# Study of Vocal Muscle Strain with Skin Deformation Tracking System

Steven Hogue*, Adrianna C. Shembel†‡, Xiaohu Guo*
*Department of Computer Science, University of Texas at Dallas, Richardson, USA
†Department of Speech, Language, and Hearing, University of Texas at Dallas, Richardson, USA
‡Department of Otolaryngology-Head & Neck Surgery, UT Southwestern, Dallas, USA

2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS) | 978-1-6654-6770-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/CBMS55023.2022.00026

***Abstract*—Vocal strain can have a profound effect on a person's life and livelihood. However, methods to identify and quantify vocal strain presumed to originate in the laryngeal muscles severely lack. We aim to address this shortcoming. Using motion capture with consumer RGBD cameras, we track skin deformation of perilaryngeal anterior neck regions in participants with and without vocal strain. Neck movement variability differences between the two groups provides insight into extrinsic laryngeal vocal muscles that may underlie symptoms of vocal strain.**

***Index Terms*—Vocal strain, motion capture, key-point tracking, skin deformation, laryngeal muscle**

## I. Introduction

Vocal strain in the neck muscles that connect the jaw, larynx, and sternum during voice productions occurs in 40 percent of occupational voice users and results in difficulties speaking and loss of income [32], [36], [37]. Considering 25-35 percent of the US population is dependent on their voice for their career, the high prevalence of vocal strain has profound impact on both individual and societal levels [3], [5], [9], [21]. Although vocal strain has significant consequences, there are no well-vetted, validated physiologic metrics to identify strain in the vocal muscles; there are also no objective metrics to monitor treatment progress. This gap has resulted in ineffective trial-and-error therapies, which require significantly more voice therapy sessions, time off work, and increased allocation of medical staff and resources. This project aims to address these gaps.

Using motion capture (MoCap) technology, we developed objective metrics for the study and identification of vocal muscle strain thought to originate in the extrinsic laryngeal muscles [22]. Currently, there are no motion capture systems designed for tracking skin deformations on a small scale. Optical motion capture systems are either marker-less or marker-based systems. Marker-less systems typically capture the entire body or the face. These systems are able to leverage a large number of features around joints and on faces to track movement [2]. Our system tracks small-scale skin deformations associated with neck movement using consumer level RGBD cameras to record short sequences. Because of the lack of features or textural differences on necks we mark key-points with green stickers. The stickers enable a marker-based motion capture that can be done quickly and easily. Once recorded, the sequences can then be directly compared against other sequences to study the neck movement.

## II. Related Works

Although the presumption that the extrinsic laryngeal muscles are involved in vocal strain is ubiquitous, there are currently no objective metrics to identify musculoskeletal pathophysiology of vocal strain. The majority of methods to assess vocal strain involve acoustic vocal output. But these methods do little to elucidate how movements in and around the laryngeal muscles needed for voice and speech result in aberrant acoustic vocal output. Methods that focus on the vocal production process are needed.

MoCap has previously been used in the limb muscles to study gait and locomotion and inform muscle overuse and strain injuries in athletes [7], [10], [35]. However it has not been applied to the vocal muscles.

The use of RGBD cameras has become wide-spread with the introduction of consumer level cameras. These cameras enable methods that once required expensive setups. Single-camera and multi-camera methods for working with RGBD datasets reach across many different fields and problems. These RGBD cameras enable accessible motion capture in many forms.

Reconstruction of the human body for the creation of 3D human avatars is one such method. Deriving from KinectFusion [24], then DynamicFusion [25], many methods have been developed to take RGBD video as input to recreate full body human avatars [11], [42], [43]. These methods leverage both the color and depth information to track a subject throughout a sequence.

Similarly, the capture of exclusively the face and head has been used to create retargetable talking heads [15], [19], [39]. These face tracking methods utilize RGBD cameras to track the face with marker-less motion capture.

Motion capture has also seen applications such as operating room assistance [6], [13], physical therapy and rehabilitation [8], [18], [29], [33], and detecting of falls [20], [26], [38], [44].

