Have We Solved Glottis Segmentation? Review and Commentary

[*]Andreas M. Kist, and [†]Michael Döllinger, *†*Erlangen, Germany*

Abstract: Quantification of voice physiology has been a key research goal. Segmenting the glottal area to describe the vocal fold motion has seen increased attention in the last two decades. However, researchers struggled to fully automatize the segmentation task. With the advent of deep learning, fully automated solutions are within reach and have been proposed. Are we then done here? This commentary highlights the open construction sites and how glottis segmentation can be still of scientific interest in this decade.

Key Words: Deep learning—Glottis segmentation—Vocal folds—Quantification—Glottal area—Image processing—Image analysis—Deep neural networks.

The glottal area is defined as the opening between the vocal folds. Through vocal fold vibration, the glottis opens and closes relative to the vocal fold oscillation cycle (Figure 1). This changing area can be used as a proxy for vocal folds' oscillation behavior. We typically refer to this changing area over time as the glottal area waveform (GAW), a key signal used in many downstream computations of clinically important quantitative parameters.[1] A key challenge is the segmentation of the glottis in endoscopic images, crucial for computing the GAW (Figure 1).

In very high-contrast conditions, simple thresholding is sufficient to separate the glottis (dark pixels) from the background (rather bright pixels). High-contrast images are not always available in a clinical setting, so researchers searched for more robust image processing algorithms. Early on, this repetitive and rather "simple" task, finding a very distinct area in an image, was thought to be largely automated. Lohscheller and colleagues proposed the semi-automated multi-threshold segmentation of keyframes with interpolation, significantly reducing the manual effort.[2] Many works focused on different computer vision techniques to segment the glottal area. Also, further fully automatic segmentation techniques were proposed using sophisticated workflows.[3]

With the advent of deep learning, multiple labs utilized deep learning methods for glottis segmentation, for example,[4–6] as these allow an end-to-end application with a single processing step (Figure 2). The application of multiple general architectures has been evaluated by Laves et al,[5] including the famous U-Net architecture introduced by Ronneberger et al.[7] The U-Net architecture (Figure 2) is an encoder-decoder neural architecture that consists of a contracting and an expanding path. In the contracting path, high-level information is extracted and represented in the latent space. The latent space is used as an entry point in the expanding path to generate the segmentation mask. Recent works showed that the latent space for glottis segmentation can be indeed very small: A single latent space image is actually sufficient for glottis segmentation.[8] Fehling and colleagues investigated a battery of U-Net modifications, including preprocessing steps, that yield the best glottis segmentation on their in-house dataset. They found that incorporating temporal context through bidirectional convolutional Long Short-Term Memory layers together with Red-Green-Blue color space yielded the best segmentation results.[4] In agreement, other works similarly showed the applicability of the U-Net architecture for glottis segmentation[9] also applicable in connected speech.[10]

For training deep neural networks, such as the U-Net, it is common sense that a large body of data is needed. The Benchmark for Automatic Glottis Segmentation (BAGLS) provides 59 250 pairs of endoscopic images and their respective glottis segmentation mask. In total, 640 videos from seven hospitals were sourced and acquired with a variety of technical equipment.[6] The BAGLS dataset is not only sufficient to work on data acquired with the same equipment but is also suited as training data for novel hardware devices.[11] It is further ideally suited for deep neural networks that are actively used at the point of care over a long period of time,[12] especially when more data is incorporated through a continual learning scheme. The expansion of the BAGLS dataset with more data (BAGLS-RT, additional 21 050 images from 267 HSV videos from eight hospitals) results in even more robust deep neural networks.[13]

As glottis segmentation has been one major bottleneck bringing high-speed videoendoscopy into the clinic,[14] fully automatic solutions that work reliably with clinical data were desired. Hence, clinically optimized deep neural networks have been proposed that perform on par with large baseline models; however, can yield in combination with inexpensive hardware accelerators, such as the Edge Tensor Processing Unit, 79 times speed-ups.[15] For example, a 1000-frame-long recording would be processed in less than one minute. This setup has been intensively validated across a 24-month duration[12] showing the applicability of glottis segmentation in a clinical context.

Accepted for publication November 20, 2024.

From the *Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91052 Erlangen, Germany; and the †Division of Phoniatrics and Pediatric Audiology, Department of Otorhinolaryngology Head and Neck Surgery, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91054 Erlangen, Germany.

