ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 102 (2016) 745 – 751

DOI 10.3813/AAA.918990

Validating Stereo-Endoscopy with a Synthetic Vocal Fold Model

Kimberly A. Stevens[1], Ryo Shimamura[2], Hiroshi Imagawa[3], Ken-Ichi Sakakibara[4], Isao T. Tokuda[2]

[1] Department of Mechanical Engineering, Brigham Young University, Provo, Utah 84602, USA.
[2] Department of Mechanical Engineering, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan. isao@fc.ritsumei.ac.jp
[3] Department of Otolaryngology, University of Tokyo Hospital, 7-3-1 Hongo, Tokyo 113-8655, Japan.
[4] Department of Communication Disorders, Health Sciences University of Hokkaido 2-5 Ainosato, Hokkaido 002-8072, Japan.

Summary

Endoscopy has long been an important technique for vocal fold visualization, but it is limited to providing only two-dimensional information. To extend its usage, stereo-endoscopy has been developed to provide two views of the larynx from different angles, enabling three-dimensional viewing of the laryngeal structure. Stereo-matching has recently been applied to reconstruct three-dimensional vocal fold motion from stereo-endoscopic measurement [1, 2]. Further information regarding its accuracy and applicability would make it more clinically relevant. To validate the stereo-endoscopy system, a setup of multiple high-resolution cameras and the stereo-endoscopy system simultaneously recorded the surface of synthetic, self-oscillating vocal fold models. The estimated error in the measurement of the three-dimensional location by the stereo-endoscopy system is low in the transverse plane, whereas the error is relatively large in the inferior-superior direction. Although further improvement is needed to provide reliable information on the inferior-superior movement of the vocal folds, we conclude that the stereo-endoscope is applicable to *in vivo* measurements of absolute distances of the glottis in the transverse plane such as glottal length, width, and area.

PACS no. 43.70.Jt, 43.70.Gr

1. Introduction

Stroboscopy and high-speed video endoscopy have been used for decades in both research and clinical settings to gain insight into the dynamics of the vocal folds and to diagnose voice disorders [3, 4, 5]. However, such tools are limited to providing information regarding glottal movement in the transverse plane. Yet, information outside of that plane would be of benefit to both clinicians and researchers, allowing the observation of left-right vertical asymmetry, measurement of absolute lengths, and providing further insight into the motion of the mucosal wave, which has been shown to be very important in phonation [6].

Various techniques have been employed to obtain three-dimensional information of the vocal fold dynamics. Three-dimensional measurement of vocal fold vibration has been obtained by observation of the medial surface of the vocal folds in a hemi-larynx setup, where half of the larynx is removed to allow visual access to the medial surface. Such a technique is limited to excised larynges, or at best *in vivo* canine [7, 8, 9]. Excised larynges lack the muscle activities present in *in vivo* vocal folds, and the material properties of the tissue begins to change upon excision. Although valuable insight into the motion of the mucosal wave is gained, this technique cannot completely capture the dynamics of real human vocal fold vibration. Moreover, these studies have attached small markers to the vocal folds in order to track their movement, which is not possible with human subjects.

Non-invasive techniques with high-frame-rate ultrasound [10] or stroboscopic x-ray laminagraphy [11] have allowed the observation of bulk motion, such as body-layer movement of the vocal folds. However, these techniques are not sufficiently sensitive to capture detailed vocal fold surface movements. Other researchers have used various optical techniques to augment traditional endoscopy systems to obtain three-dimensional information [1, 12, 13, 14, 15, 16]. Luegmair *et al.* [15] used a high-speed camera and a laser projection system to reconstruct the surface of the vocal folds during vibration. Computational method for the derivation of the vocal fold surface was further improved to quantify the vocal fold kinematics [16]. A miniaturized hardware for the set-up of [15, 16],

Received 19 May 2015,
accepted 19 April 2016.

Author's complimentary copy

which slightly exceeds the size of a rigid standard laryngoscope, was recently introduced to human *in vivo* recordings [17]. Disadvantage of the laser projection system is that a tedious and complicated calibration procedure is required whenever it is combined with the highs-peed camera. Furthermore, laser systems provide only point-wise information and cannot trace the medial edge when it falls between the laser dots, rendering the capture of surface dynamics, such as tracing the moving location of the medial edge, impossible. Optical coherence tomography (OCT) has also been utilized to reconstruct *in vivo* vocal fold motion at a point [18]. The advantage of the OCT methods is that it can penetrate into the vocal fold tissue approximately 2 mm and show the movement of both the epithelium and underlying tissue. Although triggered OCT has been successful in reconstructing three-dimensional displacement of excised calf larynges, it requires scanning of the vocal fold tissue over many periodic cycles, and artifacts can be introduced if the signal is not perfectly periodic [19].

The present work focuses on a stereo-endoscopic observation of the vocal folds as an alternative approach to non-invasive three-dimensional measurements [20, 21, 22].