## III. Method

The goal of the use of MoCap capabilities is to transform the participant's recording into a sequence that is comparable with a whole collection of data. The recording is processed by first converting RGBD images into RGB point clouds and then extracting the points within the markers. These points are clustered and tracked throughout the sequence and relabelled for consistency giving a set of key-points for each

2372-9198/22/$31.00 ©2022 IEEE
DOI 10.1109/CBMS55023.2022.00026

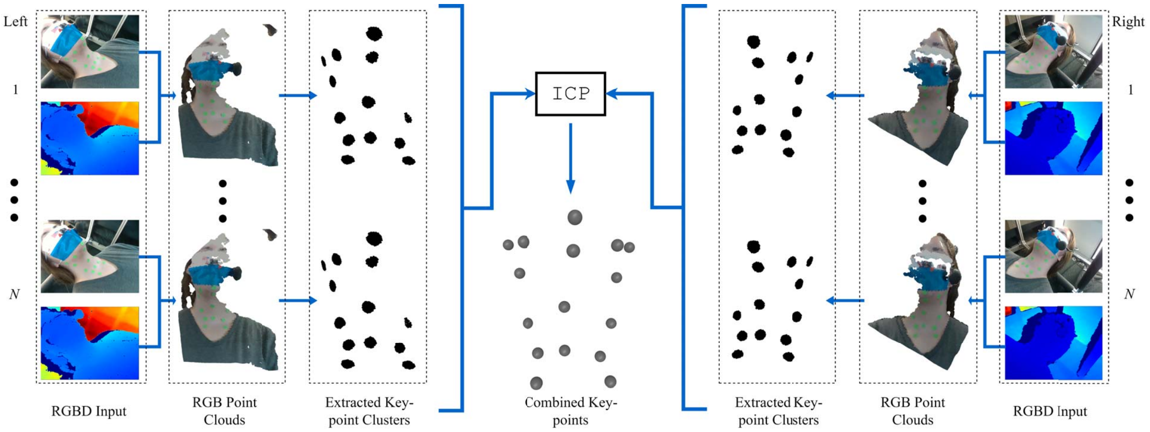


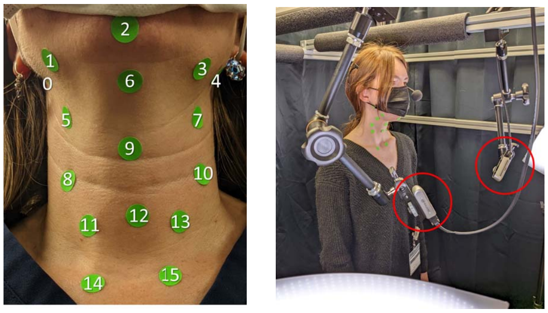
Fig. 1. Overview of the data processing pipeline.

frame of the recording. The lengths of the edges between key-points are then measured for each frame. To compare different sequences directly we need to ensure they are aligned. We align the sequences using dynamic time warping on the audio waveforms. The aligned sequences are then directly comparable.

### A. Data Collection

A total of 13 subjects with and without vocal strain were recruited for the study. Subjects with vocal strain (defined as greater than 11 on the Voice Handicap Index-10 [31], greater than 24 on Part 1 of the Vocal Fatigue index [23], and a clinical diagnosis of muscle tension dysphonia) are recruited for the experimental group. Subjects without vocal strain (less than 5 on the Voice Handicap Index-10, less than 24 on Part 1 of the Vocal Handicap Index, and no voice complaints over the past 6 months) are also recruited for the control group. 16 green neon stickers and headset microphone are placed on each subject prior to video and audio recordings. All subjects complete four speech tasks: (1) a repetitive diadochokinetic articulation rate task for 30 seconds on *pataka*, (2) standard reading passage *(Rainbow Passage)*, (3) vocal range task (*pitch glide* from lowest to higher note on /a/), and (4) vocal intensity task (*Hey you!* as loud as possible).

Data is recorded using a headset microphone and two Intel Realsense D435 cameras. Cameras are placed in close proximity to the subject and pointed at an upward slant towards one half of the front of the subject's neck. Each camera captures approximately half of the front of the subject's neck with some overlap between views. The cameras record RGBD images, with a resolution of 640x480, at 30 FPS. Cameras are mounted on a moving frame and adjustable arms to ensure sufficient viewpoints for a range of subjects. Subjects have 16 key-points on their neck area marked with green stickers as shown in Fig. 2. The labelling of these key-points can also be seen in Fig. 3. Several key-points have specific anchors such as either clavicle (14, 15), the chin (2), and along either side of the jaw (0, 1, 3, 4). These key-points are on rigid parts, that is bones, of the neck area. The remaining key-points are on soft-tissue (cartilagenous laryngeal framework and extrinsic laryngeal muscle).



