Automatic Glottis Detection and Segmentation in Stroboscopic videos using Convolutional Networks

Divya Degala[1], *Achuth Rao M V*[1], *Rahul Krishnamurthy*[2], *Pebbili Gopikishore*[3], *Veeramani Priyadharshini*[3], *Prakash T K*[3], *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering, Indian Institute of Science, Bangalore 560012, India
[2]Kasturba medical college, Manipal Academy for Higher Education, Mangalore 575001, India
[3]All India Institute of Speech and Hearing, Mysuru, 570006, India

divya.degala517@gmail.com, achuthr@iisc.ac.in, prasantg@iisc.ac.in

Abstract

Laryngeal videostroboscopy is widely used for the analysis of glottal vibration patterns. This analysis plays a crucial role in the diagnosis of voice disorders. It is essential to study these patterns using automatic glottis segmentation methods to avoid subjectiveness in diagnosis. Glottis detection is an essential step before glottis segmentation. This paper considers the problem of automatic glottis segmentation using U-Net based deep convolutional networks. For accurate glottis detection, we train a fully convolutional network with a large amount of glottal and non-glottal images. In glottis segmentation, we consider U-Net with three different weight initialization schemes: 1) Random weight Initialization (RI), 2) Detection Network weight Initialization (DNI) and 3) Detection Network encoder frozen weight Initialization (DNIFr), using two different architectures: 1) U-Net without skip connection (UWSC) 2) U-Net with skip connection (USC). Experiments with 22 subjects' data reveal that the performance of glottis segmentation network can be increased by initializing its weights using those of the glottis detection network. Among all schemes, when DNI is used, the USC yields an average localization accuracy of 81.3% and a Dice score of 0.73, which are better than those from the baseline approach by 15.87% and 0.07 (absolute), respectively.
Index Terms: Glottis segmentation, stroboscopy, U-Net

1. Introduction

During voice production, the vocal folds play an essential role by regulating airflow from lungs, through its quasiperiodic vibrations [1]. Space between the two vocal folds is called the glottis. Sulcus Vocalis (SV) is a particular type of vocal fold condition, in which a groove is formed in the vocal fold, leading to an incomplete glottic closure during phonation, which is called the glottic chink. At present, examination of vocal fold motion and classification of severity of glottic chink in SV are done by clinical experts subjectively by directly observing the endoscopic video. As the vocal fold structures vary anatomically from an individual to another, the reliability of subjective analysis in making an accurate clinical decision of the pathology is questionable. The clinicians strongly believe that stroboscopy is the most crucial session before voice assessment and it would be the gold standard technique for the next 10 to 20 years [2]. However, the video stroboscopy has some challenges while recording in a real clinical environment:1) images can be affected by uneven illumination 2) images may be taken at the wrong instants due to uncontrolled movements by the patient or strong gag reflex by the patients' tongue base 3) camera movements which cause glottis to appear in different angles 4) laryngeal image size may vary when the distance between endoscope and vocal folds changes. Fig. 1 shows sample stroboscopic video images from several subjects. From the figure it is clear that the shape of the glottis varies with respect to the camera position. Few images have been obstructed by supraglottic structures, which makes the glottis invisible as seen in the examples shown in Fig. 2. Hence, a quantified glottis image could assist the Speech-Language Pathologists (SLPs) in diagnosing voice disorders in a more objective manner. In the literature, there are only few algorithms that have been implemented for fully automatic glottis segmentation [3]-[10] and less work is done using stroboscopic videos [3] [5] [7]. Rao et al. [11] used a deep neural network (DNN) based model, in which, they posed the glottis segmentation as a pixel wise classification problem and considered the RGB values from a 3×3 neighborhood of a pixel as a feature vector, to classify whether the pixel belongs to inside or outside the glottis region. Lin et al. [12] had used a bounding box for the detection of region of interest followed by a fully convolutional network (FCN) for segmentation. Gloger et al. [3] proposed a framework for automatic glottis segmentation. They used a Fourier descriptor with prior glottal shape knowledge from a large dataset consisting of different glottal shapes to

