A Deep Learning Approach To Vocal Fold Motion Tracking Under Partial Occlusion In Rats

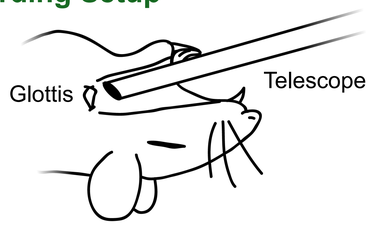
Vikram Ezhil, Robert A. Morrison[1,2], Ted Mau[2], Adrianna C. Shembel[1,2]

1. UT Dallas School of Behavioral and Brain Sciences, 2. UT Southwestern Department of Otolaryngology - Head and Neck Surgery

Introduction

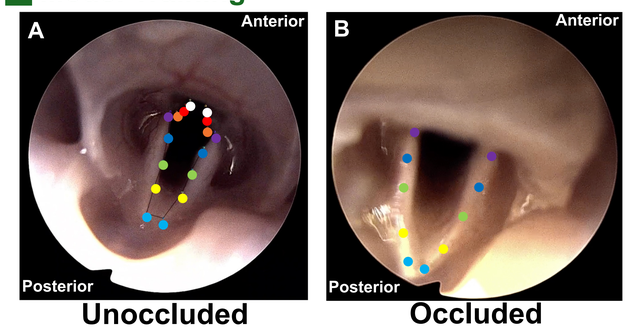
Vocal fold motion assessment relies on subjective clinical interpretation, leading to inter-observer variability and potential diagnostic inconsistencies. Quantitative approaches can improve reproducibility and speed of assessment through automation, but many current methods fail when visualization is impaired (due to epiglottic occlusion, etc.), a common clinical challenge. We developed a deep learning model that accurately tracks vocal fold motion even under partial occlusion, addressing a critical gap in quantitative vocal fold motion assessment.

Recording Setup



Laryngoscopy was performed on anesthetized rats placed in the supine position using a 30° 4mm rigid endoscope (Olympus WA96202A). Videos were recorded at 1920x1080p, 60 Hz, via an Ecleris HD camera processor and captured using an Elgato HD60 digital converter. The tongue was retracted using forceps to allow for better visualization of the glottis.

Model Training



Unoccluded **Occluded**

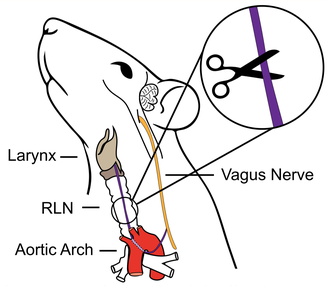
To develop a tracking model capable of handling occlusion, we utilized DeepLabCut. First, we extracted 20 nonconsecutive frames from 42 laryngoscopy videos encompassing both healthy and recurrent laryngeal nerve (RLN) injured rats with varying degrees of visual obstruction and vocal fold impairment. Each frame was manually annotated with up to 16 anatomical landmarks: five points along each arytenoid's medial border and three along each vocal fold edge (A). Occluded landmarks were intentionally left unlabeled to explicitly train the network to recognize and handle partial visibility of target structures (B).

The DeepLabCut neural network then underwent iterative refinement. Initial training on the base dataset of 1000 frames was followed by retraining on 160 frames where poor prediction confidence was identified, manually corrected, and reincorporated. This approach yielded a model that maintained tracking accuracy even when key anatomical structures were partially obscured.

The final model was visually inspected and displayed robust performance, successfully tracking vocal fold dynamics in conditions where portions of the vocal folds were obscured.

