Long-term performance assessment of biomedical image segmentation at the point of care

René Groh[1] RENE.GROH@FAU.DE
[1] *Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*

Stephan Dürr[2] STEPHAN.DUERR@UK-ERLANGEN.DE
[2] *Division of Phoniatrics and Pediatric Audiology at the Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Erlangen, Friedrich–Alexander-Universität Erlangen–Nürnberg, Erlangen, Germany*

Anne Schützenberger[2] ANNE.SCHUETZENBERGER@UK-ERLANGEN.DE

Marion Semmler[2] MARION.SEMMLER@UK-ERLANGEN.DE

Andreas M. Kist[1] ANDREAS.KIST@FAU.DE

Editors: Under Review for MIDL 2022

Abstract

Deep Learning has a large impact on medical image analysis and lately has been adopted for clinical use at the point of care. However, there is only a small number of reports of long-term studies that show the performance of deep neural networks (DNNs) in such a clinical environment. In this study, we measured the long-term performance of a clinically optimized DNN for laryngeal glottis segmentation. We have collected the video footage for two years from an AI-powered laryngeal high-speed videoendoscopy imaging system and found that the footage image quality is stable across time. Next, we determined the DNN segmentation performance on lossy and lossless compressed data revealing that 9% of recordings contain segmentation artefacts. We found that lossy and lossless compression are on par for glottis segmentation, however, lossless compression provides significantly superior image quality. Lastly, we employed continual learning strategies to continuously incorporate new data to the DNN to remove aforementioned segmentation artefacts. With modest manual intervention, we were able to largely alleviate these segmentation artefacts by up to 81%. We believe that our suggested deep learning-enhanced laryngeal imaging platform consistently provides clinically sound results, and together with our proposed continual learning scheme will have a long-lasting impact in the future of laryngeal imaging.

Keywords: Endoscopy, Larnygology, Image Segmentation, Deep Neural Network

1. Introduction

Laryngeal videoendoscopy is a major assessment tool to evaluate voice physiology qualitatively and quantitatively (Fig. 1). Especially for quantifying voice physiology, high-speed videoendoscopy (HSV) (Deliyski et al., 2008; Zacharias et al., 2018) is an emerging technique that is able to visualize each glottal cycle with high spatial and temporal resolution. As the vocal folds, the main source of our voice, are vibrating hundreds of times each second, high-speed recordings with at least 4,000 frames per second (fps) are needed to accurately record this motion (Kunduk et al., 2010; Deliyski et al., 2008).

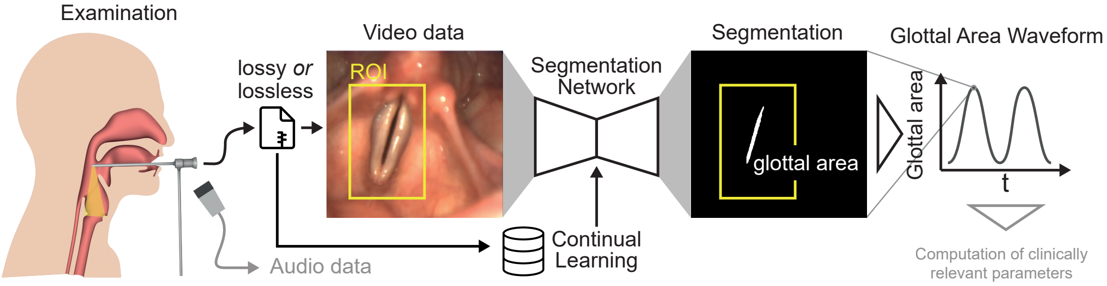

Figure 1: Data acquisition and analysis workflow. Each examination yields video and audio data, where the video data is stored either lossy or lossless compressed. In this study, we evaluate which data source and if cropping the video data to an ROI is sufficient for reliable glottis segmentation using previously proposed clinically optimized segmentation networks. The glottal area is computed for each video frame and plotted across time yielding the glottal area waveform (GAW), which is crucial for computing clinical relevant parameters. We investigate in this study how collected data can be used to allow constant fine-tuning of the segmentation network using continual learning schemes.

