Interspeech 2018
2-6 September 2018, Hyderabad

Automatic glottis localization and segmentation in stroboscopic videos using deep neural network

Achuth Rao M V[1], *Rahul Krishnamurthy* [2†], *Pebbili Gopikishore*[3], *Veeramani Priyadharshini*[4], *Prasanta Kumar Ghosh*[5]

[1,5]Electrical Engineering, Indian Institute of Science, Bangalore 560012, India
[3,4]All India Institute of Speech and Hearing, Mysuru, 570006, India
[2] Kasturba medical college, Manipal Academy for Higher Education, Mangalore 575001, India
{[1]achuhtr,[5]prasantg}@iisc.ac.in, [2]rahul.k@manipal.edu, {[3]gopiaslp ,[4]jayceechan7}@gmail.com

Abstract

Exact analysis of the glottal vibration patten is vital for assessing voice pathologies. One of the primary steps in this analysis is automatic glottis segmentation, which, in turn, has two main parts, namely, glottis localization and the glottis segmentation. In this paper, we propose a deep neural network (DNN) based automatic glottis localization and segmentation scheme. We pose the problem as a classification problem where colors of each pixel and its neighborhood is classified as belonging to inside or outside the glottis region. We further process the classification result to get the biggest cluster, which is declared as the segmented glottis. The proposed algorithm is evaluated on a dataset comprising of stroboscopic videos from 18 subjects where the glottis region is marked by the three Speech Language Pathologists (SLPs). On average, the proposed DNN based segmentation scheme achieves a localization performance of 65.33% and segmentation DICE score of 0.74 (absolute), which is better than the baseline scheme by 22.66% and 0.09 respectively. We also find that the DICE score obtained by the DNN based segmentation scheme correlates well with the average DICE score computed between annotation provided by any two SLPs suggesting the robustness of the proposed glottis segmentation scheme.

Index Terms: Glottal segmentation, DNN, stroboscope.

1. Introduction

In human speech production, vocal folds, through its quasi-periodic vibration, play a critical role in modulating airflow from lungs [1]. Glottis is the area between the vocal fold which allows the air to pass. Changes in the muscle properties or overall geometric shape of the vocal folds can cause voice alteration like hoarseness and dysphonia [2]. There is a particular class of vocal fold condition called Sulcus vocalis (SV), where a groove is formed in the vocal fold which extends from epithelium till the vocalis muscle of the vocal folds. Based on the cadaver studies, its prevalence rate has been reported to vary from 0% to 9% [4]. The groove in SV leads to reduced mass of the vocal folds and an incomplete glottic closure during voice production (glottic chink). The laryngeal functioning and classification of the severity of the glottic chink in SV is primarily done by the clinical experts through visual inspection of endoscopic video. This makes it difficult for the speech language pathologist (SLP) to consistently classify glottic chink and to document the change across the phases of interventions.

Apart from expensive high-speed camera to capture fast rate of vibration (70-500Hz), the standard clinical routine for such visualization has been stroboscopy over last decades and is likely to remain the gold standard for the next 10 to 20 years [3]. While there is no doubt among clinicians that stroboscopy is an essential part of medical voice assessment, the clinical parameters obtained from stroboscopy are highly subjective and often show little inter-rater reliability. There are several problems inherent to the technique itself: 1) some images can be incorrectly illuminated 2) images may not be taken at the right instant 3) rotation of the camera that causes the glottis to appear in different orientation posing challenges to segmentation. Stroboscopy video images (corresponding to different subjects) used in this work are illustrated in Fig. 1. It is clear from the figure that there is high variability among the glottal shapes of different subjects. In some cases, the illumination is really poor (e.g., s_8, s_9). From the figure, it is also clear that the camera position relative to glottis changes across subjects. We also observe in the data that the glottis often gets occluded by the supraglottic structures while stroboscopic recording during sustained phonation, thus posing a challenge to automatic glottis localization and segmentation algorithm as shown in Fig. 2.

