UCLA
UCLA Previously Published Works

Title

Correlating Perceptual Voice Quality in Adductor Spasmodic Dysphonia With Computer Vision Assessment of Glottal Geometry Dynamics.

Permalink
https://escholarship.org/uc/item/1r85r7s2

Journal
Journal of Speech Language and Hearing Research, 65(10)

ISSN
1092-4388

Authors
Peterson, Quinn A
Fei, Teng
Sy, Lauren E
et al.

Publication Date
2022-10-17

DOI
10.1044/2022_jslhr-22-00053

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Research Article

Correlating Perceptual Voice Quality in Adductor Spasmodic Dysphonia With Computer Vision Assessment of Glottal Geometry Dynamics

Quinn A. Peterson,[a] Teng Fei,[b] Lauren E. Sy,[b] Laura L.O. Froeschke,[c] Abie H. Mendelsohn,[d] Gerald S. Berke,[d] and David A. Peterson[e]

[a] Department of Computer Science and Software Engineering, California Polytechnic State University, San Luis Obispo [b] Department of Cognitive Science, University of California, San Diego, La Jolla [c] Department of Communication Sciences and Disorders, Elmhurst University, IL [d] Department of Head and Neck Surgery, David Geffen School of Medicine, University of California, Los Angeles [e] Institute for Neural Computation, University of California, San Diego, La Jolla

ARTICLE INFO

Article History:
Received January 24, 2022
Revision received May 5, 2022
Accepted June 18, 2022

Editor-in-Chief: Cara E. Stepp
Editor: Susan L. Thibeault

https://doi.org/10.1044/2022_JSLHR-22-00053

ABSTRACT

Purpose: This study examined the relationship between voice quality and glottal geometry dynamics in patients with adductor spasmodic dysphonia (ADSD).
Method: An objective computer vision and machine learning system was developed to extract glottal geometry dynamics from nasolaryngoscopic video recordings for 78 patients with ADSD. General regression models were used to examine the relationship between overall voice quality and 15 variables that capture glottal geometry dynamics derived from the computer vision system. Two experts in ADSD independently rated voice quality for two separate voice tasks for every patient, yielding four different voice quality rating models.
Results: All four of the regression models exhibited positive correlations with clinical assessments of voice quality (R^2s = .30–.34, Spearman rho = .55–.61, all with $p < .001$). Seven to 10 variables were included in each model. There was high overlap in the variables included between the four models, and the sign of the correlation with voice quality was consistent for each variable across all four regression models.
Conclusion: We found specific glottal geometry dynamics that correspond to voice quality in ADSD.

Adductor spasmodic dysphonia (ADSD) is the most common subtype of laryngeal dystonia. ADSD is characterized by task-specific phonatory breaks that vary according to speaking context (Hintze et al., 2017). One of the goals of spasmodic dysphonia (SD) treatment is to improve patients' overall voice quality (A. Hu et al., 2016). The standard of care for SD is botulinum neurotoxin (Botox) injections into the intrinsic muscles of the larynx. The effectiveness of Botox treatment in alleviating ADSD symptoms and improving voice quality is commonly tracked and measured by various subjective assessments of severity. These subjective techniques include clinical auditory assessments (Cannito et al., 2004), clinical visual assessments, and patient self-reports. The clinical auditory assessments are the most common but are multidimensional and exhibit high interrater variability attributed to difficulties isolating individual attributes in voice patterns and instability of raters' internal standards for different voice qualities (Kreiman et al., 2007). Assessing ADSD is also multidisciplinary, involving ear-nose-throat (ENT) physicians, speech-language pathologists, and neurologists. The subjective, multidimensional, and multidisciplinary aspects of severity assessment decrease intra- and interrater reliability (Ludlow et al., 2018). An increase in the reliability of ADSD severity ratings is imperative to understanding ADSD pathophysiology and for assessing treatments intended to improve voice quality.

To improve the reliability of severity ratings, computational analyses can be used to objectively quantify

Correspondence to David A. Peterson: qap2001@gmail.com. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

voice quality. Prior studies have done automated acoustic analysis to measure dysphonia severity in, for example, connected speech and sustained vowels (Roy et al., 2014). Although the automated acoustic approach is objective, there are mixed results regarding its relationship to subjective assessments. The automated acoustic approach has been found to have poor correlation with patient self-reports (Fujiki & Thibeault, 2021; Gillespie et al., 2014) and expert perceptual evaluation (Fujiki & Thibeault, 2021; Rabinov et al., 1995). Conversely, it has also been found to have a strong correlation with auditory perceptual ratings of overall ADSD severity (Buckley et al., 2020) and to be significantly correlated with listener ratings (Roy et al., 2014). Regardless, these approaches do not directly evaluate one of the primary structures responsible for voice quality: the larynx.

Furthermore, many studies have applied machine learning methods to patients with SD. These methods include using machine learning on brain imaging to predict the risk of SD development (Khosravani et al., 2021), using an artificial neural network on a high-dimensional features space from audio recordings to rate SD voice severity (Suppa et al., 2020), using a support vector machine, a Gaussian mixture model, and a deep neural network on high-dimensional acoustic features to distinguish between normal voice and several voice disorders including SD (Fang et al., 2019), and using a convolutional neural network on spectrograms from voice recordings to distinguish between normal voice and several voice disorders including SD (Reid et al., 2022). None of these approaches use images of the larynx.

ENT physicians regularly view the larynx using nasolaryngoscopy. In laryngeal video recordings, several computer vision and machine learning methods have been used to locate and identify the boundaries of the glottis and/or vocal folds, a process known as segmentation. Once segmented, various geometric measures that characterize the shape can be easily computed. These segmentation methods have included edge detection and region growing algorithms (Alku et al., 2019; Chen et al., 2013), pixel thresholding (Wurzbacher et al., 2006), a watershed transform and a linear predictor (Osma-Ruiz et al., 2008), a convolutional neural network–based semantic segmentation (Laves et al., 2019), and a deep convolutional long short-term memory network (Fehling et al., 2020). Many of these studies use high-speed video (HSV) recordings of the larynx. Because these HSV recordings typically have frame rates between 1000 and 10000 Hz (Gómez et al., 2020), which are well above the Nyquist rate for speech fundamental frequencies in the range of 50–300 Hz (Osma-Ruiz et al., 2008), they can register the whole vibratory cycle of the larynx. However, because of the higher frame rate, studies that use HSV can produce large quantities of data. This may be why they tend to analyze less than a second of video data per patient. These short time periods capture many vibratory cycles of the larynx but are not long enough to capture phonation of several syllables or a sentence. As a result, many studies that use HSV of the larynx focus solely on the segmentation performance of their computational method and not data analyses relevant to voice disorders (Fehling et al., 2020; Laves et al., 2019).

Although several methods exist to segment the glottis and/or vocal folds, an emphasis on the glottis is critical to study the voice. Although the voice is complex and there are many factors that influence it, the dominant factor that determines voice quality is the shape of the glottis. This is because the pressure wave produced by the larynx is a function of the glottal opening during vibration, and therefore, sound characteristics are more closely related to dynamics of the glottal opening than to dynamics of the vocal folds or other surrounding structures. Also, for many years, voice scientists have used characteristics of the glottal opening to describe resultant sound characteristics, such as the open quotient, the speed quotient, and the Liljencrants–Fant model, which consists entirely of measurements of changes in the glottis. An additional reason to focus solely on the glottis and not also the vocal folds is that computational segmentation of the glottis is much easier than the vocal folds. This is because the vocal folds, relative to the glottis, are more frequently occluded, have less color gradient along their edges, and have more variability in many attributes such as size, shape, and color (Fehling et al., 2020). Also, the false vocal folds often obscure the true vocal folds in patients with ADSD.

