Real-Time 3D Reconstruction of Human Vocal Folds via High-Speed Laser-Endoscopy

Jann-Ole Henningson[1], Marc Stamminger[1], Michael Döllinger[2], and Marion Semmler[2]

[1] Friedrich-Alexander-University Erlangen-Nuremberg, Germany
[2] Division of Phoniatrics and Pediatric Audiology at the Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Erlangen, Friedrich-Alexander-University Erlangen-Nuremberg, 91054 Erlangen, Germany

Abstract. Conventional video endoscopy and high-speed video endoscopy of the human larynx solely provides practitioners with information about the two-dimensional lateral and longitudinal deformation of vocal folds. However, experiments have shown that vibrating human vocal folds have a significant vertical component. Based upon an endoscopic laser projection unit (LPU) connected to a high-speed camera, we propose a fully-automatic and real-time capable approach for the robust 3D reconstruction of human vocal folds. We achieve this by estimating laser ray correspondences by taking epipolar constraints of the LPU into account. Unlike previous approaches only reconstructing the superior area of the vocal folds, our pipeline is based on a parametric reinterpretation of the M5 vocal fold model as a tensor product surface. Not only are we able to generate visually authentic deformations of a dense vibrating vocal fold model, but we are also able to easily generate metric measurements of points of interest on the reconstructed surfaces. Furthermore, we drastically lower the effort needed for visualizing and measuring the dynamics of the human laryngeal area during phonation. Additionally, we publish the first publicly available labeled in-vivo dataset of laser-based high-speed laryngoscopy videos. The source code and dataset are available at henningson.github.io/Vocal3D/.

Keywords: Human Vocal Folds · Endoscopy · Structured Light

1 Introduction

Human interaction is fundamentally based on the ability to communicate with each other [2]. Over the last decades, communication-based professions have increased drastically and up to 10% of the Western worlds workforce are now classified as heavy occupational voice users [22]. Hence, a lasting impairment of our oral expression is necessarily accompanied by severe social and economic limitations [5,14], increasing the significance of diagnosing laryngeal and voice-related disorders. Laryngeal disorders, impairing speech production, result from different causes ranging from functional abnormalities in the dynamic process as well as morphological alterations in the anatomical structures. Conventionally,

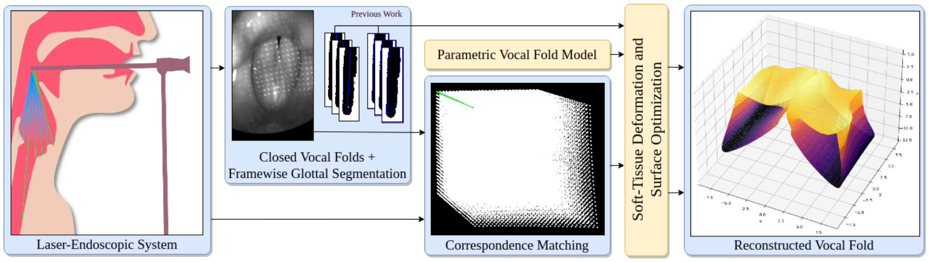

Fig. 1: Our pipeline for real-time reconstruction of vocal folds during phonation. First, we segment images into vocal fold and glottal area. Next, we estimate laser correspondences by taking the systems epipolar constraints into account. Lastly, we generate dense reconstructions by optimizing a parametric vocal fold model using a combined soft-tissue deformation and least squares surface fitting step.

