Reconstructing three-dimensional vocal fold movement via stereo matching

Isao T. Tokuda1,*, Masahiro Iwawaki2, Ken-Ichi Sakakibara3, Hiroshi Imagawa4, Takaharu Nito5, Tatsuya Yamasoba6 and Niro Tayama7

1Department of Mechanical Engineering, Ritsumeikan University, 1–1–1 Nojihigashi, Kusatsu, 525–8577 Japan
2Dai Nippon Printing Co., Ltd., 1–1–1 Ichigayaakagoya, Tokyo, 162–8001 Japan
3Department of Communication Disorders, Health Sciences University of Hokkaido, 2–5 Ainosato, Kita-ku, Sapporo, 002–8072 Japan
4Department of Otolaryngology, University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–8655 Japan
5Department of Otolaryngology, Head and Neck Surgery, National Center for Global Health and Medicine, 1–21–1 Toyama, Shinjuku-ku, Tokyo, 162–8655 Japan

(Received 1 April 2013, Accepted for publication 13 May 2013)

Keywords: Vocal folds, Stereo-endoscope, Stereo matching, Three-dimensional observation

PACS number: 43.70.Jt, 43.70.Gr [doi:10.1250/ast.34.374]

1. Introduction

Endoscopic high-speed video systems, which enable real-time observation of vocal fold vibration, have contributed significantly to the understanding of the dynamical features and functions of the vocal folds [1–4]. The characterization of the high-speed video by videokymography [5] or model-based parameter estimation [6] provides strong support for the practical use of the system in medicine. The limitation of the current high-speed technique, however, is that the vocal fold structure is observed on a two-dimensional (transverse) plane, where no information on the three-dimensional geometry in the inferior-superior direction is provided. Most of the current techniques utilize relative coordinates, which lack quantitative (absolute) measurements of the vocal folds. A possible alternative approach to three-dimensional quantitative measurement is medial surface observation of the vocal folds [7]. This approach, however, has been applied mainly to excised larynx experiments [8], in which normal muscle activities, such as contraction of the thyroarytenoid (TA) muscle, are absent. Döllinger et al. [9], on the other hand, made a breakthrough by observing medial surface dynamics of an in vivo canine with selective activation of the TA muscle. Such an invasive approach is, however, not applicable to humans.

As a noninvasive approach, we focus on a stereoeendoscopic observation of the vocal folds [10–12]. The stereo-endoscope provides two views of the larynx from different angles, enabling three-dimensional viewing of the laryngeal structure. With its compact design with dimensions comparable to those of normal endoscopes, it does not prevent normal muscle activities for natural voice production. Towards three-dimensional quantitative observation, we combine the stereo high-speed filming technique with a stereo matching technique [13] to automatically recover three-dimensional dynamics of the vocal folds in vivo. The stereo matching technique has been developed in the field of computer vision to reconstruct three-dimensional objects from a pair of left and right images using a pattern matching algorithm. By applying the standard algorithm of stereo matching, it is shown that the basic features of the vocal fold vibration can be observed in three dimensions.

2. Stereo high-speed recording

The stereo-endoscope used in a previous study [12], manufactured by Nagashima Medical Instrument Corporation (Fig. 1(a)), was employed in this study. The stereo-endoscope includes two independent ordinary rigid optical systems with a diameter of 9 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 a 70° oblique-angled view, with a field angle of 40° (Fig. 1(b)). The distance between the optical axes of the tips was 10 mm. The stereo-endoscope was attached to a CCTV lens of 50 mm that was connected to a high-speed digital camera (Photron Fastcam 1024PCI).