The glottal area, the opening between the two vocal folds, is a good proxy for assessing the oscillation behavior (Deliyski et al., 2008; Andrade-Miranda et al., 2020). Therefore, many works have focused on the segmentation of the glottal area (Fig. 1), especially avoiding any manual intervention (Andrade-Miranda et al., 2020; Gloger et al., 2014). This critical step is one of the main bottlenecks of the data analysis pipeline and has been the reason why HSV is barely applied in the clinic, because fully automatic data analysis solutions have not been around (Deliyski et al., 2008). Lately, it has been shown that deep neural networks (DNNs) are highly suitable for solving this task (Kist and Döllinger, 2020; Gómez et al., 2020; Fehling et al., 2020; Laves et al., 2019). These glottis segmentation DNNs could be further optimized towards clinical applicability with barely sacrificing segmentation accuracy (Kist and Döllinger, 2020). Together with the recent development of an open source HSV system, namely OpenHSV, latest hardware and software components were introduced to the clinic (Kist et al., 2021a) that features these optimized DNNs for clinical use. However, there is no record how these DNNs perform actually in a clinical environment, as they have been validated on limited and selected data.

In this work, we report the performance of the AI-powered OpenHSV system together with the DNN in a two-year clinical environment. Our main contributions are summarized as follows: (1) we describe for the first time the overall image quality distribution during a two-year period of clinical use, and if lossy data compression is suitable for data storage and subsequent data analysis, (2) give an unbiased performance evaluation of previously proposed optimized DNNs for clinical use for different data origins and (3) employ a continual learning and fine-tuning strategy to allow continuous integration of novel data into the DNN. Taken together, we strongly believe that our study provides trust and shows reliability for the OpenHSV system positively impacting future clinical adaptation.

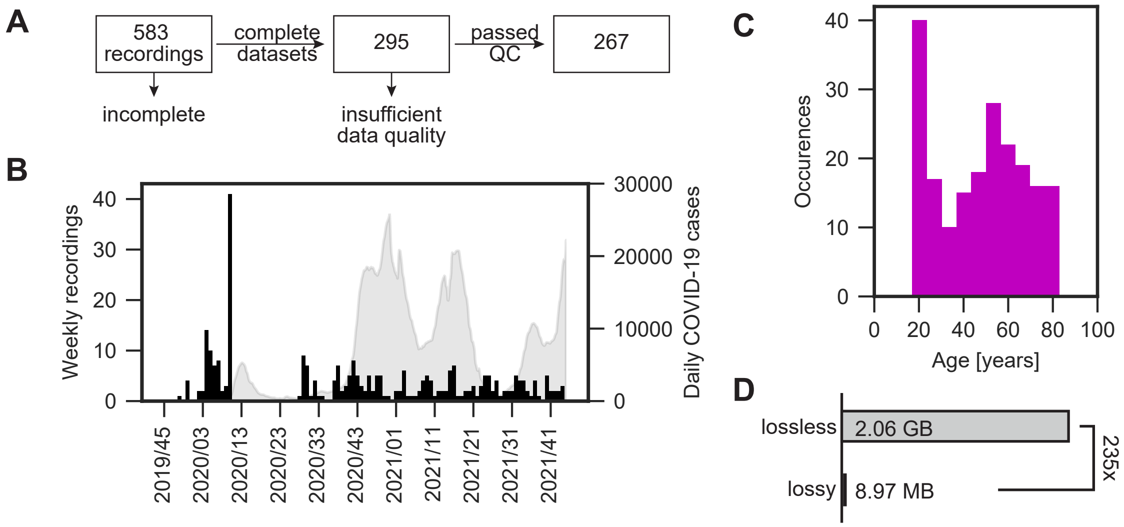

Figure 2: Acquired OpenHSV data overview. A: From initial 583 recordings, we excluded datasets that are missing any kind of data. From these, we excluded ones with insufficient data quality, such as occluded glottis. B: Frequency of recordings across the two year time frame is shown in black. The Germany COVID-19 cases are shown as a 7-day-average in gray. C: Age distribution across selected recordings from A. D: Comparison of file size between lossy and lossless compression.

2. Methods

2.1. OpenHSV system

We are using the OpenHSV platform introduced in (Kist et al., 2021a). Patients are routinely examined using a rigid endoscope equipped with a high-speed camera (IDT CCmini-1540) running at 4,000 fps. Illumination is granted by a high-power LED light source (Storz 300 W LED). Each recording is at least 1,000 frames long and contains synchronously acquired video and audio data. We further record patient metadata consisting of the patient's age, gender, and condition. For each video, we save two files encoded with the h.264 codec: (1) lossy compression using common settings for video and (2) lossless compressed data to recover the original recorded raw data. For lossy compression, we use the libx264 codec, the yuv420p pixel format and set the quality to 5 resulting in a varying bitrate. For lossless compression, we used the libx264rgb codec, the rgb24 pixel format, set the '-crf' flag to 0 and used the ultrafast preset.

