Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2013) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

The glottaltopogram: A method of analyzing high-speed images of the vocal folds☆,☆☆

Gang Chen[a,*], Jody Kreiman[b], Abeer Alwan[c]

[a] Department of Electrical Engineering, University of California Los Angeles, 63-134 Engr IV, Los Angeles, CA 90095-1594, United States
[b] Department of Head and Neck Surgery, University of California Los Angeles, School of Medicine, 31-24 Rehab Center, Los Angeles, CA 90095-1794, United States
[c] Department of Electrical Engineering, University of California Los Angeles, 66-147G Engr IV, Los Angeles, CA 90095-1594, United States

Received 21 April 2013; received in revised form 6 November 2013; accepted 28 November 2013

Abstract

Laryngeal high-speed videoendoscopy is a state-of-the-art technique to examine physiological vibrational patterns of the vocal folds. With sampling rates of thousands of frames per second, high-speed videoendoscopy produces a large amount of data that is difficult to analyze subjectively. In order to visualize high-speed video in a straightforward and intuitive way, many methods have been proposed to condense the three-dimensional data into a few static images that preserve characteristics of the underlying vocal fold vibratory patterns. In this paper, we propose the "glottaltopogram," which is based on principal component analysis of changes over time in the brightness of each pixel in consecutive video images. This method reveals the overall synchronization of the vibrational patterns of the vocal folds over the entire laryngeal area. Experimental results showed that this method is effective in visualizing pathological and normal vocal fold vibratory patterns.

Keywords: High-speed videoendoscopy; Vocal fold vibration; Principal component analysis

1. Introduction

Clinicians and speech scientists have developed a number of techniques to observe vocal fold vibrations, including electroglottography (Baken, 1992), photoglottography (Sonesson, 1959), stroboscopy (Kitzing, 1985), and videokymography (Švec and Schutte, 1996). Recently, high-speed video (HSV) of the larynx has emerged as the state of the art in laryngeal imaging, due to increased recording frame rates, improved image resolution, and the decreasing cost of high-speed recording devices.

The study of HSV remains limited, however, by the large amount of 3-dimensional data produced (Fig. 1), so that images are inherently difficult to interpret visually and usually require subjective assessment. Because humans are better at discriminating characteristics of static than of dynamic images (which impose a memory load), many methods have been proposed to reduce the dimensionality of spatial–temporal HSV data and condense the time-varying video

☆ Parts of this article appeared in the Proceedings of ICASSP 2012.
☆☆ This paper has been recommended for acceptance by R.K. Moore.
[*] Corresponding author. Tel.: +1 310 993 7106.
E-mail addresses: gangchen@ucla.edu, gangchen.ucla@gmail.com (G. Chen).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.11.006

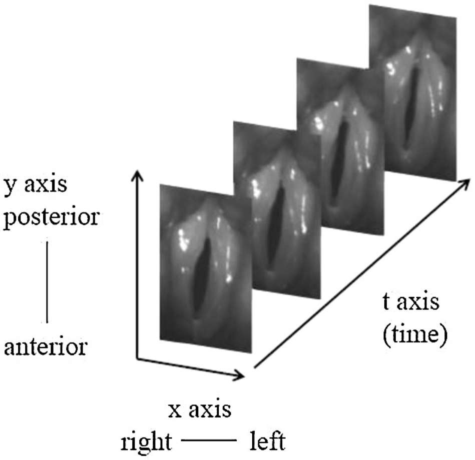

Fig. 1. The 3 dimensions of variability in high-speed video data: left-right (*x*), posterior–anterior (*y*), and time (*t*).

into a few static images that preserve the most important characteristics of the vibratory patterns. In this study, we propose a new computationally-efficient method—the glottaltopogram—to compactly summarize the overall spatial synchronization pattern of vocal fold vibration for the entire glottal area, in a manner that can be intuitively interpreted. Such a method may produce plots that are spatially similar to the original images, and which can be easily interpreted by physicians and clinicians during diagnosis.

Many previously described methods for analyzing HSV data depend on glottal area segmentation (Lohscheller et al., 2008; Karakozoglou et al., 2012; Döllinger et al., 2011; Yan et al., 2005). Automatic segmentation of the glottal area from HSV is in itself a challenging task, and a number of methods have been proposed. The most straightforward is thresholding, in which pixels with brightness lower than a certain threshold are treated as part of the glottis (e.g., Mehta et al., 2010, 2011). The threshold is typically specified based on a histogram of the image, where several peaks are assumed to exist due to clustering of glottal and non-glottal regions. However, this method is unsatisfactory when contrast is low, because segmentation performance is sensitive to threshold selection. In addition, this method is not fully automatic because it typically requires manual adjustment of thresholds over time. Other approaches to glottal area segmentation apply seeded region-growing algorithms. After manually selecting seeds from the image, neighboring pixels are examined to decide whether they should be added to the region, subject to criteria that vary from implementation to implementation (Adams and Bischof, 1994; Yan et al., 2006; Lohscheller et al., 2007). This method typically requires clear glottal edges to produce a correct result.

The segmented glottal area can subsequently be analyzed to reveal spatial and/or temporal variations in glottal vibratory patterns. For example, in phonovibrography (PVG; Lohscheller et al., 2008), the segmented glottal area is transformed into a geometric pattern representing the distance from the glottal edges to the glottal center line axis. In terms of the representation in Fig. 1, PVG condenses the *x* and *y* axes into one axis by mapping along the glottal edge trajectory, so that temporal resolution is perfectly maintained but spatial resolution is limited to the glottal edge trajectory. This method is sensitive to detection of the glottal center line axis, which strongly depends on the geometry of the detected glottal area (Karakozoglou et al., 2012) and can be difficult to identify accurately in the presence of a posterior glottal chink (glottal gap). A visual representation termed the "glottovibrogram" extends the PVG method (Karakozoglou et al., 2012; Döllinger et al., 2011). Glottovibrograms measure the distance between vocal fold contours instead of the distance to the glottal center-line axis, but visualization and interpretation of alterations in subsequent cycles remain unintuitive. Recently, Unger et al. (2013) proposed a PVG-wavegram to reveal inter-cycle characteristics of vocal fold vibrations across long sequences, where individual cycles of a PVG are segmented, normalized for cycle duration, and concatenated over time. Yan et al. (2005) applied a Hilbert transform to glottal area waveforms to analyze perturbation and periodicity. However, analyses of the glottal area waveforms do not preserve spatial information about vocal fold vibration, limiting applicability for interpreting spatial vibratory features such as asymmetry.

