 frames were used for training the neural network architectures for glottis detection. For glottis segmentation, a subset of 2723 images [1696 (nodules), 477 (cysts), and 550 (polyps)] were randomly extracted and categorized into four folds. Out of these four folds, two folds were used for the initial training, one-fold for validation and, the remaining one for testing, in a round-robin fashion to form a four-fold cross validation setup. For each type, the fold structure is described in table 1 below.

Condition (*N*)	Fold	Number of subjects	Number of images
Nodules (*N=34*)	Fold 1	9	450
	Fold 2	8	400
	Fold 3	8	400
	Fold 4	9	446
Cysts (*N=10*)	Fold 1	3	150
	Fold 2	2	100
	Fold 3	2	100

	Fold 4	3	127
Polyps (*N=11*)	Fold 1	3	150
	Fold 2	3	151
	Fold 3	2	100
	Fold 4	3	149

Table 1: *Showing four fold cross validation setup used for training, and testing the neural network architectures.*

Identifying the best performing neural network architecture

The best performing neural network architecture was identified based on the following evaluation metrics, which quantify the overall quality and precision of the obtained segmentation results. 1) ***Localization accuracy (L%):*** was calculated by the percentage of the test glottal images where the centroid of predicted segment falls inside ground truth glottis boundary. 2) ***Dice score (D):*** originally introduced by Dice (1945) has been widely used to measure the segmentation of quality for medical images (Crum, Camara, & Hill, 2006), and was used to evaluate the segmentation quality of the architectures in the present study.

Chapter III – Review of Literature

A primary challenge with stroboscopy is that it yields only uninterpreted images. However, for clinical diagnostics and, even more so, for clinical research, adequately quantified parameters with a clear meaning are required. To overcome this several authors (Bless, Hirano, & Feder, 1987; Poburka, 1999; Stemple, Gerdemann, & Kelchner, 1998) have developed subjective rating methods to quantify the stroboscopic images. These efforts have not been well received by both the practicing as well as the research community. Further, in day to day clinical practice stroboscopy is being widely used to monitor the vocal fold status as an effect of treatment procedures being used. This requires a much more quantitative measure of the changes that are induced as a result of ongoing treatment which has necessitated clinician – researchers to adopt newer methods of quantifying these changes. There are a fewer attempts made towards quantifying stroboscopic images using techniques of visual image processing (Osma et al., 2008) which have used machine learning to make better clinical decisions, but these studies are dwarfed by several inherent methodological concerns reviewed below. Gloger, Lehnert, Schrade and Volzke (2015) attempted to use machine learning to automatically segment glottal images from endoscopic recording of phonation.

Marendic, Galatsanos, Bless (2001) proposed a new active contour algorithm working based on *minimization of the energy* which was later inculcated in vocal fold tracking from the videostroboscopic videos, in a slower rate. The proposed algorithm introduces two new terms in internal energy function and the external energy (energy forced by the structure itself) manipulations were also done to improve the accuracy of tracing the Vocal folds. The stretching energy was familiarized to compensate for uneven distribution of external and internal energies down the glottal opening, and for channelizing the snaxels to the regions having high external or internal energies. They also verified the efficiency of this algorithm by tracing 39 complete Vocal fold cycles using snakes. They found that the stretching energy used enhanced the tracing capability of snakes. The modified energy terms and the Canny edge detector they used reduced the snaxels holding onto incorrect edges mostly at the bottom during closing phase. They concluded that the active contours with modified energies outperformed the classical approaches that have been used in the literature. Allin, Galeotti & Stetten (2004) assessed a system using the videos from videostroboscopy of three individuals wherein, an attempt was made to augment the existing segmentation methods by focussing on the external energy of the snake based segmentation rather than on the internal image energy. This was done by component of the system using colour transformation which discriminates the pixels of the trachea from those of the vocal fold pixels which fabricates a crude segmentation of vocal fold boundaries. The second component of the system discriminated between the surface of vocal folds and pixels around them. This assessment system extracted contours from manually completed points identified as vocal fold edges. This showed that the segmentation can be simply obtained by focusing on building of image energy. And this segmentation can be used in cases of gross laryngeal movement disorders.

