DOI: 10.61782/fa.2023.1124

forum acusticum 2023

APPLICATIONS OF MACHINE LEARNING FOR VOCAL FOLD MOTION ANALYSIS USING LARYNGEAL HIGH-SPEED VIDEOENDOSCOPY

Maryam Naghibolhosseini[1*] **Trent Henry**[1] **Ahmed M Yousef**[1]
Mohsen Zayernouri[2] **Stephanie RC Zacharias**[3,4] **Dimitar D Deliyski**[1]

[1] Department of Communicative Sciences and Disorders, Michigan State University, East Lansing, MI, USA

[2] Department of Mechanical Engineering, and Statistics and Probability, Michigan State University, East Lansing, MI, USA

[3] Head and Neck Regenerative Medicine Program, Mayo Clinic, Scottsdale, AZ, USA

[4] Department of Otolaryngology-Head and Neck Surgery, Mayo Clinic, Phoenix, AZ, USA

ABSTRACT

Laryngeal imaging is widely used to investigate anatomical and physiological aspects of voice production. Laryngeal high-speed videoendoscopy (HSV) is a laryngeal imaging technique and a powerful tool enabling us to capture the vibratory details of vocal folds in each cycle of vibration. HSV is specifically useful for studying voice production in connected speech, where non-stationary and transitory behaviors of vocal folds are consistently observed. This capability of HSV becomes more valuable when studying voice disorders, which involve more irregular and non-stationary behaviors of vocal folds. In this work, HSV is used to study neurogenic voice disorders and compare them with normophonic voices. The HSV data were obtained from the speakers during production of connected speech. The data were collected using a monochrome high-speed camera coupled with a flexible nasolaryngoscope. The dataset for each participant contains hundreds of thousands of images, therefore, machine learning is used for the analysis of this big dataset. The results show that the machine learning approach is successful in the analysis of HSV data with high levels of accuracy. This tool can be used to capture different vibratory features of vocal folds during different instances of voice production helping us characterize the normal and disordered function of voice.

Keywords: *machine learning, artificial intelligence, high-speed videoendoscopy, neurogenic voice disorders, connected speech*

1. INTRODUCTION

Production of the human voice is a unique task that involves the interactions of multiple speech subsystems, the most important of which being the phonatory subsystem [1]. The phonatory system, also referred to as the larynx is the origin of vocalization during speech [2]. Videostroboscopy is a widely-used technique to visualize the larynx and the vibrations of the vocal folds in clinical settings [3]. This technique consists of an endoscope coupled with a stroboscopic light and a video camera. While videostroboscopy is useful in the observation of vocal fold vibrations, it fails to capture the intracycle vibrations of the vocal folds and relies on periodic vibrations of the vocal folds [4]. In the observation of voice disorders, which frequently involve irregular cycles of vocal fold vibration, the limitation of videostroboscopy is more noticeable.

Laryngeal high-speed videoendoscopy (HSV) is an advanced endoscopy tool that overcomes the limitations of videostroboscopy [5-6]. The HSV system allows for the

[*] ***Corresponding author***: *naghib@msu.edu.*

observations of thousands of frames per second (fps) from the vocal fold vibrations whereas the frame rate of videostroboscopy is only 30 fps. The high frame rate of the HSV system allows for detailed images of the vocal folds during their vibrations [7-8]. Recent advancements have allowed for the coupling of HSV with a flexible endoscope, enabling us to observe the vocal fold vibrations during connected speech [9-10]. While the HSV recording allows for a large dataset of images, reviewing hundreds of thousands of HSV images can be a tedious task without a fast and automated tool. To this end, there is a need to develop automated tools based on artificial intelligence or machine learning for the analysis of the HSV images.