2.2. Patients

We assessed a total of 583 recordings acquired between November 2019 and November 2021. All recordings are done routinely in the clinic and are performed according to local regulations (Ethikkommission FAU Erlangen-Nürnberg, #290_15). We first selected only those recordings that featured a complete set of data, such that 295 recordings remained. Next, we manually investigated the data quality. We ranked each video on an ordinary scale: 0 (insufficient), 1 (okay), 2 (excellent). Videos ranked 0 were showing insufficient data quality, such as non-visible glottis or foggy videos, and were discarded. Finally, 267 recordings from 202 unique patients remained and were subjected to further analyses (Fig. 2A). We report the frequency of recordings across the last two years in Fig. 2B. Additionally, COVID-19 cases for Germany were provided as a reference how data generation was affected by lock-downs. The number of recordings were higher before the first lock-down, but their fluctuation remained constant when the general clinical activity was restored. The age distribution of the patients is largely spread (Fig 2C), where the mean age is 47.2 ± 20.2 (std) years. The reported gender for the analyzed subjects is 31.2% male, 66.8% female and 2.0% had no further specified gender. The average file size for lossless and lossy recordings was 2.06 ± 1.27 GB and 8.97 ± 5.16 MB, respectively (see Fig. 2D). As the lossy files are around 235-times lower than the lossless compressed counterparts, we investigated in this study if the lossy compression has an impact on the segmentation performance.

2.3. Image quality assessment

As we saved the data in two different compression modes, namely lossy and lossless, we evaluated if there are any compression artefacts. First, we investigated the dynamic range of the image, where 0 is no dynamic range and 1 is full dynamic range by computing the normalized absolute difference of the minimum and maximum pixel value. We determined the no-reference, completely blind image quality score ”Natural Image Quality Evaluator” (NIQE) (Mittal et al., 2012) to provide an unbiased score for objective image quality assessment. We further assessed full reference metrics, such as peak signal to noise ratio (PSNR) and structural similarity index metric (SSIM) (Wang et al., 2004) to compare lossy compression against the lossless stored data. We use respective implementations of the metrics in scikit-video (NIQE) and scikit-image (PSNR, SSIM).

2.4. Deep Neural Network

The OpenHSV system is shipped with a deep neural network (DNN) based on the U-Net architecture (Ronneberger et al., 2015), that was optimized for clinical use as described previously (Kist and Döllinger, 2020). The used DNN is openly available on the OpenHSV Github account at `https://github.com/anki-xyz/openhsv`. The training process is described in (Kist et al., 2021a). Briefly, the DNN is setup in TensorFlow 2.2 and the high-level Keras package. The DNN was pretrained in a supervised fashion on the full training dataset (55,750 images) of the open benchmark for automatic glottis segmentation (BAGLS, (Gómez et al., 2020)). The pretrained network has never been exposed to OpenHSV data during training.

2.4.1. Region of interest

A rectangular region of interest (ROI) was drawn manually for each recording. We save the ROI coordinates for further use in JSON format. Each ROI is adjusted as such that the width and height is divisible by 32 to ensure proper DNN propagation.

2.4.2. Segmentation

Lossy or lossless compressed endoscopic frames are first converted to grayscale by extracting the luminance channel using standard procedures, as it has been shown that color information is not essential for glottis segmentation (Gómez et al., 2020). In some experiments, only an ROI around the vocal folds is used for inference. The input image intensity is normalized between -1 and 1. The segmentation mask gained from the DNN provides values in the range of 0 (background) and 1 (glottis) by a sigmoid function in the output layer. For further use and due to memory limitations, we multiplied the predicted segmentation masks by 255 in order to save the data in uint8 data format. The glottal area waveform (GAW) is computed by summing the segmented pixels in each frame for every timepoint of a given recording (see Fig. 1).

2.4.3. Continual training

We performed retrospective continual training on the original OpenHSV segmentation DNN. We preprocessed new images as described above and used two continual learning strategies: We either selected a fixed time period for data collection (7, 14, 30 days) or a fixed video quantity (every 10, 20 or 40 videos, see also Fig. 5A). At each continual learning point, we trained the model for ten epochs using the previous predicted segmentation masks as ground truth for the training process. We chose a low learning rate of 10^{-6} combined with the Adam optimizer to fine-tune the model. After each continual training step, we evaluated the occurrence of artefacts by visual assessment and calculated the achieved IoU.