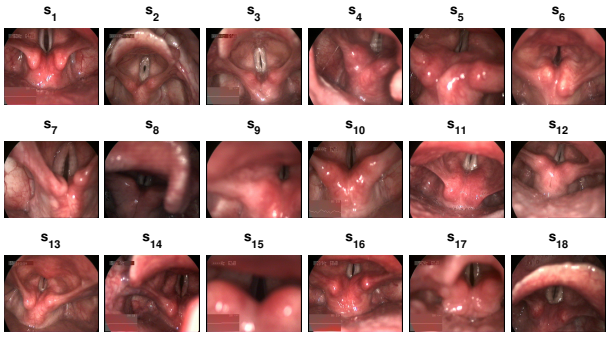

Figure 1: *Example images from all 18-subjects show the inter subject variability in terms of glottal shape, lighting, and camera position.*

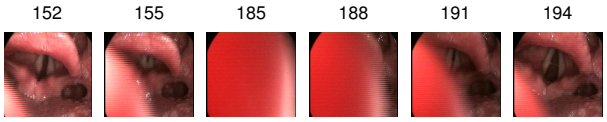

Figure 2: *Image frames (with associated frame numbers) from a video where the glottis is blocked by the supraglottic structures.*

The first step in a stroboscopic video based automatic voice assessment is the segmentation of the glottis region. The shape and the vibration pattern of the glottal region can be used to infer the modification in the muscle properties and the vocal fold geometry. There are only few algorithms in the literature de-

[†]During this work, Rahul Krishnamurthy was in All India Institute of Speech and Hearing, Mysuru, India

10.21437/Interspeech.2018-2572

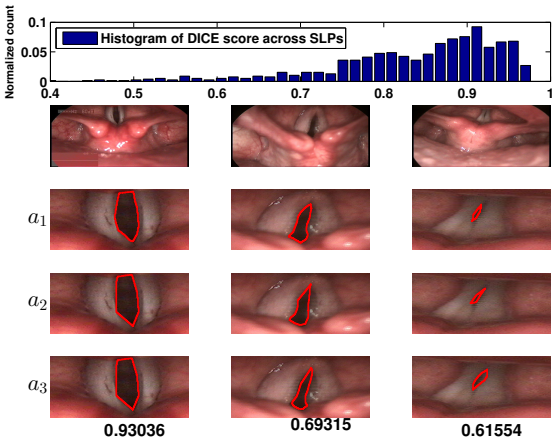

Figure 3: *Histogram of average DICE score [13] between the pair of SLP annotations and three example images and corresponding annotations by the three SLPs to show the degree of inter SLP agreement.*

signed for fully automatic segmentation [2, 4–12]. Only some of them are designed for the stroboscopic videos [2,5,7]. An automatic glottic segmentation followed by quantifying the area of glottis could facilitate the SLP in the classification of the severity of the glottic chink in an objective manner. For example, Gloger et al. have used prior shape knowledge using Fourier descriptor from a large amount of different glottal shapes to localize the glottis and the glottal shape is tracked from one image to the next using levelset segmentation and Probability Image Generation [2]. Cerrolaza et al. have used region growing with the thresholding to localize the glottis and the active shape model (ASM) to find the glottis boundary [5]. But the ASM could fail when the glottal shape during testing is much different from those in training. Osma et al. have used watershed transform with just noticeable difference criteria to localize the glottis [7]. Latter methods were found to be sensitive to the threshold selection.

In this paper, we focus on the segmentation of the glottis using its color structure and its neighboring color pattern. We hypothesize that the problem can be posed as a classification problem and the glottal region color and its neighborhood has a pattern different from that outside the glottal region. We use a 3×3 neighborhood around a pixel and its color in RGB to form the feature vector. We train a DNN to classify whether each pixel belongs to the glottal region or not. We evaluate the algorithm on stroboscopic videos from 18 subjects, where the glottis regions are marked by three SLPs. We use a 4-fold cross validation setup. We train the DNN with the first SLP annotation and evaluate on all three SLP annotations. The results show that the localization accuracies are 60%, 73.1% and 63.2% for three SLP annotations. The segmentation DICE scores [13] on the correctly localized images are 0.69, 0.62 and 0.66 for the three SLP annotations. We show that the proposed method is better than the region growing initialization used in [5] both in terms of the localization accuracy and the DICE score. We also find that the proposed method performs well on the images with high agreement among the SLPs.

2. Dataset

All the stroboscopic videos, used in this work, were recorded as a part of the evaluations that were performed by an Otorhinolaryngologist using Xion Endostrob E from Xion with 70 degree rigid scope. The Digital Video Archive Software (DiVAS) version 2.5 from XION Medical was used for the video recording purposes. The LED light source from the hardware XION Xenon R-180 was utilized for illumination.