Deep neural networks (DNNs) are artificial intelligence algorithms that have been used for the detection of the edges of vocal folds in sustained phonation [11-15]. However, voice disorders should be more thoroughly studied during production of connected speech and not sustained vocalization [16-17]. Hence, our previous research utilized DNNs to detect vocal fold edges using HSV images during connected speech [18-20]. In the most recent work, the DNN was tested for the detection of the vocal fold edges during connnected speech for a subject with adductor laryngeal dystonia and showed highly accurate results [21]. Adductor laryngeal dystonia is a neurological voice disorder in which the laryngeal muscles are impacted causing a strained, strangled, and broken vocal quality in patients [22-23]. Due to its similarity to other voice disorders, e.g., vocal tremor, laryngeal dystonia misdiagnosis is common [24]. As the effetcs of laryngeal dystonia are most often elicited during connetced speech [25], the use of DNNs may lead to more accurate diagnosis of the disoder and proper trearment outcomes and better patient satisfaction in future [26].

The goal of this study is to use the DNN from the previous study [21] on other normophinc subjects, patients with adductor laryngeal dystonia, and other neurogenic voice disorders including unilateral vocal fold paralysis, vocal fold paresis, and and vocal tremor. Vocal tremor is marked by modulations of the laryngeal muscles in a periodic manner that cause fluctuations in the fundamental frequency of the voice during speech [27]. Individuals with unilateral vocal fold paralysis often suffer from an increased vocal effort and changes in vocal quality during speech [28-29]. Those with vocal fold paresis suffer from hypomobility of the vocal folds, which might present as increased vocal effort, dysphonia, or even loss of higher registers [30]. While the underlying etiologies of these disorders may differ, machine learning could provide a useful tool to understand the mechanisms of these neurogenic voice disorders [31-33]. The results of this study will aide in the generalization of the developed machine learning methods (using DNNs) to study these voice disorders using HSV in connected speech.

2. METHODS

2.1 Data Collection

In this study, the data were collected from 34 participants between the ages of 35-76 years old. The participants consisted of 19 normophonic subjects (11 female and eight male) and 15 patients with neurogenic voice disorders (11 female and four male). Among the patients, four (three female and one male) had unilateral vocal fold paralysis, six (five female and one male) had adductor laryngeal dystonia, four (three female and one male) had vocal tremor, and one male had vocal fold paresis. A Photron FASTCAM mini AX200 monochrome high-speed camera (Photron Inc., San Diego, CA), coupled with a flexible nasolaryngoscope was used for the data collection. The HSV recording was done at 4,000 fps (resolution of 256x224 pixels) while the participants were reading the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) [34] sentences and the Rainbow Passage [35].

2.2 Data Analysis

This study used a trained DNN model based on the U-Net architecture in MATLAB 2020b [21]. The block diagram of the DNN is shown in Figure. 1. The orignal DNN was trained using HSV data from three normophonic and three patients with adductor laryngeal dystonia. The training was done to classify the pixels in each frame into the background or glottal area pixel. As the block diagram of the DNN model in Figure 1 shows, the inputs to the model included a series of manually labeled frames as performed by two experienced raters. In total, the raters manually segmented 738 frames of HSV images that were representative of different laryngeal maneuvers (i.e., abducted and adducted vocal folds and vocal fold obstruction by different tissues). After the manual segmentation, 80% of the frames were used for training, and 20% were used for validation. Furthermore, the trained DNN was visually evaluated on a testing set of images by two raters. The output of this diagram represents the automatically generated masks obtained by the network.

The U-Net model encoded the HSV frames to create image-based feature maps through several convolutional layers followed by rectified linear unit activation functions and 2x2 max pooling for downsampling (with stride two). These features were then used in the decoder path

consisting of transposed convolutional layers followed by rectified linear unit activation functions and a soft-max function to create a segmentation mask for each frame. The segmentation mask was a binary image in which the glottal area pixels have the intensity of one and the background pixels have the intensity of zero. The training of the network was done using Adam optimizer. This network was trained and retrained several times by considering a different number of layers in the decoder-encoder parts of the DNN and using different batch sizes, epochs, and learning rates for retraining. Retraining was performed to ensure that the best performing network will be used for the segmentation of the vocal folds and creation of the segmentation masks.