Model Application

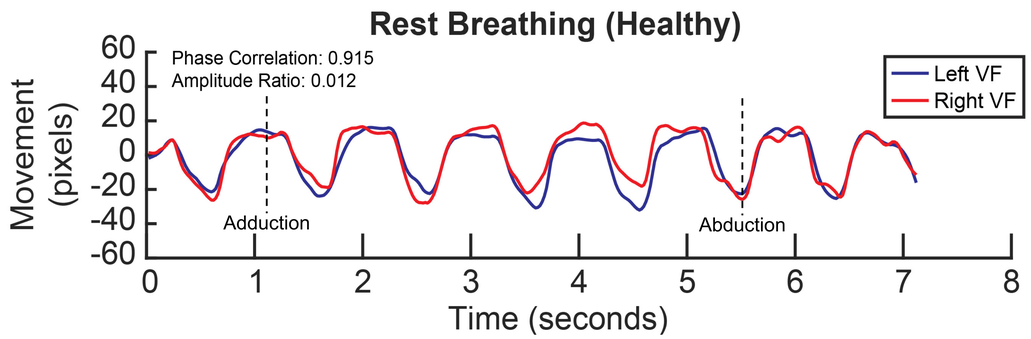
To evaluate the model's ability to effectively track and quantify vocal fold motion under a diverse range of conditions, we utilized a right-RLN injury rat model. 8 rats underwent complete right RLN transection at the level of the sixth tracheal ring, resulting in right vocal fold paralysis. Vocal fold motion under anesthesia during respiration was recorded via laryngoscopy and subsequently analyzed using the deep learning model at pre-injury and post-injury timepoints.

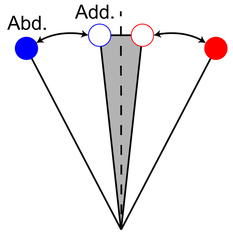


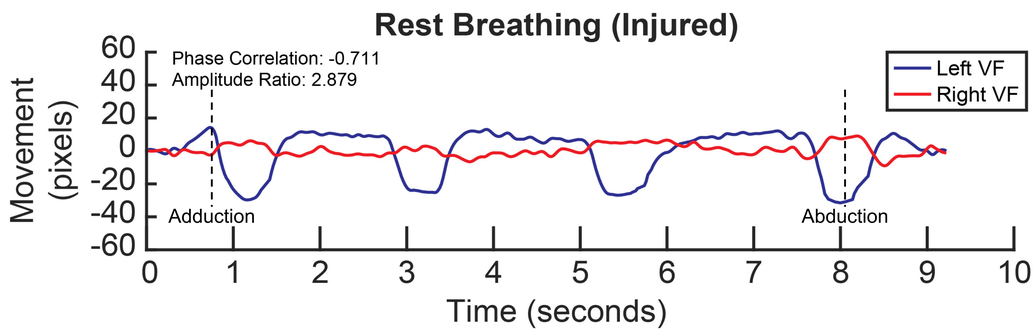
Using the model tracking output, for each recording we quantified two metrics:

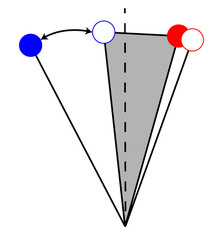
1) The relative direction of motion of the left and right vocal folds to each other (phase), where a value of +1 indicates healthy simultaneous adduction or abduction of the folds, and a value of -1 indicates paradoxical motion where one fold adducts while the other abducts
2) The relative magnitude of motion of the vocal folds to each other (amplitude), where a negative value indicates a left-sided paresis, a positive score indicates a right-sided paresis, and a value of 0 indicates equal distance travelled of the folds during movement.

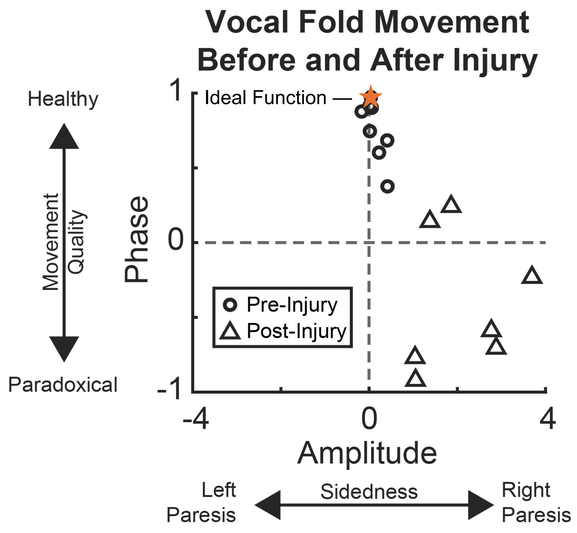
Model Tracking Results











Rats at the pre-injury timepoint cluster heavily near the ideal section of the graph representing healthy vocal fold movement where phase is +1, indicating simultaneous adduction and simultaneous abduction of the vocal folds over the course of the recording. Pre-injury, rats also have an amplitude metric of 0, indicating that the right and left vocal folds travel the same distance during adduction and abduction.