human vocal folds have been examined using standard video endoscopy. However, when the dynamics of the vocal folds are of interest (i.e. the vocal folds during **phonation**), High-Speed Video Endoscopy (HSV) is generally used, as the frequency of the vibration necessitates a high temporal resolution of the recording camera [10,11,18,7]. Döllinger et al. [4] have shown that moving vocal folds have a significant vertical expansion in its motion, leading to the assumption that not only the lateral and longitudinal deformation of vocal folds during phonation is of interest, but also their vertical deformation. However, classical video-endoscopy and HSV-Imaging can not resolve the vertical deformation. In common literature, two systems are used for reconstructing the surface of the vocal folds during phonation. Either **a)** stereo-endoscopy systems [26,23], which suffer from a lack of feature points on the smooth tissue inside the larynx [25] or (as in our case) **b)** structured light supported endoscopy which projects a symmetric pattern onto the surface of the vocal folds [20,13,21,16], which can then be used for stereo triangulation. To reach clinical applicability, it is necessary that these systems work robustly without any human input. However, the methods proposed in [20,13,21,16] do require human input in form of a time-consuming manual labeling step. Thus, based upon a laser-based endoscopic high-speed video system [20], we propose the first fully-automatic real-time capable pipeline for reconstructing human vocal folds during phonation. An overview of our pipeline is shown in Fig. 1. Furthermore, we publish a dataset containing laser-endoscopy videos of 10 healthy subjects that can be used to drive further research in this area. We believe this work is taking a big step towards integrating 3D video endoscopy into the clinical routine.

2 Method

Our reconstruction pipeline can be properly divided into three significant parts, as shown in Fig. 1. It receives high-speed video images as input, that are taken

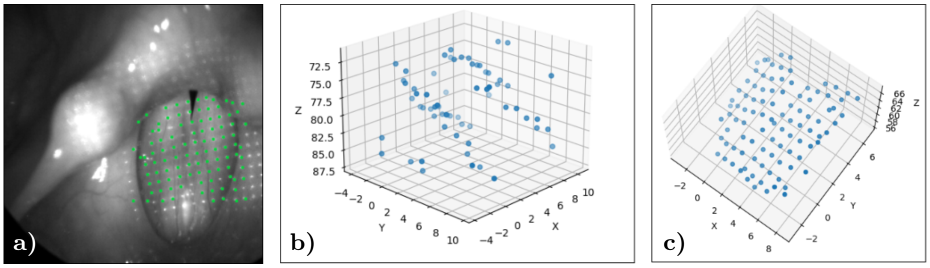

Fig. 2: **a)** Extracted local maxima lying on the surface of the vocal folds **b)** Reconstructed points by Mask Sweeping generated using Epipolar Constraints, **c)** Globally aligned and optimized triangulation using RHC.

with a calibrated laser-endoscopy system consisting of a high-speed video camera (4000 Hz, resolution 256 × 512) and an LPU projecting a symmetric grid of laser rays. A short video sequence (∼100-200 ms) is recorded and passed to our pipeline. First, vocal folds and the glottal area are segmented and a region of interest is extracted. We use this segmentation to find a frame where the vocal folds are closed and quasi-planar. Secondly, we generate first laser ray correspondences in this frame by utilizing epipolar constraints given by the laser-endoscopy system. We globally align and refine these correspondences, using a novel RANSAC-based discrete hill climbing (RHC) approach. Third, by stereo triangulation, we generate per frame 3D points on the surface of the vocal folds. These are used to fit a novel 3D B-Spline model of the vocal folds, based on the M5 model [19] to generate realistically moving vocal folds. For step 1 we apply the technique of Koç et al. [12]. Steps 2 and 3 are presented in the following sections. Our pipeline is designed to be robust and real-time capable, such that immediate feedback is given to the user about the success of the recording. -Time

2.1 Correspondence Matching via Epipolar Constraints

Based on a segmentation of the glottis and the vocal folds, we extract the laser points lying on the superior area of the vocal folds. To this end, we use dilatation filtering on the ROI of the vocal folds to find local maxima corresponding to the projected laser rays and weight the estimated local areas by their intensities for a sub-pixel accurate centroid calculation. Note that we only apply this part of the pipeline to the estimated frame in which the glottal area is minimal, as we then use a temporal nearest neighbor search to label all of the remaining frames. The goal now is to find the proper correspondences between laser rays and their projected points in the image.