Stereo-endoscopy provides two views of the larynx from different angles, enabling three-dimensional viewing of the laryngeal structure. With its compact design, the stereo-endoscope has a relatively low impact on articulators and controlling musculature of the vocal folds, thus enabling a natural vocalization. Once the camera is calibrated, it does not require further adjustment between other measurement devices, *e.g.*, alignment of lasers or unique calibration for each subject. Imagawa *et al.* [23] applied the stereo-endoscopy for estimating the absolute glottal dimensions such as the glottal width, length, and area functions of human subjects. To recover three-dimensional dynamics from the stereo-endoscopic measurement of vocal folds *in vivo*, Tokuda *et al.* [1] and Sommer *et al.* [2] applied stereo-matching, a widely developed technique in a computer vision research. They extracted quantitative information regarding vocal fold movement in three-dimensions. However, their work has not yet been extensively validated. The purpose of the work described here was to evaluate the accuracy of the stereo-endoscopy technique by comparing it with a standard stereo-vision technique using two high-resolution cameras. In a controlled environment, three-dimensional motion of a synthetic vocal fold model was simultaneously recorded by the two imaging techniques and compared to evaluate the efficacy of the stereo-endoscopy system.

Synthetic vocal fold models are ideally suited for the task of evaluating the stereo-endoscope. Their motion is similar to the vibration of *in vivo* vocal folds, but their location outside of the larynx provided a more controlled environment and permitted the use of a secondary imaging system to evaluate the accuracy of the stereo-endoscope. Without limitations on the size or location of cameras, the secondary imaging system was able to have a much higher resolution and wider baseline, making it more accurate than the stereo-endoscope. Because of this, the secondary imaging system was considered "ground truth" in evaluating the stereo-endoscope. The accuracy of the stereo-endoscope is heavily influenced by the distance from the imaging surface, but patient comfort often precludes ideal placement of the endoscope during an exam. However, synthetic models allow for fixed and ideal placement of the endoscope. Additionally, models allow a controlled lighting system without the spectral reflections commonly encountered in an *in vivo* environment. Correspondence between left and right camera views is a major challenge encountered in using stereo-vision inside the larynx. With models, tracking markers can be used, eliminating this challenge, as well as allowing precise comparison of points between the stereo-endoscope system and the evaluating system. With many of the challenges of stereo-vision eliminated by the use of synthetic models, the accuracy of the stereo-endoscope was more easily evaluated.

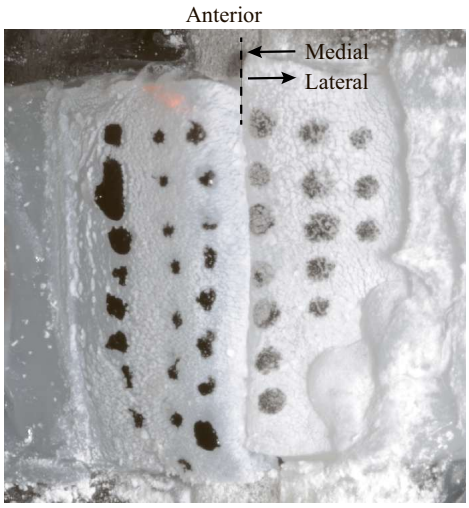

Figure 1. Grid of dots painted on the model to eliminate error due to the correspondence problem.

2. Method

2.1. Experimental Setup

A synthetic vocal fold model (two-layer M5 model) was created, based on the geometry of Drechsel and Thomson [24] and following the methods of Murray and Thomson [25, 26]. A grid of black dots was painted onto the surface in order to match points between recording systems, as shown in Figure 1. The model vibrated at an onset pressure of 0.74 kPa with a frequency of 108 Hz, comparable to that of Drechsel and Thomson [24]. The length of the duct upstream from the vocal folds was 70 cm. Further details regarding the experimental setup can be found in [27]. The vibrating surface of the model was recorded simultaneously with the stereo-endoscope (Nagashima Medical Instrument Corporation) and two digital single lens reflex (DSLR) cameras (Nikon D5100) as shown in Figure 2.

Author's complimentary copy

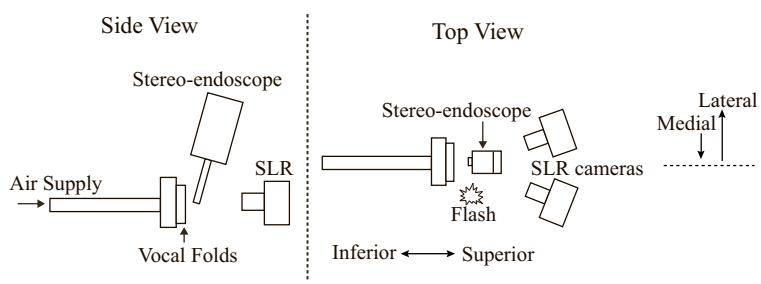

Figure 2. The surface of the vocal folds was recorded simultaneously with 2 DSLR cameras and the stereo-endoscope.