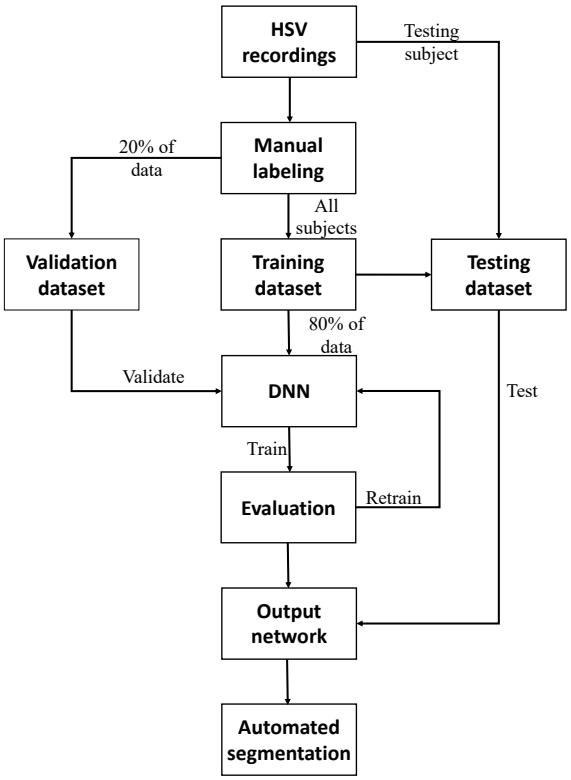

Figure 1. Block diagram of the DNN showing the process of training, validation, and testing of the network.

Based on the DNNs with different structures, the best performing network was selected. The network was tested on HSV data of a subject with adductor laryngeal dystonia and the model's performance was evaluated. The trained DNN model was then applied to subjects with other voice disorders that were not included during the training of the network. The performance of the DNN was then evaluated visually for these additional subjects. Two experienced raters analyzed the automatically created segmentation masks and compared them with the HSV frames t hat were adjusted using a Gaussian filter and by changing the brightness and gamma values.

3. RESULTS

A sample of the HSV frames capturing different laryngeal configuration/postures are shown in Figure 2. This figure shows 12 frames, selected from subjects with different voice disorders. The frames from the left to right columns belong to the following disorders, respectively: adductor laryngeal dystonia, vocal tremor, unilateral vocal fold paralysis, and vocal fold paresis. Each row shows a different vocal configuration.

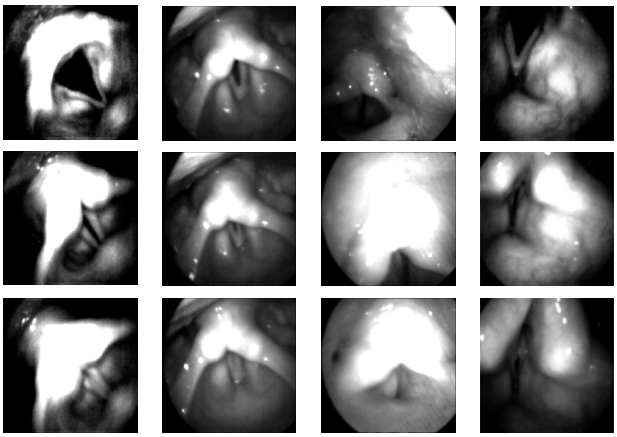

Figure 2. A sample of HSV frames extracted during different laryngeal maneuvers in connected speech. Each column represents a different disorder (adductor laryngeal dystonia, vocal tremor, unilateral vocal fold paralysis, and vocal fold paresis, from the left column to the right, respectively). A wide glottal area, small glottal area, and no glottal area are represented in rows one, two, and three for each subject, respectively.

Using the same frames as in column two of Figure 2, Figure 3 shows the result of the automated segmentation using the DNN, the frames in left panels show the images after the brightness and gamma values of the frames were adjusted

and Gaussian filtering was done. The frames in middle show the automatically segmented area (in white) and the right panels show the automatically segmented area on the adjusted HSV frames (in red).