Epipolar Constraint-based Search Space reduction We assume that we have a calibration between high-speed camera and LPU and that their relative position is static. Thus, we can project each point on the laser ray $r_{x,y}(t) = o_l + t * d_{x,y}$ with distance t from the rays origin o_l to the point $u^t_{x,y}$ in image space, where

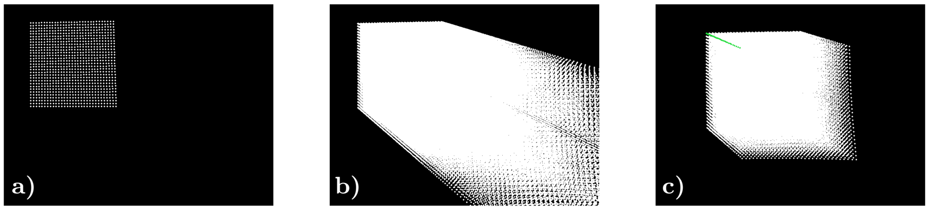

Fig. 3: **a)** Laser rays sampled at $u^{\infty,\infty}_{x,y}$ **b)** Uniformly sampled laser rays from $u^{0,\infty}_{x,y}$ **c)** Uniformly sampled laser rays from $u^{r,s}_{x,y}$ where $u^{r,s}_{0,0}$ is highlighted in green. By taking the epipolar constraints into account, the search space can be drastically reduced and laser correspondences can be found by sampling from $u^{r,s}_{x,y}$.

$(x, y) \in [0 \dots M-1, 0 \dots N-1]$ are the indices of the ray's position in the laser grid. Thus, we can restrict the search space for the laser dots p_i to the epipolar line $u^{0,\infty}_{x,y} := u^{0}_{x,y} \rightarrow u^{\infty}_{x,y}$. However, due to the high density of the lasergrid, we have to restrict the search space further, to be able to disambiguate all correspondences. In general, one can generate first depth estimations of fronto-parallel surfaces by measuring the laser points extent in image space. As a laser ray can be assumed to be collimated for close distances, the projected laser points circumference is inversely proportional to the surfaces depth. In our setup however, the image resolution does not allow for such a depth estimation. Instead, we use the observation of Semmler et al. [21], that states that an endoscopes working distance is in between 50mm to 80mm above the vocal folds, so we can confine the search space to this depth range. We refer to the projection of the reduced search space as $u^{r,s}_{x,y}$. A visual representation of this is given in Figure 3. Note that, as the LPU projects a symmetric laser grid onto the vocal folds surface, neighboring search spaces overlap. Thus, a disambiguation is still necessary.

Estimating Laser Grid Correspondences Let $P = \{p_i\}$ be the set of extracted local laser points. We then want to generate initial laser point - laser ray mappings, i.e. we need to know which p_i corresponds to which laser ray $r_{x,y}$. To this end, we rasterize the line $u^{r,s}_{x,y}$ for each ray $r_{x,y}$ into the image, whereas the mask of $u^{r,s}_{x,y}$ is directly dependent on the radius of the projected laser dots. Whenever a probable dot p_i is hit, we map it to $r_{x,y}$ and remove the local maximum from P. Note that in this stage, we will still have several wrong correspondences. After this initialization, we can use stereo triangulation to reconstruct the 3D world coordinates of the laser points. As can be seen in Figure 2, the triangulated points do not have any global alignment, due to overlapping search spaces. Thus, many correspondences are mislabeled by a small offset. To globally align the correspondences, we randomly select a single local maximum $p_i \in P$ and use a recursive grid-based search to label the remaining ones. Based on the selected starting point, all of the consecutively labeled local laser points may now be mislabeled by a discrete static offset. To find the correct labeling we propose RHC, a method specifically designed for labelling symmetric laser grids.