The two DSLR cameras with 105 mm lenses (AF Micro Nikkor) comprised the secondary imaging system used to evaluate the stereo-endoscope. The centers of the two DSLR cameras were approximately 20 cm apart and 60° apart. The stereo-endoscope included two independent, ordinary, rigid optical systems with diameters of 4 mm, a fiber-optic light guide, an optical connector, a light source, and a camera. The tips of the optical systems house objective lenses with prisms designed for 70° oblique-angled view, with a field angle of 40°. The distance between the optical axes of the tips was 6 mm. The stereo-endoscope was attached to a CCTV lens with a focal length of 50 mm, which was connected to the high-speed digital camera (Photron FASTCAM 1024PCI). The high-speed recordings were made at 10,000 frames per second with a shutter speed of 1/1,000,000 second. A flash (Bolt VX-760N) was used to synchronize timing between the two imaging systems. While the stereo-endoscope was recording, the shutter on both DSLR cameras was opened, the flash was fired at 1/8 power, and the shutters were closed. The camera settings and ambient lighting were adjusted such that only the light from the flash was visible in both systems.

2.2. Stereo-endoscopy System

A representative stereo-endoscopic image from the high-speed camera is shown in Figure 3. Pixel-based coordinates x and y are defined, with the origin located at the top left of the stereo image. With respect to a feature point indicated by diamonds in both left and right images, the main quantities needed to recover its three-dimensional location are the vertical distance, D_V, and horizontal distances, D_L and D_R, from the corresponding centers of the optical field (circles). Locations of the centers of the optical field, (x_{LC}, y_{LC}) and (x_{RC}, y_{RC}), in the left and right images are determined through calibration, and they remain constant for a given optical set-up. Note that the lowercase coordinates x and y correspond to pixel locations in the image, whereas the uppercase coordinates X, Y, and Z represent physical dimensions relative to the endoscope probe head as defined in Figure 4. X and Y are in the plane normal to the stereo-endoscope and Z is orthogonal to that plane. In terms of physiological coordinates, X is in the medial-lateral direction, Y is in the anterior-posterior direction, and Z is in the inferior-superior direction.

Given the pixel coordinates of the feature in the left and right images, (x_L, y_L) and (x_R, y_R), the quantities D_L,

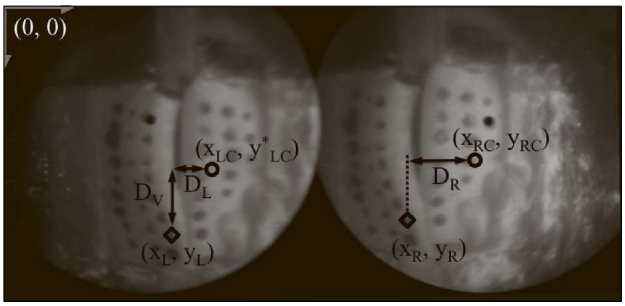

Figure 3. Representative image from the stereo-endoscope.

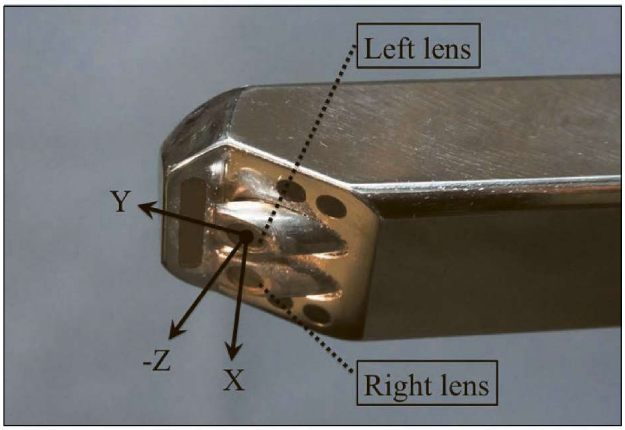

Figure 4. Stereo-endoscope.

D_R, and D_V are given by:

$$D_L = x_L - x_{LC}, \quad (1)$$
$$D_R = x_R - x_{RC}, \quad (2)$$
$$D_V = y_{LC} - 0.5(y_L + y_R). \quad (3)$$

The left-right disparity is defined as $\Delta = D_L - D_R$. The Cartesian coordinates of the feature point are then estimated as [28]:

$$\tilde{Z} = 1/(k_1\Delta + k_2), \quad (4)$$
$$\tilde{X} = k_3\tilde{Z}D_L + k_4, \quad (5)$$
$$\tilde{Y} = k_5\tilde{Z}D_V + k_6, \quad (6)$$

where the constants k_1 through k_6 are calibrated for a given optical set-up. Equation (4) is, in general, valid for an idealistic situation. To account for optical distortion inherent in the lens and camera systems, a modified version

Author's complimentary copy

of Equation (4) having higher-order polynomials [29] is used as

$$\tilde{Z} = 1/\left\{k_1\left[\Delta + c_1 D_L^2 + c_2 D_L + c_3 D_V^2 + c_4 D_V\right] + k_2\right\}. \quad (7)$$