Despite these efforts, segmentation of the glottal area remains a non-trivial task. Results depend on the quality of the HSV data, including image contrast and the clarity of the glottal edge. Manual interactions are typically needed, such as initial seed assignment or threshold selection, and the segmented glottal area sequence requires inspection. In addition, segmentation of the glottal area typically requires processing the HSV data on a frame-by-frame basis, and the long computational time required for image processing limits the applicability of glottal-area based approaches under clinical conditions, where prompt results are preferred.

Other HSV analysis tools do not rely on glottal area segmentation. The most common of these, kymography (Tigges et al., 1999; Larsson et al., 2000), reduces data dimensionality by selecting pixels with a given value on the *y* axis (anterior–posterior dimension; Fig. 1)—or several values in multiplane kymography—usually chosen near the glottal midpoint. By limiting resolution along the *y* axis, kymography essentially collapses image analysis along the anterior–posterior dimension, so that temporal resolution is lossless but spatial resolution is limited to at most a few points. In a second method, temporal oscillation patterns across the entire laryngeal area are visualized by applying a Fourier transform to the light intensity time sequences from sequential high-speed images (Granqvist and Lindestad, 2001). The resulting signal contains amplitude and phase information as a function of frequency, and is displayed as color saturation on top of a single image selected from the original sequence, to characterize vibrational characteristics of the entire laryngeal area. On the basis of this work, Sakakibara et al. (2010) proposed a third method they called "laryngotopography" to visualize spatial characteristics of the Fourier spectra of the pixel-wise brightness curves (e.g., the frequency component that has the maximum amplitude in the Fourier spectra), which they claimed was effective in visualizing various vibrational modes of the vocal folds of patients with paralysis and cysts. Laryngotopography compresses the time axis by mapping the pixel-wise brightness scale time course into several transformed coefficients, where temporal information is condensed but spatial resolution is fully preserved. In other words, while kymography has limited spatial resolution, laryngotopography maintains the spatial characteristics of the entire image but focuses only on a single frequency component of the spectrum of the vibrational pattern.

In this paper, we propose the "glottaltopogram" to visualize HSV data. In this method, principal component analysis (PCA) is applied to light intensity time sequences from consecutive high-speed images and PCA coefficients are visualized. The proposed method reveals the overall spatial synchronization pattern of the vocal fold vibrations for the entire laryngeal area, rather than focusing on a specific location or frequency. Full spatial resolution is maintained, although the time axis is not preserved. Further, the proposed method does not rely on segmentation of the glottal area, and is robust to perturbations of video quality that might result in artifacts during glottal area detection. With minimal user interaction and fast processing time, glottaltopography provides an automatic way of finding the region of interest from the entire image and is suitable for clinical application. Comparisons between analyses of pathological and normal data, described in the next sections of this paper, show that the proposed method is effective in visualizing a wide variety of vocal fold vibrational patterns. Additional comparisons between glottaltopograms and kymograms show the manner in which these two analysis techniques (one that compresses the time axis, and one that compresses area) can complement each other in understanding glottal vibration.

2. Data and methods

2.1. Subjects and equipment

High-speed images were recorded at 4000 frames/s using a 70° rigid laryngoscope (KayPentax, Lincoln Park, New Jersey) with a 300 W Xenon light source (KayPentax, Lincoln Park, New Jersey) and a Color High-Speed Video System, Model 9710 (KayPentax, Lincoln Park, New Jersey). The image resolution was 512 pixels × 256 pixels and the color mode was 8 bit RGB. Audio signals were synchronously recorded with a Brüel & Kjær microphone (1.27 cm diameter; type 4193-L-004) and directly digitized at a sampling rate of 40 kHz, with a conditioning amplifier (NEXUS 2690, Brüel & Kjær, Denmark). Four subjects (3 males, denoted by M1–M3, and 1 female, denoted by F1) without voice disorders were recorded saying the vowel /i/ with breathy, modal, and pressed voice qualities (although for the male speakers only the modal voice samples were examined in this paper). Similar to Chen et al. (2013), normal subjects were phonetically knowledgeable and voice quality was demonstrated by a phonetician prior to each recording. Four additional male subjects with voice disorders (denoted by PM1–PM4) were also recorded while saying /i/ using their habitual pitch and loudness. All subjects were asked to sustain the phonation for at least one second during rigid endoscopy.

Please cite this article in press as: Chen, G., et al., The glottaltopogram: A method of analyzing high-speed images of the vocal folds. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.11.006

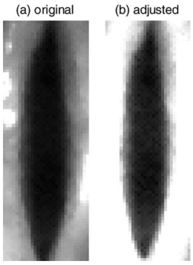

Fig. 2. (a) The original image of the glottis. (b) The image after brightness adjustment. The posterior glottis is shown at the top of the images, and the anterior glottis is at the bottom.

2.2. Image preprocessing

High-speed images were first converted from RGB to brightness scale. Due to illumination conditions, brightness of some glare spots needed to be adjusted before subsequent pixel brightness scale analysis (Fig. 2), because their brightness did not reflect actual vocal fold movement. Histogram equalization was performed manually (through an interactive graphical user interface) to enhance edge contrast of the vocal folds and remove the glare spots as much as possible. Compared to the original image in panel (a), the glare spots in the posterior glottis have been removed after the brightness adjustment in panel (b). Note that although the overall brightness increased after the adjustment, the contrast between glottal and non-glottal areas in the image was enhanced. The brightness of vocal folds approaches its maximum value and the brightness of the glottal open area approaches 0 (a non-linear transformation from physical position to light intensity), so that brightness curves better represent movements of the vocal folds.