In this study, we analyzed clinically relevant tasks in laryngoscope videos previously recorded at 30 Hz from 78 patients with ADSD across multiple sites affiliated with the Dystonia Coalition (Kilic-Berkmen et al., 2021; Ludlow et al., 2018). The goals of this study were (a) to develop a computer vision/machine learning system that automatically provides an objective visual assessment by segmenting the glottis and calculating associated glottal geometries and their dynamics, (b) to evaluate the system's correlation with subjective voice quality ratings (VQRs), and (c) to identify dynamics of glottal geometries associated with pathological voice quality.

Materials and Method

Patients

Patients with SD were recruited and evaluated at four sites affiliated with the Dystonia Coalition (Kilic-Berkmen et al., 2021; Ludlow et al., 2018): James Madison University, Emory University, Washington University in St. Louis, and Medical College of Wisconsin. All patients provided written informed consent prior to participating in the original study conducted in accordance with the Declaration of Helsinki. The patients were assessed at least 2 months after their last Botox injections, by which time much of the effect would have worn off. The selection criteria were as follows: age of 18 years and older; no medical condition precluding nasolaryngoscopy; no known cause for the voice disorder; no lesions; no surgery for SD; and diagnosis of ADSD, abductor spasmodic dysphonia, voice tremor, muscle tension dysphonia (MTD), vocal fold paralysis, or other voice disorders. The data in this study included a subset of 91 patients who were diagnosed with ADSD by the clinicians at the respective sites. Every patient was asked to do a sequence of speech tasks during their laryngoscope video recording (Yan et al., 2015).

The voice manifestations of ADSD have been shown to be different for task-oriented phonation versus meaningful propositional speech (Somanath & Mau, 2016). Therefore, we analyzed two steps in the video procedure specifically designed to evoke symptoms of ADSD, the sustained vowel /i/ (longE), and a sentence with glottal stops, "we eat eels everyday" (sentA). One trial of each step was obtained per patient.

Of the 91 patients, four were missing video recordings, four were missing both the longE and sentA step, and five were excluded because of comorbid voice tremor (per Berke and Froeschke), leaving 78 final patients (see Table 1). This is a suitable set, given that prior studies that use machine learning on patients with SD have included a subset of 17 patients with SD (Fang et al., 2019), a subset of 30 patients with ADSD (H.-C. Hu et al., 2021), and a total set of 60 patients with ADSD (Suppa et al., 2020). We created subsets from these 78 patients based on whether the videos contained certain procedural steps. There were 73 videos that contain a longE, 70 videos that contain a sentA, and 65 that contain both a longE and a sentA.

Annotations and VQRs

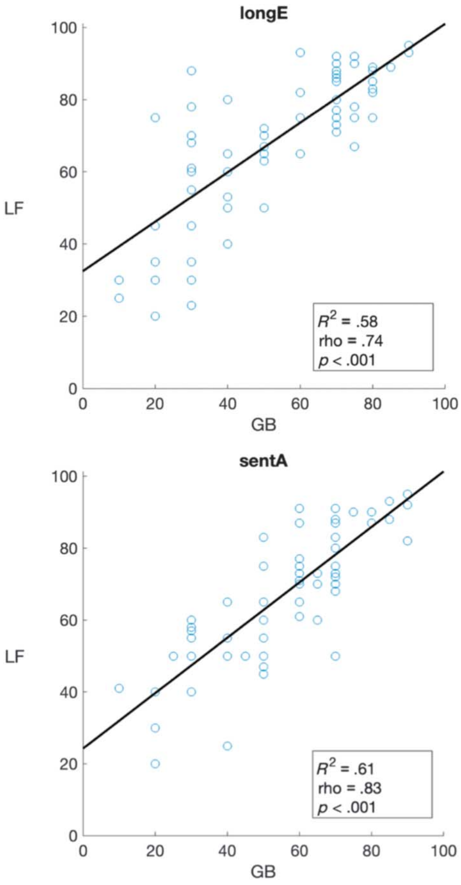
An individual (Sy) blinded to the algorithm and the VQRs labeled the time periods of the longE and sentA steps in patients' videos with ELAN linguistic software (ELAN, Version 4.9.4, Max Planck Institute for Psycholinguistics, 2016). Some patients have multiple occurrences of a given step. For example, a patient might pronounce sentA twice instead of just once. For each patient, Sy annotated all occurrences of the longE and sentA steps.

Table 1. Patient demographics.

Variable	Range	*M*	*SD*
Age (years)	27–77	60.78	10.33
Disease duration (years)	0–42	15.11	10.11
Sex (female/male)		56/17	

Two nationally recognized physician experts in the field of SD (Berke [Rater 1] and Froeschke [Rater 2]) independently rated voice quality for the first occurrence of each patient's longE and sentA steps. Figure 1 shows the correlation between the raters for each step. Whereas the interrater reliability is strong, no intrarater reliability was assessed. This is suitable given the focus of this study, which is concerned with relating clinical ratings to objective computer measures rather than evaluating reliability of clinical rating assessments themselves. The raters were blinded to the video channel of the recording and only heard the audio channel. They were allowed to listen to each step multiple times and were subject to no particular order for rating the two steps. They rated the voice quality on a scale of 0–100, where 0 is a minimally usable voice (worst voicing) and 100 is normal voice (best voicing). The scale is based on and similar in multiple regards to the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V), which has been used in assessment of ADSD (A. Hu et al., 2016). The CAPE-V scale

Figure 1. Correlations between raters' (GB and LF) voice quality ratings for the two different time steps. Each data point represents a patient. rho = Spearman rho.



uses a range from 0 to 100. Also, the scale is based on assessment of the following well-defined set of perceptual attributes: overall severity, roughness, breathiness, strain, pitch, and loudness. Similarly, the two raters performed auditory-perceptual evaluation of the voice by listening for overall voice quality, as a gestalt, encompassing roughness, breathiness, strain, pitch, loudness, voice breaks, and other characteristics. It should be noted that these individual characteristics are likely not independent; they are substantially overlapped in SD in terms of both their presence and their response to treatment. Also, steady-state long vowel phonations do not normally elicit voice breaks but rather can be best assessed from loudness, hoarseness, roughness, strain, and so forth. Meanwhile, sentences with many vowels, such as sentA, tend to include voiceless fricatives and plosives, which induce voice breaks. Thus, although both steps were graded by a gestalt of voice quality, the raters focused more on perception of patient strain and voice harshness in the longE step and focused more on a composite multifactorial impression of impaired fluency associated with either spasmodic voice breaks or voice quality that is strained, strangled, and/or stressed in the sentA step.