3. Stereo matching

Stereo matching is one of the standard techniques in the field of computer vision [13]. The idea of stereo matching is to find the corresponding points in the left and right images to measure depth. Let \( L(i, j) \) and \( R(i, j) \) be the left and right images, respectively, where \( i \) and \( j \) represent the horizontal and vertical indices of each pixel point. With respect to one point, e.g., \( L(x_l, y_l) \), in the left image as a reference, its corresponding point in the right image \( R(x_r, y_r) \) is sought by the area-based stereo matching technique [13] combined with landmarks [14]. In the area-based stereo matching technique [13], matching error between \( L(x_l, y_l) = \sum_{j=0}^{m} \sum_{i=0}^{m} |L(x_l + i, y_l + j) - R(x_r + i, y_r + j)| \) (1)
Here, the difference is measured with a window size of $(2m + 1) \times (2m + 1)$ centered at $(x_r, y_r)$ and $(x_l, y_l)$ coordinates. In addition to the area-based SAD function $E_1(x_l, y_l)$, a constraint is provided by a distance from landmarks [14]. Namely, $N$ pairs of corresponding points, $L(x_i(n), y_i(n))$ and $R(x_i(n), y_i(n))$ $(n = 1, 2, \ldots, N)$, were selected in the left and right images, respectively, prior to stereo matching. Supposing $L(x_i(\alpha), y_i(\alpha))$ to be the nearest landmark to $L(x_l, y_l)$ in the left image, a constraint on the corresponding right point is given by

$$E_2(x_r, y_r) = \frac{1}{(x_l - \tilde{x}_c) + k_2}.$$

Combining the SAD function with the constraint, the cost function to be minimized is given as $E(x_l, y_l) = p E_1(x_l, y_l) + (1 - p) E_2(x_l, y_l)$, where $p$ ($0 < p < 1$) is a balancing coefficient. For each reference point $(x_l, y_l)$ in the left image, the corresponding point $(x_r, y_r)$ that minimizes the cost function is found in the right image.

Once a pair of matching points, $(x_l, y_l)$ and $(x_r, y_r)$, are selected, its location in three-dimensional coordinates $(X, Y, Z)$ is given by the following formulas:

$$Z = \frac{1}{k_3(D_L - D_R) + k_2},$$

$$X = k_3 D_L,$$

$$Y = k_3 D_R,$$

where $D_L = x_l - \tilde{x}_c$ and $D_R = x_r - \tilde{x}_c$, stand for the horizontal distances of the matching points from the centers of the left and right optical fields, $(\tilde{x}_L, \tilde{y}_L)$ and $(\tilde{x}_R, \tilde{y}_R)$, respectively, and $D_V = y_l - 0.5(y_l + y_r)$ represents the vertical distance of the center of the left optical field from the middle of the matching points. $k_1$, $k_2$, and $k_3$ are calibration constants. Note, however, that these formulas are valid in an ideal situation. In reality, a photographic lens causes optical distortion, which violates the ideal equations. To correct such distortion, modified formulas with higher-order polynomials were utilized:

$$Z = \frac{1}{[k'_1(D_L - c'_1 D_R + c'_2 D_V) + k'_2]},$$

$$X = k'_3 f(Z, D_L) + k'_4,$$

and $Y = k'_5 Z D_V + k'_6$, where the function $f$ is in the form of $f(Z, D_L) = (D_L - c'_2 Z^2 + c'_3 Z + c'_4)/(c'_5 Z^c'6)$. The calibration constants $k'_1$ and $c'_1$ were optimized as follows. First, $D_L$ and $D_R$ were measured on Cartesian graph paper with $5\text{ mm} \times 5\text{ mm}$ resolution. The distance between the graph paper and the tips of the endoscope was changed from 14 mm to 84 mm. The calibration constants were then determined by the least-squares method to fit the modified equations to the measured points. As a result of calibration, we found that the distribution of errors between the real coordinates and the estimated coordinates in the three-dimensional Euclidean space had a median of 0.55 mm (5% within 0.15 mm, 95% within 2.96 mm).

4. Results

4.1. Recording

One male subject with no laryngeal pathology was instructed to produce a modal voice with a sustained vowel, which was observed by stereo-endoscopic high-speed digital imaging. The recording was carried out with an image resolution of 768 (horizontal) $\times$ 352 (vertical) pixels, a frame rate of 3,750 fps, and a recording duration of 10.12 s. Figure 2 shows an example of static laryngeal views from left and right angles during the phonation.