2.3. PCA implementation

One PCA was performed for each HSV recording. A rectangular window was manually selected to isolate the image region containing the vocal folds (Fig. 3). To ensure the representativeness of each function, the brightness

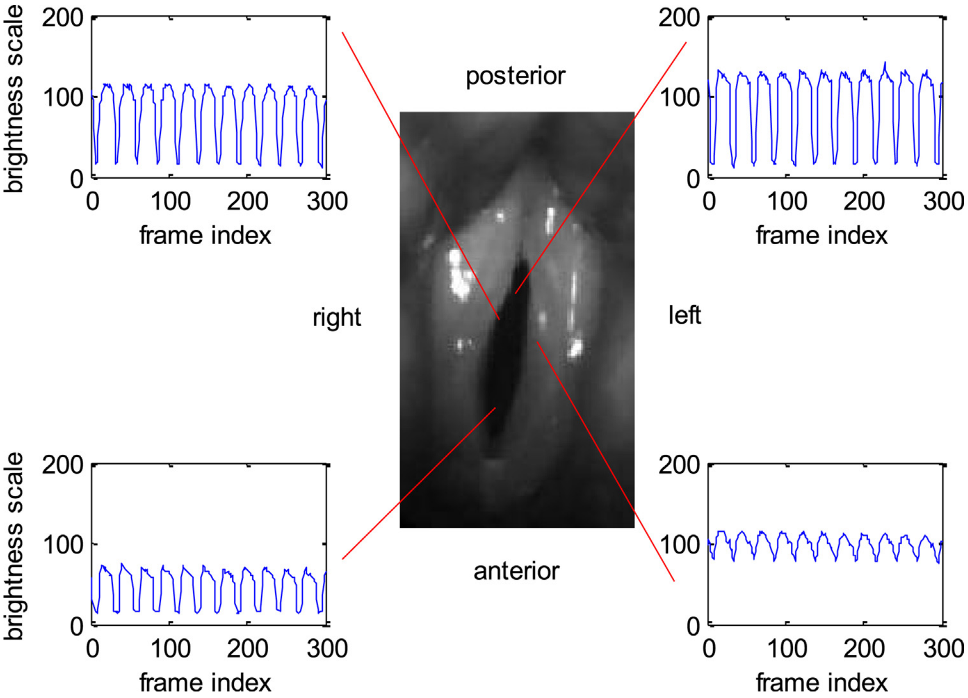

Fig. 3. (Color online) Center: image selected for analyses. Surrounding panels: Brightness scale time functions of pixels at different locations in and around the glottis.

scale time course was extracted across 300 consecutive frames (roughly 8–15 glottal cycles depending on the speaker's fundamental frequency) for each pixel inside the rectangular window. The number of pixels included in each analysis differed across recordings, ranging roughly from 5000 to 10,000, depending on the distance of the laryngoscope from the glottis.

The amplitude values for the brightness scale time course for each pixel served as input to the PCA, which was implemented using the Matlab Toolbox for Dimensionality Reduction (van der Maaten, 2011). Specifically, for a given HSV, if $g_{(i,j)}(t)$ is a 1-by-N vector and contains the glottal vibration information at pixel location (i, j), then:

$$g_{(i,j)}(t) = [b_{i,j}(1), b_{(i,j)}(2), \ldots, b_{(i,j)}(N)] \tag{1}$$

denotes the brightness time sequence (from frame 1 to frame N) at pixel location (i,j), where $1 \leq i \leq W,\quad 1 \leq j \leq H$, W is the image width, H is the image height, N is the total number of image frames, t is the frame index, and $b_{(i,j)}(k)$ denotes the brightness value of pixel (i, j) at frame index k. Examples of $g_{(i,j)}(t)$ are shown in the panels surrounding the central image in Fig. 3.

After performing a mean subtraction (for each frame) to ensure each frame has a zero mean for the brightness scale, a PCA was conducted. PCA models the brightness time sequence $g(t)$, treating each spatial pixel as a "repetition" of the experiment and each frame as a "feature". The matrix G in Eq. (2) was thus built and used as the input to PCA. This $W \times H$-by-N matrix G was constructed by concatenating all the brightness scale time sequences across all pixels in the video:

$$G_{W\times H,N} \begin{bmatrix} g_{(1,1)}(t) \\ g_{(1,2)}(t) \\ \vdots \\ g_{(1,H)}(t) \\ g_{(2,1)}(t) \\ g_{(2,2)}(t) \\ \vdots \\ g_{(2,H)}(t) \\ \vdots \\ g_{(W,H)}(t) \end{bmatrix} \tag{2}$$

This matrix G losslessly contains all the glottal vibration information from the video under study. Each brightness time sequence $g_{(i,j)}(t)$ can be decomposed as:

$$g_{(i,j)}(t) = \alpha_{i,j} \cdot \mathrm{PC1}(t) + \beta_{i,j} \cdot \mathrm{PC2}(t) + e_{i,j}(t) \tag{3}$$

where $\mathrm{PC1}(t)$ and $\mathrm{PC2}(t)$ are the first two principal components (orthogonal bases), $\alpha_{i,j}$ and $\beta_{i,j}$ are projections on the principal components, and $e_{i,j}(t)$ is the error term. Unlike conventional PCAs which are applied to model multiple images in other studies (e.g., face recognition), the PCA used in this study was applied to model the brightness time sequence, treating a spatial pixel's sequence as a "repetition". One PCA was conducted to model the brightness scale time sequences from all spatial points within a single recording. Thus, the basis of the PCA (principal component) was the same for all spatial points within that recording. That is, a single matrix G was derived for each individual video, so that PC1 and PC2 did not depend on pixel locations (i, j).

2.4. Analysis and visualization

For each brightness scale time sequence $g_{(i,j)}(t)$, the first two PCA coefficients $\alpha_{(i,j)}$ and $\beta_{(i,j)}$ (projections on the first two principal components, PC1 and PC2) were calculated. The coefficients were normalized to an 8 bit (0–255) scale and visualized at the original pixel location (i, j) in terms of color saturation to facilitate interpretation. The brightness scale curve was then reconstructed using the first two coefficients and principal components. Mean square reconstruction errors (mean square of $e_{(i,j)}(t)$) were calculated and visualized in the same way. In the final stage, the percentage of variance explained by the first two principal components (eigenvalues, or energy, corresponding to the orthogonal bases) was calculated, which partially reflects the energy compactness of PCA (synchronization of the glottal vibration).