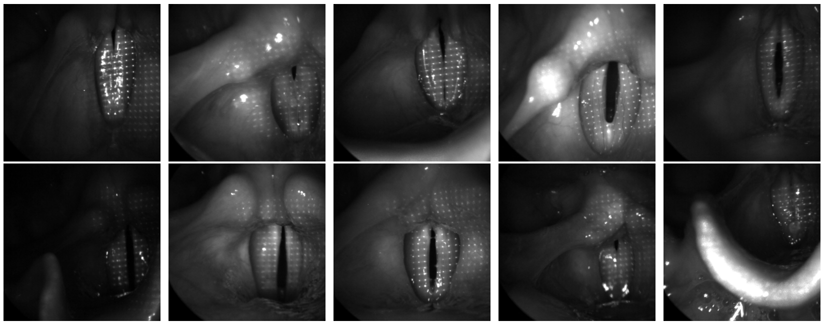

Fig. 4: The set of diverse videos contained in the labeled HLE dataset.

RANSAC-based Hill Climbing for discrete Correspondence Optimization In this step of the pipeline, the found correspondences are globally aligned, but based on the chosen starting point for the grid-based search, the correspondences may only be close to their optimal solution. Thus, we search a static grid-based offset $a, b \in \mathbb{Z}$ such that the reprojection error $\mathcal{E}(x, \tilde{x})$ is minimized, i.e. $\min_{a,b} \sum_{p_i \in P} \mathcal{E}(p_i, r_{x+a,y+b})$. Where $\mathcal{E}(x, \tilde{x})$ is the Euclidean reprojection error of point x in image space and the reprojection of the intersection between the laser ray $\tilde{x}$ and the camera ray stemming from x. Instead of just naively brute-forcing a global optimum for a and b that might be inaccurate due to outliers, we propose a RANSAC-based [8] recursive hill climbing algorithm to find optimal parameters a and b. RHC optimizes the labels in such a manner that a local grid-based minimum is found. First, to be robust against outliers, e.g. local maxima stemming from specular reflections, we take a random subset $\tilde{P} \subseteq P$ of local maxima and their corresponding labels $\tilde{x}, \tilde{y} \in \tilde{X} \subseteq X$ and calculate their reprojection error $e = \mathcal{E}(\tilde{P}, r_{\tilde{x},\tilde{y}})$. We then calculate the reprojection error of the labels inside the 4-neighborhood of $r_{\tilde{x},\tilde{y}}$ and recurse in the direction of the smallest error $\hat{e} = \arg\min_{a,b} \mathcal{E}(\tilde{P}, r_{\tilde{x}+a,\tilde{y}+b})$. If $e < \tilde{e}$, we stop the recursion, otherwise, we repeat this process with $e = \mathcal{E}(\tilde{P}, r_{\tilde{x}+a,\tilde{y}+b})$ until a local minimum has been found. This algorithm is then repeated for different subsets $\tilde{P}$, until a convergence criterion or the maximum amount of iterations has been reached. Next, we update all of the labels based on the discrete labeling-offset a, b that produced the smallest reprojection error.

Finally, as the videos are of very short duration, we can assume the camera to be almost static. Thus, we can use a temporal nearest neighbor search on consecutive frames to label all of the remaining images and compute frame-wise point clouds of the superior vocal fold surface using stereo triangulation.

2.2 Surface Reconstruction

Goal of this step is to generate dense moving vocal folds to better guide practitioners in diagnosing laryngeal disorders, while simultaneously improving comprehensibility and plausibility of the data. To achieve real-time performance for