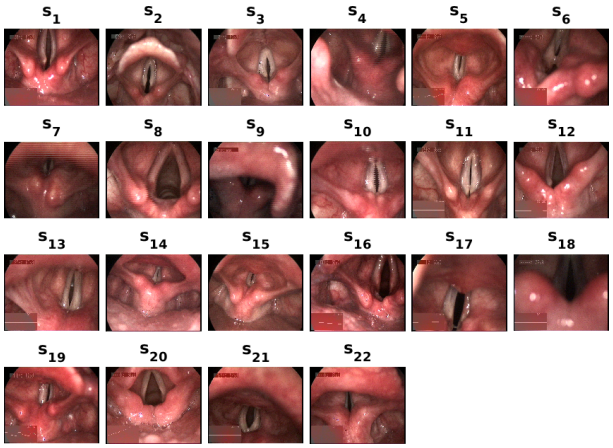


Figure 1: *Sample images from 22 subjects describe the variations in terms of illumination, glottis shape and orientation in video stroboscopy recording.*

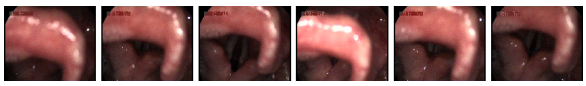


Figure 2: *Example images to show the supraglottic structures covering the glottis region.*

localize the glottis and levelset segmentation followed by Probability Image Generation for glottal shape tracking. Cerrolaza et al. [5] had used threshold based region growing approach for glottis localization and the active shape model (ASM) for glottis boundary detection. However, it failed on the testing data when the glottal shape is much different from those in the training set. Osma et al. [7] used a watershed transform followed by a region merging and linear prediction for glottis segmentation.

Image segmentation and localization play a key role in medical imaging applications. Recently, many deep learning methods have achieved a great success in image segmentation. Fully Convolutional Networks (FCNs) is one of those methods that have shown good results in segmentation [13]-[16]. In this work, we address the problem of glottis segmentation using two types of FCNs called U-Net [13] and Segnet [15]. U-Net [13] is especially designed for the segmentation of biomedical images. It has an encoder-decoder architecture with skip connections. U-Net [13] uses the skip connections to connect each pair of encoder layer and the corresponding decoder layer, which passes the spatial information directly to the much deeper layers of the network and, in turn, gives a more accurate segmentation. Similar to U-Net [13], Segnet [15] also has encoder-decoder architecture without skip connection.

In this study, we experiment the U-Net with and without skip connection (Segnet), using three different weight initialization schemes followed by fine-tuning to achieve good results for glottis segmentation. For glottis detection, we train a fully convolutional network, which acts as a binary classifier to detect the presence or absence of glottis in a given image. The network is trained using 24970 frames, which are extracted from 22 stroboscopic videos, the sample frames of which are shown in Fig. 1. The trained detection network weights are used for better initialization in U-Net to get more precise segmented output compared to the random initialization. For glottis segmentation, we train U-Net with 921 images, which are extracted from 18 subjects (among 22 subjects mentioned earlier), where glottis boundaries are marked by 3 SLPs. On the other hand, for glottis detection network, all images are grouped into two sets, one having glottis and the other where glottis is absent. We use a 4-fold cross validation setup for both the detection as well as the segmentation tasks. We perform experiments to compare the results of U-Net with the three different initialization schemes. It is shown that the performance of the glottis segmentation network is increased when appropriate weight initialization scheme is used although it is fine-tuned with less labeled data. Here the proposed method achieves an average localization accuracy of 81.3% and a Dice score [17] of 0.73, which is significantly better than the DNN based baseline approach for glottis segmentation by 15.87% and 0.07 respectively.