By performing PCA, the glottal vibratory pattern represented by the brightness scale time courses is presumably “mapped” to a two-dimensional space captured by PC1 and PC2, given that PC1 and PC2 can account for the majority of the variance in the time-varying data. That is, glottaltopography compresses the time axis by mapping the pixel-wise brightness scale time course into the PCA coefficients, where temporal information is condensed into a single static image but spatial resolution is fully preserved. Pixels with similar brightness scale time courses should have similar PCA coefficients, which are represented in the glottaltopogram as similar colors. Recall that the PCA for each HSV recording was based on brightness scale time sequences from all spatial points within this video, which ensures homogeneity across the spatial points within one HSV recording. Thus, if the left and right vocal folds are vibrating symmetrically, the pixels on the two folds should also exhibit similar brightness scale time sequences. This similarity should be captured by the first two PCA coefficients and the derived images should exhibit symmetric color patterns. If the left and right vocal folds are vibrating asymmetrically, as might occur in a vocal fold paralysis, this asymmetry should result in a glottaltopogram with asymmetric color patterns. Similarly, a glottal region with highly aperiodic vibrations will appear with a distinct color pattern with respect to the remaining steady-vibrating region. When vibration of the two vocal folds is synchronized, the variance accounted for by the principal components should be higher (more compact energy concentration) than when vibrations are unsynchronized, because synchronization results in similar pixel-wise brightness scale time sequences. Similarly, the pixel-wise mean square reconstruction error should be generally low and (roughly) evenly distributed across pixels when glottal vibration is synchronized, while higher reconstruction errors should be observed in laryngeal regions exhibiting unsynchronized glottal vibrations.

3. Results

In the following, results of the glottaltopographic visualization approach are presented for both normal speakers and subjects with voice disorders. Each HSV recording was visualized using a glottaltopogram to determine the underlying glottal vibratory pattern. In some cases, kymograms are also presented, to highlight the complementary information available from each type of display.

3.1. Variations in voice quality within and across normal subjects

In this subsection, we apply glottaltopography first to samples of modal voice from three normal male subjects (speakers M1, M2, and M3) and secondly to modal, breathy, and pressed voice samples from a normal female subject (speaker F1). These relatively simple cases demonstrate the manner in which glottaltopograms can be interpreted, and how these analyses can augment information available from existing analysis approaches.

3.1.1. Variations in modal quality among normal subjects

We first examined modal voice as produced by 3 male speakers (M1, M2, and M3) without voice disorders.[1] Fig. 4 shows the glottaltopograms from each speaker. Variance accounted for by each analysis is given in Table 1. The first principal coefficient distributions ((a) panels) display symmetric patterns, roughly representing the means of the pixels' light-intensity time courses, which are predominantly determined by the average shape of the time-evolving glottal area (the glottal area generally has lower brightness than the non-glottal area). Recall that a mean subtraction was conducted (for each frame) before performing PCA to ensure each frame has a zero mean brightness, whereas each pixel's brightness scale time sequence was not normalized.

The color differences between the left and right folds in the second principal coefficient are shown in Fig. 4(b), and reflect the difference between folds in vibratory pattern. Fig. 5 shows the corresponding kymogram from the first speaker (M1), and reveals a phase difference between the left and right vocal folds but no obvious differences between folds in the amplitude or frequency of vibration. In this case, the asymmetric vibrational patterns are captured in both

[1] HSV and audio recordings are included in the supplemental material associated with this paper.

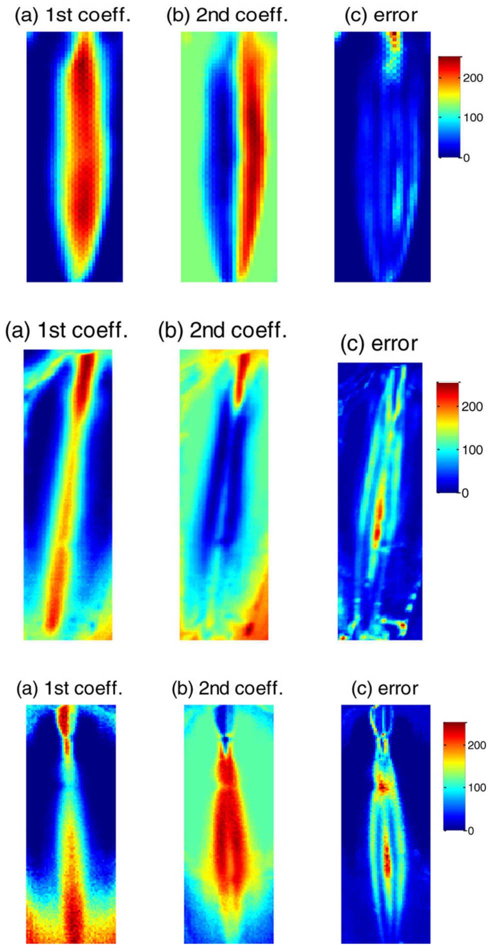

Fig. 4. Glottaltopograms of modal voice produced by three males without voice disorders. (a and b) The first and second principal coefficients, displayed in terms of color saturation. (c) Reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The first row represents speaker M1; the second row represents speaker M2; and the third row represents speaker M3. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Variance accounted for by the first and second principal components for each speaker.

	Percent Variance Accounted For			
Speaker	PC1	PC2	Total variance	Comment
M1	72	19	91	Asymmetric vibrations
M2	74	17	91	–
M3	79	14	93	–
F1	77	11	88	Modal phonation
F1	71	13	84	Breathy phonation
F1	70	21	91	Pressed phonation
PM1	66	14	80	Complex and asymmetric vibrations; creaky
PM2	72	15	87	Phase difference, anterior glottis; breathy
PM3	76	12	88	Phase difference, whole glottis; breathy
PM4	78	7	85	Hyperfunctional quality

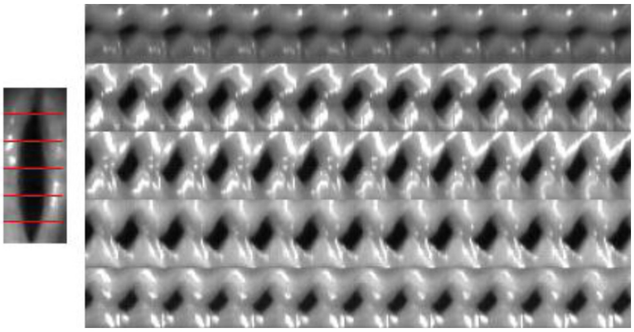

Fig. 5. Multi-line kymogram of a modal voice from a normal subject (speaker M1). The *x* axis represents time, and the *y* axis represents the amplitude of vocal fold vibration. Each row of images corresponds to movement of the folds at one glottal location (indicated by the red lines through the frame at the left of the figure). Movements of the right vocal fold are shown at the top of the kymogram, and those of the left vocal fold are shown at the bottom. (For interpretation of the references to color in text, the reader is referred to the web version of this article.)