The calibration constants $\{k_i\}$ and $\{c_i\}$ were optimized for the current camera setup with the following procedure. Images on a Cartesian graph paper with a grid of 5 mm by 5 mm were captured at distances ranging from Z = 32 mm to Z = 52 mm. At each distance, the X-, Y-, and Z-coordinates of the grid points with known dimensions D_L, D_R, and D_V were obtained. The calibration constants were then determined by the least squares method to best fit the data. The constant values were obtained as $c_1 = 1.4{\cdot}10^{-4}$, $c_2 = -2.3{\cdot}10^{-3}$, $c_3 = 6.2{\cdot}10^{-5}$, $c_4 = 2.6{\cdot}10^{-3}$, $k_1 = 3.197{\cdot}10^{-4}$, $k_2 = -3.248{\cdot}10^{-3}$, $k_3 = 1.713{\cdot}10^{-3}$, $k_4 = -2.069$, $k_5 = 1.724{\cdot}10^{-3}$, $k_6 = -5.083{\cdot}10^{-2}$, $x_{LC} = 160.5$, $y_{LC} = 141.0$, $x_{RC} = 534.0$.

The average calibration errors were $\langle|X - \tilde{X}|\rangle = 0.11$ mm, $\langle|Y - \tilde{Y}|\rangle = 0.088$ mm, and $\langle|Z - \tilde{Z}|\rangle = 0.18$ mm in X-axis, Y-axis, and Z-axis, respectively, where $\langle\cdot\rangle$ denotes averaging over all calibration points. The average calibration error in all three dimensions was $\langle\sqrt{(X-\tilde{X})^2 + (Y-\tilde{Y})^2 + (Z-\tilde{Z})^2}\,\rangle = 0.27$ mm.

2.3. DSLR System

The image space for the DSLR system was calibrated using a direct linear transformation (DLT), as described by Abdel-Aziz and Karara [30]. A checkerboard pattern (10 cm by 10 cm) was placed in view of both cameras, a picture was taken, and then the target was moved forward by 2 mm. This was repeated 7 times. The centers of the white portions of the checkerboard were found automatically using connected components, a common computer vision technique. In the field of view of each image, 28 centers were visible. On each of the 7 images located in different depth at increments of 2 mm, 28 points were extracted. As a result, a total of 196 points were collected and used as the calibration points.

Calibration points from the left and right camera images are denoted as (u_L, v_L) and (u_R, v_R), respectively, whereas the corresponding three-dimensional location is represented by (X, Y, Z). For all calibration points, the Moore–Penrose pseudo-inverse method determined the DLT, which maps (u_L, v_L) and (u_R, v_R) to $(\hat{X}, \hat{Y}, \hat{Z})$. In order to ascertain the accuracy of the transform, the average error was calculated for all calibration points. The programs were written and executed in MATLAB (MathWorks, Inc., V7.13 (R2011b)). The obtained errors were $\langle|X - \hat{X}|\rangle = 0.031$ mm, $\langle|Y - \hat{Y}|\rangle = 0.016$ mm, $\langle|Z - \hat{Z}|\rangle = 0.073$ mm, and $\langle\sqrt{(X-\hat{X})^2 + (Y-\hat{Y})^2 + (Z-\hat{Z})^2}\,\rangle = 0.0808$ mm.

2.4. Application to synthetic vocal fold model

For each point in time, three-dimensional locations of the markers on the synthetic vocal fold model were estimated as follows. In the DSLR system, the locations of the markers were automatically selected using connected components. The three-dimensional locations of the marked points were then computed using the DLT algorithm. In the stereo-endoscopy system, the locations of the corresponding points were selected manually. To improve the accuracy, cross-correlation was computed between the left and right images when selecting the corresponding points. Because the marked points were located not exactly on the pixel line but somewhere in the middle, a *Gaussian* 3-point curve, which assumes a *Gaussian* distribution for the cross-correlation function in the region around the peak, was fitted to the surface to achieve sub-pixel accuracy. Three points (the peak and two points directly adjacent to the peak) were used to fit a curve to the cross correlation function in the vertical and horizontal directions, as described in [31, 32]. The three-dimensional locations of the points were then estimated by Equations (1)–(7).