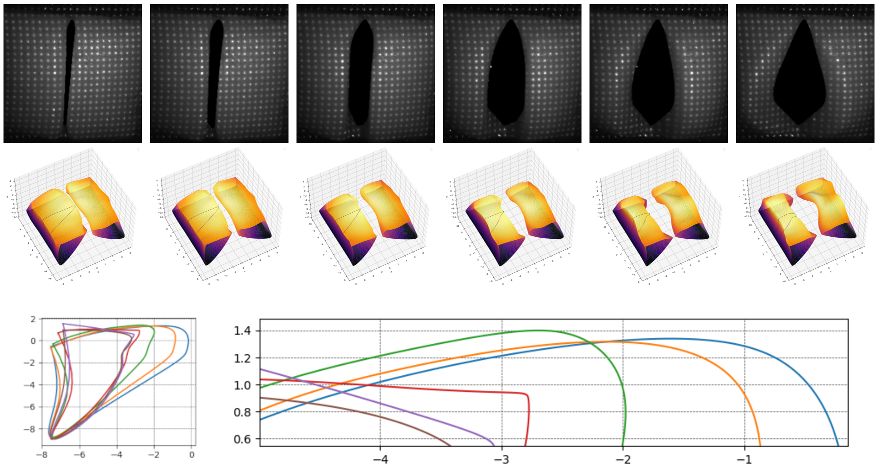

Fig. 5: First Row: Input to our pipeline. Second Row: Framewise reconstruction of a vocal fold model and visualization of the geodesic curvatures to measure. Third Row: Acquired cross sections and geodesic curvatures.

reconstructing dense vocal fold models, we extend the M5 Model by Scherer et al. [19] to 3D using B-splines, such that we can reduce the amount of parameters to optimize. Formally, a B-spline surface is a piecewise polynomial function, where a surface point parameterized by (u, v) is defined by $\mathbf{S}(u, v) = \sum_{i=0}^{n} \sum_{j=0}^{m} N_i^p(u) N_j^q(v) \mathbf{P}_{ij}$. Here, $N_i^p(u)$ is the i-th polynomial basis function, where p is the degree in u direction and $\mathbf{P}_{ij}$ the set of control points building the surfaces convex hull. In case of NURBS and B-Splines, they are commonly computed using the Cox-De-Boor recursion formula [17]. Given the piecewise definition of the 2D-based M5 model in [19], we propose a B-Spline surface based M5 vocal fold model (BM5). To generate the BM5 we first subdivide each piece of the parametric function n times and generate points p_i lying on the M5's surface. We define the knots $u \in U$ to be uniformly sampled in the interval $[0, 1]$. Next, we extrude the parametric surface in z-direction. We define the knots of the knot vector $v_i \in V$ similar to U. Lastly, we define the control points of the BM5 to be exactly the points P lying on the extruded surface.

Surface Optimization from Sparse Samples Let T be the set of triangulated points $t_i \in \mathbb{R}^3$. We then fit a plane to the triangulated points t_i and project the extreme points of the glottal midline as well as the glottal outline into the point cloud, generating $\hat{T}$. Next, we align T and $\hat{T}$ such that the glottal midline lies on the z-axis to generate a BM5 that lies directly below T. Next, we compute $\arg\min_k ||\hat{P}_{ij} - \hat{t}_k||_2$, i.e. the nearest triangulated point $\hat{t}_k \in \hat{T}$ to every control point $\hat{P}_{ij}$ defining the superior surface of the vocal fold. We then use as-rigid-as-possible deformation [24] using the pre-computed mapping for each $\hat{P}_{ij}$ and $\hat{t}_k$ while restricting the deformation of the inferior control points of the BM5. Next,

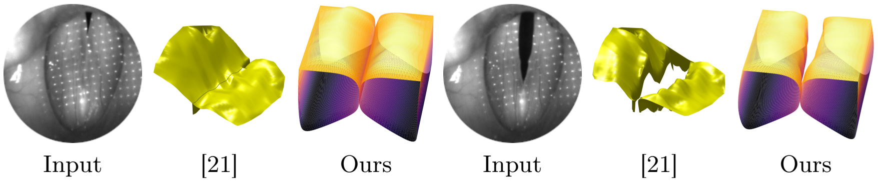

Fig. 6: Qualitative comparison of the approach by Semmler et al. [21] and our method using images of the HLE Dataset as input.

we optimize $\arg\min_P \mathcal{L}(\mathbf{S}(u,v), t_i)$, i.e. minimize the distance between specific points on the parameterized surface $\mathbf{S}(u,v)$ and their nearest neighbor t_i, where $\mathcal{L}(\mathbf{S}(u,v), t_i)$ is the metric to be minimized.