Figs. 4 (first row (b)) and 5. Speakers M2 and M3 have symmetric vocal fold vibratory patterns, which are visualized in Fig. 4 (second and third rows).

3.1.2. Comparing phonation types within a single subject

Fig. 6 shows three glottaltopograms for subject F1, representing modal, breathy, and pressed phonation, respectively. Variance accounted for by each analysis is included in Table 1. As shown in the first column of this figure, the first principal coefficient distributions for modal (first row) and pressed (third row) phonation display highly symmetric patterns, although more movement is apparent in the anterior glottis than in the posterior (see panel (b) in the third row of Fig. 6) when phonation is pressed, possibly due to a recording artifact.[2] In contrast, breathy phonation in this speaker (middle row) is characterized by some irregularity in the posterior glottis (presumably representing a glottal gap), which is symmetric for both modal and pressed phonation. Similar roughly symmetric patterns are also observed in the second principal coefficient distributions, with slight asymmetries at the posterior end for modal and breathy phonation. Reconstruction error distributions are visualized in the third column of the figure. These show that reconstruction error is consistently highest in the posterior glottis, presumably due to variability in glottal gap configurations and to the region's small vibration amplitude and slight phase lag compared to the middle portion of the vocal folds.

3.2. Patients with voice disorders

In this subsection, glottaltopography is applied to visualize more complex phonatory patterns in four patients with voice disorders. Three patients (PM1–PM3) had asymmetric vocal fold vibrations to different degrees, while one patient (PM4) had symmetric vibrations. As expected, analyses for these speakers accounted for significantly less variance in the underlying HSV data than did analyses for normal speakers (one-way ANOVA; $F(1, 8) = 13.95$, $p = .006$, $R^2 = 0.64$), due to greater irregularity in vibratory patterns.

3.2.1. A patient with a creaky voice and asymmetric glottal vibration

Figs. 7 and 8 show a kymogram and a glottaltopogram, respectively, for a male patient (speaker PM1) exhibiting complex asymmetry in vibrational amplitude between the left and right vocal folds. The percent of variance accounted for by each principal component is given in Table 1. As Fig. 7 shows, the right fold vibrated with larger amplitude than the left, and the vibrating amplitude of the left fold alternated cycle by cycle. Both left and right vocal folds have

[2] Examination of the HSV for the pressed example suggests that this apparent movement may be due in part to a recording artifact, which resulted in a poor view of the most anterior part of the glottis. The high amount of variance accounted for by the second PC may also be due to this effect. Note, however, that the difference between normal and pathological speakers in variance accounted for by the glottaltopograms remains significant when values from the pressed case were excluded from analysis [$F(1, 7) = 10.68$, $p = .01$, $R^2 = 0.60$].

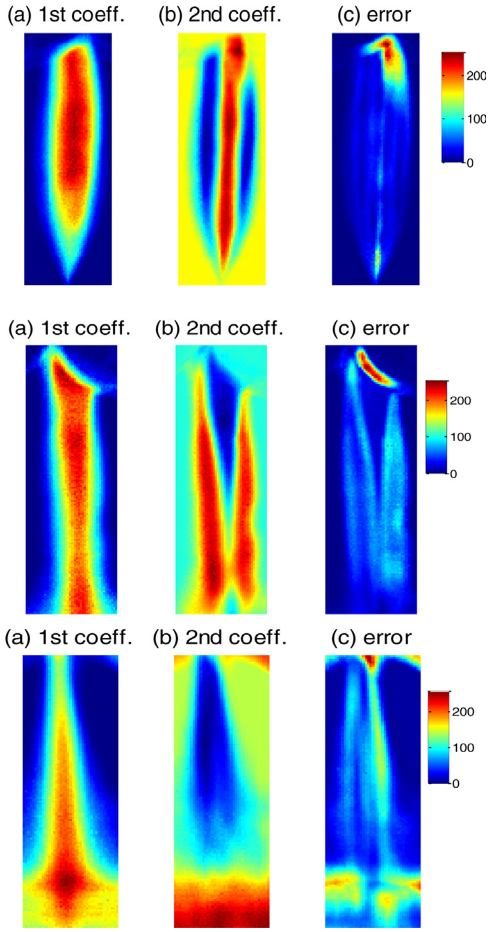

Fig. 6. Glottaltopograms for modal, breathy, and pressed phonation produced by a normal subject (speaker F1). (a and b) The first and second principal coefficients, displayed in terms of color saturation. (c) Reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The first row represents modal phonation; the second row represents breathy phonation; and the third row represents pressed phonation. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

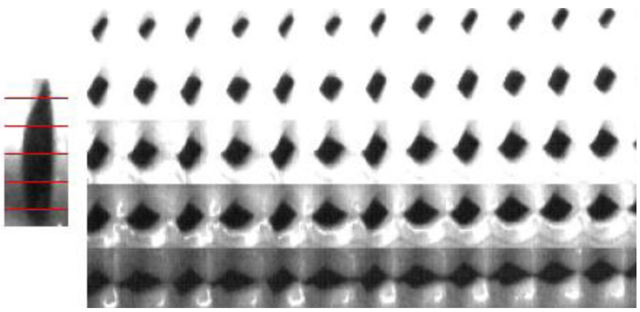

Fig. 7. Multi-line kymogram of a patient (speaker PM1) with creaky voice. The x axis represents time, and the y axis represents the amplitude of vocal fold vibration. Each row of images corresponds to movement of the folds at one glottal location (indicated by the red lines through the frame at the left of the figure). Movements of the right vocal fold are shown at the top of each frame, and those of the left vocal fold are shown at the bottom. (For interpretation of the references to color in text, the reader is referred to the web version of this article.)

Please cite this article in press as: Chen, G., et al., The glottaltopogram: A method of analyzing high-speed images of the vocal folds. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.11.006

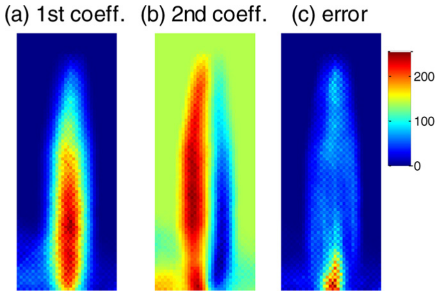

Fig. 8. The glottaltopogram of a patient (speaker PM1) with creaky voice. (a and b) The first and second principal coefficients, displayed in terms of color saturation. (c) Reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

approximately the same vibratory frequency, although the frequency of phonation appears to alternate in an A-B-A-B pattern. The corresponding acoustic signal sounds creaky.