Mendez, Garcia, Ruiz, Iturricha (2008) proposed a method to obtain the glottal space segmentation without the user interfering. The authors used Gabor filter segmentation and the 261 images obtained from 6 videos including the normal and pathological cases (vocal polyps and vocal nodules) were used for examining the developed algorithm and the Glottal area segmentation was done and the Glottal area waveform was drawn in which they found that this algorithm had some incorrect segmentations though, it still would serve as a good objective measure to diagnose vocal pathologies. Videostroboscopy and Videokymography

(VKG) with the new 2-point laser projection device was used to obtain dynamic images which used simulated vocal folds phonating on a laboratory bench for verifying if complete width of glottis could be measured and the stability, vibratory periodicity and the reliability. The results revealed that the VKG with the even with its extensive time consumption, showed an average variation of 3.10% in maximum glottal width measurements compared to stroboscopic data. It was concluded that the simulated-cyclic illustration of vocal fold vibrations obtained with videostroboscopy was equivalent to the full-cycle information obtained with VKG to within 7%, for stability, periodicity. So, the stroboscope which is more regularly available and less expensive can be used for Vocal measurements.

Wang, Yu, Zhang & Xu (2010) used high-speed digital imaging, voice signal processing and quantified the voice productions by extracting the periodic and aperiodic vocal fold vibrations and tried to provide an automatic image processing technique. They found that the images of vocal fold vibrations with lower resolution were effectively extracted. This study provided a new way of combining both digital imaging and voice signal processing. The new algorithm was said made the diagnosis more accessible by analyzing thousands of high speed images. Feng Jeffrey Kuo, Hsiang Chu, Chun Wang, Yu Lai, Lin Chu, ShingLeu & Won Wang (2013) intended to design an automatic vocal cord image selection system. The study used a rigid videolaryngoscopy for recording (RLS 9100B, Kay elemetrics, NJ). Initial part of the study consisted of the automatic selection of the vocal cord opening to its largest extent by applying the color space conversion and image processing, it included enrichment of the glottis region, removal of object boundaries, conversion of image obtained into a binary image, area filling and region deletion, intending to emphasize the features of the vocal area image. Next part of the study included selecting the automatic image of the vocal cord opening to the largest extent wherein, a screening happens in which the three images with larger glottal gap will automatically selected and are to be arranged in descending order. The final part of the study again used the image processing by calculating the statistics of the pixels of the obtained binary images and automatically 15 images with the smallest vocal cord closing were selected. This system automatised the image choice for vocal cord opening to the largest extent and vocal cord closing to the smallest extent. This study enhanced the technique of image selection and facilitated efficient diagnosis.

Pinheiro, Dajer, Hachiya, Montagnoli, & Tsuji (2014) developed a method for computing the glottal area using the images obtained by high-speed videolaryngoscopy. Eight clinically normal participants were asked to sustain vowel /a/ when subjected to videolaryngoscopy. They found that the developed algorithm was advantageously able to process the low contrast images in-vivo. The use of such computational methods for segmentation of Glottal area and the edge of the vocal fold on the basis fixed, numerical criterion improve the rising by increasing the objectivity of the analysis which would help in reducing the practical issues of segmentation.

Kopczynski, Strumillo, and Bogusz (2015)'s objective was to apply computer image processing and analysis methods for quantification of vocal folds' phonatory movement using the frames obtained by Videostroboscopy of the larynx. Thirty individuals with no voice problems and 15 individuals diagnosed as having Vocal nodules' samples were considered for the study. During the phonatory task, the glottal area estimation/computation was done using Image pre-processing and image segmentation algorithms. For visualising the spatio-temporal aspect of vocal vibrations, the glottovibrograms were constructed. Few indices were introduced for quantifying which include the closed index and speed index. In the results, they found that the closed index could help in differentiating the individuals with no pathology from the individuals having vocal nodules. And the speed index had negative values for most of the individuals with pathology which indicates that the opening time of the vocal folds is slower than the closing time.