Address correspondence and reprint requests to Andreas M. Kist. E-mail: andreas.kist@fau.de

Journal of Voice, Vol 39, No 3, pp. 574–576
0892-1997

https://doi.org/10.1016/j.jvoice.2024.11.037

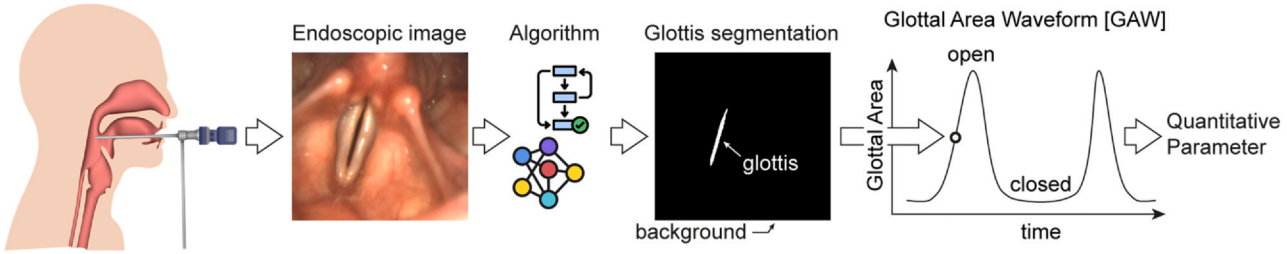

FIGURE 1. *Analysis workflow.* Endoscopic footage will be analyzed frame-by-frame or in a batch of frames via an algorithm that yields ideally fully-automatically a binary glottis segmentation. Each frame contributes as a single data point to the glottal area waveform (GAW). The GAW is information-rich and is used for downstream quantitative parameter computation.

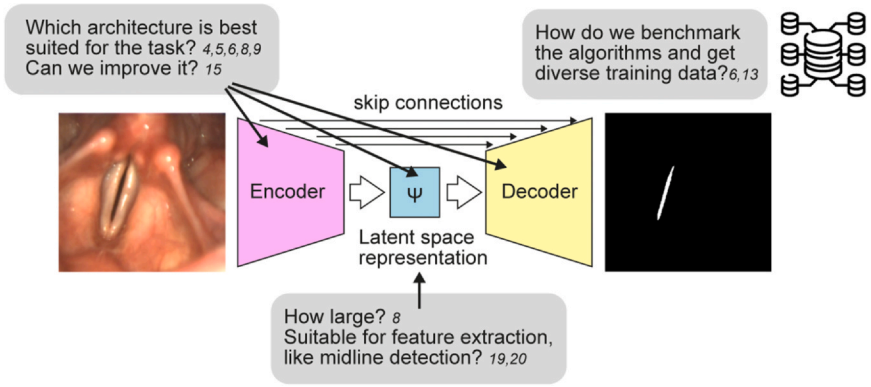

FIGURE 2. *Schematic overview of U-Net-like deep neural networks for glottis segmentation.* The encoder (pink) extracts features from the endoscopic image, yielding in a latent space representation (blue). This representation is used by the decoder to spatially reconstruct and classify any glottis-containing pixel. This is thought to be improved with spatial information stemming from the encoder via skip connections. We highlight important scientific questions and indicate the appropriate literature as cited in the main text. For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.

One question, however, still remains: there is no real and perfect ground truth for glottis segmentation. Especially at the border close to the vocal folds, it is debatable where to draw the border between the "glottis" and "not glottis," that is, background or trachea (Figure 1). A recent study has shown that manual glottis area segmentation yields very consistent results in downstream quantitative parameters.[1] This is in line with the other findings that show that optimized deep neural networks yield no relevant differences in clinical parameters.[15] However, how can we judge if an automatic method is actually producing high-quality segmentations or failed because of any unexpected circumstances, such as artifacts and ill-illuminated footage? This question has also been addressed recently: by predicting the Intersection over Union score, a metric to determine the segmentation quality ranging between 0 (very poor segmentation) to 1 (perfect segmentation), we can determine for each segmented frame the overall quality. With a low average error of less than 0.1, one can reliably identify failed segmentations and not consider these segmentations in downstream processing. The authors also showed that we do not need to reach a perfect Intersection-over-Union score. Values around 0.7 on unseen data are sufficient and competitive with human inter-rater and intra-rater reliability,[16] being in line with previous reports.[6]

Taken together, contemporary research found fully automatic methods to segment the glottal area using deep neural networks, compiled a large open dataset to train these deep neural networks, and evaluated their performance in a clinical environment, showing that this technology is ready and available for broad research and clinical use, for example.[11,17] In combination with an internal evaluation by predicting the Intersection-over-Union score for each frame, we made a giant leap forward towards the application of HSV in the clinic.