2. Dataset

For this work, 22 stroboscopic videos were recorded from 22 subjects (14 males and 8 females) by an Otorhinolaryngologist. For the recording, Xion Endostrob E with 70 degree rigid scope and Digital Video Archive Software (DiVAS) version 2.5 were used as the hardware and software setup, respectively. During the process of recording, each subject was asked to sit on a stool facing the examiner, and Xylocaine solution was sprayed to the participant's oropharyngeal region to avoid gag reflex. The participant was asked to extend the tongue out and instructed to phonate the vowel /i/ (as in word 'bit') for 4-5 seconds. The subject was asked to repeat the phonations until the examiner could get a clear picture of the glottis.

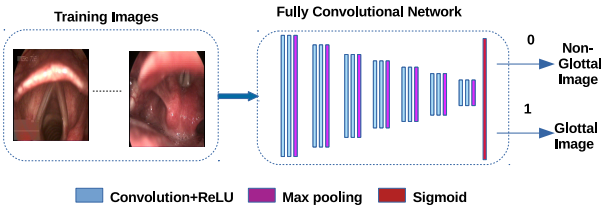
In this work, we consider 22 stroboscopic videos (one video from each of the 22 patients with SV) for experiments, denoted by $S_i, i = 1, ...22$. All videos have recordings of multiple phonations and all are recorded in an avi format with a resolution of 720×576 and a frame rate of 25 frames/sec. Duration of a video varies from 11s to 84s. The average duration of a video is 44s. The subjects who have vocal folds mass lesions or any other voice disorders (apart from SV) are excluded in this work. We train glottis detection network with a set of 24970 images, which are extracted from 22 stroboscopic videos and labeling is done by manually segregating the glottal and non-glottal images into two groups. Out of 24970 images only 921 randomly selected images from 18 subjects are used for glottis segmentation in this work. A graphical user interface is developed using MATLAB, to mark the boundaries of the glottis region. Each of the 921 images has been annotated by three SLPs

3. Proposed Approach

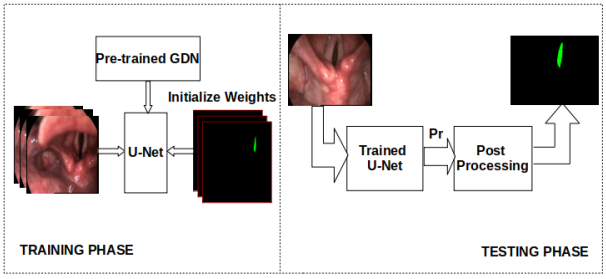
The proposed approach mainly involved two steps: 1) Given the sequence of frames, detect the frames with glottis images. 2) Given the glottis image, segment the glottis.

3.1. Glottis Detection Network

The network architecture of the glottis detection network (GDN) is shown in Fig. 3(a). The glottis can be present in any part of the images and can vary drastically from training set to test set. To make the predictions robust to translation we use fully convolutional network based classifier. A fully convolutional network consists of convolution, batch normalization, relu activation followed by the max pooling of (2, 2). Finally we have used a sigmoid layer to classify if the image contains glottis or not. We give $224 \times 224 \times 3$ RGB image as an input to the GDN which classifies the presence or absence of glottis in each frame of the video. We consider images with clear visible glottis as one class (Label = 1) and remaining all images as the other class (Label = 0). We have experimented GDN with and without batch-normalization layers to achieve a good classification accuracy.