3. Results

3.1. Recording quality is consistent across time

To evaluate the performance of the segmentation DNN, we first assessed the overall image quality for both, lossless and lossy recordings, as this is a major confounding source for segmentation success. First, we determined the maximum dynamic range of each given image. We found that the maximum dynamic range is constantly high for both, lossless and lossy recordings (Fig. 3A, left panel), however, the dynamic range is significantly lower for lossy recordings as for lossless recordings (paired Student's t-test, $p < 0.01$, Fig. 3A, right panel). Using a complete blind, non-reference metric, the NIQE score (Mittal et al., 2012), we could show that lossless recordings have lower, therefore better NIQE scores until beginning of 2021 (Fig. 3B), afterwards the NIQE scores were highly overlapping. In (Kist et al., 2021a), a mean NIQE for the OpenHSV system of 13.19 was reported, showing that image quality has been consistent since the clinical introduction of the system.

To compare images retrieved from lossy recordings to their lossless counterparts, we relied on two reference metrics, peak-signal-to-noise-ratio (PSNR) and structural similarity index measure (SSIM). PSNR has a constant high value above 90 dB, suggesting a high-quality compression (Fig. 3C). In contrast to PSNR, SSIM also takes the perceptual change in structural information into account. In Fig. 3D we show that the lossy compressed videos nevertheless are in almost perfect agreement with the lossless reference.

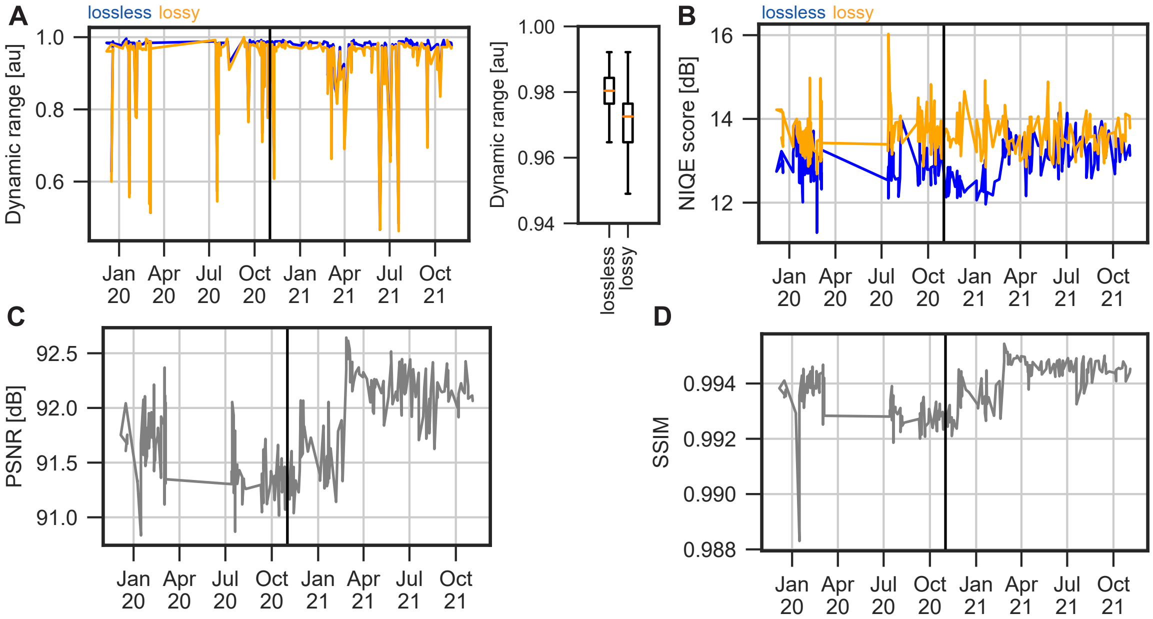

Figure 3: Image quality is relatively stable across time. A: Dynamic range per recording. Lossless compressed data in blue, lossy in orange. Boxplots show the 95-percentile of the data. B: NIQE score for lossy (orange) and lossless (blue) compressed videos across time. C: Peak-Signal-To-Noise-Ratio (PSNR) across time. D: Structural Similarity Index Measure (SSIM) across time.

In summary, we found that the image quality has single outliers, but overall is highly consistent across time. Further, high PSNR and SSIM values indicate an overall accurate conversion from lossless to lossy image content and high quality in lossy compressed video.