3. Results

The coordinate spaces of the two systems, *i.e.*, DSLR cameras and stereo-endoscope, were shifted to the same origin and rotated to match the anatomical directions. As an example, locations of the marker points in three-dimensions are plotted in Figure 5. The error in each direction for the three-dimensional locations of the points was computed using Equations (8)–(10). The total error is obtained by Equation 11.

$$\epsilon_X = \frac{1}{N}\sum_{i=1}^{N}\left|\tilde{X}_i - \hat{X}_i\right|, \quad (8)$$

$$\epsilon_Y = \frac{1}{N}\sum_{i=1}^{N}\left|\tilde{Y}_i - \hat{Y}_i\right|, \quad (9)$$

$$\epsilon_Z = \frac{1}{N}\sum_{i=1}^{N}\left|\tilde{Z}_i - \hat{Z}_i\right|, \quad (10)$$

$$\epsilon_{XYZ} = \frac{1}{N}\sum_{i=1}^{N}\left\{\left(\tilde{X} - \hat{X}\right)^2 + \left(\tilde{Y} - \hat{Y}\right)^2 + \left(\tilde{Z} - \hat{Z}\right)^2\right\}^{1/2}. \quad (11)$$

Here, $(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)$ and $(\hat{X}_i, \hat{Y}_i, \hat{Z}_i)$ represent three dimensional location of the i-th marker estimated by stereo-endoscopy and DSLR camera systems, respectively. The errors are summarized in Table I for four different images, taken at varying points in the periodic cycle of the vocal fold model vibration. The phase of the glottal cycle is indicated by the variable ϕ, where the zero phase corresponds to maximum glottal pressure. Because of the differences in lighting between the left and right sides of the models, the two sides were analyzed and summarized separately. The last column of Table I indicates the number of the points estimated on each side. In each direction, relative errors in percentage are indicated in parentheses. For anterior-posterior direction, the errors were normal-

Author's complimentary copy

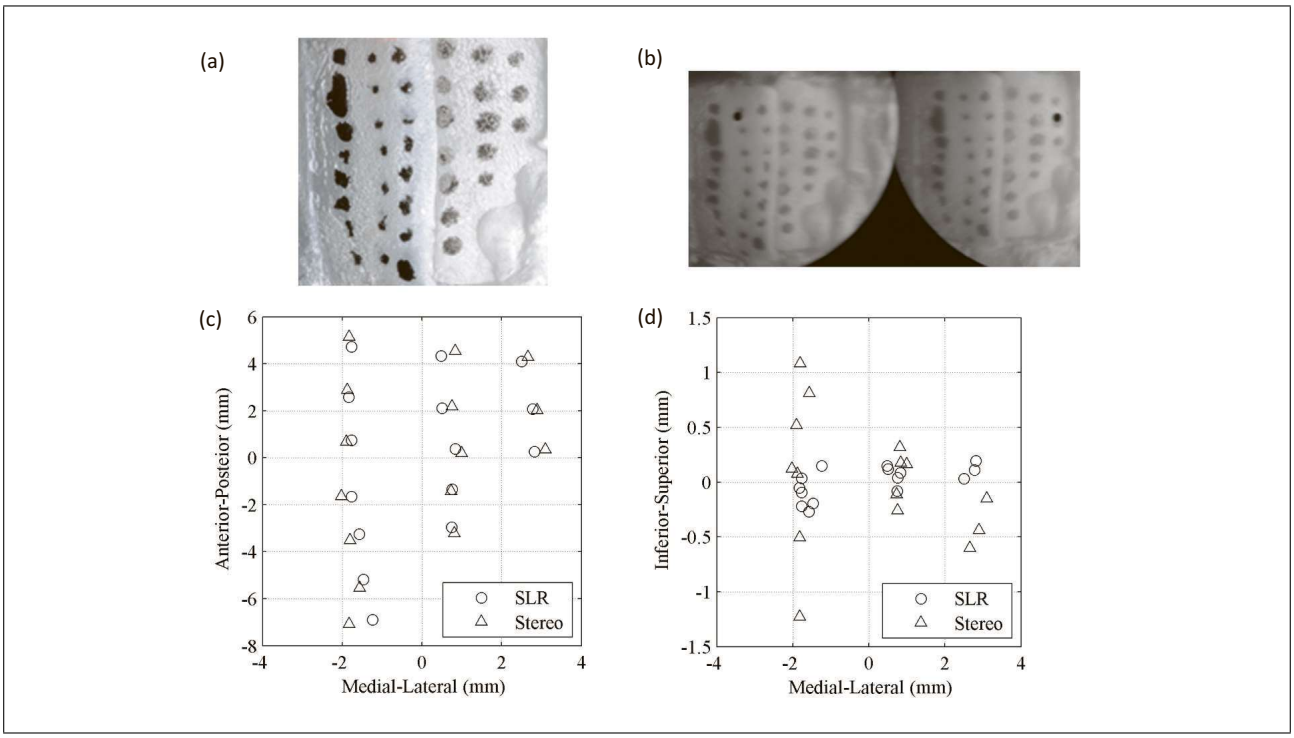

Figure 5. Images (No. 4) of the vocal fold model surface, taken with the DSLR (a) and stereo-endoscope (b). The estimated locations of the markers are plotted in (X, Y)-coordinate (c) and (X, Z)-coordinate (d).

ized by the glottal length of 17 mm. For medial-lateral direction, the errors were normalized by the maximum displacement of any of the points in that direction, which was 3.46 mm. For inferior-superior direction, the error was normalized by the maximum displacement of any of the points in that direction, which was 1.78 mm. The average errors in the X- and Y-directions are relatively small compared to the error in the Z-direction. This is not surprising, given the small baseline between the cameras on the stereo-endoscope (6 mm). Such a small baseline is necessary for the endoscope to be small enough for use *in vivo*. The mean error of 0.557 mm in the Z-direction is comparable to those calculated by Sommer *et al.* [2], in which average errors of 2.1, 2.0, and 0.74 mm were reported for three data sets. The error obtained by the present study is smaller than those of the previous study, because the error source included calibration error, mismatch of stereo points, and measurement error in [2]. In the present study, the mismatch error has been greatly reduced because of the well-defined markers.