(a)



(b)

Figure 3: *(a) Blocks summarizing the architecture of glottis detection network (GDN) (b) U-Net based glottis segmentation network.*

3.2. Glottis Segmentation Network

The architecture of the segmentation network is shown in Fig. 3(b). Glottis segmentation has been performed on the images selected by GDN with the glottis. We use a variant of U-Net architecture for the segmentation: 1) U-Net with skip connection (USC) having 702,977 parameters 2) U-Net without skip connection (UWSC) having 666,113 parameters. U-Net [13] has been shown to work well for various bio-medical image segmentation problems. But in the current work, the main limitation of the U-Net is that the number of annotated images (921) are significantly less compared to the number of parameters of the U-net (666k). Hence, we consider three different kind of initialization schemes for U-net (it's variant) so that segmentation can be done accurately using a few samples of annotated images: 1) Random weight Initialization (RI): U-Net models are trained by initializing weights randomly. 2) Detection Network weight Initialization (DNI): we use GDN weights to initialize the U-Net encoder weights, which has an architecture similar to that of the first six layers of GDN, and fine-tuned both encoder and decoder of the U-Net with the labeled data. 3) Detection Network encoder frozen weight Initialization (DNIFr): weight initialization is done in a manner similar to that of the DNI except that encoder network weights are not updated during fine tuning. The segmentation problem is converted into pixel wise classification problem. In our experiments, we use a mini U-Net with the same architecture as in [13], differing only in the way of weight initialization. Each image and corresponding annotation image of size 720×576 are resized to 224×224. We give this image as an input to the U-Net. The skip connection in U-Net is used for glottis localization by preserving features that are learned in the contracting path. The final layer of the U-Net is used to classify whether each pixel of the image belongs to inside or outside the glottis. The network is trained with a weighted binary cross entropy loss function. As the number of pixels inside the glottis region is less compared to the number of pixels outside the glottis region, the weight ratio used for objective function is 200:1 from inside to outside the glottis region. We use adam [19] optimizer with default parameters. The best model is saved using early stopping criteria by monitoring validation loss. U-Net is implemented by using deep learning libraries called keras [20] and theano [21].

U-Net consists of a contracting path with two encoders and an expansive path with two decoders. The contracting path follows a typical architecture of a convolutional network. The first encoder consists of 2 convolutional layers each having 64 filters with a filter size of 3×3, followed by a rectified linear unit (ReLU). This is followed by a 2×2 max pooling layer with stride 2 for down-sampling. The second encoder has 2 convolutional layers each having 128 filters with a filter size of 3×3, followed by a ReLU activation function. This is followed by a 2×2 max pooling layer with stride 2. The first decoder in the expansion path consists of an up-sampling layer followed by 2 convolutional layers, each having 128 filters with a filter size of 3×3, followed by a ReLU activation. In the second decoder, a concatenation layer is added after up-sampling layer with the correspondingly cropped feature map from the first encoder, which is called as "Unet with skip connection (USC)". Its absence is termed as "Unet without skip connection (UWSC)". This is followed by two convolution layers, each having 64 filters with 3×3 filter size, followed by a rectified linear unit (ReLU). At the end, a convolution layer with a filter size of 1×1, followed by sigmoid activation is used to classify whether each pixel belongs to inside or outside the glottis region.

3.2.1. Post Processing

From the output (Pr) of U-Net, we obtain an image where each pixel corresponds to a probability of being inside or outside the glottis region. In order to construct a binary image, we choose a threshold of $\max(\text{Pr}) - 0.05$. Here, label '1' is assigned for pixels inside the glottis region and '0' for outside the glottis region.

4. Experiments and Results

4.1. Experimental Setup

The whole dataset is divided into four folds namely, fold1: $[S_1, S_2, S_3, S_4, S_5]$, fold2: $[S_6, S_7, S_8, S_9, S_{10}, S_{11}]$, fold3: $[S_{12}, S_{13}, S_{14}, S_{15}, S_{16}, S_{17}]$, fold4: $[S_{18}, S_{19}, S_{20}, S_{21}, S_{22}]$. Among the four folds, 2 folds are used for training, 1 fold for validation and, the remaining 1 for testing, in a round-robin fashion to form a 4-fold cross validation setup. U-Net is trained with annotations from one of the SLPs and evaluated on all the three annotations. We consider the work done by Rao et al. [11] as a baseline. Both baseline and the proposed method are evaluated with the same dataset and fold structure.