3 Results

We implemented our pipeline in Python, using the NumPy[9], PyTorch[15] and OpenCV[1] libraries. We achieve real-time performance (around 25 fps) on an Intel Core i7-6700K CPU and NVidia Quadro RTX 4000 GPU. For optimizing the BM5, we use the NURBS-Diff module proposed by Prasad et al. [3]. We use the approach by Koç et al. [12] for segmenting the vocal folds and glottal area. Lastly, for calibration we use the method proposed in [21], the systems error measures can be inferred from that work as well.
We propose the **human laser-endoscopic (HLE)** dataset, a labeled dataset consisting of 10 in-vivo monochromatic recordings using a laser-based endoscopic recording setup of 10 healthy subjects (Figure 4). The videos were recorded using a 4000FPS camera with a spatial resolution of 256×512 pixels and labeled manually using [21], whereas a 18 by 18 symmetric laser pattern is projected into the laryngeal area [20]. Subjects were ordered to make a sustained /i/ vowel during recording. The dataset consists of high-quality to lower-quality recordings containing slight camera movement and under-exposed imagery. An excerpt of the HLE Dataset is given in Figure 4.

RHC Evaluation We evaluate the RHC algorithm on 21 videos of silicone M5 vocal folds under distinct viewing angles (-15° to +15°) and distances (50mm to 80mm), in which a 31 by 31 laser grid is projected onto the vocal fold model [21]. We are interested in the general labelling error of our Mask Sweeping algorithm, the global alignment pass and the RHC algorithm. To show the validity of each step, we compute the averaged per label L1-Norm for 20 randomly generated regions of interest lying inside the laser pattern per video. As we use a grid-based labelling, the L1-Norm shows the general accuracy of our method. For example, in case of a 31 by 31 laser grid, we need to discern 961 different labels. If an algorithm now estimates (n, m) for every label, whereas $(n+1, m)$ would be correct, the averaged L1-Norm is one. The ground-truth labels were generated using the semi-automatic approach proposed in [21] for a direct quantitative

Table 1: L1 Error and Standard Deviation of grid offsets using the Mask Sweeping (MS), Global Alignment (GA) and RHC step for different viewing angles of silicone M5 vocal folds. Ground-truth data was generated manually using [21].

α	-15°	-10°	-5°	0°	5°	10°	15°	Average
MS	1.54 ± 1.25	3.16 ± 1.97	3.84 ± 1.81	3.31 ± 1.39	3.34 ± 1.39	3.21 ± 1.63	2.86 ± 1.78	3.26 ± 1.60
GA	2.29 ± 1.29	3.15 ± 1.65	2.60 ± 1.70	2.65 ± 1.41	2.31 ± 1.50	2.54 ± 1.44	3.30 ± 1.57	2.57 ± 1.51
RHC	1.49 ± 1.29	1.67 ± 1.27	1.39 ± 1.26	1.55 ± 0.54	1.26 ± 0.93	1.85 ± 1.33	1.36 ± 1.07	1.51 ± 1.10

comparison between the methods. The results can be seen in Table 1. It shows that the Mask Sweeping algorithm finds solutions close to a local optimum. However, as can be seen, globally aligning the points generally reduces the error, while RHC drastically minimizes the offset of the laser grid. We also tested RHC with an 8- and 12-neighborhood look-up, but couldn't observe any differences in labelling accuracy. In general, our approach works robustly in cases where the laser grid's edge is included in the ROI. However, in cases where the labelling is misaligned, RHC finds a mapping that lies on the epipolar lines $u_{x,y}^{r,s}$, thus still finding a solution capable for visualization purposes as the relative proportions of the triangulation stay intact.