The relatively low error in the measurement of absolute distances in the transverse (X, Y)-plane suggests that a potential application for the stereo-endoscope would be to obtain measurements of distances in that plane, such as glottal length, width, and area. It is impossible to measure absolute distances with a single camera without a mechanism for calibrating, such as a ruler or object of known size located in the same plane as the measurements. However, the placement of such an object is very difficult *in vivo*. Even though the absolute error in the inferior-superior (Z) direction is greater than in the transverse plane, the stereo-endoscope may still provide useful information regarding the relative heights of features along the surface of the vocal folds (*e.g.*, which of the left and right vocal fold surfaces is higher than the other) as well as the surface movement during one cycle of the vocal fold oscillation (*e.g.*, [2]).

4. Conclusions and Discussions

This paper utilized a synthetic vocal fold model as an idealistic object to evaluate accuracy of stereo-endoscopy to make three-dimensional measurements. The results were compared with those of two DSLR cameras, used as a secondary imaging system to provide a more accurate geometrical description of the object. Reconstruction of the three-dimensional coordinates using stereo-vision techniques is fraught with potential sources of error, including error from calibration and correspondence. Typically, finding good correspondence between left and right images is the most challenging aspect of three-dimensional imaging. This problem is compounded inside the glottis; poor lighting, smooth, continuous, and monochromatic tissue surface, non-distinct edges, spectral reflections from the mucosa, *etc.* make finding good correspondence especially challenging. Sommer *et al.* [2] have focused on this aspect of the problem, significantly improving the accuracy of the results. The present study focused on the triangulation part of the process, evaluating the error that comes mainly from hardware limitation and calibration of the stereo-endoscope. In order to significantly reduce any error due to correspondence between images, specific points were marked on the surface of the vocal folds; this work focused on finding the location of those specific points, rather than on a dense reconstruction of the entire surface.

Author's complimentary copy

Table I. Averaged estimation error in each direction (mm) measured by distance between stereo-endoscope and DSLR cameras. For anterior-posterior direction, the relative errors in square parentheses were normalized by the glottal length. For medial-lateral and inferior-superior directions, the errors were normalized by the maximum displacement of the markers in the corresponding direction. The last column indicates number of the marker points estimated on each side. ϕ indicates phase of the glottal cycle, where the zero phase corresponds to maximum glottal pressure.

Data #	side	Medial-lateral (X)	Anterior-posterior (Y)	Inferior-superior (Z)	Total	# points
# 1	right	0.175 mm [5.1%]	0.232 mm [1.4%]	0.443 mm [24.9%]	0.599 mm	$N = 15$
($\phi = 0\,°$)	left	0.287 mm [8.3%]	0.280 mm [1.6%]	0.574 mm [32.2%]	0.778 mm	$N = 24$
# 2	right	0.199 mm [5.8%]	0.254 mm [1.5%]	1.266 mm [71.1%]	1.366 mm	$N = 15$
($\phi = 43.5\,°$)	left	0.252 mm [7.3%]	0.347 mm [2.0%]	0.331 mm [18.6%]	0.575 mm	$N = 8$
# 3	right	0.161 mm [4.7%]	0.229 mm [1.3%]	0.375 mm [21.1%]	0.530 mm	$N = 15$
($\phi = 73.1\,°$)	left	0.312 mm [9.0%]	0.245 mm [1.4%]	0.320 mm [18.0%]	0.583 mm	$N = 8$
# 4	right	0.188 mm [5.4%]	0.183 mm [1.1%]	0.527 mm [29.6%]	0.636 mm	$N = 15$
($\phi = 92.6\,°$)	left	0.325 mm [9.4%]	0.284 mm [1.7%]	0.618 mm [34.7%]	0.815 mm	$N = 21$
Mean		0.237 mm [6.9%]	0.257 mm [1.5%]	0.557 mm [31.3%]	0.735 mm	

In addition to reducing the correspondence challenge (matching locations in left and right images), synthetic models provided the ideal imaging situation. The images were recorded at an optimal distance from the stereo-endoscope to the surface of the vocal folds. In an *in vivo* environment, optimal placement of the endoscope is rarely possible, and the challenges with correspondence are still present. Such ideal conditions should be kept in mind when considering the present results. By focusing on the hardware capability (*i.e.*, sensitivity of the stereo lenses and resolution of the high-speed camera) as the sources of error in the process of three-dimensional reconstruction, the present results highlight the potential and limitations of the current stereo-endoscope, suggesting directions for future development and areas in which the current endoscope may be of utility. The fine precision of the absolute distance measurements in the transverse plane is of particular note.