Kuo, Joseph Kuo, Hsiao, Lee, Lee & Ke (2016) developed a glottis appreciation laser projection marking module system for converting the physiological parameters of glottis and for providing scale conversion reference parameters for glottal imaging. The vital regions of the larynx were separated using image processing obtained by videostroboscopy. Image pre-processing procedure used the histogram equalization to increase the contrast of glottal image. Center weighted median filter helped in preserving the texture features and eliminating noise. Statistical threshold selection determines the threshold for image segmentation automatically and dilation and erosion, opening and closing for eliminating the noise post segmentation enhance the precision in interpretation. Region filling was done to fill-up the region formed by a group of adjacent points in the image with characteristic values. They also used image processing to automatically identify an image of vocal fold region in order to quantify information from the glottal image, such as glottal area, vocal fold perimeter, vocal fold length, glottal width, and vocal fold angle.

Chapter IV - Results

Ground truth segmentation

In the present study, three SLPs independently identified (annotated) glottis on a set of 2723 images. Dice score (D) was used to quantify the agreement of the glottis boundary annotation of three SLPs, a Dice score of 1 would indicate a perfect agreement between the annotations. We computed the DICE score across each pair of annotations for every image and a mean DICE score was computed by taking average across all pairs.

The first row of the image 1 shows target image (to be annotated) by the three SLPs. Second row to last row of image 1 shows the three images on which annotations were performed by the three SLPs. The mean DICE score for three illustrative examples in image 1(a) are given at end of respective columns. We observed that the annotations often vary across SLPs, particularly when the glottal opening is small as shown in the first and second columns of image 1. We also observed poor inter SLP agreement when there were variations in terms of illumination, glottis shape, and orientation of LVS recordings, as shown in image 2.

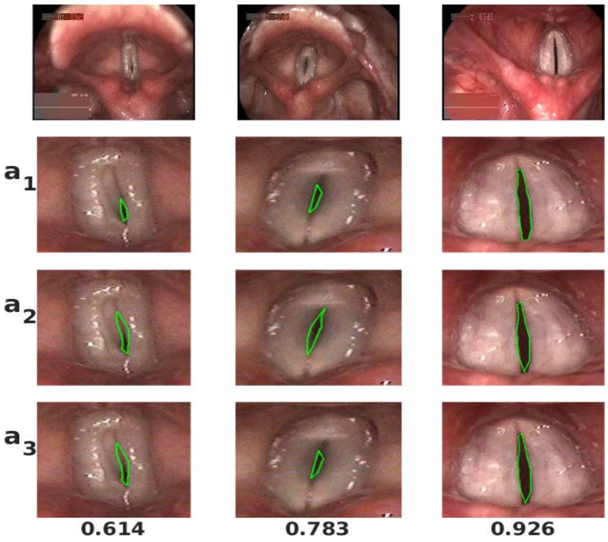

Image – 1: Illustration of three sample annotated images by three SLPs (*a1* , *a2, a3*) to show the inter SLP agreement.