Surface Reconstruction As there does not exist any real-world ground-truth data for vocal folds during phonation, we're showing image based comparisons of our reconstructions using a silicone based M5 vocal fold model and the HLE dataset. We set the anchor weight of the ARAP algorithm to 10^5 and the iterations to 2. In the surface fitting step we use the Chamfer Distance [6] as a loss function, as we want to maximize the point cloud similarity between the points lying on the superior surface of the vocal folds and the triangulated points. We refine the control points for 5 iterations with a learning rate of 0.5. In Figure 6 we show a qualitative comparison of the approach by Semmler et al. [21] and our method. It can be seen that in previous works, only the superior surface of the glottis could be reconstructed, while our method also takes the deformation of the inferior part of the vocal folds into account. In Figure 5 a reconstructed glottal opening cycle is visualized depicting the temporal coherence of our method, as well as measurements taken of its geodesic curvature over time.

4 Conclusion

In this work, we proposed the first fully automatic pipeline for reconstructing dynamic vocal folds during phonation based on a laser-endoscopy system that records high-speed videos. We achieve this through highly specialized algorithms for correspondence estimation between symmetric laser grids and their projection in image space. These algorithms enable the triangulation of thousands of frames in a matter of seconds. Unlike other approaches, we do not only visualize and measure the upper surface of the vocal fold, but instead use a parametric reinterpretation of the M5 vocal fold model for a dense surface reconstruction, that also takes the inferior part of the vocal folds into account. Based on an ARAP-deformation and Least Squares Optimization, we can generate visually appealing reconstructions of vocal folds in real-time. Furthermore, we proposed a dataset that can be used to drive further research in this area.

Acknowledgements We thank **Florian Güthlein** and **Bernhard Egger** for their valuable feedback. This work was supported by Deutsche Forschungsgemeinschaft (DFG) under grant STA662/6-1 (DFG project number: 448240908).

References

1. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
2. Cummins, F.: Voice, (inter-)subjectivity, and real time recurrent interaction. Frontiers in Psychology **5** (2014). https://doi.org/10.3389/fpsyg.2014.00760, https://www.frontiersin.org/article/10.3389/fpsyg.2014.00760
3. Deva Prasad, A., Balu, A., Shah, H., Sarkar, S., Hegde, C., Krishnamurthy, A.: Nurbs-diff: A differentiable programming module for nurbs. Computer-Aided Design **146**, 103199 (2022). https://doi.org/https://doi.org/10.1016/j.cad.2022.103199, https://www.sciencedirect.com/science/article/pii/S0010448522000045
4. Döllinger, M., Berry, D.A., Berke, G.S.: Medial surface dynamics of an in vivo canine vocal fold during phonation. The Journal of the Acoustical Society of America **117**(5), 3174–3183 (2005). https://doi.org/10.1121/1.1871772, https://doi.org/10.1121/1.1871772
5. FAAP, R., Ruben, R.: Redefining the survival of the fittest: Communication disorders in the 21st century. The Laryngoscope **110**, 241 – 241 (02 2000). https://doi.org/10.1097/00005537-200002010-00010
6. Fan, H., Su, H., Guibas, L.: A point set generation network for 3d object reconstruction from a single image (12 2016)
7. Fehling, M.K., Grosch, F., Schuster, M.E., Schick, B., Lohscheller, J.: Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional lstm network. PLOS One (2) (2020), http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-87208-2
8. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (jun 1981). https://doi.org/10.1145/358669.358692, https://doi.org/10.1145/358669.358692
9. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (Sep 2020). https://doi.org/10.1038/s41586-020-2649-2, https://doi.org/10.1038/s41586-020-2649-2
10. Kist, A., Dürr, S., Schützenberger, A., Döllinger, M.: Openhsv: an open platform for laryngeal high-speed videoendoscopy. Scientific Reports **11** (07 2021). https://doi.org/10.1038/s41598-021-93149-0
11. Kist, A., Gómez, P., Dubrovskiy, D., Schlegel, P., Kunduk, M., Echternach, M., Patel, R., Semmler, M., Bohr, C., Dürr, S., Schützenberger, A., Döllinger, M.:

A deep learning enhanced novel software tool for laryngeal dynamics analysis. Journal of Speech, Language, and Hearing Research **64**, 1–15 (05 2021). https://doi.org/10.1044/2021_JSLHR-20-00498