(a) (b)

Fig. 2. (a) Positioning of key-point markers on the neck. 0-1 = right jaw, 2 = chin, 3-4 = left jaw, 5, 7 = hyoid, 6 = base of tongue, 9 = thyroid notch, 8,11 = right sternocleidomastoid, 12 = sternothyroid, 10, 13 = left sternocleidomastoid, 14, 15 = clavicle, (b) Setup used for recording, cameras are circled in red.

### B. Key-point Extraction

Given a sequence of RGBD images $\{I_1, I_2, ..., I_n\}^v$, for each viewpoint $v$, we create a combined sequence of key-points $\{K_1, K_2, ..., K_n\}$.

For each frame, $I_i = [r, g, b, d]$, we back-project the depth values to the camera's coordinate space to obtain a point cloud $\mathbf{x}_i = (x, y, z)^\top$:

$$\mathbf{x}_i(\mathbf{u}) = D_i(\mathbf{u})\mathbf{K}^{-1}\mathbf{u}, \tag{1}$$

where $\mathbf{u} = (u, v)^\top$ is a pixel of the image $I_i^v$, $D_i(\mathbf{u})$ is the depth of the pixel, and $\mathbf{K}$ is the camera's calibration matrix.

We then reduce the point cloud to include only the points that are in the marked areas to get $P_i$. This is done with a color threshold,

$$P_i = \{\mathbf{x}_i(\mathbf{u}) | T_{lower} \leq C_i(\mathbf{x}_i(\mathbf{u})) \leq T_{upper}\}, \tag{2}$$

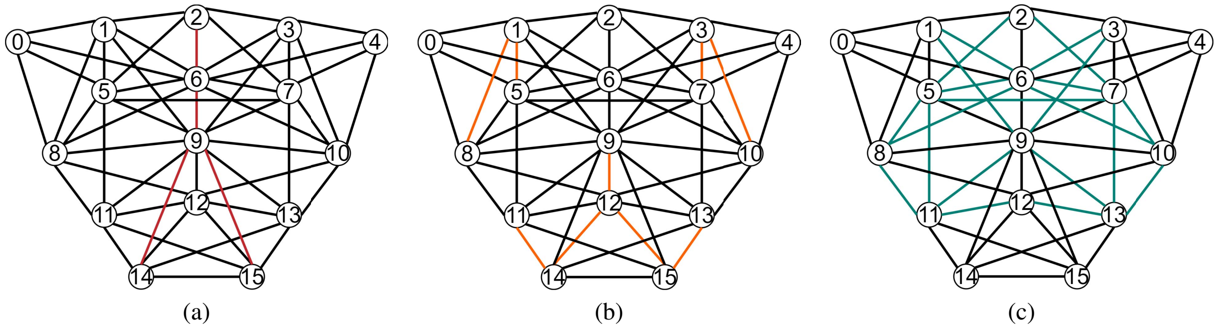


Fig. 3. Grouping of edges based on movement, (a) shows the edges with the most movement, (b) the edges with a moderate amount of movement, and (c) the edges with light movement.

where $C_i$ is the color of the points, and $T_{lower}$ and $T_{upper}$ are the lower and upper thresholds, respectively.

The points in $P_i$ are clustered with distance $d$. Clusters are formed for points within distance $d$ of each other. The value of $d$ is dependent on the subject as the distance between markers varies; if this distance is smaller than $d$ then two markers can be clustered together. Averaging the points in each cluster gives a set of key-points $K_i$.

Arbitrary labelling from clustering can differ for each frame in the sequence, so key-points are tracked throughout a sequence. Tracking is done by finding smallest pairwise distances between $K_i$ and $K_{i+1}$ and labels of $K_{i+1}$ are updated to match labels of $K_i$.

