Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 28 (2014) 1156–1169
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
Available online 7 December 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

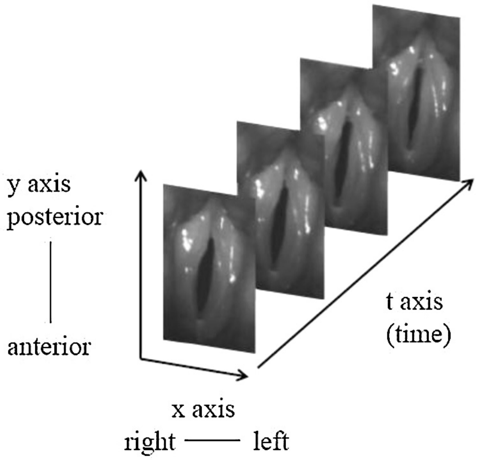

Fig. 1. The 3 dimensions of variability in high-speed video data: left-right (x), posterior–anterior (y), and time (t).

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,