We obtained frame annotations of the glottis. For a frame where the glottis is visible, its glottis mask is a labeled region of its glottal opening. Although one might expect the glottis to always be visible during certain phonatory steps, the glottis is frequently out of view in many videos. The glottis visibility is often affected by camera orientation changes and/or obstructions from surrounding structures such as the epiglottis or false vocal folds. An individual (T.F.), blinded to the algorithm and the VQRs, annotated the videos by labeling each frame with the MATLAB R2020b (MathWorks) Video Labeler. We used a three-phase approach to maximize reliability of these annotations. In the training phase, Q.A.P. trained T.F. on the glottis mask annotations, checking the accuracy of the annotations and critiquing errors on random subsets of training data. Once confident in T.F.'s ability to accurately perform the annotations, in the subsequent phase with the data ultimately used in our system, validation was performed by Q.A.P. in which random frames were selected and inspected for accuracy. If any frame was inaccurate, Q.A.P. reviewed how to accurately annotate the frame with T.F., and the patient's step was redone and rechecked. Finally, we used MATLAB to reassemble the phonatory step videos, with each frame's glottis mask annotation overlaid. This was done for all patients and for all annotated frames and facilitated a quick visual assessment of the accuracy of all the glottis annotations. There were three classes of annotations for each frame: (a) glottis visible: Label one point at each of the top left of the visible glottis, at the bottom of the visible glottis, and at the top right of the visible glottis to define the glottis mask (this triangle mask is used as opposed to a more precise outline for annotation efficiency; in cases of a very narrow glottal opening, the triangle approximates a line segment subtending the major axis of the glottis); (b) unsure: Label 2 points at the top right corner of the frame; and (c) glottis not visible: Label 2 points at the top left corner of the frame. Glottis masks were only annotated within the first instance of a given step for patients that had several occurrences of the given step. The approximate range of duration for the sentA step is 1–3 s, and that for the longE step is 3–10 s. Given that the nasolaryngoscopic recordings are at 30 Hz, this yields an approximate range of 2,340–7,020 and 7,020–23,400 frames annotated for the sentA and longE steps, respectively.

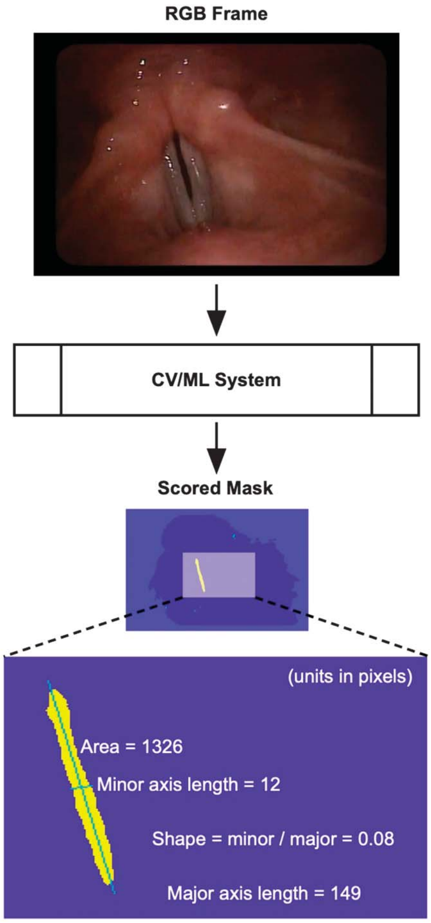
Computer Vision/Machine Learning System for Deriving Glottal Geometry

We developed a computer vision/machine learning system to calculate glottal geometries (for more detail, see the Appendix). As shown in Figure 2, the system's input is a nasolaryngoscopic image, and its output is a segmented glottis and glottal geometries. Within the system is a sequence of steps. First, the system uses a shallow neural network to predict a luminance threshold for the given frame. For further clarification, a shallow neural network is a computer neural network with a single layer of nodes or, essentially, a single layer of artificial neurons. Like most artificial neural networks, during its training, it learns a mapping from inputs to outputs. Once trained, the network outputs predictions for a given input, based on the mapping it learned during training (again, for more detail, see the Appendix). Second, the system uses this luminance threshold to segment dark regions from within the frame. Essentially, the threshold bifurcates pixels within the frame into those with value greater than the threshold and those with value less than the threshold. Third, the system uses another shallow neural network to score the segmented regions and identify the glottis. We used the glottis mask annotations to train both of the shallow neural networks. Finally, the system calculates geometries from the segmented glottis region such as area and axis lengths.

Validation of the system's accuracy was not performed at the precise pixel level outline of the glottis because obtaining these for each frame would be extremely tedious. However, our system was validated against our glottis masks, which provided ground truth for the neural networks used in the system to identify the presence, location, and orientation of the glottis.

It should be acknowledged that variation in the distance between the scope and the glottis during recording has a direct effect on measures such as area of the glottis. However, within a recording's step, any variation in the scope–glottis distance is largely accounted for because our

Figure 2. An example input and output of the computer vision/machine learning (CV/ML) system. The input is a frame from a nasolaryngoscopic video, and the output is a segmentation of the glottis with calculated geometries. (For a more detailed visualization of the system, see Figure A1 in the Appendix.)



metrics are normalized (see the Metrics and Models section below) both spatially and temporally.

Metrics and Models

We derived two types of metrics from a single frame. The first is the confidence that the frame contains the glottis (glottis_conf). The second type are geometric features of the glottis, what we refer to as glottal geometries: area, minor and major axis lengths, and shape (for more detail, see the Appendix). In total, this yields five metrics for each frame. For each of these metrics, we computed nonparametric statistics that characterize average values and temporal variability across a patient's frames: median, median absolute deviation (mad), and median absolute deviation normalized by the median (madm). Table 2 shows that crossing the glottal metrics with the statistics yields 15 total variables (hereafter referred to in **bold**) that are calculated separately for each patient's longE and sentA steps. Collectively, because these variables encapsulate information about changes in glottal attributes over time, they represent what we refer to as "glottal geometry dynamics."

To assess which variables are related to ADSD voice quality, we used JMP Pro Version 16 statistical software from SAS to build general regression models using adaptive elastic nets with Akaike information criterion stopping criterion. These models aimed to predict the VQRs using the glottal variables. We created four models for our four VQRs, which will be referred to as the GB longE model, LF longE model, GB sentA model, and LF sentA model. For clarity, each model name refers to both the rater and the step from which the ratings were done. For example, the GB longE model is the model that correlates with the ratings done by G.S.B. for the longE step. We kept the models independent for the two raters and for the two steps to see which models best predicted the scores based on the glottal geometry dynamics.

Results

Figure 3 shows the linear regressions of the four predictor models. The R^2s ranged between .30 and .34, and the Spearman rho ranged between .55 and .61, all with $p < .001$.

Table 3 shows which variables were included in each of the four models and whether their coefficients were positive or negative, as well as their corresponding *p* values. Of the 15 total variables entered into each of the four models, the number of variables that were selected in each of the four models were seven in GB longE, eight in LF longE, eight in GB sentA, and nine in LF sentA. The majority of each models' subset of chosen variables were similar across models. All of the variables included in the GB longE model were also included in the LF longE model, and only one variable was unique to the LF longE model between the two models: **madm_minor**. Similarly, most of the variables included in the GB sentA model and the LF sentA model were similar, where **median_minor** was the only variable unique to the GB sentA model and **mad_glottis_conf** and **median_area** were the only variables unique to the LF sentA model.

The top 6 rows of Table 3 represent the six variables that were included in all four predictor models. The coefficient sign for each variable was consistent across all four models. For the five variables that had consistently significant *p* values, as the quality of the voice decreased:

- **madm_glottis_conf**, the normalized variability in the glottis visibility, *decreased*, that is, less variability when the glottis is visible correlates with a worse voice quality.