3.2. Segmentation performance is not affected by lossy data compression

To further evaluate the performance of the segmentation DNN, we manually annotated the glottal area in the first 100 frames of 20 randomly selected videos, half of which were rated as quality 1 (okay) and the other half as quality 2 (excellent). Using this ground truth and the predicted segmentation masks by the DNN, we computed the Intersection over Union (IoU) for each frame across all videos. We found that lossy and lossless saved videos achieve a similar performance with a median IoU of 0.756/0.742 and 0.770/0.768 for segmentation masks computed with and without ROI, respectively (Fig. 4B). These IoU values are comparable to other works (Kist and Döllinger, 2020), where IoU scores between 0.741 and 0.769 were achieved, and are sufficient for clinical reliability.The use of an ROI enhances the segmentation speed as smaller images are used, however, this leads to worse results. We hypothesize that this is caused due to the loss of global spatial information. We further mined the computed IoU scores to determine why very low IoU scores are obtained. Fig. 4C shows that the low IoU scores emerge with a small segmented area, i.e. when the glottis is closed. This is in line with previous reports (Gómez et al., 2020), and has

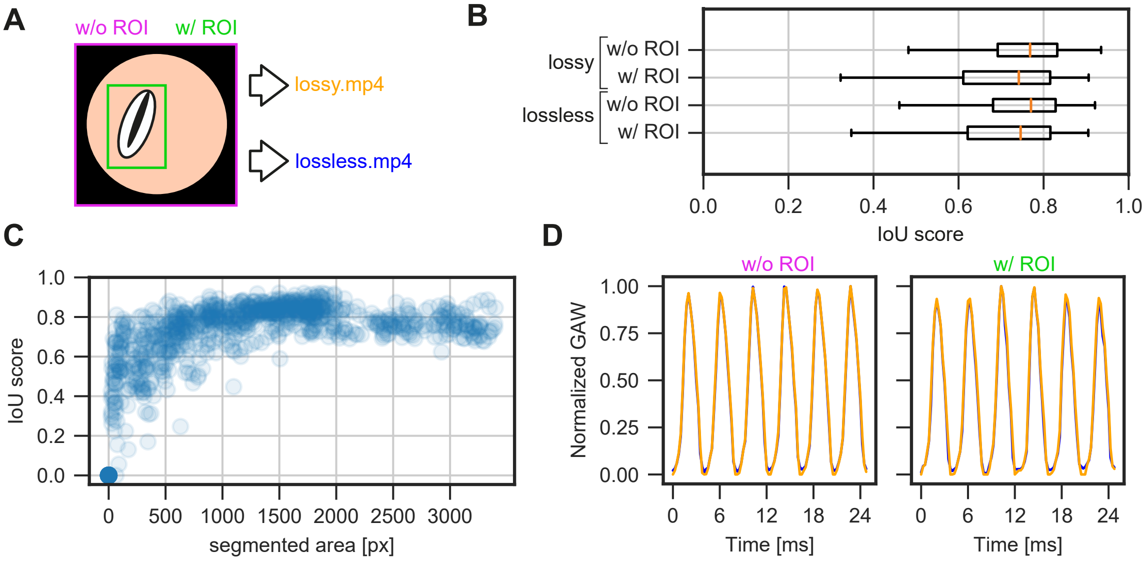

Figure 4: Lossy compression does not affect segmentation performance. A: Input data for DNN inference. Either full frame (magenta) or ROI-constrained (green) image data were used for inference. Image data were stored either lossless or lossy compressed as .mp4-files (see Methods). B: Intersection over Union (IoU) score per configuration shown in A. C: Dependency of IoU scores on the segmented glottal area. D: Exemplary glottal area waveform for full frame (w/o ROI) and ROI-based (w/ ROI) segmentations. Lossy and lossless compressed data are plotted in orange and blue, respectively.

a negligible effect on the data analysis. We next investigated if any configuration has an impact on the clinically relevant glottal area waveform (GAW) signal. We were able to confirm that all combinations despite their deviation in the IoU score have little to no affect on the GAW, as they do not deviate (Fig. S1A), and almost perfectly correlate (Fig. S1B), important for downstream computation of quantitative parameters.

3.3. Continual training for DNN fine-tuning improves performance

Despite the fact that we gained mostly successful and accurate segmentations, we found for a minority of videos (9%) two common issues: artefacts in the segmentation masks and empty segmentation masks (Fig. S2A). We hypothesized that continuous integration of new, system-specific data using continual training (Parisi et al., 2019; Lee and Lee, 2020) is increasing the DNN performance. Here, we evaluated two strategies for integrating new data (Fig. 5A). We either used a fixed time period or fixed quantities of videos. We used artefact-free, full-frame, lossy compressed videos for continual training, as full frames resulted in higher IoU scores (Fig. 4A). We found that all strategies were able to reduce the number of artefacts after only the first iteration of continual training and removed artefacts on average by 64-81% (Fig. 5B). Additionally, the more data is used for each training step, the higher the impact on artefact removal, and that a fixed quantity is preferable to