For stroboscopic recording, each participant was asked to relax and sit on a metallic stool facing the examiner. The Xylocaine solution was sprayed to the participant's oropharyngeal region to eliminate gag reflex. He/she was instructed to protrude the tongue out and to phonate vowel /i/ (as in word 'heed') for 4-5 seconds whenever indicated verbally. The participant was instructed to repeat the phonations until the Otolaryngologist could obtain an appropriate view of the laryngeal structures. The inbuilt recording option was used for recording the laryngeal structures initially at rest, followed by recording over the inhalation and phonation tasks.

Stroboscopic videos from 18 patients (one video per patient having multiple phonations)(12 males and 6 females) with SV are considered in this work. They are denoted by S_i, $i = 1, ..., 18$. Sample glottis images from stroboscopic video for each of these patients are shown in Fig. 1. The average age of a patient is 30.72 years. Each video is converted to avi format with resolution of 720× 576 and 25 frames per second. A video in this corpus has a minimum of 3 and a maximum of 15 phonations. Each video was chosen to ensure that it contains audible recording events with adequate view of the laryngeal inlet and glottis. Subjects considered in this work were reported to have no associated mass occupying vocal fold lesions, neurological conditions, or any other speech language disorders. The duration of a video varies from a minimum of 11s to a maximum of 84s with an average duration of 44(± 20) seconds. A subset of 921 randomly selected image frames from all 18 videos are for experiments in this work.

A MATLAB based graphical user interface is created to annotate the images, using which the SLPs mark the boundary of the glottis region. Three SLPs (a_1, a_2, a_3) annotated each of 921 images. Second row to the last row of Fig. 3 illustrates three images for which annotations performed by three SLPs. In last three rows the image is zoomed near the glottis to clearly show the boundary depiction. We observe that the annotations often vary across SLPs, particularly when the glottal opening is small as shown in the third column of Fig. 3. Similarly, poor inter-SLP agreement has been observed when the illumination is poor, e.g., S_8, S_{15} and S_6 as shown in Fig. 1. In order to quantify the agreement of the glottis boundary annotation across SLPs, DICE score [13] is used. We compute the DICE score among each pair of annotations for every image. Mean DICE score is computed by taking average across all pairs. Mean DICE score for three illustrative examples in Fig. 3 are given at end of respective columns. It is clear that the DICE score is low for images with small glottal opening. A histogram of mean DICE score for 921 images is shown in the first row of Fig. 3. It is clear from the figure that the majority of the images the DICE score is greater than 0.8, although there is disagreement between the SLPs for few images.

3. DNN based glottis localization and segmentation

The proposed DNN based method consists of two main steps as shown in the Fig. 4 (red box). In the first step, we classify each pixel in the image to predict whether it belongs to inside or outside glottis region. In the second step, we cluster the pixels which are classified as inside glottis regions and filter them

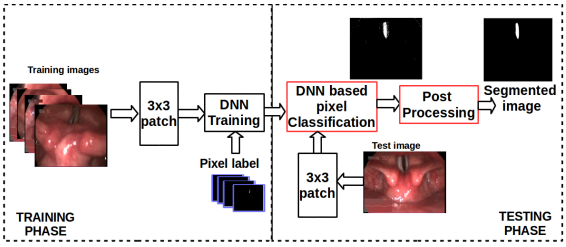

Figure 4: *block diagram of the proposed approach.*

based on eccentricity and its orientation to find the final glottis segment. Each of these steps is explained in detail below.

3.1. DNN based pixel classification

We pose the problem of glottis segmentation given an endoscopic RGB color image (I) of size $M \times N \times 3$ as a classification problem, where each pixel of the image is classified as belonging to inside or outside the glottis region (pixel label). We consider w neighborhood around a pixel to construct a feature of length $(3(2w+1)^2)$. We use this feature as input to the DNN with L layers and the output is a pixel label. Given an input vector x at the first layer of DNN, we obtain the predicted pixel label y_L at the output layer. The output of the l-th hidden layer y_l, given the weight matrix W_l and hidden bias b_l is given by $y_l(x) = \phi(W_l y_{l-1}(x) + b_l), l = 2, \ldots, L-1$. where, ϕ is the activation function. We define d to be the desired pixel label for training the DNN. We define the objective function to be minimized as the binary cross entropy error between the d and predicted label y_L. The weights of the DNN are learnt using the back-propagation algorithm. The weights are updated using ADAM [14]. DNN is implemented by using keras [15] and theano [16] libraries. Given the predicted labels (y_L) for each pixel, we construct a binary image B with pixel value of '1' for those which are predicted as inside the glottis regions.