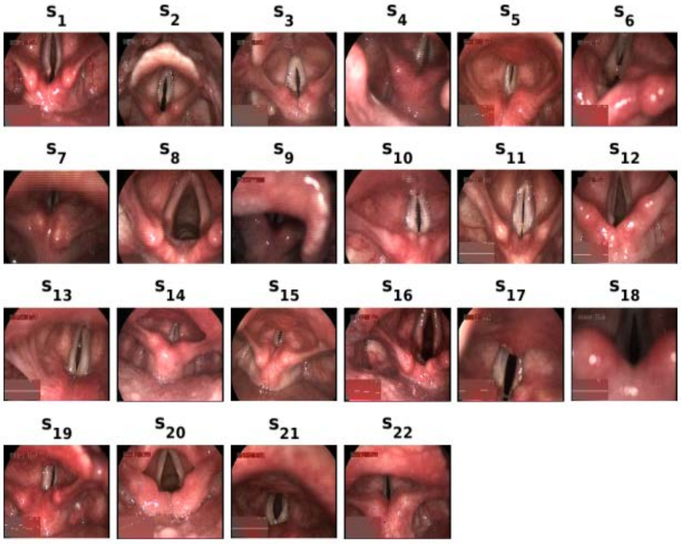

Image - 2: Sample images randomly selected from 22 subjects to show variations in terms of illumination, glottis shape, and orientation of LVS recordings.

Objective 1: Identifying the best performing NN architecture

Table 2 shows fold-wise localization accuracy (%) across three SLPs for DNN and U – Net based approaches. Table 3 shows corresponding fold-wise Dice score (D) calculated only on the correctly localized images. The three values in each cell shows the Localization accuracy (or Dice score) calculated with respect to corresponding SLPs' annotations (*a1, a2, and a3*). Average Dice score / Localization accuracy of all folds has been shown in the last row of Tables 2, and 3. Figure 4 shows mean Dice score, and localization accuracy of correctly localized images achieved by DNN and the U-Net based architecture across three SLPs' annotations (*a1*,*a2*,*a3*).

Table 2: Fold-wise Localization accuracy (%) of correctly localized images achieved by DNN and the U-Net approaches across three SLPs' annotations *(a1,a2,a3*).

	Localization accuracy					
	DNN			***U - Net***		
Fold	***SLP 1***	***SLP 2***	***SLP 3***	***SLP 1***	***SLP 2***	***SLP 3***
	(a1)	***(a2)***	***(a3)***	***(a1)***	***(a2)***	***(a3)***
Fold 1	89.9	78.7	98.5	90.6	91.2	90.6
Fold 2	38.0	69.0	40.0	61.4	63.6	62.3
Fold 3	72.2	68.6	72.8	85.9	87.1	65.3
Fold 4	39.9	76.2	41.4	93.9	91.2	91.7
Mean	60	73.1	63.2	83	83.3	77.5
(SD)	(25.36)	(5.01)	(27.99)	(14.73)	(13.25)	(15.85)