Table 2. Model variables.

Variable	Median	Median absolute deviation (mad)	Median absolute deviation from the median (madm)
glottis_conf (glottis confidence)	**median_glottis_conf**: the median visibility of the glottis	**mad_glottis_conf**: the variability in the visibility of the glottis	**madm_glottis_conf**: the normalized variability in the visibility of the glottis
shape (minor/major)	**median_shape**: the median shape of the glottis	**mad_shape**: the variability in the shape of the glottis	**madm_shape**: the normalized variability in the shape of the glottis
major	**median_major**: the median major axis length of the glottis	**mad_major**: the variability in the major axis length of the glottis	**madm_major**: the normalized variability in the major axis length of the glottis
minor	**median_minor**: the median minor axis length of the glottis	**mad_minor**: the variability in the minor axis length of the glottis	**madm_minor**: the normalized variability in the minor axis length of the glottis
area	**median_area**: the median area of the glottis	**mad_area**: the variability in the area of the glottis	**madm_area**: the normalized variability in the area of the glottis

- **median_shape**, the median value of the shape, *increased*. Because shape is defined as the minor axis length over the major axis length, this means that as shape increases, the minor axis length must be coming relatively closer to the major axis length; the glottis object is becoming more circular and less elongated, that is, a more circular and less elongated glottis correlates with a worse voice quality.
- **mad_shape**, the variability in the general shape of the glottis, *decreased*, that is, less variability in the shape of the glottis correlates with a worse voice quality.
- **mad_minor**, the variability in the minor axis length of the glottis, *decreased*, that is, less variability in the approximate glottis width correlates with a worse voice quality.
- **madm_area**, the normalized variability in area, *increased*, that is, more variability in glottis area correlates with a worse voice quality.

Discussion

We developed a computer vision/machine learning system that automatically provides an objective visual assessment of ADSD severity by calculating glottal geometry dynamics. We found reasonable correlation between these metrics and expert clinical assessments. Finally, as described below, analyses of these models are informative of potential relationships between glottal geometry dynamics and pathological voice quality.

Relevance for Rating Voice Quality in ADSD

The system's agreement with expert clinical perceptions suggests that, once further validated, the system could be combined with acoustic analyses in an overall objective system to rate ADSD severity. These objective measurements of severity could be used to compare treatments across centers, which would otherwise suffer from interrater variability associated with subjective perceptual ratings. Furthermore, the system could act as a supplement to patient self-perceptual ratings in quantifying patient response to Botox over time. Indeed, in our post hoc analysis, glottal geometry dynamics did not correlate with patient self-reports, suggesting that our system provides information that's complementary to rather than redundant with patient impressions of their severity. Ultimately, the system could also be implemented as a network-based software service accessible through either a web- or app-based interface. The interface could accommodate simultaneous uploads of clinical and/or patient perceptual assessments and provide a consolidated report back to the clinician.

The computer vision/machine learning system is efficient and flexible. It is computationally efficient as it can process an approximately 500-kB image in an average of 0.186 s (with a MacBook M1 with 16-GB RAM). It is flexible as it only requires a single RGB image as input. This means that the system can be used to analyze laryngeal images from almost any type of recording equipment, regular speed, or HSV. Furthermore, the exact size of the image and the phonatory step within the laryngoscope video do not matter. On the contrary, many computer vision/machine learning algorithms that aim to segment the glottis or vocal folds have several requirements, such as a specific image resolution (usually around 256 × 256 pixels; Alku et al., 2019; Fehling et al., 2020; Wurzbacher et al., 2008), the high frame rates associated with HSV, and sometimes manual intervention to handpick frames that depict the glottis or certain laryngeal characteristics. Our system is not limited by strict specifications regarding video quality or time within a nasolaryngoscopic video. The flexibility of the system differentiates it from other computer vision/machine learning systems that have been developed for the glottis and/or vocal folds. Also, this

Figure 3. Relationship between model predictions and clinical ratings of voice quality. Each scatter plot represents a different model (a model built to predict the given rating). For example, the bottom scatter plot represents the model built to predict LF's sentA voice quality ratings. Each data point represents a patient. VQR = voice quality rating; rho = Spearman rho.

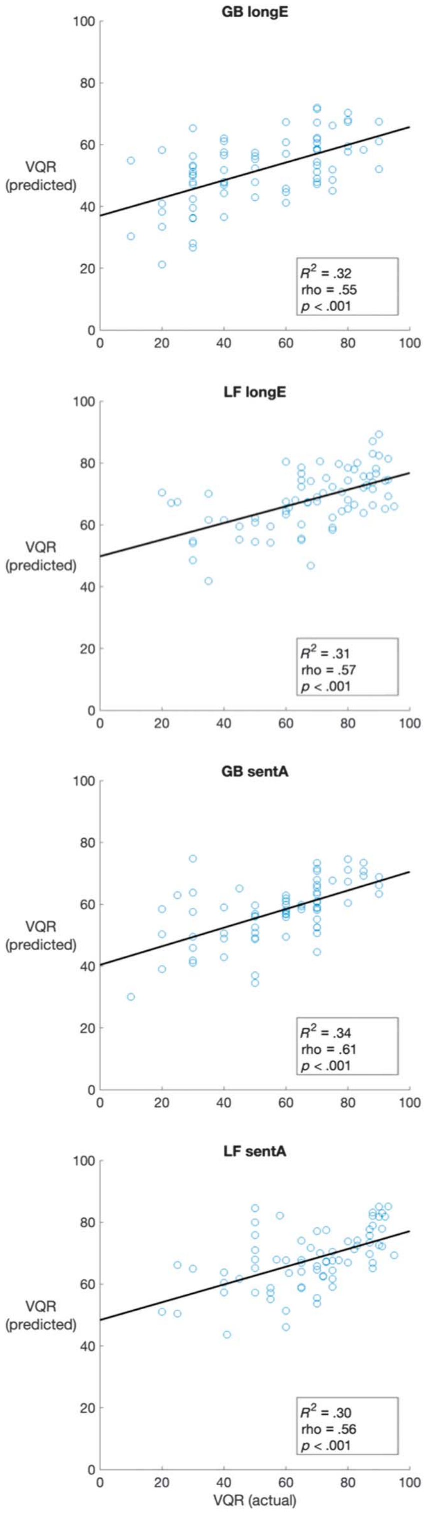


Table 3. Metrics and their *p* values.

	longE		**sentA**	
Metrics	**GB**	**LF**	**GB**	**LF**
Madm_glottis_conf	< .001	< .001	< .001	.004
Mad_shape	< .001	.011	.012	.032
Madm_major	.122	.294	.189	.086
Mad_minor	< .001	< .001	< .001	< .001
Median_shape	< .001	.004	.004	012
Madm_area	.001	.070	.012	.008
Mad_glottis_conf	/	/	/	.402
Median_minor	/	/	.017	/
Madm_minor	/	.434	.328	.211
Median_area	.028	.206	/	.012

Note. Each column is one of the four models. **bold** = positive correlations; *italic* = negative correlations; / = variable not in the model.

flexibility better facilities study of the voice with respect to the glottis, as the system can be applied at any level of temporal granularity and at virtually any decent level of video quality. Furthermore, as referenced in the introduction, many studies that use machine learning with respect to SD analyze audio data rather than videos, and many studies that use machine learning with respect to the larynx focus solely on segmentation performance and not also on voice disorders. Thus, our system is unique, as it bridges the gap between the development of computer vision/machine learning of the larynx and the study of ADSD.