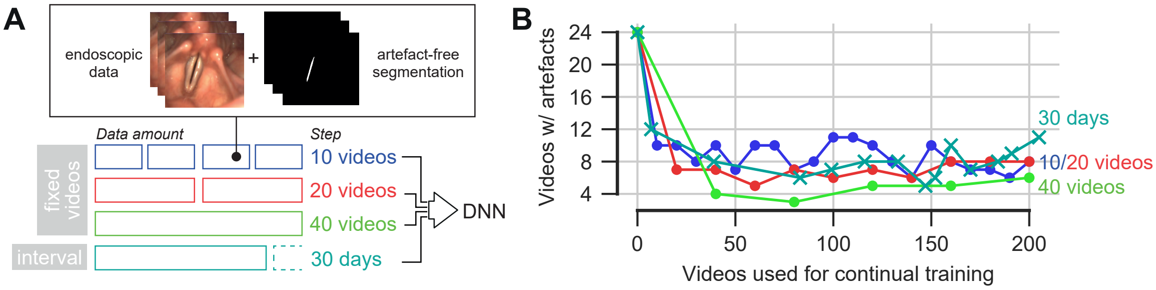

Figure 5: Continual training alleviates segmentation artefacts. A: Continual training strategies. B: Reduction of artefacts depending on continual learning strategy.

a fixed time period (Fig 5B). To further evaluate this, we annotated the first 30 frames of each artefact video to calculate the achieved IoU scores during continual learning (Fig. S3). Each strategy was able to increase the IoU score already after the first continual learning iteration. Looking at the strategy of using fixed time periods, it is clear that the IoUs do not change significantly during the first COVID-19 wave, which again shows that the use of fixed quantities is preferable to fixed time periods. Taken together, our data shows that the segmentation DNN is not only able to quickly adapt to new data, but also that continual learning is an important feature in using DNNs in a clinical context.

4. Discussion

Laryngeal high-speed videoendoscopy is a major tool in quantifying laryngeal physiology. In this study, we show that clinically optimized DNNs have an overall high performance due to a constantly high data quality and a well pre-trained, generalized DNN. Although the DNN has never be trained on this system's data, we still gain in most (91%) cases decent results. We further show that lossy compressed videos are on par with lossless compressed videos in terms of segmentation performance (Fig. 4). Using continuous integration of novel data, we can also show that the DNN is able to adapt such that previously identified artefacts are reduced and ensure a constantly improving segmentation environment (Fig. 5).

Other, especially offline imaging analysis platforms, such as the Glottis Analysis Tools (Kist et al., 2021b) (GAT), can serve as a reference for segmentation performance. The OpenHSV DNN has similar performance as the smallest GAT neural network, however, larger and more elaborate networks have superior performance (Tab. S1), that is largely compensated by our continual training scheme (Fig. 5). Nevertheless are the average IoUs obtained in this study (0.742-0.770) in an acceptable range that do not impact the clinical soundness of downstream quantitative parameter computation (Kist and Döllinger, 2020; Kist et al., 2021a). Our proposed continuous integration of more data is straightforward and effortless. However, we have not investigated how more sophisticated human-in-the-loop strategies (Budd et al., 2021), such as manual segmentation and retraining, perform. With a wider adoption of OpenHSV, we also believe that federated learning techniques (Xu et al., 2021) will further boost the DNN segmentation performance.

Acknowledgments

This work was supported by the German Research Council (DFG) under grant no. SCHU 3441/3-2.

References

Gustavo Andrade-Miranda, Yannis Stylianou, Dimitar D Deliyski, Juan Ignacio Godino-Llorente, and Nathalie Henrich Bernardoni. Laryngeal image processing of vocal folds motion. *Applied Sciences*, 10(5):1556, 2020.

Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71: 102062, 2021.

Dimitar D Deliyski, Pencho P Petrushev, Heather Shaw Bonilha, Terri Treman Gerlach, Bonnie Martin-Harris, and Robert E Hillman. Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. *Folia Phoniatrica et Logopaedica*, 60(1): 33–44, 2008.

Mona Kirstin Fehling, Fabian Grosch, Maria Elke Schuster, Bernhard Schick, and Jörg Lohscheller. Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional lstm network. *Plos one*, 15(2):e0227791, 2020.

Oliver Gloger, Bernhard Lehnert, Andreas Schrade, and Henry Völzke. Fully automated glottis segmentation in endoscopic videos using local color and shape features of glottal regions. *IEEE Transactions on Biomedical Engineering*, 62(3):795–806, 2014.

Pablo Gómez, Andreas M Kist, Patrick Schlegel, David A Berry, Dinesh K Chhetri, Stephan Dürr, Matthias Echternach, Aaron M Johnson, Stefan Kniesburges, Melda Kunduk, et al. Bagls, a multihospital benchmark for automatic glottis segmentation. *Scientific data*, 7 (1):186, 2020.