4.2. Evaluation metrics

We have used Classification accuracy, Localization accuracy and Dice score [17] as an evaluation metrics in this work.

Classification accuracy is used to measure the performance of GDN. It is the ratio of number of correctly classified glottal or non-glottal images to the total number of images.

Localization accuracy is calculated by the percentage of the test glottal images where the centroid of predicted segment falls inside ground truth glottis boundary. Dice score is used to measure the segmentation quality of the methods. Crum et al. [18] present a framework to summarize the results of segmentation studies by computing the overlap region between the predicted pixels and the ground truth pixels. Dice score (D) is calculated using the formula: $D = \frac{2 \times N(U_p \cap L_p)}{N(U_p) + N(L_p)}$, where N denotes the number of pixels inside the glottis region. U_p and L_p represent predicted pixels and ground truth pixels, respectively.

4.3. Results and Discussion

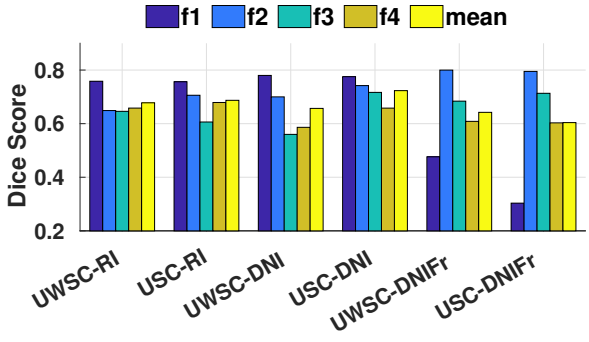


Figure 4: *Bar graph of fold-wise Dice score and mean Dice score across all folds for all U-Net models.*

We have experimented GDN with and without batch-normalization layer between each convolutional layer. It yields a classification accuracy of 87% and 68% with batch-normalization and without batch-normalization, respectively. Performance of GDN is highly affected by the images which

have small glottis or supraglottic structure blocked the glottis, which leads to misclassification of glottis as non-glottis. We compare the results of U-Net with three weight initialization schemes. A total of six experiments are done with three weight initialization schemes (RI, DNI, DNIFr) using two networks (UWSC, USC). Fig. 4 illustrates the fold-wise Dice score of correctly localized images calculated from one of the SLPs for all the six experiments and the last bar graph in each case indicates the mean Dice score across all folds. The UWSC-RI, USC-RI, UWSC-DNI, USC-DNI, UWSC-DNIFr and USC-DNIFr achieve a mean Dice score of 0.678, 0.687, 0.657, 0.723, 0.642 and 0.604, respectively. Here among all the U-Net models, USC-DNI shows the best performance, whereas USC-DNIFr yields the worst performance. It shows that the performance of U-Net can be increased by appropriately initializing the weights of the encoder network. Hence, we consider USC-DNI as our proposed method and it is used for comparison with the baseline. In case of RI and DNI, among all the folds, the Dice score for fold1 is observed to be high as it contains clearly visible glottis images. Similarly, it is low for fold3 except in case of USC-DNI. This is because subjects S_{14} and S_{15} (as shown in Fig. 1) have very small glottis opening. USC-DNI yields the best results even though the glottis opening is small in fold3.

Table 1: *Fold-wise Localization accuracy (%) and corresponding Dice score of correctly localized images achieved by baseline and the proposed approach.*

	(a) Localization accuracy (%)		(b) Dice score	
	Baseline	Proposed method	Baseline	Proposed method
fold1	89.9,78.7,98.5	90.6,91.2,90.6	0.76,0.76,0.81	0.78,0.77,0.76
fold2	38.0,69.0,40.0	61.4,63.6,62.3	0.67,0.69,0.64	0.74,0.76,0.74
fold3	72.2,68.6,72.8	85.9,87.1,65.3	0.62,0.63,0.52	0.72,0.75,0.72
fold4	39.9,76.2,41.4	93.9,91.2,91.7	0.56,0.66,0.55	0.66,0.64,0.65
Average	**60.0,73.1,63.2**	**83.0,83.3,77.5**	**0.66,0.69,0.63**	**0.73,0.73,0.72**