3.2. Post processing

The previous step independently classify whether the pixel is inside or outside the glottis region. There is no constraint that the pixels with label '1' should form a single region. Hence, we propose post processing step on the binary image B to get the final segmented glottis region.

Given the binary image, where we scan the image column wise and assign the cluster number to the pixel with label '1'. The pixel is assigned to a cluster number based on its majority of the neighbor's cluster numbers. If none of the neighbours of the pixel are assigned to a cluster, we assign a new cluster number to the pixel. This method is also called as run length implementation of the local table method [17]. Given the clustered pixels, we measure the area, eccentricity and the orientations of each of the cluster [17]. It can be observed from Fig. 1 that the glottis shape is similar to an ellipse and the orientation of the major axis is greater than 35^o with respect to the horizontal line. Hence, we retain the clusters which have eccentricity greater than 0.2 and orientation greater than 35^o. The region with the highest area is declared as the segmented glottis region. The proposed method is indicated by DNN_S.

4. Experiments and results

4.1. Experimental Setup

We divide the entire data into four folds, namely, fold1: (S_1, S_2, S_3, S_4), fold2 :$(S_5, S_6, S_7, S_8, S_9)$, fold3: $(S_{10}, S_{11}, S_{12}, S_{13}, S_{14})$, fold4: $(S_{15}, S_{16}, S_{17}, S_{18})$. The subjects in each fold are selected such that the number of images in each fold is approximately the same. We use three folds for training and one fold for testing in a round robin fashion. Frames corresponding to one subject from the training data is used as a validation set. The RGB values of a 3x3 neighborhood centered at each pixel is considered, resulting in a 27-dimensional feature vector with respect to the center pixel. We use the annotations from a_1 for DNN training. The label of '1' is assigned to pixels which belong to inside glottis region and '0' label for rest of the pixels. The number of pixels with '0' labels is more than that with label '1' as the number of pixels within the glottis region is less than that outside the glottis region. Hence, the pixels with label '0' are randomly subsampled to have equal number of feature vectors for both classes. Each feature in the feature vector is made zero mean and unit standard deviation prior to training.

For DNN, we have chosen a configuration with 3-hidden layers with 128 units in each layer. The relu function is used as the activation function for all layers. 27-dim feature vector is given as the input to the first layer. The output layer is a sigmoid layer with 1-dimension. The parameters for ADAM are chosen as follows: learning rate=0.001, $\beta 1 = 0.9, \beta 2 = 0.999$, $\epsilon = 10^{-8}$ and batch size of 32. DNN weights are learnt and validation loss is monitored to stop the training process.

4.2. Baseline scheme

As a baseline scheme, we use the method by Cerrolaza et al., which uses region growing with the thresholding (BL). We found that using ASM with the region growing makes the performance poor. Similarly, the scheme by Gloger et al. [2] does not work well and, hence, not used as a baseline, as the 921 images chosen for annotation did not cover all types of glottal shapes which is critical for it to work well. We combine the proposed method with the baseline by considering the pixels which are classified as belonging to the glottal region by both the proposed DNN based method and the baseline scheme. This is indicated by $BL + DNN_S$.

4.3. Evaluation

We use two evaluation metrics. The first one measures the localization accuracy and the second one measures the accuracy of the segmentation. We evaluate the performance of the algorithms using annotations by all three SLPs.

Localization accuracy $L(\%)$: We evaluate the localization accuracy by the percentage of the test images where the centroid of the predicted glottal region falls within the ground truth glottis contour.

Dice Score (D): To evaluate the segmentation quality of the methods, the DICE score [13] is used, which has been established as a reliable segmentation quality measure for medical images [18]. The DICE score is computed based on the equation: $D = \frac{2 \times N(M_T \cap M_S)}{N(M_T)+N(M_S)}$, where M_T and M_S represent manually annotated and automatically predicted glottal regions, respectively, and $N()$ stands for the number of pixels in a region. Thus, higher the DICE score, better is the glottis segmentation.