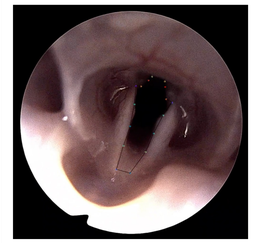
Rats at the post injury timepoint score near 0, indicating that the vocal folds do not move in unison either during adduction or abduction (i.e. one fold is immobile), or phase is negative, indicating paradoxical movement where one vocal fold abducts while the other adducts, moving towards the midline. Rats post-injury also universally display positive amplitude metrics, indicating that the right vocal fold travels a shorter distance during movement (i.e. right vocal fold paresis/paralysis).

Together, these results demonstrate that the model successfully captured and quantified a shift from pre-injury healthy movement to right vocal fold paralysis post-injury.

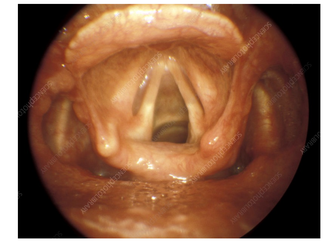
Conclusion

Our deep learning model successfully captured the shift from healthy vocal fold motion to vocal fold disfunction after recurrent laryngeal nerve transection. This change was captured under conditions where previous methods have struggled – in situations where target structures are occluded by the epiglottis, saliva, or other obstructions. Automated analysis holds promise as a more efficient means of assessing vocal fold function by reducing inter-observer variability and enabling consistent, reproducible evaluation. The model is capable of tracking vocal fold dynamics even in instances where visualization is impaired due to partial occlusion, thereby improving reliability in challenging scenarios.

Clinical Application



Rat



Human

The anatomical similarities between rat and human larynges facilitate straightforward model translation. Applying the model to human laryngoscopy could enable rapid integration into clinical practice.

References

1. Douglas C, Menon R, Montgomery J, et al. Vocal cord movement: can it be accurately graded? Ann R Coll Surg Engl. 2024;106(1):36-40. doi:10.1308/rcsann.2022.0091

2. Fehling MK, Grosch F, Schuster ME, Schick B, Lohscheller J. Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. Wang Y, ed. PLOS ONE. 2020;15(2):e0227791. doi:10.1371/journal.pone.0227791

3. Pennington-FitzGerald W, Joshi A, Honzel E, Hernandez-Morato I, Pitman MJ, Moayedi Y. Development and Application of Automated Vocal Fold Tracking Software in a Rat Surgical Model. The Laryngoscope. 2024;134(1):340-346. doi:10.1002/lary.30930

4. Adamian N, Naunheim MR, Jowett N. An OPEN-SOURCE Computer Vision Tool for Automated Vocal Fold Tracking From Videoendoscopy. The Laryngoscope. 2021;131(1). doi:10.1002/lary.28669

5. Wang YY, Hamad AS, Palaniappan K, Lever TE, Bunyak F. LARNet-STC: Spatio-temporal orthogonal region selection network for laryngeal closure detection in endoscopy videos. Comput Biol Med. 2022;144:105339. doi:10.1016/j.compbiomed.2022.105339

6. Mathis A, Mamidanna P, Cury KM, et al. DeepLabCut: markerless pose estimation of user- defined body parts with deep learning. Nat Neurosci. 2018;21(9):1281-1289. doi:10.1038/s41593-018-0209-y

7. Hernández-Morato I, Valderrama-Canales FJ, Berdugo G, et al. Reorganization of laryngeal motoneurons after crush injury in the recurrent laryngeal nerve of the rat. J Anat. 2013;222(4):451-461. doi:10.1111/joa.12031

8. Mor N, Naggar I, Das O, et al. Quantitative Video Laryngoscopy to Monitor Recovery from Recurrent Laryngeal Nerve Injury in the Rat. Otolaryngology Neck Surg. 2014;150(5):824-826. doi:10.1177/0194599814521572

Ethics Statement

All procedures involving animals in this study were reviewed and approved by the Institutional Animal Care and Use Committee (IACUC) of University of Texas at Dallas (IACUC protocol number: 21-09).