Andreas M Kist and Michael Döllinger. Efficient biomedical image segmentation on edgetpus at point of care. *IEEE Access*, 8:139356–139366, 2020.

Andreas M Kist, Stephan Dürr, Anne Schützenberger, and Michael Döllinger. Openhsv: an open platform for laryngeal high-speed videoendoscopy. *Scientific Reports*, 11(1):1–12, 2021a.

Andreas M Kist, Pablo Gómez, Denis Dubrovskiy, Patrick Schlegel, Melda Kunduk, Matthias Echternach, Rita Patel, Marion Semmler, Christopher Bohr, Stephan Dürr, et al. A deep learning enhanced novel software tool for laryngeal dynamics analysis. *Journal of Speech, Language, and Hearing Research*, 64(6):1889–1903, 2021b.

Melda Kunduk, Michael Döllinger, Andrew J McWhorter, and Jörg Lohscheller. Assessment of the variability of vocal fold dynamics within and between recordings with high-speed imaging and by phonovibrogram. *The Laryngoscope*, 120(5):981–987, 2010.

Max-Heinrich Laves, Jens Bicker, Lüder A Kahrs, and Tobias Ortmaier. A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *International journal of computer assisted radiology and surgery*, 14(3):483–492, 2019.

Cecilia S Lee and Aaron Y Lee. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, 2020.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5 (1):1–19, 2021.

Stephanie RC Zacharias, Dimitar D Deliyski, and Terri Treman Gerlach. Utility of laryngeal high-speed videoendoscopy in clinical voice assessment. *Journal of Voice*, 32(2):216–220, 2018.

Appendix A. Correlation and MAE

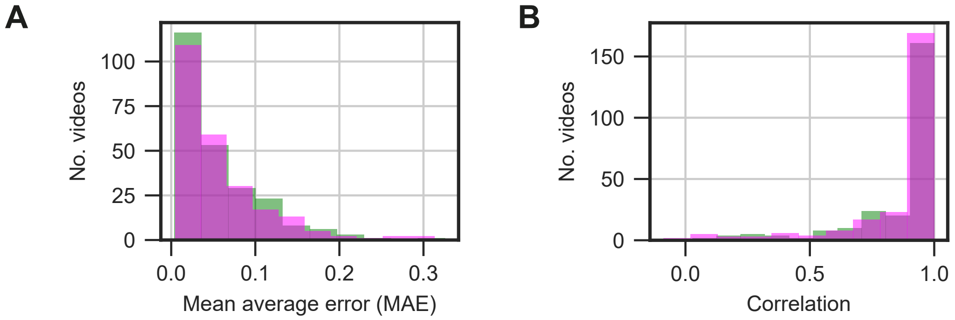

Figure S1: GAWs are highly correlated across compression methods. A: Distribution of mean absolute error (MAE) across compression modes segmented w/ ROI (green) or w/o ROI (magenta). B: Distribution of the Pearson's correlation coefficient across compression modes segmented w/ ROI (green) or w/o ROI (magenta).