12. Koc, T., Çiloglu, T.: Automatic segmentation of high speed video images of vocal folds. Journal of Applied Mathematics **2014** (06 2014). https://doi.org/10.1155/2014/818415
13. Luegmair, G., Mehta, D., Kobler, J., Döllinger, M.: Three-dimensional optical reconstruction of vocal fold kinematics using high-speed videomicroscopy with a laser projection system. IEEE transactions on medical imaging **34** (06 2015). https://doi.org/10.1109/TMI.2015.2445921
14. Merrill, R.M., Roy, N., Lowe, J.: Voice-related symptoms and their effects on quality of life. Annals of Otology, Rhinology & Laryngology **122**(6), 404–411 (2013). https://doi.org/10.1177/000348941312200610, https://doi.org/10.1177/000348941312200610, pMID: 23837394
15. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
16. Patel, R., Donohue, K., Lau, D., Unnikrishnan, H.: In vivo measurement of pediatric vocal fold motion using structured light laser projection. Journal of voice : official journal of the Voice Foundation **27**, 463–72 (07 2013). https://doi.org/10.1016/j.jvoice.2013.03.004
17. Piegl, L., Tiller, W.: The NURBS Book. Springer-Verlag, Berlin, Heidelberg (1995)
18. Schenk, F., Urschler, M., Aigner, C., Roesner, I., Aichinger, P., Bischof, H.: Automatic glottis segmentation from laryngeal high-speed videos using 3d active contours (01 2014)
19. Scherer, R.C., Shinwari, D., De Witt, K.J., Zhang, C., Kucinschi, B.R., Afjeh, A.A.: Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees. The Journal of the Acoustical Society of America **109**(4), 1616–1630 (2001). https://doi.org/10.1121/1.1333420, https://asa.scitation.org/doi/abs/10.1121/1.1333420
20. Semmler, M., Kniesburges, S., Birk, V., Ziethe, A., Patel, R., Döllinger, M.: 3d reconstruction of human laryngeal dynamics based on endoscopic high-speed recordings. IEEE Transactions on Medical Imaging **35**(7), 1615–1624 (2016). https://doi.org/10.1109/TMI.2016.2521419
21. Semmler, M., Kniesburges, S., Parchent, J., Jakubaß, B., Zimmermann, M., Bohr, C., Schützenberger, A., Döllinger, M.: Endoscopic laser-based 3d imaging for functional voice diagnostics. Applied Sciences **7** (06 2017). https://doi.org/10.3390/app7060600
22. Snyder, T., Dillow, S.: Digest of education statistics, 2010. nces 2011-015. National Center for Education Statistics (01 2011)
23. Sommer, D.E., Tokuda, I.T., Peterson, S.D., Sakakibara, K.I., Imagawa, H., Yamauchi, A., Nito, T., Yamasoba, T., Tayama, N.: Estimation of inferior-superior vocal fold kinematics from high-speed stereo endoscopic data in vivo. The Journal of the Acoustical Society of America **136**(6), 3290–3300 (2014). https://doi.org/10.1121/1.4900572, https://doi.org/10.1121/1.4900572

24. Sorkine, O., Alexa, M.: As-rigid-as-possible surface modeling. pp. 109–116 (01 2007). https://doi.org/10.1145/1281991.1282006
25. Stevens Boster, K., Shimamura, R., Imagawa, H., Sakakibara, K.I., Tokuda, I.: Validating stereo-endoscopy with a synthetic vocal fold model. Acta Acustica united with Acustica **102**, 745–751 (07 2016). https://doi.org/10.3813/AAA.918990
26. Tokuda, I., Iwawaki, M., Sakakibara, K.I., Imagawa, H., Nito, T., Yamaosba, T., Tayama, N.: Reconstructing three-dimensional vocal fold movement via stereo matching. Acoustical Science and Technology **34**, 374–377 (09 2013). https://doi.org/10.1250/ast.34.374