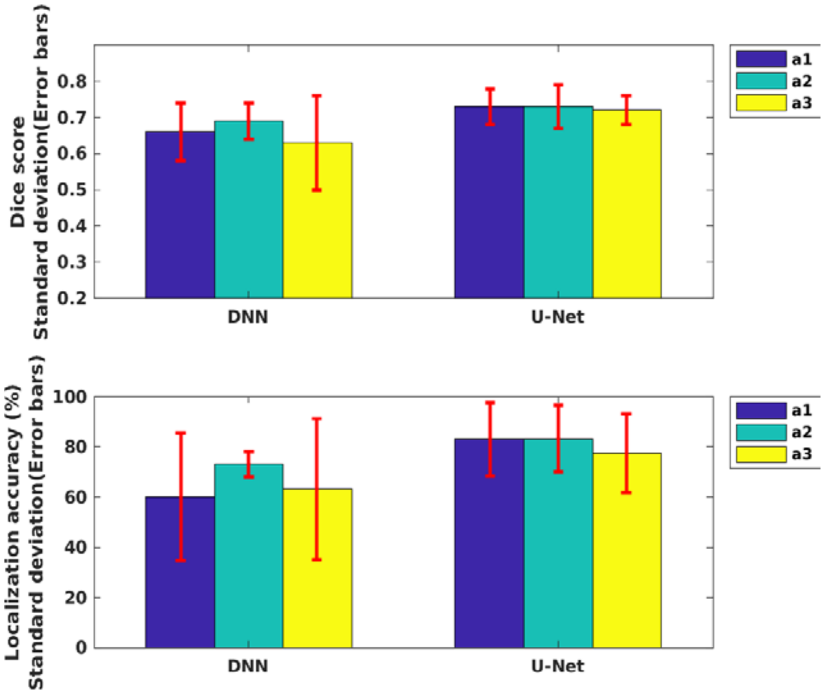

Figure - 4: Mean Dice score (SD as error bar), and Localization accuracy (SD as error bar) of correctly localized images achieved by DNN and the U-Net based architecture across three SLPs' annotations (a1,a2,a3).

As seen from the tables 2, and 3, the U- Net based architecture performs better than the DNN architecture. When a fold wise interpretation of results was carried out, we observed that both localization accuracy, and Dice score for fold – 1 was high, and may be due to the clear visibility of glottis in subjects, as illustrated in examples (*S1, S2, S3, S4, and S5*) as shown in image 2. For fold 2, we observed that the localization accuracy was low, but the corresponding Dice score was high. This difference may be due to the poor illumination in recordings as illustrated in examples S6, S7, and S9 of image 2, and the corresponding Dice score was high because of the skip connection used in the U – Net based architecture enables exact detection of glottis boundary. In conditions, where the supraglottic structures obstructed the view of glottal opening; we observed high localization accuracy but low Dice score, as shown in the results of fold 4.

Table 3: Fold-wise Dice score of correctly localized images achieved by DNN and the U-Net approaches across three SLPs' annotations *(a1*,*a2*,*a3*).

	Dice score					
Fold	DNN			U - Net		
	SLP 1 *(a1)*	SLP 2 *(a2)*	SLP 3 *(a3)*	SLP 1 *(a1)*	SLP 2 *(a2)*	SLP 3 *(a3)*
Fold 1	0.76	0.76	0.81	0.78	0.77	0.76
Fold 2	0.67	0.69	0.64	0.74	0.76	0.74
Fold 3	0.62	0.63	0.52	0.72	0.75	0.72
Fold 4	0.56	0.66	0.55	0.73	0.73	0.72
Mean (SD)	0.66 (0.08)	0.69 (0.05)	0.63 (0.13)	0.73 (0.05)	0.73 (0.06)	0.72 (0.04)

Objective 2: Performance of U – Net based architecture in localization of glottis in some pathological conditions

The data set used in experiment 2 consisted a fresh set of images (those not used in experiment 1) from three pathological conditions (vocal nodules: *N = 34*, Cysts: *N = 10*, Polyps: *N = 11*). The three values in each cell shows the Localization accuracy (or Dice score) calculated with respect to corresponding SLPs' annotations (*a1, a2, and a3*). Figure 5 shows mean Dice score, and localization accuracy of correctly localized images achieved by the U-Net based architecture across for nodules, polyps, and cysts, respectively. It can be observed from table 3, and figure 5 that both localization accuracy and Dice scores were poor.