4.2. Three-dimensional reconstruction

The stereo matching technique was applied to the pair of stereoscopic images in Fig. 2. For the area-based function, a window size of $11 \times 11$ ($m = 5$) was utilized. $N = 35$ landmarks were empirically selected as the constraint, where the balancing coefficient was set to $p = 0.5$. Figure 3(a) shows the result of the three-dimensional reconstruction of the vocal folds. Smooth surfaces of the left and right vocal folds can be clearly recognized. The glottal opening area is also seen in the middle as the deep blue area.

The application of the present technique to a series of stereoscopic images yielded moving images of the vocal fold vibration in a three-dimensional space, as shown in Fig. 4 (the corresponding movie can be viewed at [15]). From $t = 0$ ms to $t = 5.3$ ms, the left and right vocal folds move inwards and they collide with each other at about $t = 6$ ms. After the complete closure of the glottis ($t = 6.93$ ms), the vocal folds...
start to reopen at $t = 7.7$ ms. At $t = 9.07$ ms, an opening structure of the glottis similar to that at the starting point ($t = 0$ ms) is recovered. This dynamic pattern represents one glottal cycle. It should be noted that, in the closing phase, the lower edge of the vocal fold precedes the upper edge. This produces a phase shift between the lower and upper edges of the vocal fold, which is well known as the mechanism of efficient energy transfer from the airflow to the vibrating vocal folds [16]. To quantify the phase shift, the vocal fold movement was further viewed in a frontal section, where the glottal opening width was measured on both upper and lower edges. Figure 3(b) shows that the closing movement of the lower edge clearly precedes that of the upper edge. The phase difference of approximately 80 degrees was estimated between the upper and lower edges, agreeing quite well with the results of a previous study [17]. Because the lower edge is hidden by the upper edge during the opening phase, the phase shift is discernible mainly during the closing phase.

5. Conclusions and discussion

The stereo matching technique has been applied to the stereo high-speed filming data to recover three-dimensional dynamics of the vocal folds. Compared with alternative approaches, such as the medial surface observation of the vocal folds [7,8], the present approach is fairly noninvasive, allowing for an \textit{in vivo} measurement of natural phonation without interfering with the normal muscle activities. Our technique well reconstructed a laryngeal mechanism that is in good agreement with the known results, such as the phase

![Fig. 3](image_url)  
Fig. 3 (a) Three-dimensional configuration of the vocal folds reconstructed from the stereoscopic images of Fig. 2. (b) Dynamical change in glottal opening width measured at upper (solid line) and lower (dotted line) edges of the vocal folds.

![Fig. 4](image_url)  
Fig. 4 Successive images of the moving vocal folds reconstructed in a three-dimensional space.
shift between the lower and upper edges of the vocal folds. To detect further details of vocal fold movement, especially in the vertical (inferior-superior) direction, improved precision will be needed. Such improvement should be realized by developing a more photosensitive lens to capture detailed features of the vocal fold surface with higher resolution. The stereo matching technique has several inevitable difficulties such as the problem of occlusion, in which some of the corresponding points on one side of the image can be hidden, e.g., by a raised surface of the vocal fold. The vocal fold surface, which does not show too many characteristic features, is also a source of a certain amount of matching error. To deal with such problems, further improvement based on more advanced techniques of computer vision should be made to the present algorithm. Making good use of the continuity of the moving points between the successive image frames should also be utilized as an additional constraint. Stereoe-ndoscopic observation of inspiratory phonation [18] will furthermore compensate for the hidden features of the medial surface dynamics during the closing phase.

Despite the remaining technical issues, the present approach has the strong advantages of a noninvasive nature and lowcost, which lead to a high potential for widespread use of the stereo-endoscopic system in medical practices. Our future works include the application of the present approach to various subjects to examine individual differences. Quantitative three-dimensional measurement of pathological voices as well as different types of voice registers may provide new insight into the dynamics and functions of the vocal folds.