Though further development might be needed before highly accurate information regarding the dynamics of the mucosal wave or of small features on the surface of the vocal folds can be extracted from the present stereo-endoscope, stereo-endoscopy still could potentially be used to measure absolute distances *in vivo*.

Using the stereo-endoscopy, Imagawa *et al.* [23] measured absolute glottal dimensions of human larynx *in vivo*. For vocalization of a female subject (fundamental frequency of 230 Hz), the maximal glottal length, width, and area were estimated as 7 mm, 2 mm, and 11 mm^2, respectively. The averaged errors of 0.257 mm and 0.237 mm obtained in the present study for the anterior-posterior and medial-lateral directions imply that the estimation errors would be less than 4%, 12%, and 8% for the glottal length, width, and area, respectively. We therefore conclude that the present technique is of practical use for measuring the vocal fold dimension in the transverse plane, allowing a certain amount of estimation error. Since the vertical range of the vocal fold movement is reported to be around 1.76 to 2.75 mm [8, 15], the averaged error of 0.557 mm obtained for the present hardware in the inferior-superior direction is insufficient to provide a reliable estimate of the vertical dynamics. Application of the present technique to *in vivo* human larynx may increase the error even more, because the problem of stereo-mismatch, which has been avoided in the present study, should arise. In order for the present technique to be applicable to the measurement in inferior-superior direction, further improvement is needed. Image resolution of the high-speed video should be much more increased and at the same time sensitivity of the lenses used for the endoscopy should be strongly enhanced. Calibration error caused by the optical distortion can be further reduced by introducing a machine learning technique, which optimizes nonlinear function to approximate the distortion. An alternative approach, *e.g.*, optical technique [17], might be also efficient for the measurement of the inferior-superior glottal dynamics.

Acknowledgment

This work was partially supported by the JSPS-NSF Summer Program (No. SP14059), National Science Foundation (No. IIA-1414741), and Grant-in-Aid for Scientific Research (No. 25540074, No. 23300071, No. 25861530) from Japan Society for the Promotion of Science (JSPS). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] I. T. Tokuda, M. Iwawaki, K.-I. Sakakibara, H. Imagawa, T. Nito, T. Yamasoba, N. Tayama: Reconstructing three-dimensional vocal fold movement via stereo matching. Acoustical Science and Technology **34** (2013) 374–377.

[2] D. E. Sommer, I. T. Tokuda, S. D. Peterson, K.-I. Sakakibara, H. Imagawa, A. Yamauchi, T. Nito, T. Yamasoba, N. Tayama: Estimation of inferior-superior vocal fold kinematics from high-speed stereo endoscopic data in vivo. The Journal of the Acoustical Society of America **136** (2014) 3290–3300.

Author's complimentary copy

[3] P. Kitzing: Stroboscopy - a pertinent laryngological examination. The Journal of Otolaryngology **14** (1985) 151–157.

[4] S. Kiritani, H. Hirose, H. Imagawa: High-speed digital image analysis of vocal cord vibration in diplophonia. Speech Communication **13** (1993) 23–32.

[5] T. Baer, A. Löfqvist, N. S. McGarr: Laryngeal vibrations: A comparison between high-speed filming and glottographic techniques. The Journal of the Acoustical Society of America **73** (1983) 1304–1308.

[6] D. Bless, M. Hirano, R. Feder: Videostroboscopic evaluation of the larynx. Ear, Nose, & Throat Journal **66** (1987) 289–296.

[7] J. Van den Berg, T. Tan: Results of experiments with human larynxes. Pract Otorhinolaryngol (Basel) **21** (1959) 425–450.

[8] M. Döllinger, D. A. Berry, G. S. Berke: Medial surface dynamics of an in vivo canine vocal fold during phonation. The Journal of the Acoustical Society of America **117** (2005) 3174–3183.

[9] D. A. Berry, D. W. Montequin, N. Tayama: High-speed digital imaging of the medial surface of the vocal folds. The Journal of the Acoustical Society of America **110** (2001) 2539–2547.

[10] S. Tang, Y. Zhang, X. Qin, S. Wang, M. Wan: Measuring body layer vibration of vocal folds by high-frame-rate ultrasound synchronized with a modified electroglottograph. The Journal of the Acoustical Society of America **134** (2013) 528–538.

[11] S. Saito, H. Fukuda, S. Kitahara, Y. Isogai, T. Tsuzuki, H. Muta, E. Takayama, T. Fujioka, N. Kokawa, K. Makino: Pellet tracking in the vocal fold while phonating - experimental study using canine larynges with muscle activity. In: Vocal fold physiology. I. Titze, R. Scherer (eds.). Denver Center for the Performing Arts, Denver, CO, USA, 1985, 169–182.