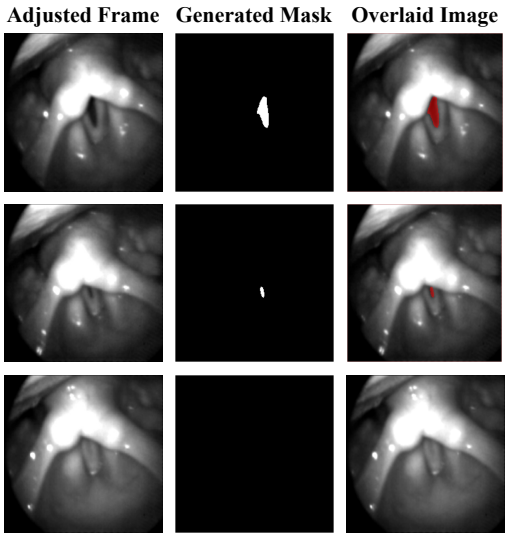

Figure 3. A selection of HSV frames for a participant with vocal tremor, showing the HSV images post adjustment (left panels), the automatically created masks (middle panels), and the two images overlaid on top of each other (right panels). The figure illustrates the ability of the network to generate masks for wide (first row, frame #38,750 at 9.69 s) and small glottal areas (second row, frame #41,000 at 10.25 s), as well as the successful identification of no glottal area (third row frame #34,143 at 8.54 s).

The glottal area waveform, calculated for a patient with vocal tremor is shown in Figure 4. The location of the vertical red lines in the waveform correspond to the three HSV frames shown in Figure 3. The higher frequency oscillations in the waveform represent different vocalizations in connected speech. The low frequency motions indicate the transitions of the vocal folds between vocalizations.

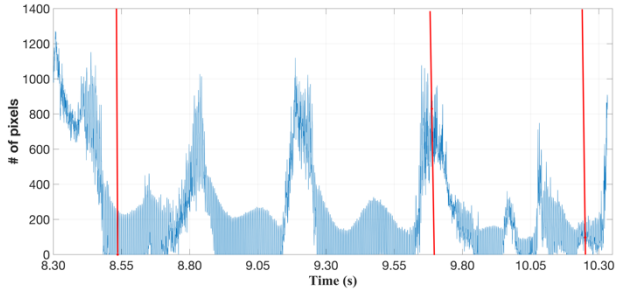

Figure 4. Glottal area waveform for a patient with vocal tremor. The waveform is only shown between 8.3 s and 10.3 s of the recording. The red vertical lines indicate the glottal area for the frames displayed in Figure 3 at 8.54, 9.69, and 10.25 s.

The segmented frames using DNN were analyzed visually by two raters using a specially designed MATLAB script, which displayed the created masks alongside the adjusted HSV frames. Since the developed DNN was found to have 80% accuracy on the testing data (for the subject with adductor laryngeal dystonia), the raters ensured that the masks were accurate for at least 80% of the frames for each subject.

4. DISCUSSION

DNN for the analysis of HSV data obtained from individuals with normophonic voices and several different neurogenic voice disorders. This study used the DNN created in our previous work and implemented it on new HSV data during connected speech. Although the DNN was tested using a subject with adductor laryngeal dystonia, the DNN was successful in segmentation of the vocal fold edges in HSV images for subjects with different neurogenic disorders. This was found based on the visual analysis of the data by the two raters to make sure the DNN was able to detect the correct edges at least 80% of the times. It should be noted that the raters did not view all the DNN generated masks, but a predetermined number based on the length of each recording. As each subject had between 200,000-400,000 frames, raters viewed more than 20,000 frames for each participant. Furthermore, the raters viewed these frames multiple times to determine accuracy of each generated mask. The developed DNN showed difficulties in segmenting the glottal area in some HSV frames, especially for those subjects' recordings on which the network was not trained due to the different image quality and VF configurations. This can be avoided, in future, by including extra frames from those subjects in the training process.

This will further increase the generalizability of the created DNN for the automatic segmentation of the vocal folds for laryngeal dystonia, vocal tremor, unilateral vocal fold paralysis, and paresis. One of the limitations of this work is that the performance of the model was visually assessed due to the costly process of manual labeling to conduct a quantitative performance assessment. Developing automated methods for the evaluation of the performance of the DNN in future will help address this limitation. Overall, the results of this study indicate that artificial intelligence may serve as a useful tool for the objective analysis of HSV data for different neurogenic voice disorders in connected speech.

5. ACKNOWLEDGMENTS

The authors would like to acknowledge the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders (NIDCD) K01DC017751, R21DC020003 and R01DC019402, and Michigan State University Discretionary Funding Initiative for supporting this research, also Maruf Md Ikram for his help with the data analysis.