Appendix B. Effect of continual learning

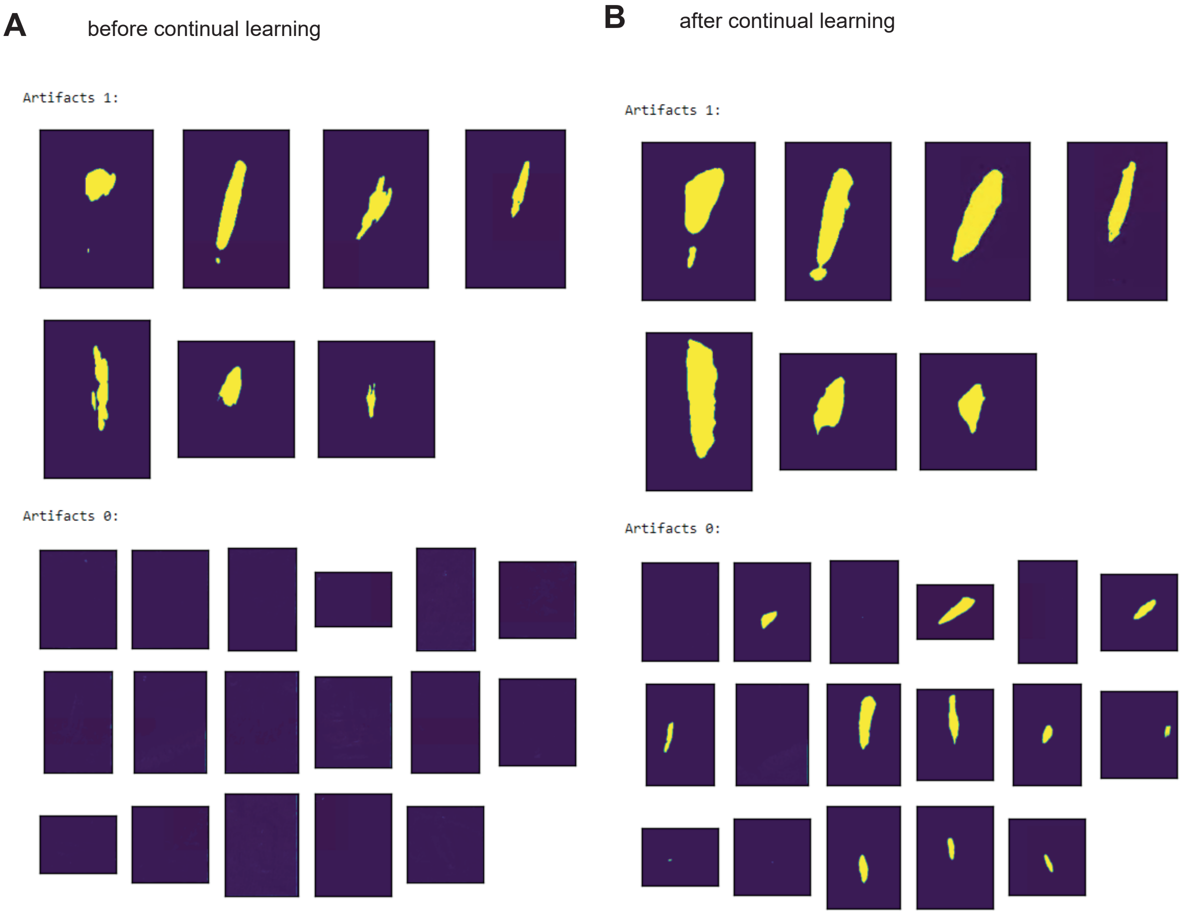

Figure S2: Continual learning has an effect on erroneous segmentations (upper panels) and on failed segmentations (lower panels).

Appendix C. Fixed data amount for continual training is better than fixed time intervals in terms of IoU.

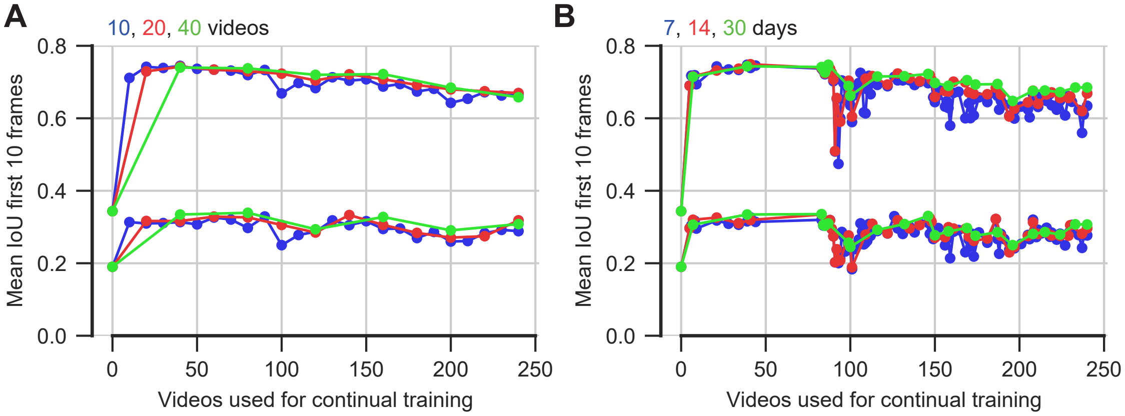

Figure S3: Average Intersection over Union (IoU) score of the first 30 frames relative to incorporated videos during continual training with (A) fixed data amount steps and (B) fixed time intervals.

Appendix D. Larger pre-trained DNNs perform slightly better in glottal segmentation before fine-tuning our model.

	Artefacts Rating 0		**Artefacts Rating 1**	
	# Artefacts	IoU	# Artefacts	IoU
Custom UNet	14/17	0.21	4/7	0.47
EfficientNetB0	10/17	0.35	1/7	0.57
ResNet50	8/17	0.45	2/7	0.54
Reference	17/17	0.19	7/7	0.34

Table S1: Number of Artefacts and Intersection over Union (IoU) for several deep neural networks.