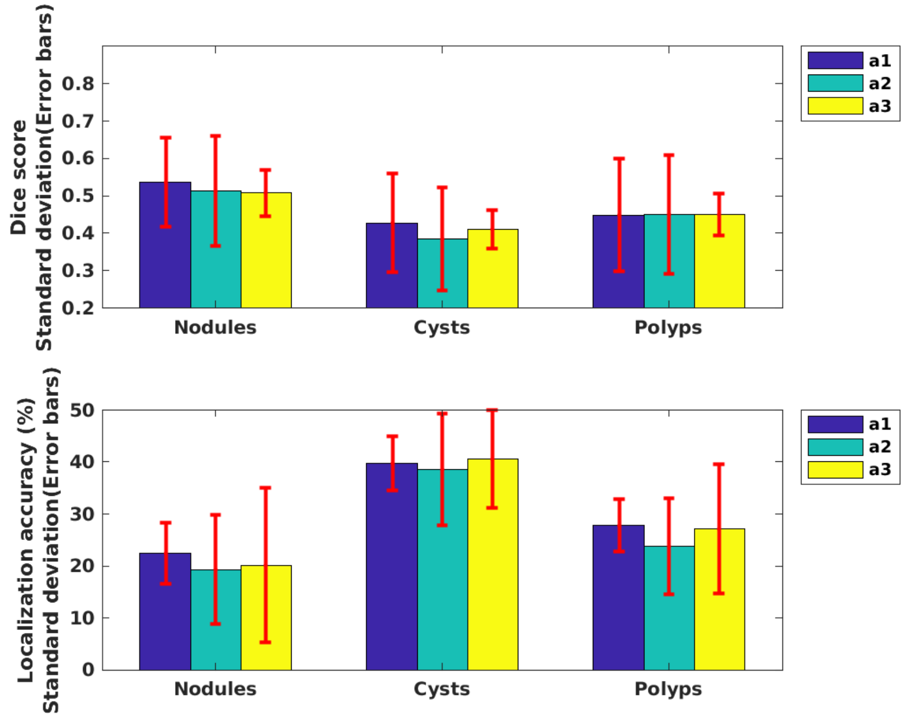

Figure - 5: Mean Dice score (SD as error bar), and Localization accuracy (SD as error bar) of correctly localized images achieved by the U-Net based architecture for nodules, cysts, and polyps, respectively.

Chapter V - Discussion

Image segmentation and localization play a key role in medical imaging applications. Accurate localization and segmentation of glottis is the essential first step towards developing of a fully automatic quantification solution to analysis of LVS recording. In the current paper, we carried out experiments to identify and evaluate the performance of two deep learning methods (neural networks) in automatic localization and segmentation of glottis from LVS recordings.

In our first experiment, we aimed at identifying the best performing neural network architecture by comparing two NNs, namely, the DDN and the U – Net based architecture in automatic localization and segmentation of glottis from LVS recordings. The results of our first experiment revealed that U – Net with and without Segnet based architecture performed better than the DNN based architecture The DNN based approach used in the current study relies on a 3x3 pixel neighborhood for glottis localization and segmentation, and the choice of 3x3 neighborhood context is ad-hoc. However, in the datasets used in the current study, the size and location of glottis keeps changing across images. In such conditions, a larger context may be required for efficient localization and segmentation. For a DNN based architecture, determining the context window adaptively depending on the size and location of the glottis becomes challenging, and thus resulting in relatively poor performance.

The superior performance of U-Net based architecture can be attributed to the fully convolutional nature of the architecture. The U-Net based architecture used in the current study is a type of fully convolutional network (FCN), which has been specially designed for the segmentation of biomedical images. It consists of a FCN consisting of convolution, relu activation followed by the max pooling. For glottis segmentation, we have used a U-Net consisting of a contracting path with two encoders and an expansive path with two decoders. U-Net with skip connection is an architecture where a concatenation layer is added after an up-sampling layer with corresponding cropped features from the encoder, which helps in combining high resolution features from the contracting path. A successive convolution layer can then learn to assemble a more precise segmentation based on this information, making the U-Net based architecture superior compared to DNN based architecture.

Our second experiment involved testing the performance of U-Net based architecture in localization and segmentation of glottis on an untrained data set. It was observed that the performance of U-Net based architecture was poor on untrained data set, and maybe due to the following reasons. 1) Limited number of images on which the training was performed, and 2) variations in LVS image frames in terms of illumination, glottis shape, and orientation of LVS recordings, as shown in image 2. It was also observed that for difficult frames (those with poor illumination, shape, and orientation) the inter SLP agreement was also poor. FCNs such as U-Net based architecture require training on a very large data set. Even though relatively large data set (24970 individual images) was used in the current study, it appears that a much larger data set may be necessary to achieve better performance. 3) The U-Net architecture used in the current study was trained using three different weight initialization schemes. Perhaps, the use of different loss functions can improve the performance of the algorithm.