6. REFERENCES

[1] R. T. Sataloff, Y. D. Heman-Ackah, and M. J. Hawkshaw, "Clinical Anatomy and Physiology of the Voice," *Otolaryngologic Clinics of North America*, vol. 40, no. 5, pp. 909–929, Oct. 2007.

[2] J. Van Den Berg, "Myoelastic-Aerodynamic Theory of Voice Production," *Journal of Speech and Hearing Research*, vol. 1, no. 3, pp. 227–244, Sep. 1958.

[3] K. A. Kendall and R. J. Leonard, Laryngeal evaluation: indirect laryngoscopy to high-speed digital imaging. New York: Thieme, 2010.

[4] D. D. Mehta and R. E. Hillman, "Current role of stroboscopy in laryngeal imaging" *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 20, no. 6, pp. 429–436, Dec. 2012.

[5] R. Patel, S. Dailey, and D. Bless, "Comparison of High-Speed Digital Imaging with Stroboscopy for Laryngeal Imaging of Glottal Disorders," *Ann Otol Rhinol Laryngol*, vol. 117, no. 6, pp. 413–424, Jun. 2008.

[6] S. R. C. Zacharias, C. M. Myer, J. Meinzen-Derr, L. Kelchner, D. D. Deliyski, and A. De Alarcón, "Comparison of Videostroboscopy and High-speed Videoendoscopy in Evaluation of Supraglottic Phonation," *Ann Otol Rhinol Laryngol*, vol. 125, no. 10, pp. 829–837, Oct. 2016.

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

[8] A. M. Kist, S. Dürr, A. Schützenberger, and M. Döllinger, "OpenHSV: an open platform for laryngeal high-speed videoendoscopy," *Sci Rep*, vol. 11, no. 1, p. 13760, Jul. 2021.

[9] D. D. Mehta, D. D. Deliyski, S. M. Zeitels, M. Zañartu, and R. E. Hillman, "Integration of transnasal fiberoptic high-speed videoendoscopy with time-synchronized recordings of vocal function," K. Izdebski, Y. Yan, R. R. Ward, B. J. F. Wong, and R. M. Cruz, Eds., San Francisco: Pacific Voice & Speech Foundation, 2015, pp. 105–114.

[10] M. Naghibolhosseini, D. D. Deliyski, S. R. C. Zacharias, A. De Alarcon, and R. F. Orlikoff, "Temporal Segmentation for Laryngeal High-Speed Videoendoscopy in Connected Speech," *Journal of Voice*, vol. 32, no. 2, p. 256.e1-256.e12, Mar. 2018.

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

[12] P. Gómez *et al.*, "BAGLS, a multihospital Benchmark for Automatic Glottis Segmentation," *Sci Data*, vol. 7, no. 1, p. 186, Jun. 2020.

[13] A. M. Kist, J. Zilker, P. Gómez, A. Schützenberger, and M. Döllinger, "Rethinking glottal midline detection," *Sci Rep*, vol. 10, no. 1, p. 20723, Nov. 2020.

[14] A. M. Kist and M. Dollinger, "Efficient Biomedical Image Segmentation on EdgeTPUs at Point of Care," *IEEE Access*, vol. 8, pp. 139356–139366, 2020.

[15] A. M. Kist *et al.*, "A Deep Learning Enhanced Novel Software Tool for Laryngeal Dynamics Analysis," *J Speech Lang Hear Res*, vol. 64, no. 6, pp. 1889–1903, Jun. 2021.

[16] Edwin Yiu, Linda Worrall, Jennifer, "Analysing vocal quality of connected speech using Kay's computerized speech lab: a preliminary finding," *Clinical Linguistics & Phonetics*, vol. 14, no. 4, pp. 295–305, Jan. 2000.

[17] B. Halberstam, "Acoustic and Perceptual Parameters Relating to Connected Speech Are More Reliable Measures

of Hoarseness than Parameters Relating to Sustained Vowels," *ORL*, vol. 66, no. 2, pp. 70–73, 2004.