Image noise and the color of subject's clothing can cause spurious key-points or clusters that are split. Additionally, the automatic and arbitrary labelling contribute to key-point labels that differ between views of the same sequence and between sequences. To ensure a consistent labelling, remove spurious key-points, and combine split clusters, the first frame of each sequence is labelled manually and then propagated throughout the entire sequence. Manual labelling can be done for frames where tracking is lost, which usually happens because of quick motions by the subject.

The key-points are extracted and labelled for each viewpoint separately and need to be combined. Camera calibration done during the data recording gives the transformation, $\mathbf{T}$, between cameras and can be used to combine the views. However, this calibration has some significant and visible error. We reduce this error by using the iterative closest point (ICP) algorithm [4]. The entire sequence of key-points from each view are matched and used to do this correction. The positioning of the cameras discussed in section III-A is important here as the overlap of certain key-points is crucial for the correction. Key-points 2, 6, 9, 11, 14, 15 are the minimal needed overlap for good correction to occur.

The updated transform, $\mathbf{T}'$ is used to transform the clusters of each view and combine them to get new key-points. Because some clusters are only partially captured, we average the points of each cluster from each view rather than their key-points to prevent skewing the key-point heavily towards one view.

### C. Measurement

Once the views are merged and labelling of key-points is consistent between sequences, we produce measurements for each of the sequences that are then comparable. We take a subset of the edges produced by pairwise connections of each key-point, as shown in Fig. 3. The lengths of these edges are then normalized to the lengths of a canonical frame for the sequence, giving a sequence of normalized edge length changes for each edge. This canonical frame represents an at rest frame for the subject.

Differences in the speed of speech, timings of breaths, and other natural speech variations cause each sequence to be of different lengths and misaligned and create a meaningless direct comparison between sequences. To rectify these differences, the sequences are aligned to a template sequence using dynamic time warping on the audio waveform [34]. After audio alignment, we warp the measurement sequences using linear interpolation. Specifically, using the timestamps of the audio frames and image frames as an audio-to-image alignment, we sample the measurements at each audio frame. The sampling is done by linearly interpolating between the measurements:

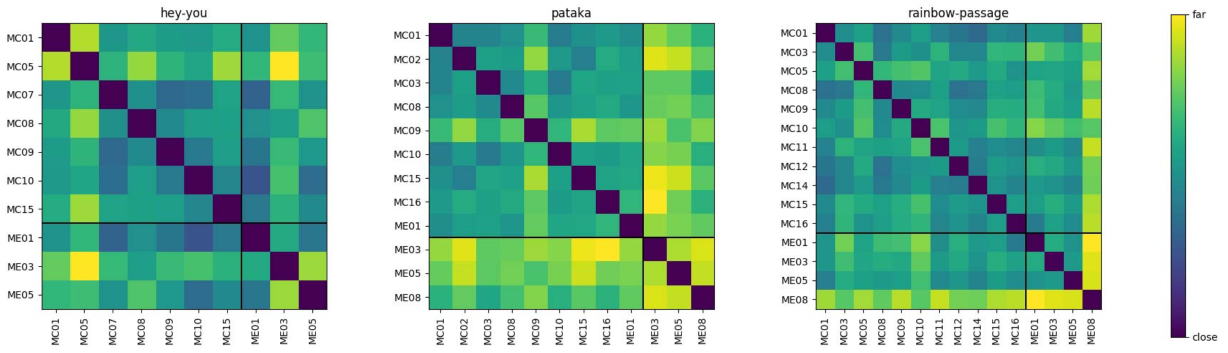
$$f_i^{resampled} = (1 - w) * f_j + w * f_{j+1}, \tag{3}$$

where $f_j$ is the measurements at frame $j$, $f_i^{resampled}$ corresponds to the measurements warped to the audio frame, $i$, and $w$ is the weight calculated as $w = (t_i^{audio} - t_{f_j})/(t_{f_{j+1}} - t_{f_j})$. Here $t_i^{audio}$ is the audio frame's timestamp and $t_{f_j}$ is the measurement's timestamp.

The aligned sequences are directly comparable. We compare sequences pairwise using the Euclidean distance of each normalized edge length.