Does this mean we have solved glottis segmentation? One could argue we solved the greatest challenges, such as a reliable, robust, fast, and fully automatic glottis segmentation as well as segmentation quality assessment. Also, by having two large and publicly available datasets,[6,13] all future algorithms with the goal of improving current approaches can be compared to these objective benchmarks. For example, techniques such as federated learning[18] can be explored to unify the footage acquired around the world to foster and deploy generalized and robust deep neural networks to the point of care. Also, a reliable and robust glottis midline detection has not been suggested yet, although first promising approaches were suggested.[19,20]

Nevertheless, we observe that especially in low-lit conditions, as well as semi-occluded areas around the glottis, deep neural networks still have issues in reliably detecting

glottal pixels. In addition, the aforementioned datasets contain mostly rigid endoscope-derived footage, falling short of footage derived from nasal, flexible endoscopes, which require additional preprocessing.[21] The diversity of organic and functional disorders may have implications for the overall performance, which should be assessed and systematically studied in the future.

Taken together, we believe the community's strong works have pushed the field dramatically in the last years, and we are looking forward to seeing these last issues being solved.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Maryn Y, et al. Intersegmenter variability in high-speed laryngoscopy-based glottal area waveform measures. *Laryngoscope.* 2020;130:E654–E661.
2. Lohscheller J, Toy H, Rosanowski F, et al. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Med Image Anal.* 2007;11:400–413. https://doi.org/10.1016/j.media.2007.04.005.
3. Gloger O, Lehnert B, Schrade A, Völzke H. Fully automated glottis segmentation in endoscopic videos using local color and shape features of glottal regions. *IEEE Trans Biomed Eng.* 2015;62:795–806. https://doi.org/10.1109/TBME.2014.2364862.
4. Fehling MK, Grosch F, Schuster ME, et al. Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional LSTM network. *PLOS One.* 2020;15:e0227791. https://doi.org/10.1371/journal.pone.0227791.
5. Laves M-H, Bicker J, Kahrs LA, Ortmaier T. A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *Int J CARS.* 2019;14:483–492. https://doi.org/10.1007/s11548-018-01910-0.
6. Gómez P, et al. BAGLS, a multihospital benchmark for automatic glottis segmentation. *SciData.* 2020;7:186. https://doi.org/10.1038/s41597-020-0526-3.
7. Ronneberger O, Fischer P, BroxT. U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv: 1505.04597 [cs], 2015, Zugegriffen: 9. November 2018. [Online]. Available at http://arxiv.org/abs/1505.04597,.
8. Kist AM, Breininger K, Dörrich M, et al. A single latent channel is sufficient for biomedical glottis segmentation. *Sci Rep.* 2022;12:14292. https://doi.org/10.1038/s41598-022-17764-1.
9. Ding H, Cen Q, Si X, et al. Automatic glottis segmentation for laryngeal endoscopic images based on U-Net. *Biomed Signal Process Control.* 2022;71:103116.
10. Yousef AM, Deliyski DD, Zacharias SRC, et al. Spatial segmentation for laryngeal high-speed videoendoscopy in connected speech. *J Voice.* 2023;37:26–36. https://doi.org/10.1016/j.jvoice.2020.10.017.
11. Kist AM, Dürr S, Schützenberger A, Döllinger M. OpenHSV: an open platform for laryngeal high-speed videoendoscopy. *Sci Rep.* 2021;11:13760. https://doi.org/10.1038/s41598-021-93149-0.
12. Groh R, Dürr S, Schützenberger A, et al. Long-term performance assessment of fully automatic biomedical glottis segmentation at the point of care. *Plos One.* 2022;17:e0266989.
13. Döllinger u. a M. Re-training of convolutional neural networks for glottis segmentation in endoscopic high-speed videos. *Appl Sci.* 2022;12:9791.
14. Deliyski DD, Petrushev PP, Bonilha HS, et al. Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. *FPL.* 2008;60:33–44. https://doi.org/10.1159/000111802.
15. Kist AM, Döllinger M. Efficient biomedical image segmentation on edgeTPUs at point of care. *IEEE Access.* 2020;8:139356–139366. https://doi.org/10.1109/ACCESS.2020.3012722.
16. Kist AM, Razi S, Groh R, *et al*, Predicting semantic segmentation quality in laryngeal endoscopy images, 15. November 2024, bioRxiv. doi: 10.1101/2024.11.14.623604.
17. Wevosys, lingWAVES 4 High Speed Videoendoscopy (HSV). [Online]. Available at https://www.wevosys.com/products/lingwaves4/lingwaves4_high_speed_videoendoscopy.html,.
18. Nguyen DC, et al. Federated learning for smart healthcare: a survey. *ACM Comput Surv (Csur).* 2022;55:1–37.
19. Kist AM, Zilker J, Gómez P, et al. Rethinking glottal midline detection. *Sci Rep.* 2020;10:1–15.
20. Kruse E, Döllinger M, Schützenberger A, Kist AM. Glottisnetv2: temporal glottal midline detection using deep convolutional neural networks. *IEEE J Trans Eng Health Med.* 2023;11:137–144.
21. Echternach M, et al. Biomechanics of sound production in high-pitched classical singing. *Sci Rep.* 2024;14:13132.