Fig. 8 shows a glottaltopogram corresponding to this kymogram. While the glottaltopogram does not reveal the alternations in amplitude and period that are apparent in the kymogram, it does show that the vibrational patterns are distinct between the left and right vocal folds. Note that PC1 accounts for relatively little variance compared to the other cases listed in Table 1, possibly reflecting the complex synchronization of this example. Compared to the kymogram, the glottaltopogram provides better spatial resolution in visualizing the different vocal fold vibratory patterns, in a display that includes only 3 images (PC1, PC2, and reconstruction error) rather than the 60 frames included in the kymogram.

3.2.2. *A patient with a breathy voice and unsynchronized glottal vibration*

The glottaltopogram of a second male patient (speaker PM2) with a breathy voice is shown in Fig. 9. Variance accounted for is included in Table 1. The first principal coefficient distribution (panel (a)) displays a symmetric pattern, representing the means of the pixels' brightness scale time sequences (roughly dependent on the average shape of the glottal area). Frame-by-frame visual inspection of the video recording shows that the left anterior portion of the vocal folds is in opposite phase with respect to the rest of the vocal folds. This is manifested in panel (b) as two distinct portions in the second PCA coefficient distribution: the left anterior portion (lower right of the image) versus the rest of the vocal folds. In (c), the left middle portion of the vocal folds has the largest reconstruction error. This is due to the fact that the first two PCA coefficients poorly model this part. Thus the vibration pattern is more complex than a synchronous pattern or a pattern with perfectly opposite phase. The left middle portion is the border where normal phase and opposite phase meet, which produces an irregular vibratory pattern.

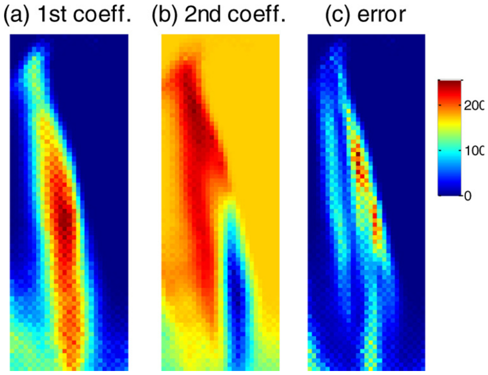

Fig. 9. The glottaltopogram of a patient (speaker PM2) with breathy voice. (a and b) The first and second principal coefficients, displayed in terms of color saturation. (c) Reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

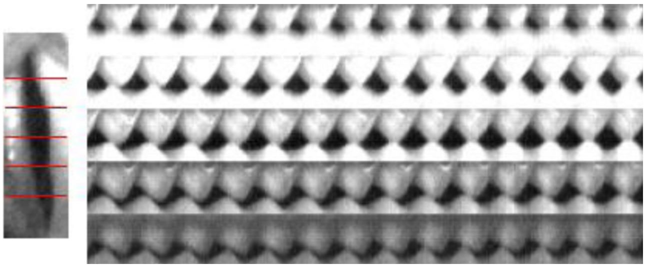

Fig. 10. Multi-line kymogram of a patient (speaker PM2) with breathy voice. The *x* axis represents time, and the *y* axis represents the amplitude of vocal fold vibration. Each row of images corresponds to movement of the folds at one glottal location (indicated by the red lines through the frame at the left of the figure). Movements of the right vocal fold are shown at the top of the kymogram, and those of the left vocal fold are shown at the bottom. (For interpretation of the references to color in text, the reader is referred to the web version of this article.)

Fig. 10 shows a multi-line kymogram of speaker PM2, where the anterior portion shows the phase difference between the left and right vocal folds. However, vocal fold activity in the anterior–posterior direction is not well captured in the kymogram. The glottaltopogram in Fig. 9(b) clearly shows that the "phase-unsynchronized" region is the anterior portion of the left vocal fold. The size and position of this problematic region are also visualized, but the actual degree of phase-difference can only be accessed from the kymogram.

3.2.3. *A patient with a breathy voice and unsynchronized glottal vibration*

The glottaltopogram of a third male patient (speaker PM3) with a breathy voice is shown in Fig. 11. Variance accounted for is included in Table 1. Frame-by-frame visual inspection of the HSV recording shows that most of the left vocal fold has a phase lag of about 90° relative to the right fold. This manifests in (b) as two distinct portions in the second PCA coefficient distribution: the left fold (with the exception of the posterior-most segment, near the arytenoids) versus the rest of the vocal folds. The symmetric pattern of the first principal coefficient distribution (panel (a)) illustrates the means of the pixels' brightness scale time sequences, roughly showing the average shape of the time-evolving glottal area.

Fig. 12 shows a multi-line kymogram of speaker PM3, where the anterior (but not the posterior) glottis shows the phase difference between the left and right vocal folds. Similar to Section 3.2.2, the glottaltopogram in Fig. 11 clearly shows the "phase-unsynchronized" region, providing better spatial information about the overall vocal fold vibrational pattern than does kymography.