Identifying Glottal Geometry Dynamics Associated With Voice Quality in ADSD

The models' R^2s, which are around .3, are reasonable given variable laryngoscope frame borders, variable focus quality, obstructions of the glottis or vocal folds, saliva on the lens, and so forth. Image segmentation of the larynx is challenging due to many factors, such as camera rotation, movement of the laryngoscope, movement of the patient, and variable illumination (Osma-Ruiz et al., 2008).

Table 3 shows that each included variable had a consistent correlation sign across all four models. This consistency across models lends support to each variable's implied relationship with voice quality, whether it be a positive or negative correlation. Four of the five variables that played a significant role across all four of the models had meaningful relationships with voice quality. Generally, these relationships are consistent with the possibility that a persistently tight configuration of the glottis correlates with poorer voice quality. As voice quality decreases:

- There is less variability in when the glottis is visible. In more severe patients, this may reflect reduced ability of the true folds to transiently open the glottis.

This may also reflect reduced activity in superglottal structures, such as the false vocal folds and the epiglottis, which could otherwise provide compensatory mechanisms that ameliorate pathological voice.

- The glottis is more circular and less elongated. This suggests that patients' with a tight configuration in both dimensions tend to have poorer voice.
- There is less variability in the shape of the glottis. This suggests that patients with a persistently tight or dysphonic larynx have less change in the shape of the glottis. On the contrary, patients with a better voice may be better able to go between normal and tight configurations.
- There is less variability in the width of the glottis. This suggests that less variability in the glottal width dynamics is associated with poorer voice.

The one consistently significant variable whose results contradict the trends found above is the normalized variability of the glottal area (**madm_area**), for which an increase in variability in the glottal area correlates with poorer voice. This contradictory trend can be partially explained by the fact that the glottal area measure is not directly correlated with the glottal axis length or shape measures. For example, the area of an object could change, whereas the axis lengths and the shape of the object both remain fixed. Furthermore, this contradictory trend might be better understood by investigating temporal glottal metrics, such as the glottal area waveform or the open quotient (Chen et al., 2013; Fehling et al., 2020; Wurzbacher et al., 2008). Relating these temporal glottal metrics to voice quality could give more insight into how dynamics of the glottal configuration correlate with voice quality.

The high overlap among variables chosen for each model across the two steps (longE and sentA) suggests some similarity in the glottal dynamics associated with ADSD severity between the two steps. The expert raters aimed to rate a patient's two steps independently, but subconscious bias from rating a longE step could have affected the rating of the same patient's subsequent sentA step. This step rating procedure could have increased the similarity between the steps' VQRs and therefore could help explain the similarity between the different steps' predictor models. Although the potential for rating bias from one step to the next is noted, the perceptual assessment process used in this study is consistent with standard clinical procedures, which incorporate a set of non–speech-based and speech-based tasks that a patient performs in the same evaluation session and are evaluated by the same examiners.

ADSD is classified as a task-specific focal dystonia (Hintze et al., 2017; Ludlow, 2011; Roy et al., 2005). Various vocal tasks may have different likelihoods of eliciting spasmodic overclosure of the vocal folds (Roy et al., 2005). Anecdotal reports suggest that sustained vowels evoke less symptoms than contextual speech (Ludlow & Connor, 1987). Given this, one might expect a large difference between the predictor models for each of the two steps: the sustained vowel /i/ (longE) and the contextual phrase "we eat eels every day" (sentA). Contrary to this, the similarities found between the models that predict the VQRs for the separate steps suggest that the ADSD symptoms, which are task specific, may only partially contribute to an overall assessment of severity. It should be further noted that the speaking condition of oral sentence repetition ("we eat eels everyday"), though a type of running speech, does not simulate the same type of task demands associated with spontaneous, meaningful, speaker-composed speech in daily life. As both the longE and sentA tasks are phonatory tasks that are dissociated from purposeful and propositional communication, speech conditions and types may help further explain model similarity in the current study.

Glottal geometry dynamics and the associated pattern activations of the laryngeal muscles are the final outputs of speech-related networks in the central nervous system. ADSD and other laryngeal dystonias are characterized by broad changes in those structural and functional neural networks, including inferior/middle frontal gyri, superior/middle temporal gyri, and parietal operculum (Kostic et al., 2016; Simonyan et al., 2021; Simonyan & Ludlow, 2012). These cortical and many subcortical areas of interest reflect the complex interaction among speech-related brain regions and laryngeal motor pathways in ADSD.

Limitations

The computer vision/machine learning system does not perform perfectly on all nasolaryngoscopic images. One such case is when the glottis region is overilluminated. This phenomenon limits the performance of the system because the glottis segmentation is contingent upon the assumption that the glottis region is relatively dark within the frame. From a cursory review, this is the case in about 8% of patients. Another limitation of this study is the focus on glottal dynamics and not also vocal fold dynamics. Some laryngeal computer vision/machine learning studies have segmented and analyzed data from both the glottis and vocal folds (Fehling et al., 2020; Kist et al., 2020). However, segmentation of the vocal folds is more difficult than the glottis because of frequent occlusion by other structures and high variability in shape, size, and color (Fehling et al., 2020).

For any computer vision system applied to assessing voice quality in ADSD, evaluating its convergent validity with clinical ratings is sensitive to the reliability of those ratings. In this study, the raters represented the disciplines of ENT and speech-language pathology but not neurology.

It is possible that even neurologists who manage many patients with ADSD may have a different approach to quantifying voice quality. Nevertheless, this study's raters from ENT and speech-language pathology, who conducted ratings independently and without any prior consensus training, had highly correlated ratings (see Figure 1).

Future Work

This study lays the foundation for evaluating this computer vision system in the context of measuring the effect of treatment by comparing the system's measured change in voice quality to clinicians and patients' assessments of changes in voice quality. Analyses could also be done to identify specific vocal fold features associated with voice quality. The computer vision/machine learning system could be retrained, using vocal fold masks and/or a glottis mask that more accurately outlines the glottal opening than the 3-point masks used in this study. This would further validate the system and likely improve investigation of glottal-specific dynamics that correlate with voice quality. Regarding glottis segmentation, the glottis features found in this study could be compared to clinical ratings of other disorders that are sometimes also present in ADSD, such as MTD. One can expect that because MTD patients do not exhibit voice breaks but share some of the strained characteristics of ADSD that there would be both differences and similarities in the role of the glottal geometry variables in models that predict severity in the two disorders. In order to make this comparison, the analyses with other disorders, including MTD, could use the same set of glottal geometry dynamics entered into the models of this study. Then, comparison between the glottal geometry dynamics included in other models and those consistent to the models in this study could inform on similarities and differences between different voice characteristics and/or different voice disorders. Though the age range of our cohort is representative of patients with ADSD, analyses could also be carried out in patients with ADSD of advanced age (> 65 years old), who may present with comorbid presbyphonia, a condition that, though distinct from ADSD, may co-occur and therefore exert influence on glottal geometry dynamics. Some patients with ADSD also have comorbid voice tremor. With videos recorded at sufficiently high frame rates, the computer vision system developed in this study may be able to detect voice tremor that would otherwise be undersampled at 30 Hz. Future studies could also use the system developed in this study to observe glottal configurations across a variety of speech and phonation tasks that approximate natural, spontaneous speech behaviors and simulate the interaction between cognitive–linguistic processing and laryngeal motor pathways.