Table 1(a) shows fold-wise Localization accuracy across 3 SLPs using the baseline and proposed approach. Table 1(b) shows corresponding fold-wise Dice score calculated only on the correctly localized images. The three values in each cell indicates the Localization accuracy (or Dice score) calculated with respect to the three SLPs' annotations. Average Dice score / Localization accuracy of all folds has been shown in the last row of Table 1. It is clear from the table that when DNI is used, the USC performs better than the baseline. Both localization accuracy and Dice score are high for fold1. It is due to the clear visible glottis images present in the subjects S_1, S_2, S_3, S_4 and S_5 as shown in Fig. 1. Localization accuracy is low for fold2 using both the baseline and proposed method due to the poor illumination in the subjects S_6, S_7 and S_9 and the corresponding Dice score is high because of the skip connection used in the proposed approach is able to detect the exact boundary of the glottis. On the other hand Localization accuracy for fold4 is high, but the Dice score obtained is not very high. The reason for low Dice score is due to the supraglottic structures that cover the glottis openings.

We observe that both Localization accuracy and Dice score is improved for all the annotators using the proposed scheme compared to the baseline. We also observe a low variance in the averaged Localization accuracy and Dice score across 3 SLPs, which shows the robustness of the proposed method. The usage of skip connection and DNI are the main reasons for getting high Dice score using the proposed approach. The average Localization accuracy of 81.3% and a Dice score of 0.73 from the proposed approach, outperforms the baseline approach by 15.87% and 0.07, respectively. Fig. 5 illustrates the cases where the proposed method performs better than the baseline. It can be observed from the figure that the proposed approach does better than the baseline even though the glottis boundary is not visible due to low illumination in all the cases. The proposed method performs better than the baseline even when the glottis opening is small in case of S_{14}. In this case, the clustering method in the DNN based baseline might select some dark region which is bigger than the glottis leading to poor localization. It is observed that the proposed approach is not performing well for the image in the last row in Fig. 5 where the glottis opening is blocked by supraglottic structures.

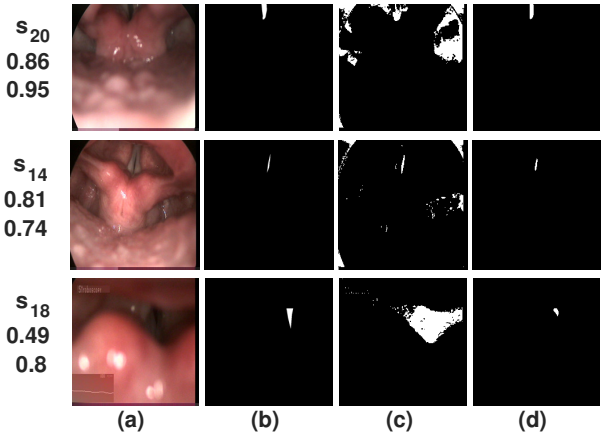


Figure 5: *Column (a): Three sample images with corresponding subject number, Dice score achieved by proposed approach and mean Dice score across pair of SLPs annotations. Column (b): corresponding ground truth image. Column (c): corresponding segmented image by baseline. Column (d): corresponding segmentation by the proposed approach.*

5. Conclusions

In this work, we use two architectures, namely, GDN, to detect the frames with glottis in a given sequence of videostroboscopic images and U-Net using transfer learning approach for glottis segmentation. We convert the problem of glottis segmentation into a classification problem. Threshold based post processing technique is used, in order to reconstruct the glottis image from the predictions. We experiment with two U-Net architectures by considering three different weight initialization schemes. It is shown that the performance of the glottis segmentation network is increased by using glottis detection network weight initialization although fine-tuned with less labeled data. We obtain an average Localization accuracy of 81.3% and a Dice score of 0.73 from the proposed method, which outperforms the baseline scheme by 15.87% and 0.07, respectively. The low variance in the averaged Localization accuracy and Dice score across 3 SLPs suggest the robustness of the method. As a part of future work, we want to improve the performance of U-Net using different loss functions. Furthermore, we would like to extend this algorithm for all types of voice disorders.