### D. Implementation

The data collection system is implemented using the Robotic Operating System (ROS) [28], Intel's ROS Wrapper for Intel RealSense Devices for the cameras [16], and ROS audio_capture package for the microphone [17]. The wrapper handles aligning depth and color images from the RGBD cameras. Because there are two cameras we utilize the the *ApproximateTime* policy from the message_filters package of



(a) (b) (c)

Fig. 4. Direct distance comparison matrices, (a) *Hey you!*, (b) *pataka*, and (c) the *rainbow passage*. Each row and column is labelled by a subject ID, with MC indicating a control and ME indicating an experimental.

ROS to synchronize the cameras in software [12]. The markers used on the subject are 1/2-inch green circular stickers.

Calibration of camera transforms is done using ARPG's Vicalib [1]. We do a correction on this calibration using ICP with PyTorch3D's [30] implementation of Umeyama's method [41]. We implement this correction as an iterative approach but in practice only one iteration is necessary. Additionally, while calibration is done at recording time, the correction is robust enough to not need this initial calibration.

The data processing system is written in Python using PyTorch [27] for CUDA to speed up the processing. We use a distance threshold default of 750 mm for point cloud conversion to exclude background pixels. The default color thresholding uses an HSV range of [40, 70, 70] to [70, 255, 255]. The standard key-point clustering distance used is 5 mm. These are default values used for all subjects. Some subjects required slight modification, for example if markers are placed too close together, a smaller clustering distance would be required.

To align the audio sequences the dtw-python package is used for a dynamic time warping implementation [14]. We use an open end and open beginning with an asymmetric step pattern [40]. Before warping, audio is resampled from 16,000 Hz to 160 Hz. Resampling enables the warping algorithm to run with a reasonable computation time and within memory constraints.

## IV. Results and Experiments

The goal of the experiments performed is to differentiate between the control and experimental subjects. In all experiments, each sequence is first warped to a template sequence, and then direct comparisons of the warped sequences are done as described in Sec. III-C.

### A. Direct Comparison by Distance

We first compare sequences of each task directly by taking the Euclidean distance between two pairs. Following the alignment of the sequences to a template, the movement of each corresponding edge should be roughly similar.

Fig. 4 shows, specifically in the *pataka* sequences, that the controls have a smaller distance between them than the experimental sequences and the experimentals have greater variance in distances among themselves. This does not hold true in the other tasks.

Greater variance in distance between key-points and edges in the experimental group compared to the control group suggests extrinsic laryngeal muscles of vocalization move differently (i.e., with greater variance) in those with vocal strain.

### B. Comparison by Variability

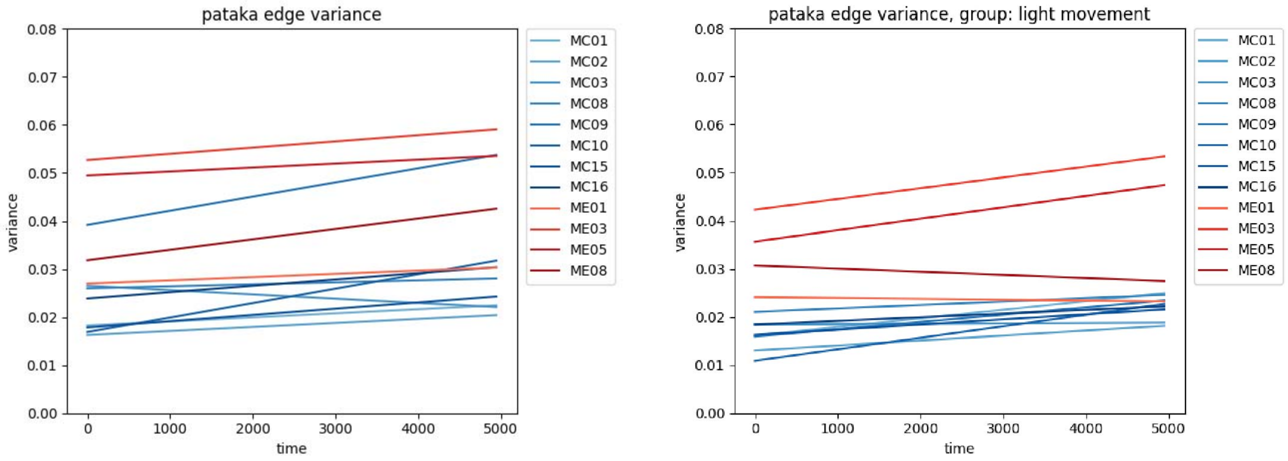
To gain further insight into how the subjects are moving, we look at the variability in their movements. This is done in two ways: variability of movement in each time frame and variability of each edge across the entire sequence. For both we use the standard deviation as a measure of variability.