3.2.4. *A patient with pressed voice and synchronized glottal vibration*

Fig. 13 shows the glottaltopogram of a fourth male patient (speaker PM4) with vocal hyperfunction. Variance accounted for is included in Table 1. A multi-line kymogram for this speaker is shown in Fig. 14, where synchronized

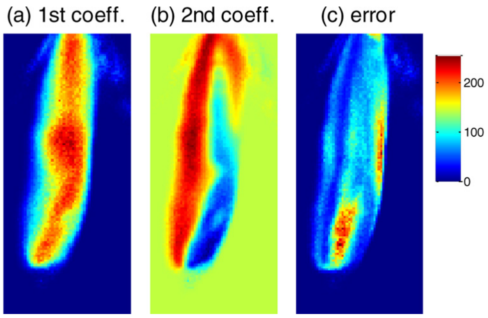

Fig. 11. The glottaltopogram of a patient (speaker PM3) with breathy voice. (a and b) The first and second principal coefficients, displayed in terms of color saturation. (c) Reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

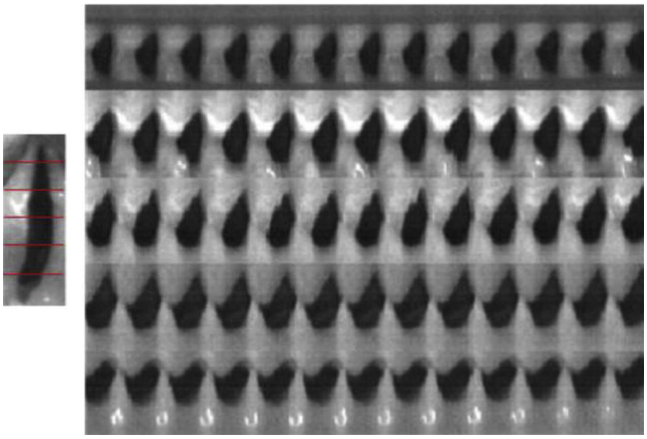

Fig. 12. Multi-line kymogram of a patient (speaker PM3) with breathy voice. The *x* axis represents time, and the *y* axis represents the amplitude of vocal fold vibration. Each row of images corresponds to movement of the folds at one glottal location (indicated by the red lines through the frame at the left of the figure). Movements of the right vocal fold are shown at the top of the kymogram, and those of the left vocal fold are shown at the bottom. (For interpretation of the references to color in text, the reader is referred to the web version of this article.)

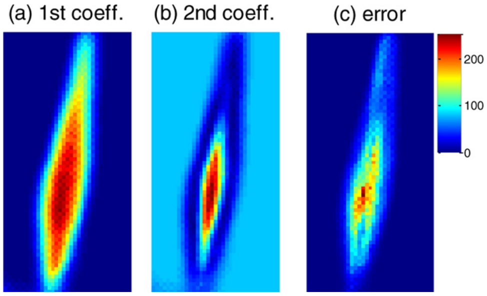

Fig. 13. The glottaltopogram of a patient (speaker PM4) with vocal hyperfunction. (a and b) The first and second principal coefficients, displayed in terms of color saturation. (c) Reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

vibrations can be observed for the left and right vocal folds. This symmetric vibrational pattern is also captured in Fig. 13 as a left-right symmetric color distribution. In this case of highly symmetrical vibration, glottaltopography illustrates the spatial synchronization pattern, while kymography visualizes the temporal synchronized evolution within pre-selected lines.

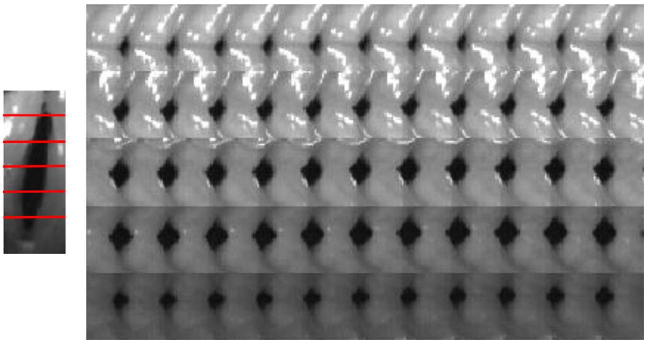

Fig. 14. Multi-line kymogram of a patient (speaker PM4) with vocal hyperfunction. The *x* axis represents time, and the *y* axis represents the amplitude of vocal fold vibration. Each row of images corresponds to movement of the folds at one glottal location (indicated by the red lines through the frame at the left of the figure). Movements of the right vocal fold are shown at the top of the kymogram, and those of the left vocal fold are shown at the bottom. (For interpretation of the references to color in text, the reader is referred to the web version of this article.)

Please cite this article in press as: Chen, G., et al., The glottaltopogram: A method of analyzing high-speed images of the vocal folds. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.11.006

4. Discussion and conclusions

Data reduction methods like glottaltopography all reduce HSV data from 3 dimensions to 2, which inevitably leads to loss of information, either temporal or spatial. In this sense, glottaltopography, kymography, and laryngotopography visualize different aspects of HSV data, by maintaining information from different dimensions, but no one method "outperforms" the others. However, the results presented here show how methods can be combined to analyze and interpret HSV data while overcoming the limitations inherent in each individual visualization approach.

Two attributes of glottaltopography make it a particularly useful addition to the set of methods available for working with HSV data. First, glottaltopography is robust (especially when compared to methods like PVG requiring glottal area segmentation) when used with HSV data with variations in contrast levels, random noise during recordings, and multiple glottal gaps, where detection of the glottal edges is inherently difficult. Because some subjects have difficulty tolerating a rigid endoscope, it can be impractical to create multiple high-speed recordings of the same subject in clinical application, and the ability to adjust focus and illumination levels during recording may be limited by the need to complete an exam quickly. As a result, recorded HSV data are often suboptimal in quality (Lohscheller et al., 2007), so that robustness is an important advantage of the method described here.

Secondly, the computational complexity of the glottaltopogram is much lower than that of methods based on glottal area segmentation (e.g., PVG), where the detection of glottal area has to be implemented for each image on a frame-by-frame basis. A glottaltopogram can be generated from 300 video frames in under 5 s, while calculating a PVG typically takes a few minutes and involves visual inspection of (at least a few) key frames to ensure the accuracy of glottal area detection.

The first PCA coefficient describes the projection on the dimension that represents the maximum variance in the underlying HSV data. In the present data, this first coefficient always roughly represents the mean of the pixel's brightness scale time sequence, which predominantly depends on the average shape of the glottal area. The second PCA coefficient shows more variability in vibrational pattern across pixel locations, and thus differed more from speaker to speaker. For both synchronized and unsynchronized vocal fold vibrations, the first two PCA coefficients accounted for an average of almost 88% of the variance, largely due to the prevalent quasi-periodic shapes of the brightness scale time sequences among pixels that resulted from quasi-periodic vocal fold vibrations (Table 1). This also indicates that the mapping into PCA coefficients substantially maintains the characteristics of vocal fold vibration, as represented by brightness scale time sequences.