Chapter V1 – Summary and Conclusion

Accurate localization and segmentation are an important first step towards developing a fully automatic quantification solution for the analysis of LVS recordings. The results of the current study are promising, the proposed methods allow a very stable, reliable, and high-quality fully automatic localization and segmentation of glottis from LVS recordings. Upon training, the proposed NNs can localize and segment glottis from LVS recordings without any clinician (user) intervention, which is an essential prerequisite for clinical use. Even though parallelization of segmentation process was not performed in the current study, such a task is possible and is being considered in future developments. Other prospective developments include, testing the performance of our NNs on augmented data, further optimization, and generalization of current NNs on larger data sets. Further improvements to our NN algorithms itself can facilitate better segmentation performance.

References

Mehta, D. D., & Hillman, R. E. (2012). Current role of stroboscopy in laryngeal imaging. *Current opinion in otolaryngology & head and neck surgery*, *20*(6), 429.

Bless., M, B., Hirano, M., & RJ, F. (1987). Videostroboscopic evaluation of the larynx. *Ear, Nose, and Throat Journal*, *66*, 289-296.

Poburka, B. J. (1999). A new stroboscopy rating form. *Journal of Voice*, *13*(3), 403-413.

Poburka, B. J., Patel, R. R., & Bless, D. M. (2016). Voice-vibratory assessment with laryngeal imaging (VALI) form: Reliability of rating stroboscopy and high-speed videoendoscopy. *Journal of Voice*, *31*(4), 513-e1.

Stemple, J. C., Glaze, L. E., & Klaben, B. (2010). *Clinical voice pathology: Theory and management*. Plural Publishing.

Nawka, T., & Konerding, U. (2012). The interrater reliability of stroboscopy evaluations. *Journal of Voice*, *26*(6), 812-e1.

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.

Gloger, B. Lehnert, A. Schrade, and H. Volzke, "Fully automated glottis segmentation in endoscopic videos using local color and shape features of glottal regions," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 3, pp. 795–806, 2015.

Cerrolaza, V. Osma-Ruiz, N. Saenz-Lech ´ on, A. Villanueva, ´J. M. Gutierrez Arriola, J. I. Godino-Llorente, and R. Cabeza, ´"Fully-automatic glottis segmentation with active shape models." in *MAVEBA*, 2011, pp. 35–38

Marendic, B., Galatsanos, N., & Bless, D. (2001). New active contour algorithm for tracking vibrating vocal folds. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)* (Vol. 1, pp. 397-400). IEEE.

Mendez, A., Garcia, B., Ruiz, I., & Iturricha, I. (2008, December). Glottal area segmentation without initialization using gabor filters. In *2008 IEEE International Symposium on Signal Processing and Information Technology* (pp. 18-22). IEEE.

Allin, S., Galeotti, J., Stetten, G., & Dailey, S. H. (2004, April). Enhanced snake based segmentation of vocal folds. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)* (pp. 812-815). IEEE.

Pinheiro, A. P., Dajer, M. E., Hachiya, A., Montagnoli, A. N., & Tsuji, D. (2014). Graphical evaluation of vocal fold vibratory patterns by high-speed videolaryngoscopy. *Journal of Voice*, *28*(1), 106-111.Poburka, B. J. (1999). A new stroboscopy rating form. *Journal of Voice*, *13*(3), 403-413.

Kopczynski, B., Strumiłło, P., & Niebudek-Bogusz, E. (2015, September). Computer based quantification of normal and pathological vocal folds phonatory processes from laryngovideostroboscopy. In *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 269-274). IEEE.

Kuo, C. F. J., Kuo, J., Hsiao, S. W., Lee, C. L., Lee, J. C., & Ke, B. H. (2017). Automatic and quantitative measurement of laryngeal video stroboscopic images. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, *231*(1), 48-57

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.