The variability of movement in each time frame compares the movements of each edge at each point in time. This shows the range of movement of the participant across the sequence. The controls exhibit less variance across each frame as compared to the experimentals, shown for *pataka* in Fig. 5. Similarly to the direct distance comparison, this pattern is present for the *pataka* sequence but not the others.

Looking at the movement at each time frame gives one look at how a participant moves across the entire sequence, but fails to show some of the specifics about how they are moving. To look at what is moving we can examine the variability of each edge across the entire sequence. This gives a look at what edges are moving and which edges are not in each sequence.

### C. Edge Grouping and Comparisons

Using the results from the previous experiments, we can group the edges based on how much they move during a sequence. We group edges into four different groups, heavy movement, moderate movement, light movement, and little to no movement, these groups are shown in Fig. 3. We compare the direct distance, as in Sec. IV-A and we compare the variability of movement at each time frame as in Sec. IV-B.



(a)

(b)

Fig. 5. (a) Variance at each point in time for the *pataka* sequence. (b) Variance of edge movement in the light movement group for the *pataka* sequence.

In the variability of the edges at a point in time we see no grouping in all tasks and groups except for *pataka* in the light movement group. In this group and task, there is similar grouping of experimentals and controls that is present in the direct distance comparison and the variability comparison.

The reason for this separation can be seen in the grouping of edges and the physiology underlying those groupings. As shown in Fig. 3, high movement areas represented in (a) consists of edges going down the center of the neck, from the chin to either clavicle. The movement largely captures the up and down movement of the chin and the larynx. Moderate movement areas in Fig. 3(b) represent the jaw, suprahyoid extrinsic laryngeal muscles, and accessory muscles (scalenes, sternocleidomastoids). The light movement in Fig. 3(c) represents muscles that suspend the larynx as well as the accessory neck muscles. Movement in Fig. 3(c) areas were observed more consistently and prominently in the experimental group, suggesting these areas are more active in subjects with vocal strain.

## V. Conclusion and Future Work

These finds demonstrate the use of MoCap to identify physiological areas that underlie symptoms of vocal strain. Our data suggest neck movement patterns in patients with vocal strain differ from those without vocal strain during specific speech tasks (e.g., pataka). Specifically, greater variability of edge movement throughout a repetitive speech sequence was observed in the experimental group, with greater movement in muscles that suspend the larynx and aid in upper body posture. These findings suggest higher variability in this group, especially in specific muscle groups, could indicate the presence of vocal strain.

Increased variability in the *pataka* task is likely due to the prolonged, fast, and repetitive nature of the task that taxes the muscles involved in speech production, creating instability within the vocal system. Specifically, production of *pataka* requires quick and precise movement changes from the middle of the tongue (*pa*), to the tongue tip (*ta*), and back to the tongue base (*ka*), over and over again across 30 seconds. These quick tongue turnovers are not present in the *rainbow passage* or brief *pitch glide* and *Hey You!* vocal intensity task.

In future research, we hope to refine and identify new metrics to more precisely identify areas of vocal strain and identify vocal strain within a subject. We aim to further study specific edge and edge group movements, as well as movement not captured by the edges, with a larger group of participants. We will also determine inter- and intra-rater reliability across 5 additional subjects.

## Acknowledgment

This project is supported by UTD Seed Program for Interdisciplinary Research (SPIRe #09572). Hogue and Guo are partially supported by a grant from National Science Foundation (#2007661). Shembel is partially supported by a grant from the National Institute of Deafness and Other Communication Disorders (NIDCD R21DC019207).

## References

[1] Autonomous Robotics & Perception Group, "Vicalib," https://github.com/arpg/vicalib.

[2] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," 2020. [Online]. Available: https://arxiv.org/abs/2006.10204

[3] M. S. Benninger, C. E. Holy, P. C. Bryson, and C. F. Milstein, "Prevalence and occupation of patients presenting with dysphonia in the united states," *J. Voice*, vol. 31, no. 5, pp. 594–600, Sep. 2017.