[12] M. Schuster, J. Lohscheller, P. Kummer, U. Eysholdt, U. Hoppe: Laser projection in high-speed glottography for high-precision measurements of laryngeal dimensions and dynamics. European Archives of Oto-Rhino-Laryngology and Head & Neck **262** (2005) 477–481.

[13] T. Wurzbacher, I. Voigt, R. Schwarz, M. Döllinger, U. Hoppe, J. Penne, U. Eysholdt, J. Lohscheller: Calibration of laryngeal endoscopic high-speed image sequences by an automated detection of parallel laser line projections. Medical Image Analysis **12** (2008) 300–317.

[14] N. A. George, F. F. de Mul, Q. Qiu, G. Rakhorst, H. K. Schutte: Depth-Kymography: High-speed calibrated 3D imaging of human vocal fold vibration dynamics. Physics in Medicine and Biology **53** (2008) 2667–2675.

[15] G. Luegmair, S. Kniesburges, M. Zimmermann, A. Sutor, U. Eysholdt, M. Dollinger: Optical reconstruction of high-speed surface dynamics in an uncontrollable environment. IEEE Transactions on Medical Imaging **29** (2010) 1979–1991.

[16] G. Luegmair, D. D. Mehta, J. B. Kobler, M. Döllinger: Three-dimensional optical reconstruction of vocal fold kinematics using high-speed video with a laser projection system. IEEE Transactions on Medical Imaging **34** (2015) 2572–2582.

[17] M. Semmler, S. Kniesburges, V. Birk, A. Ziethe, R. Patel, M. Döllinger: 3D reconstruction of human laryngeal dynamics based on endoscopic high-speed recordings. IEEE Transactions on Medical Imaging (2016) In press.

[18] L. Yu, G. Liu, M. Rubinstein, A. Saidi, B. J. Wong, Z. Chen: Office-based dynamic imaging of vocal cords in awake patients with swept-source optical coherence tomography. Journal of Biomedical Optics **14** (2009) 064020.

[19] E. W. Chang, J. B. Kobler, S. H. Yun: Triggered optical coherence tomography for capturing rapid periodic motion. Scientific Reports **1** (2011) 48.

[20] M. Sawashima, S. Miyazaki: Stereo-fiberscopic measurement of the larynx: a preliminary experiment by use of ordinary laryngeal fiberscopes. Annual Bulletin Research Institute of Logopedics and Phoniatrics, University of Tokyo **8** (1974) 7–12.

[21] O. Fujimura, T. Baer, S. Niimi: A stereo-fiberscope with a magnetic interlens bridge for laryngeal observation. The Journal of the Acoustical Society of America **65** (1979) 478–480.

[22] M. Sawashima, H. Hirose, K. Honda, H. Yoshioka, S. Hibi, N. Kawase, M. Yamada: Stereoendoscopic measurement of the laryngeal structure. – In: Vocal Fold Physiology: Contemporary Research and Clinical Issues. D. Abbs (ed.). College-Hill Press, San Diego, CA, USA, 1983, 264–276.

[23] H. Imagawa, K.-I. Sakakibara, I. T. Tokuda, M. Otsuka, N. Tayama: Estimation of glottal area function using stereo-endoscopic high-speed digital imaging. Interspeech, 2010, 1005–1008.

[24] J. S. Drechsel, S. L. Thomson: Influence of supraglottal structures on the glottal jet exiting a two-layer synthetic, self-oscillating vocal fold model. The Journal of the Acoustical Society of America **123** (2008) 4434–4445.

[25] P. R. Murray, S. L. Thomson: Synthetic, multi-layer, self-oscillating vocal fold model fabrication. Journal of Visualized Experiments **58** (2011) e3498.

[26] P. R. Murray, S. L. Thomson: Vibratory responses of synthetic, self-oscillating vocal fold models. The Journal of the Acoustical Society of America **132** (2012) 3428–3438.

[27] T. Riede, I. T. Tokuda, J. B. Munger, S. L. Thomson: Mammalian laryngseal air sacs add variability to the vocal tract impedance: Physical and computational modeling. The Journal of the Acoustical Society of America **124** (2008) 634–647.

[28] R. Szeliski: Computer Vision: Algorithms and Applications. Springer, London, 2011.

[29] J. Fan, I. Gijbels: Local Polynomial Modelling and Its Applications. Chapman and Hall/CRC, 1996.

[30] Y. Abdel-Aziz, H. Karara: Direct linear transformation from comparator coordinates in close-range photogrammetry. ASP Symposium on Close-Range Photogrammetry in Illinois, 1971, 1–18.

[31] J. Westerweel: Fundamentals of digital particle image velocimetry. Measurement Science and Technology **8** (1997) 1379-1392.

[32] C. E. Willert, M. Gharib: Digital particle image velocimetry. Experiments in Fluids **10** (1991) 181-193.

Author's complimentary copy