6. Acknowledgements

Authors thank the Department of Science and Technology (DST), Government of India for their support in this work.

7. References

[1] I. R. Titze, "The myoelastic aerodynamic theory of phonation, national centre for voice and speech, iowa city," *ISBN*: 0-87414-122-2, pp. 197–214, 2006.

[2] T. Nawka and U. Konerding, "The interrater reliability of stroboscopy evaluations," *Journal of Voice*, vol. 26, no. 6, pp. 812–e1, 2012.

[3] O. Gloger, B. Lehnert, A. Schrade, and H. Völzke, "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, 2014.

[4] J. Demeyer, T. Dubuisson, B. Gosselin, and M. Remacle, "Glottis segmentation with a high-speed glottography: a fully automatic method," *in 3rd Adv. Voice Funct. Assess. Int. Workshop*, 2009.

[5] J. J. Cerrolaza, V. Osma-Ruiz, N. Sáenz-Lechón, A. Villanueva, J. M. Gutiérrez-Arriola, J. I. Godino-Llorente, and R. Cabeza, "Fully-automatic glottis segmentation with active shape models," in *MAVEBA*, pp. 35–38, 2011.

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

[7] V. Osma-Ruiz, J. I. Godino-Llorente, N. Sáenz-Lechón, and R. Fraile, "Segmentation of the glottal space from laryngeal images using the watershed transform," *Computerized Medical Imaging and Graphics*, vol. 32, no. 3, pp. 193–201, 2008.

[8] B. Marendic, N. Galatsanos, and D. Bless, "New active contour algorithm for tracking vibrating vocal folds," in *International Conference on Image Processing (ICIP)*, vol. 1, pp. 397–400, 2001.

[9] C. Palm, T. Lehmann, J. Bredno, C. Neuschaefer-Rube, S. Klajman, and K. Spitzer, "Automated analysis of stroboscopic image sequences by vibration profiles," in *5th Int. Workshop Advances Quantitative Laryngol., Voice Speech Res.*, 2001.

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

[11] M. A. Rao, R. Krishnamurthy, P. Gopikishore, V. Priyadharshini, P. K. Ghosh, "Automatic Glottis Localization and Segmentation in Stroboscopic Videos Using Deep Neural Network,"In *Interspeech* , pp. 3007–3011, 2018.

[12] J. Lin, E. S. Walsted, V. Backer, J. H. Hull, and D. S. Elson, "Quantification and analysis of laryngeal closure from endoscopic videos," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 4, pp. 1127–1136, 2018.

[13] O. Ronneberger, P. Fischer, and T. Brox "U-net: Convolutional networks for biomedical image segmentation," In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[14] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam,and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," In *International conference on medical imagecomputing and computer-assisted intervention.*, pp. 246–253, Springer, 2013.

[15] V. Badrinarayanan, A. Kendall, R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[16] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B.Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *International Conference on MedicalImage Computing and Computer-Assisted Intervention.*, pp. 556–564, Springer, 2015.

[17] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[18] W. R. Crum, O. Camara, and D. L. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE transactions on medical imaging*, vol. 25, no. 11, pp.1451–1461, 2006.

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

[20] F. Chollet et al.,"Keras,", 2015.

[21] T. T. D. Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Anger mueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov et al., "Theano: A python framework for fast computation of mathematical expressions," *arXiv preprint arXiv*:1605.02688, 2016.