Acknowledgments

This research was partly supported by Grants-in-Aid for Scientific Research (No. 23560446, No. 23300071) from the MEXT of Japan, SCOPE (No. 071705501) of the MIC of Japan, the JAIST Grant for Exploratory Research.

References

[1] T. Baer, A. Lögqvist and N. S. McGarr, “Laryngeal vibrations: A comparison between high-speed filming and glottographic techniques,” J. Acoust. Soc. Am., 73, 1304–1308 (1983).
[2] D. Bless, M. Hirano and R. J. Feder, “Videostroscopic evaluation of the larynx,” Ear Nose Throat J., 66, 289–296 (1987).
[3] H. Hirose, “High-speed digital imaging of vocal fold vibration,” Acta Otolaryngol., 458, 151–153 (1988).
[4] S. Kiritani, H. Hirose and H. Imagawa, “High-speed digital imaging analysis of vocal fold vibration in diplophonia,” Speech Commun., 13, 23 (1993).
[5] J. G. Švec and H. K. Schutte, “Videokymography: High-speed line scanning of vocal fold vibration,” J. Voice, 10, 201–205 (1996).
[6] M. Döllinger, U. Hoppe, F. Hettlich, J. Lohscheller, S. Schuberth and U. Eysholdt, “Vocal fold parameter extraction using the two-mass-model,” IEEE Trans. Biomed. Eng., 49, 773–781 (2002).
[7] D. A. Berry, D. W. Montequin and N. Tayama, “High-speed digital imaging of the medial surface of the vocal folds,” J. Acoust. Soc. Am., 110, 2539–2547 (2001).
[8] J. van den Berg and T. S. Tan, “Results of experiments with human larynxes,” Pract. Otorhinolaryngol. (Basel), 21, 425–450 (1959).
[9] M. Döllinger, D. A. Berry and G. S. Berke, “Medial surface dynamics of an in vitro canine vocal fold during phonation,” J. Acoust. Soc. Am., 117, 3174–3183 (2005).
[10] M. Sawashima and S. Miyazaki, “Stereo-fiberscopic measurement of the larynx: A preliminary experiment by use of ordinary laryngeal fiberscopes,” Ann. Bull. RILP, 8, 7–10 (1974).
[11] O. Fujimura, T. Baer and S. Niimi, “A stereo-fiberscope with a magnetic interlens bridge for laryngeal observation,” J. Acoust. Soc. Am., 65, 478–480 (1979).
[12] M. Sawashima, H. Hirose, K. Honda, H. Yoshioka, S. R. Hibi, N. Kawase and M. Yamada, “Stereoeendoscopic measurement of the laryngeal structure,” in Vocal Fold Physiology: Contemporary Research and Clinical Issues, D. M. Bless and J. H. Abbs, Eds. (College-Hill Press, San Diego, 1983), pp. 264–276.
[13] D. H. Ballard and C. M. Brown, Computer Vision (Prentice-Hall, Englewood Cliffs, N.J., 1982).
[14] M. S. Lew, T. S. Huang and K. Wong, “Learning and feature selection in stereo matching,” IEEE Trans. Pattern Anal. Mach. Intell., 16, 869–881 (1994).
[15] http://www.ritsumei.ac.jp/~siao/Movie.avi
[16] I. R. Titze, “The physics of small-amplitude oscillation of the vocal folds,” J. Acoust. Soc. Am., 83, 1536–1552 (1988).
[17] I. R. Titze, J. J. Jiang and T. Y. Hsiao, “Measurement of mucosal wave propagation and vertical phase difference in vocal fold vibration,” Ann. Otol. Rhinol. Laryngol., 102, 58–63 (1993).
[18] K.-I. Sakakibara, H. Imagawa, H. Yokonishi, T. Goto, M. Kimura, T. Nito and N. Tayama, “Vocal fold vibratory patterns of inspiratory phonation,” Proc. Pan European Voice Conf., p. 40 (2009).