[4] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, Feb. 1992. [Online]. Available: https://doi.org/10.1109/34.121791

[5] N. Bhattacharyya, "The prevalence of voice problems among adults in the united states," *Laryngoscope*, vol. 124, no. 10, pp. 2359–2362, Oct. 2014.

[6] A. Bigdelou, T. Benz, L. Schwarz, and N. Navab, "Simultaneous categorical and spatio-temporal 3d gestures using kinect," in *2012 IEEE Symposium on 3D User Interfaces (3DUI)*, 2012, pp. 53–60.

[7] V. Camomilla, A. Cappozzo, and G. Vannozzi, "Three-dimensional reconstruction of the human skeleton in motion," in *Handbook of Human Motion*. Cham: Springer International Publishing, 2018, pp. 17–45.

[8] C.-Y. Chang *et al.*, "Towards pervasive physical rehabilitation using microsoft kinect," in *2012 6th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, 2012, pp. 159–162.

[9] S. M. Cohen, J. Kim, N. Roy, C. Asche, and M. Courey, "The impact of laryngeal disorders on work-related dysfunction," *Laryngoscope*, vol. 122, no. 7, pp. 1589–1594, Jul. 2012.

[10] C. A. DiCesare, "A computational framework for the discovery, modeling, and exploration of task-specific human motor coordination strategies," Ph.D. dissertation, University of Cincinnati, 2020.

[11] M. Dou *et al.*, "Fusion4d," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1–13, Jul. 2016. [Online]. Available: https://doi.org/10.1145/2897824.2925969

[12] J. Faust, V. Pradeep, and D. Thomas, "ROS message filters," http://wiki.ros.org/message_filters.

[13] L. Gallo, A. P. Placitelli, and M. Ciampi, "Controller-free exploration of medical image data: Experiencing the kinect," in *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, Jun. 2011. [Online]. Available: https://doi.org/10.1109/cbms.2011.5999138

[14] T. Giorgino, "Computing and visualizing dynamic time warping alignments inir/i: Thebdtw/bpackage," *Journal of Statistical Software*, vol. 31, no. 7, 2009. [Online]. Available: https://doi.org/10.18637/jss.v031.i07

[15] P.-L. Hsieh, C. Ma, J. Yu, and H. Li, "Unconstrained realtime facial performance capture," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2015. [Online]. Available: https://doi.org/10.1109/cvpr.2015.7298776

[16] Intel, "Intel® RealSense™ ROS," https://github.com/IntelRealSense/realsense-ros.

[17] N. Koenig, "Ros audio capture," http://wiki.ros.org/audio_capture.

[18] B. Lange, C.-Y. Chang, E. Suma, B. Newman, A. S. Rizzo, and M. Bolas, "Development and evaluation of low cost game-based balance rehabilitation tool using the microsoft kinect sensor," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, Aug. 2011. [Online]. Available: https://doi.org/10.1109/iembs.2011.6090521

[19] H. Li, J. Yu, Y. Ye, and C. Bregler, "Realtime facial animation with on-the-fly correctives," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 1–10, Jul. 2013. [Online]. Available: https://doi.org/10.1145/2461912.2462019

[20] G. Mastorakis and D. Makris, "Fall detection system using kinect's infrared sensor," *Journal of Real-Time Image Processing*, vol. 9, no. 4, pp. 635–646, Mar. 2012. [Online]. Available: https://doi.org/10.1007/s11554-012-0246-9

[21] S. Misono, M. Dietrich, and J. F. Piccirillo, "The puzzle of medically unexplained symptoms-a holistic view of the patient with laryngeal symptoms," *JAMA Otolaryngol. Head Neck Surg.*, vol. 146, no. 6, pp. 550–551, Jun. 2020.

[22] M. D. Morrison, L. A. Rammage, G. M. Belisle, C. B. Pullan, and H. Nichol, "Muscular tension dysphonia," *J. Otolaryngol.*, vol. 12, no. 5, pp. 302–306, Oct. 1983.