4.4. Results and Discussion

Table 1 shows the fold-wise localization accuracy using three methods (BL, $BL + DNN_S$ and DNN_S) as well as accuracy averaged across folds when annotation from each of three SLP

Table 2: *DICE score for the frame for which the (a) only DNN_S localization is correct. (b) only BL localization is correct. (c) both BL and DNN_S localization is correct.*[1]

	(a) DNN_S only			(b) BL only			(c) $DNN_S + BL$		
	BL	$DNN_S + BL$	DNN_S	BL	$DNN_S + BL$	DNN_S	BL	$DNN_S + BL$	DNN_S
fold1	*0.79*,0.52,*0.79*	0.78,0.63,*0.83*	0.76,0.76,*0.81*	0.61,0.76,0.62	0.70,*0.79*,0.79	0.90,0.79,*0.98*	*0.80*,0.73,0.79	*0.83*,0.72,*0.83*	*0.81*,0.72,*0.81*
fold2	0.63,0.09,*0.66*	*0.72*,0.16,0.70	0.67,*0.69*,0.64	0.05,0.59,0.05	0.09,*0.69*,0.10	0.38,*0.69*,0.40	0.63,0.59,*0.66*	*0.72*,0.69,0.70	0.67,*0.69*,0.64
fold3	*0.58*,0.10,0.56	*0.65*,0.22,0.56	0.62,*0.63*,0.52	0.08,0.60,0.07	0.24,*0.68*,0.26	0.72,0.69,*0.73*	*0.64*,0.59,0.62	*0.69*,0.67,0.62	*0.68*,*0.68*,0.59
fold4	*0.74*,0.60,0.72	*0.62*,0.61,0.60	0.56,*0.66*,0.55	0.42,0.81,0.41	0.36,*0.81*,0.35	0.40,*0.76*,0.41	*0.85*,0.70,0.81	*0.85*,0.57,0.81	*0.80*,0.53,0.76
average	***0.69*,0.33,0.68**	***0.69*,0.41,0.67**	**0.66,*0.69*,0.63**	0.29,0.69,0.29	0.35,*0.74*,0.37	**0.60,*0.73*,0.63**	*0.73*,0.65,0.72	***0.77*,0.66,0.74**	***0.74*,0.66,0.70**

Table 1: *Localization accuracy (%) of the proposed algorithm, baseline and the combination of both.*[1]

L(%)	BL	$DNN_S + BL$	DNN_S
fold1	61.0,*76.0*,62.0	70.4,*79.1*,*79.1*	89.9,78.7,*98.5*
fold2	5.4, *59.4*, 5.0	9.0,*69.0*, 9.6	38.0,*69.0*,40.0
fold3	8.2,*60.4*, 7.5	23.9,*67.7*,26.0	72.2,68.6,*72.8*
fold4	42.1,*81.0*,41.4	35.5,*81.3*,35.4	39.9,*76.2*,41.4
average	29.1,*69.2*,29.0	34.7,*74.3*,37.5	**60.0,73.1,63.2**

is used separately for evaluation. It is clear from the table that the DNN_S performs better than the baseline. The accuracy for fold2 using the baseline and proposed method is poor. It could be because of the poor illumination in the subjects S_5, S_6, S_8, S_9 in fold2 as shown in Fig. 1. The glottis is clearly visible in the images for all subjects in fold1. This results in a significantly better localization accuracy in fold1 by the DNN_S over the baseline method.