Data Availability Statement

Original data available from the Dystonia Coalition upon reasonable request.

Acknowledgments

This work was supported in part by (a) grants to the Dystonia Coalition, a consortium of the Rare Diseases Clinical Research Network, including U54 TR001456 from the Office of Rare Diseases Research at the National Center for Advancing Clinical and Translational Sciences and U54 NS065701 and U54 NS116025 from the National Institute for Neurological Diseases and Stroke; (b) the National Spasmodic Dysphonia Association; and (c) the Office of the Assistant Secretary of Defense for Health Affairs, through the Peer-Reviewed Medical Research Program under Awards W81XWH-17-1-0393 and W81XWH- 19-1-0146. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense. The authors gratefully acknowledge suggestions on project planning from Christy Ludlow, Tom Bartol, Linjing Fang, and Jake Whitehill; data collection by the Dystonia Coalition investigators; and manuscript editing from Hyder (Buz) Jinnah and Joel Perlmutter.

References

Alku, P., Murtola, T., Malinen, J., Geneid, A., & Vilkman, E. (2019). Skewing of the glottal flow with respect to the glottal area measured in natural production of vowels. *The Journal of the Acoustical Society of America, 146*(4), 2501–2509. https://doi.org/10.1121/1.5129121

Buckley, D. P., Cadiz, M. D., Eadie, T. L., & Stepp, C. E. (2020). Acoustic model of perceived overall severity of dysphonia in adductor-type laryngeal dystonia. *Journal of Speech, Language, and Hearing Research, 63*(8), 2713. https://doi.org/10.1044/2020_JSLHR-19-00354

Cannito, M. P., Woodson, G. E., Murry, T., & Bender, B. (2004). Perceptual analyses of spasmodic dysphonia before and after treatment. *Archives of Otolaryngology—Head & Neck Surgery, 130*(12), 1393–1399. https://doi.org/10.1001/ARCHOTOL.130.12.1393

Chen, G., Kreiman, J., Gerratt, B. R., Neubauer, J., Shue, Y.-L., & Alwan, A. (2013). Development of a glottal area index that integrates glottal gap size and open quotient. *The Journal of the Acoustical Society of America, 133*(3), 1656–1666. https://doi.org/10.1121/1.4789931

Fang, S., Tsao, Y., Hsiao, M., Chen, J., Lai, Y., Lin, F., & Wang, C. (2019). Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice, 33*(5), 634–641. https://doi.org/10.1016/J.JVOICE.2018.02.003

Fehling, M. K., Grosch, F., Schuster, M. E., Schick, B., & Lohscheller, J. (2020). 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. https://doi.org/10.1371/JOURNAL.PONE.0227791

Fujiki, R. B., & Thibeault, S. L. (2021). Examining relationships between GRBAS ratings and acoustic, aerodynamic and patient-reported voice measures in adults with voice disorders. *Journal of Voice*. S0892-1997(21)00074-6 https://doi.org/10.1016/J.JVOICE.2021.02.007

Gillespie, A. I., Gooding, W., Rosen, C., & Gartner-Schmidt, J. (2014). Correlation of VHI-10 to voice laboratory measurements across five common voice disorders. *Journal of Voice, 28*(4), 440–448. https://doi.org/10.1016/J.JVOICE.2013.10.023

Gómez, P., Kist, A. M., Schlegel, P., Berry, D. A., Chhetri, D. K., Dürr, S., Echternach, M., Johnson, A. M., Kniesburges, S., Kunduk, M., Maryn, Y., Schützenberger, A., Verguts, M., & Döllinger, M. (2020). BAGLS, a multihospital benchmark for automatic glottis segmentation. *Scientific Data, 7,* Article No. 186. https://doi.org/10.1038/s41597-020-0526-3

Hintze, J. M., Ludlow, C. L., Bansberg, S. F., Adler, C. H., & Lott, D. G. (2017). Spasmodic dysphonia: A review. Part 1: Pathogenic factors. *Otolaryngology—Head & Neck Surgery, 157*(4), 551–557. https://doi.org/10.1177/0194599817728521

Hu, A., Hillel, A., & Meyer, T. (2016). Factors associated with patient-perceived hoarseness in spasmodic dysphonia patients. *Journal of Voice: Official Journal of the Voice Foundation, 30*(6), 769.e23–769.e26. https://doi.org/10.1016/J.JVOICE.2015.11.003

Hu, H.-C., Chang, S.-Y., Wang, C.-H., Li, K.-J., Cho, H.-Y., Chen, Y.-T., Lu, C.-J., Tsai, T.-P., & Lee, O. K.-S. (2021). Deep learning application for vocal fold disease prediction through voice recognition: Preliminary development study. *Journal of Medical Internet Research, 23*(6), e25247. https://doi.org/10.2196/25247

Khosravani, S., Chen, G., Ozelius, L. J., & Simonyan, K. (2021). Neural endophenotypes and predictors of laryngeal dystonia penetrance and manifestation. *Neurobiology of Disease, 148,* 105223. https://doi.org/10.1016/J.NBD.2020.105223

Kilic-Berkmen, G., Wright, L. J., Perlmutter, J. S., Comella, C., Hallett, M., Teller, J., Pirio Richardson, S., Peterson, D. A., Cruchaga, C., Lungu, C., & Jinnah, H. A. (2021). The dystonia coalition: A multicenter network for clinical and translational studies. *Frontiers in Neurology, 12,* 415. https://doi.org/10.3389/FNEUR.2021.660909

Kist, A. M., Zilker, J., Gómez, P., Schützenberger, A., & Döllinger, M. (2020). Rethinking glottal midline detection. *Scientific Reports, 10,* Article No. 20723. https://doi.org/10.1038/s41598-020-77216-6

Kostic, V. S., Agosta, F., Sarro, L., Tomić, A., Kresojević, N., Galantucci, S., Svetel, M., Valsasina, P., & Filippi, M. (2016). Brain structural changes in spasmodic dysphonia: A multimodal magnetic resonance imaging study. *Parkinsonism & Related Disorders, 25,* 78–84. https://doi.org/10.1016/J.PARKRELDIS.2016.02.003

Kreiman, J., Gerratt, B., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America, 122*(4), 2354–2364. https://doi.org/10.1121/1.2770547

Laves, M.-H., Bicker, J., Kahrs, L. A., & Ortmaier, T. (2019). 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. https://doi.org/10.1007/S11548-018-01910-0

Ludlow, C. L. (2011). Spasmodic dysphonia: A laryngeal control disorder specific to speech. *Journal of Neuroscience, 31*(3), 793–797. https://doi.org/10.1523/JNEUROSCI.2758-10.2011

Ludlow, C. L., & Connor, N. P. (1987). Dynamic aspects of phonatory control in spasmodic dysphonia. *Journal of Speech and Hearing Research, 30*(2), 197–206. https://doi.org/10.1044/jshr.3002.197