[23] C. Nanjundeswaran, B. H. Jacobson, J. Gartner-Schmidt, and K. V. Abbott, "Vocal fatigue index (VFI): Development and validation," *Journal of Voice*, vol. 29, no. 4, pp. 433–440, Jul. 2015. [Online]. Available: https://doi.org/10.1016/j.jvoice.2014.09.012

[24] R. A. Newcombe *et al.*, "KinectFusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE International Symposium on Mixed and Augmented Reality*. IEEE, Oct. 2011. [Online]. Available: https://doi.org/10.1109/ismar.2011.6092378

[25] R. A. Newcombe, D. Fox, and S. M. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2015. [Online]. Available: https://doi.org/10.1109/cvpr.2015.7298631

[26] B. Ni, C. D. Nguyen, and P. Moulin, "RGBD-camera based get-up event detection for hospital fall prevention," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2012. [Online]. Available: https://doi.org/10.1109/icassp.2012.6287947

[27] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[28] M. Quigley *et al.*, "Ros: an open-source robot operating system," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*, Kobe, Japan, May 2009.

[29] M. A. Rahman, A. M. Qamar, M. A. Ahmed, M. A. Rahman, and S. Basalamah, "Multimedia interactive therapy environment for children having physical disabilities," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval - ICMR '13*. ACM Press, 2013. [Online]. Available: https://doi.org/10.1145/2461466.2461522

[30] N. Ravi *et al.*, "Accelerating 3d deep learning with pytorch3d," *arXiv:2007.08501*, 2020.

[31] C. A. Rosen, A. S. Lee, J. Osborne, T. Zullo, and T. Murry, "Development and validation of the voice handicap index-10," *The Laryngoscope*, vol. 114, no. 9, pp. 1549–1556, Sep. 2004. [Online]. Available: https://doi.org/10.1097/00005537-200409000-00009

[32] A. Russell, J. Oates, and K. M. Greenwood, "Prevalence of voice problems in teachers," *Journal of Voice*, vol. 12, no. 4, pp. 467–479, Jan. 1998. [Online]. Available: https://doi.org/10.1016/s0892-1997(98)80056-8

[33] S. Saini, D. R. A. Rambli, S. Sulaiman, M. N. Zakaria, and S. R. M. Shukri, "A low-cost game framework for a home-based stroke rehabilitation system," in *2012 International Conference on Computer &; Information Science (ICCIS)*. IEEE, Jun. 2012. [Online]. Available: https://doi.org/10.1109/iccisci.2012.6297212

[34] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[35] W. S. Selbie and M. J. Brown, "3D dynamic pose estimation from marker-based optical data," in *Handbook of Human Motion*. Cham: Springer International Publishing, 2017, pp. 1–20.

[36] M. Sliwinska-Kowalska *et al.*, "The prevalence and risk factors for occupational voice disorders in teachers," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 2, pp. 85–101, 2006. [Online]. Available: https://doi.org/10.1159/000089610

[37] S. Smolander and K. Huttunen, "Voice problems experienced by finnish comprehensive school teachers and realization of occupational health care," *Logopedics Phoniatrics Vocology*, vol. 31, no. 4, pp. 166–171, Jan. 2006. [Online]. Available: https://doi.org/10.1080/14015430600576097

[38] E. Stone and M. Skubic, "Passive, in-home gait measurement using an inexpensive depth camera: Initial results," in *Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, 2012. [Online]. Available: https://doi.org/10.4108/icst.pervasivehealth.2012.248731

[39] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 1–14, Nov. 2015. [Online]. Available: https://doi.org/10.1145/2816795.2818056

[40] P. Tormene, T. Giorgino, S. Quaglini, and M. Stefanelli, "Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation," *Artificial Intelligence in Medicine*, vol. 45, no. 1, pp. 11–34, Jan. 2009. [Online]. Available: https://doi.org/10.1016/j.artmed.2008.11.007

[41] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, Apr. 1991. [Online]. Available: https://doi.org/10.1109/34.88573

[42] T. Yu *et al.*, "BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017. [Online]. Available: https://doi.org/10.1109/iccv.2017.104

[43] ——, "DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2523–2539, Oct. 2020. [Online]. Available: https://doi.org/10.1109/tpami.2019.2928296

[44] Z. Zhang, W. Liu, V. Metsis, and V. Athitsos, "A viewpoint-independent statistical method for fall detection," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 3626–3630.