Table 2(a) shows the DICE score for all folds computed using frames in which glottis is correctly localized by the DNN_S. As expected the DNN_S is performs significantly better than the baseline in these frames. Table 2(b) shows the DICE score for the all folds computed using frames in which glottis is correctly localized by the BL. It can be observed that the DICE scores using DNN_S does not drop significantly in frames where BL based localization is accurate. This is mostly due to the fact that DNN_S based localization is accurate on frames where BL accurately localizes the glottis but not vice-versa. This indicates that the DNN_S is performs as good as BL in all images. Table 2(c) shows the DICE score for all folds computed using frames in which the glottis is correctly localized by both BL and DNN_S ($\sim$26%, $\sim$25%, $\sim$21% of frames for three SLPs respectively). In this case combining both methods performs better. The Fig. 5 compares glottis segmentation results using DNN_S, BL and $DNN_S + BL$ on three example images. It can be observed from the figure that the BL scheme results in wrong segmentation in the first image due to dark region around the glottis while that does not happen for the proposed DNN_S. In the second image, there is considerable amount of reflection from the trachea tube; as a result, BL could not localize any glottal region. Since there is a disagreement among the SLP annotations for some images, we investigate the relation between the DNN_S segmentation and the degree of disagreement across SLP annotations. Fig. 6 shows the DICE score averaged across all three pairs of SLP annotations vs DICE score using DNN_S averaged across three SLPs annotations. For this plot, we consider the frames, where the DNN_S localization is correct. It can be observed from the figure that DNN_S based segmentation performs poorly in frames where there is more disagreement across SLP annotations results in a correlation of 0.39.

[1]The highest average L(%) (or DICE score) among three SLPs annotations are marked in italics and the highest average L(%) (or DICE score) is marked in bold. Three numbers in each cell indicates the L(%) (or DICE score) with respect to each SLP annotation.

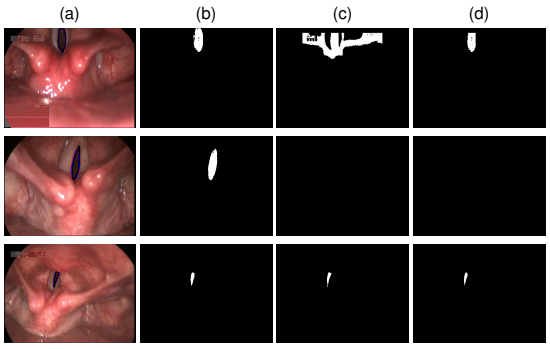

Figure 5: *Column (a): Three example images with annotation. Column (b): corresponding segmentation by DNN_S. Column (c): corresponding segmentation by BL. Column (d):: corresponding segmentation by $BL + DNN_S$.*

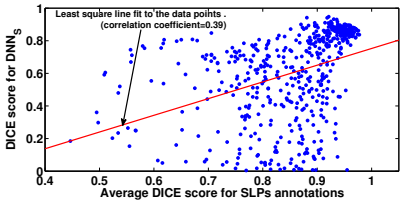

Figure 6: *DICE score averaged across all three pairs of SLP annotations vs the DICE SCORE using DNN_S averaged across three SLPs annotations.*

5. Conclusion

We propose a deep neural network based automatic glottis localization and segmentation scheme. We pose this as a classification problem where colors of each pixel and its neighborhood is classified as belonging to inside or outside the glottis region. We further process the classification result to get the biggest cluster as final segmented glottis. We evaluate the proposed scheme on a dataset comprising stroboscopic videos from 18 subjects, where the glottis region is marked by the three SLPs. On average, the proposed DNN_S scheme achieves a localization performance of 65.33% and segmentation DICE score of 0.74 (absolute), which is better than the baseline scheme by 22.66% and 0.09 respectively. We also find that the DICE score obtained by the DNN based segmentation scheme is not different from the average DICE score computed between segmentation provided by any two SLPs suggesting the robustness of the proposed glottis segmentation scheme. As a part of future work, we want use better way of clustering the DNN output, fine-tuning the predicted contours using Active Contour methods and end to end segmentation based on the DNN. Quantifying the area under glottis for diagnostic as well to use as an outcome measure and application of similar localization methods for identifying the density of redness/inflammation in conditions such as Gastroesophagea reflux disorder.

Authors thank Pratiksha Trust for their support.

6. References

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

[2] 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, 2015.

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

[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*, 2011, pp. 35–38.

[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, 2001, pp. 397–400.

[9] Y. Yan, G. Du, C. Zhu, and G. Marriott, "Snake based automatic tracing of vocal-fold motion from high-speed digital images," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 593–596.

[10] G. A. Miranda, N. Saenz-Lechón, V. Osma-Ruiz, and J. Godino-Llorente, "A new approach for the glottis segmentation using snakes," 02 2013.

[11] 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.

[12] 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.

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

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

[15] F. Chollet *et al.*, "Keras," 2015.

[16] T. T. D. Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, 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.

[17] R. M. Haralick and L. G. Shapiro, *Computer and robot vision*. Addison-wesley, 1992.

[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.