Ludlow, C. L., Domangue, R., Sharma, D., Jinnah, H. A., Perlmutter, J. S., Berke, G., Sapienza, C., Smith, M. E., Blumin, J. H., Kalata, C. E., Blindauer, K., Johns, M., Hapner, E., Harmon, A., Paniello, R., Adler, C. H., Crujido, L., Lott, D. G., Bansberg, S. F., . . . Stebbins, G. (2018). Consensus-based attributes for identifying patients with spasmodic dysphonia and other voice disorders. *JAMA Otolaryngology—Head & Neck Surgery, 144*(8), 657–665. https://doi.org/10.1001/JAMAOTO.2018.0644

Osma-Ruiz, V., Godino-Llorente, J. I., Sáenz-Lechón, N., & Fraile, R. (2008). Segmentation of the glottal space from laryngeal images using the watershed transform. *Computerized Medical Imaging and Graphics, 32*(3), 193–201. https://doi.org/10.1016/J.COMPMEDIMAG.2007.12.003

Rabinov, C. R., Kreiman, J., Gerratt, B. R., & Bielamowicz, S. (1995). Comparing reliability of perceptual ratings of roughness and acoustic measure of jitter. *Journal of Speech and Hearing Research, 38*(1), 26–32. https://doi.org/10.1044/jshr.3801.26

Reid, J., Parmar, P., Lund, T., Aalto, D., & Jeffery, C. (2022). Development of a machine-learning based voice disorder screening tool. *American Journal of Otolaryngology, 43*(2), 103327. https://doi.org/10.1016/J.AMJOTO.2021.103327

Roy, N., Gouse, M., Mauszycki, S., Merrill, R., & Smith, M. (2005). Task specificity in adductor spasmodic dysphonia versus muscle tension dysphonia. *The Laryngoscope, 115*(2), 311–316. https://doi.org/10.1097/01.mlg.0000154739.48314.ee

Roy, N., Mazin, A., & Awan, S. (2014). Automated acoustic analysis of task dependency in adductor spasmodic dysphonia versus muscle tension dysphonia. *The Laryngoscope, 124*(3), 718–724. https://doi.org/10.1002/LARY.24362

Simonyan, K., Barkmeier-Kraemer, J., Blitzer, A., Hallett, M., Houde, J. F., Jacobson Kimberley, T., Ozelius, L. J., Pitman, M. J., Richardson, R. M., Sharma, N., & Tanner, K. (2021). Laryngeal dystonia: Multidisciplinary update on terminology, pathophysiology, and research priorities. *Neurology, 96*(21), 989–1001. https://doi.org/10.1212/WNL.0000000000011922

Simonyan, K., & Ludlow, C. (2012). Abnormal structure–function relationship in spasmodic dysphonia. *Cerebral Cortex, 22*(2), 417–425. https://doi.org/10.1093/CERCOR/BHR120

Somanath, K., & Mau, T. (2016). A measure of the auditory-perceptual quality of strain from electroglottographic analysis of continuous dysphonic speech: Application to adductor spasmodic dysphonia. *Journal of Voice, 30*(6), 770.e9–770.e21. https://doi.org/10.1016/J.JVOICE.2015.11.005

Suppa, A., Asci, F., Saggio, G., Marsili, L., Casali, D., Zarezadeh, Z., Ruoppolo, G., Berardelli, A., & Costantini, G. (2020). Voice analysis in adductor spasmodic dysphonia: Objective diagnosis and response to botulinum toxin. *Parkinsonism & Related Disorders, 73,* 23–30. https://doi.org/10.1016/J.PARKRELDIS.2020.03.012

Wurzbacher, T., Döllinger, M., Schwarz, R., Hoppe, U., Eysholdt, U., & Lohscheller, J. (2008). Spatiotemporal classification of vocal fold dynamics by a multimass model comprising time-dependent parameters. *The Journal of the Acoustical Society of America, 123*(4), 2324–2334. https://doi.org/10.1121/1.2835435

Wurzbacher, T., Schwarz, R., Döllinger, M., Hoppe, U., Eysholdt, U., & Lohscheller, J. (2006). Model-based classification of nonstationary vocal fold vibrations. *The Journal of the Acoustical Society of America, 120*(2), 1012–1027. https://doi.org/10.1121/1.2211550

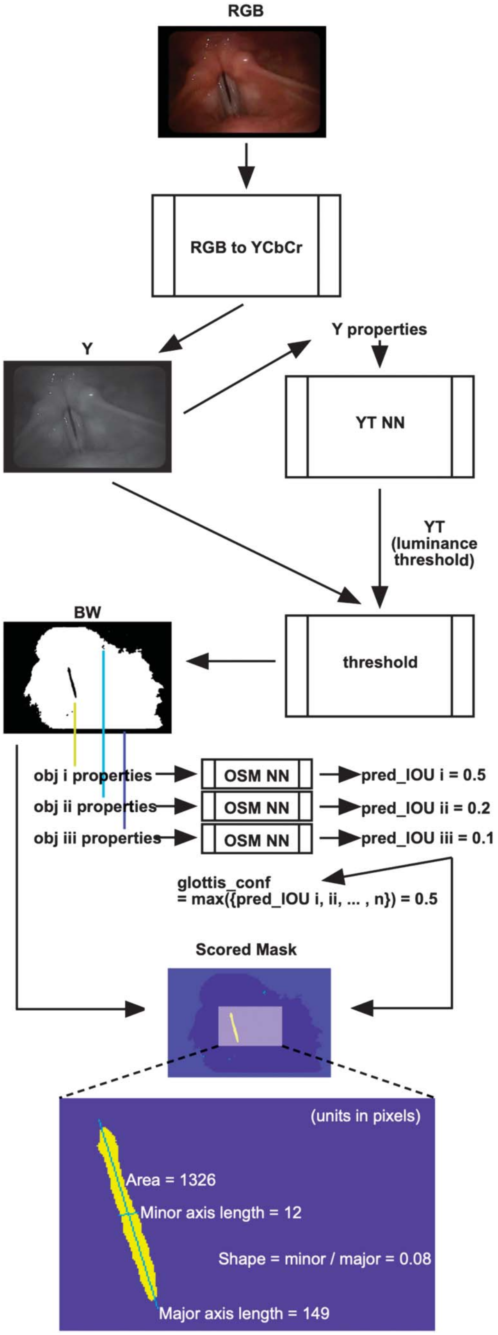
Yan, L., Hicks, M., Winslow, K., Comella, C., Ludlow, C., Jinnah, H. A., Rosen, A. R., Wright, L., Galpern, W. R., & Perlmutter, J. S. (2015). Secured web-based video repository for multicenter studies. *Parkinsonism & Related Disorders, 21*(4), 366–371. https://doi.org/10.1016/J.PARKRELDIS.2015.01.011

Appendix (p. 1 of 2)

Computer Vision/Machine Learning System for Deriving Glottal Geometry

We developed a custom computer vision/machine learning system to derive glottal geometries. The details of the system are provided in three sections. First is a description of how the overall system works (see Figure A1). Then, we describe how we trained the two primary components of the system, each a single hidden layer neural network.

Figure A1. Flowchart of the computer vision/machine learning system. Given an RGB image, the system generates a predicted glottis object and the object's subsequent glottal metrics: area, major and minor axis lengths, and shape. Two neural networks are used in the system. The YT NN (luminance threshold neural network) predicts a luminance threshold to segment the glottis from the image. The OSM NN (object score mapping neural network) gives each of the binary objects a measure of confidence that it is the glottis (pred_IOU). The object with the maximum pred_IOU is considered the glottis object. RGB = red, green, blue; YCbCr = luma blue-difference red-difference; Y = luminance; BW = black and white binary image; YT = luminance threshold; IOU = intersection over union.