[18] A. M. Yousef, D. D. Deliyski, S. R. C. Zacharias, A. De Alarcon, R. F. Orlikoff, and M. Naghibolhosseini, "A Hybrid Machine-Learning-Based Method for Analytic Representation of the Vocal Fold Edges during Connected Speech," *Applied Sciences*, vol. 11, no. 3, pp. 1179.e1-1179.e15, 2021.

[19] A. M. Yousef, D. D. Deliyski, S. R. C. Zacharias, A. De Alarcon, R. F. Orlikoff, and M. Naghibolhosseini, "Spatial Segmentation for Laryngeal High-Speed Videoendoscopy in Connected Speech," *Journal of Voice*, vol. 37, no. 1, pp. 26–36, S0892-1997(20)30408-2, 2023.

[20] A. M. Yousef, D. D. Deliyski, S. R. C. Zacharias, A. De Alarcon, R. F. Orlikoff, and M. Naghibolhosseini, "A Deep Learning Approach for Quantifying Vocal Fold Dynamics During Connected Speech Using Laryngeal High-Speed Videoendoscopy," *J Speech Lang Hear Res*, vol. 65, no. 6, pp. 2098–2113, Jun. 2022.

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

[22] N. Roy, M. Gouse, S. C. Mauszycki, R. M. Merrill, and M. E. Smith, "Task Specificity in Adductor Spasmodic Dysphonia Versus Muscle Tension Dysphonia," *The Laryngoscope*, vol. 115, no. 2, pp. 311–316, 2005.

[23] D. Torres-Russotto and J. S. Perlmutter, "Task-specific Dystonias," *Annals of the New York Academy of Sciences*, vol. 1142, no. 1, pp. 179–199, Oct. 2008.

[24] S. M. Cohen, M. J. Pitman, J. P. Noordzij, and M. Courey, "Evaluation of Dysphonic Patients by General Otolaryngologists," *Journal of Voice*, vol. 26, no. 6, pp. 772–778, Nov. 2012.

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

[26] J. M. Hintze, C. L. Ludlow, S. F. Bansberg, C. H. Adler, and D. G. Lott, "Spasmodic Dysphonia: A Review. Part 2: Characterization of Pathophysiology," *Otolaryngol.--head neck surg.*, vol. 157, no. 4, pp. 558–564, Oct. 2017.

[27] P. Warrick, C. Dromey, J. C. Irish, L. Durkin, A. Pakiam, and A. Lang, "Botulinum Toxin for Essential Tremor of the Voice With Multiple Anatomical Sites of Tremor: A Crossover Design Study of Unilateral Versus Bilateral Injection:," *The Laryngoscope*, vol. 110, no. 8, pp. 1366–1374, Aug. 2000.

[28] B. C. Spector, J. L. Netterville, C. Billante, J. Clary, L. Reinisch, and T. L. Smith, "Quality-of-Life Assessment in Patients with Unilateral Vocal Cord Paralysis," *Otolaryngol.--head neck surg.*, vol. 125, no. 3, pp. 176–182, Sep. 2001.

[29] B. E. Richardson and R. W. Bastian, "Clinical evaluation of vocal fold paralysis," *Otolaryngologic Clinics of North America*, vol. 37, no. 1, pp. 45–58, Feb. 2004.

[30] A. D. Rubin and R. T. Sataloff, "Vocal Fold Paresis and Paralysis," Otolaryngologic Clinics of North America, vol. 40, no. 5, pp. 1109–1131, Oct. 2007.

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

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

[33] A. Yousef, D. Deliyski, S. d. A. A. Zacharias, R. Orlikoff and M. Naghibolhosseini, "Automated detection and segmentation of glottal area using deep-learning neural networks in high-speed videoendoscopy during connected speech," in In 14TH International Conference Advances In Quantitative Laryngology, Voice And Speech Research (AQL), Bogotá, Colombia, 2021 June 7-10.

[34] G. B. Kempster, B. R. Gerratt, A. K. Verdolini, -Kraemer Julie Barkmeier, and R. E. Hillman, "Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol," *American Journal of Speech-Language Pathology*, vol. 18, no. 2, pp. 124–132, May 2009.

[35] Fairbanks G. “The Rainbow Passage.” *In: Voice and articulation drillbook.* New York: Harper & Row; 1960 p. 127