It is often claimed that healthy voices are characterized by symmetric, periodic vocal fold vibrations (Hertegård et al., 2003; Döllinger et al., 2003), and previous studies have sometimes found links between the presence of asymmetric vocal fold vibration and degradations in perceived voice quality in patients with voice disorders (Niimi and Miyaji, 2000; Verdonck-de Leeuw et al., 2001). The present data are not entirely consistent with this scenario. Although the manner in which vibratory asymmetries or phase lags affect perceived voice quality is far from well understood, virtually all of the glottaltopograms of phonation from normal speakers revealed at least minor asymmetries (and in some cases very large asymmetries) in vibratory patterns. We note that a recent study based on physical vocal fold models showed that left-right asymmetry in vocal fold vibration does not produce a perceivable perceptual effect unless the asymmetry is so large that it causes a change in vibratory mode (Zhang et al., 2013). The potential applicability of detecting unsynchronized vocal folds vibration via glottaltopography in clinical settings may provide the data needed to explicate which asymmetries are clinically significant, and which have little or no impact on voice quality. In this way, this method constitutes a promising aid in studying the perceptual consequences of irregular vocal fold vibrations among normal subjects and patients.

Acknowledgments

This work was supported in part by NSF Grant No. IIS-1018863 and by NIH/NIDCD Grant Nos. DC01797 and DC011300. We thank Bruce Gerratt, Robin Samlan, and Juergen Neubauer for help recording high-speed images.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.csl.2013.11.006.

References

Adams, R., Bischof, L., 1994. Seeded region growing. IEEE Trans. Pattern Anal. Mach. Intell. 16 (6), 641–647.

Baken, R.J., 1992. Electroglottography. J. Voice 6, 98–110.

Chen, G., Kreiman, J., Gerratt, B.R., Neubauer, J., Shue, Y.-L., Alwan, A., 2013. Development of a glottal area index that integrates glottal gap size and open quotient. J. Acoust. Soc. Am. 133, 1656–1666.

Döllinger, M., Braunschweig, T., Lohscheller, J., Eysholdt, U., Hoppe, U., 2003. Normal voice production: computation of driving parameters from endoscopic digital high speed images. Methods Inf. Med. 42 (3), 271–276.

Döllinger, M., Lohscheller, J., Svec, J., McWhorter, A., Kunduk, M., 2011. Support vector machine classification of vocal fold vibrations based on phonovibrogram features. In: Ebrahim, F. (Ed.), Advances in Vibration Analysis Research. InTech, Croatia, pp. 435–456.

Granqvist, S., Lindestad, P.-Å., 2001. A method of applying Fourier analysis to high-speed laryngoscopy. J. Acoust. Soc. Am. 110 (6), 3193–3197.

Hertegård, S., Larsson, H., Wittenberg, T., 2003. High-speed imaging: applications and development. Logoped. Phonatr. Vocol. 28 (3), 133–139.

Karakozoglou, S.-Z., Henrich, N., d'Alessandro, C., Stylianou, Y., 2012. Automatic glottal segmentation using local-based active contours and application to glottovibrography. Speech Commun. 54, 641–654.

Kitzing, P., 1985. Stroboscopy – a pertinent laryngological examination. J. Otolaryngol. 14 (3), 151–157.

Larsson, H., Hertegård, S., Lindestad, P.-Å., Hammarberg, B., 2000. Vocal fold vibrations: high-speed imaging, kymography and acoustic analysis: a preliminary report. Laryngoscope 110, 2117–2122.

Lohscheller, J., Eysholdt, U., Toy, H., Döllinger, M., 2008. Phonovibrography: mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics. IEEE Trans. Med. Imaging 27 (3), 300–309.

Lohscheller, J., Toy, H., Rosanowski, F., Eysholdt, U., Döllinger, M., 2007. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. Med. Image Anal. 11 (4), 400–413.

Mehta, D., Deliyski, D., Quatieri, T., Hillman, R., 2011. Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings. J. Speech Lang. Hear. Res. 54, 47–54.

Mehta, D.D., Deliyski, D.D., Zeitels, S.M., Quatieri, T.F., Hillman, R.E., 2010. Voice production mechanisms following phonosurgical treatment of early glottic cancer. Ann. Otol. Rhinol. Laryngol. 119, 1–9.

Niimi, S., Miyaji, M., 2000. Vocal fold vibration and voice quality. Folia Phoniatr. Logoped. 52, 32–38.

Sakakibara, K., Imagawa, H., Kimura, M., Yokonishi, H., Tayama, N., 2010. Modal analysis of vocal fold vibrations using laryngotopography. In: Interspeech, pp. 917–920.

Sonesson, B., 1959. A method for studying the vibratory movements of the vocal cords. J. Laryngol. Otol. 73, 732–737.

Švec, J.G., Schutte, H.K., 1996. Videokymography: high-speed line scanning of vocal fold vibration. J. Voice 10, 201–205.

Tigges, M., Wittenberg, T., Mergell, P., Eysholdt, U., 1999. Imaging of vocal fold vibration by digital multiplane kymography. Comput. Med. Imaging Graph. 23 (6), 323–330.

Unger, J., Meyer, T., Herbst, C.T., Fitch, W.T.S., Döllinger, M., Lohscheller, J., 2013. Phonovibrographic wavegrams: visualizing vocal fold kinematics. J. Acoust. Soc. Am. 133, 1055–1064.

van der Maaten, L., 2011. Matlab toolbox for dimensionality reduction (version: 0.7.2b), http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html (viewed 01.10.11).

Verdonck-de Leeuw, I.M., Festen, J.M., Mahieu, H.F., 2001. Deviant vocal fold vibration as observed during videokymography: the effect on voice quality. J. Voice 15, 313–322.

Yan, Y., Ahmad, K., Kunduk, M., Bless, D., 2005. Analysis of vocal-fold vibrations from high-speed laryngeal images using a Hilbert transform-based methodology. J. Voice 19, 161–175.

Yan, Y., Chen, X., Bless, D., 2006. Automatic tracing of vocal-fold motion from high-speed digital images. IEEE Trans. Biomed. Eng. 53 (7), 1394–1400.

Zhang, Z., Kreiman, J., Gerratt, B.R., Garellek, M., 2013. Acoustic and perceptual effects of changes in body layer stiffness in symmetric and asymmetric vocal fold models. J. Acoust. Soc. Am. 133, 453–462.

Please cite this article in press as: Chen, G., et al., The glottaltopogram: A method of analyzing high-speed images of the vocal folds. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.11.006