Appendix (p. 2 of 2)

Computer Vision/Machine Learning System for Deriving Glottal Geometry

Overall Computer Vision/Machine Learning System

Given a nasolaryngoscopic image, the system seeks to automatically segment the glottis. More specifically, the system converts an RGB image from a nasolaryngoscopic video to the YCbCr color space. Then, it extracts the luminance channel from the YCbCr image and calculates the following properties from the pixels' distribution of luminance values: mean, standard deviation, mode, skewness (γ), kurtosis (κ), and Sarle's bimodality coefficient (β), defined as:

$$\beta = \frac{\gamma^2 + 1}{\kappa} \tag{A1}$$

The system inputs these luminance distribution properties into the luminance threshold neural network (YT NN): a single hidden layer neural network that, given a vector of luminance properties, predicts a luminance threshold (Y_T). The goal of the YT NN is to predict a Y_T to successfully segment the glottis. The system can segment the glottis object from the image using a pixel threshold because the glottis is often darker than its surrounding region. The system is not contingent on the assumption that the glottis is the darkest region in the image but rather that the glottis is a locally dark region. The system uses Y_T to perform thresholding to transform the grayscale image (Y) into a binary image (BW). This thresholding works on every pixel value in the image (Y): If the pixel is less than Y_T, the system sets the pixel's value to 1; otherwise, it sets the pixel's value to 0. After thresholding, the resulting binary image (BW) contains binary objects, which are defined as contiguous regions of pixels for which the luminance values in Y were less than Y_T (contiguous regions of pixels with value 1). In general, the thresholding procedure yields objects that include not only the glottis but also nonglottal objects. These objects are referred to as shadow objects. Then, the system calculates the following properties for each object to distinguish the glottis object from the shadow objects:

- A (area): the object's number of pixels
- θ (orientation): the angle between the x-axis and the major axis of the ellipse that has the same second-moments as the object
- x_{COM}: the object's center of mass x-coordinate
- y_{COM}: the object's center of mass y-coordinate
- $P_{boundary}$: the proportion of the object's perimeter pixels that lay on the image boundary
- $PAGM$ (perimeter average gradient magnitude): the average gradient magnitude across the object's perimeter pixels
- $\overline{Y}$: the average luminance values of the object's pixels
- $\overline{Cb}$: the average blue difference values of the object's pixels
- $\overline{Cr}$: the average red difference values of the object's pixels

These properties are designed to distinguish the glottis object from the shadow objects, based on the assumptions that a typical glottis object is oriented vertically (θ), unlikely to be near the image boundary (x_{COM}, y_{COM}, and $P_{boundary}$), a dark region surrounded by relatively lighter pixels ($PAGM$), and a dark color ($\overline{Y}$, $\overline{Cb}$, and $\overline{Cr}$). For each object, the system inputs the object's properties into the object score mapping neural network (OSM NN): a single hidden layer neural network that, given a vector of object properties, predicts the object's intersection over union (IOU) ($\widehat{IOU}$). The object with the maximum $\widehat{IOU}$ is considered to be the glottis object. The system calculates glottal geometries from the glottis object: Area (A), minor and major axis lengths which are derived from the ellipse that has the same second-moments as the object, and a loosely defined measure of shape, calculated as the ratio between the minor and major axis lengths.

Training the YT NN

We used the labeled glottis annotations to create training data for the YT NN. In general, training neural networks requires many samples. The network attempts to learn a mapping between a sample's training data and the sample's training target. For training the YT NN, a sample represents a single frame. A sample's training data are its luminance distribution properties and its training target is its optimal luminance threshold ($opt(Y_T)$). Thus, the YT NN attempts to learn how to map a vector of luminance distribution properties (a vector of numbers) to an optimal luminance threshold (a single number).

We used our glottis annotations to create the samples' training targets. For a given frame, we defined its $opt(Y_T)$ as the Y_T within the set of natural numbers $S : \{Y_T \mid 16 \leq Y_T \leq 235\}$ that maximizes the IOU between the true glottis mask and the binary image segmented by that Y_T:

$$opt(Y_T) = \operatorname*{argmax}_{Y_T \in S} \mathrm{f}(Y_T) \ := \{Y_T \in S : f(Y_T) = IOU(glottis\ mask, BW|Y_T)\} \tag{A2}$$

Appendix (p. 2 of 3)

Computer Vision/Machine Learning System for Deriving Glottal Geometry

The set *S* is used because this is the range of values that the luminance channel *Y* has in the YCbCr color representation of an image. To find the *opt*(Y_T) for all frames with a 3-point glottis mask annotation (glottis visible), we tested all possible luminance thresholds in the set S. We used each threshold (Y_T) in S to segment the grayscale image (*Y*) to create an object mask *BW*. The frame's *opt*(Y_T) is the Y_T in *S*, which returns the maximum IOU between the segmented object mask *BW* and the frame's labeled glottis mask. We stored this as the sample's training target and stored the frame's luminance distribution properties as the sample's training data.

For training the YT NN, we used leave-one-out cross-validation, where one patient is left out for testing and the rest are used for training. For each cross-validation training partition, we calculated the performance of the trained network as the mean absolute error between all the test frames' predicted luminance thresholds and the test frames' optimal luminance thresholds:

$$YT\ performance = \overline{\left|\widehat{Y}_{Tf} - opt(Y_{Tf})\right|} \qquad f : all\ test\ frames \tag{A3}$$

Along with the cross-validation training loop, we searched the network's number of hidden neurons parameter. To find the optimal number of hidden neurons, we chose the value that returned the minimum mean in test performance across the validation folds. Then, to choose a single trained YT NN to use in the system, we found the mean performance of the YT NNs trained with the optimal number of hidden neurons. Finally, we chose the YT NN whose performance value was closest to the mean performance value.

Training the OSM NN

Similar to the YT NN, we used the glottis annotations to create the training data for the OSM NN. However, here, a sample represents a binary object instead of a single frame. Thus, several OSM samples can be created from a single frame.

To ensure that the image was thresholded in an optimal manner for the glottis object, we created the OSM NN data in conjunction with the YT NN data. Once a frame's *opt*(Y_T) was found, we used it to threshold the image and get a set of binary objects, one of which was the glottis object.

For each object, we stored its object properties as its training data, and its IOU with the frame's glottis mask as its training target. Thus, the glottis object received a nonzero target value and the shadow objects received a target value of 0.

Also, similar to the training of the YT NN, we used leave-one-out patient cross-validation to train the OSM NN. To choose a single OSM NN to use in the system, we used the same process as that to choose a YT NN described above, where the only difference is the testing performance metric. For each training partition of the OSM NN, the performance was calculated as the mean absolute error between all the test objects' predicted IOUs ($\widehat{IOU}$) and the test objects' actual IOUs:

$$OSM\ performance = \overline{\left|\widehat{IOU}_j - IOU_j\right|} \qquad j : all\ test\ objects \tag{A4}$$

To aid generalizability of the system, we trained the YT and OSM nets with the greatest amount of data—frames aggregated from both the longE and sentA steps.