Highspeedglottography with a flexible endoscope for the examination of the human larynx during running speech

Thomas Wittenberg, Patrick Mergell, Monika Tigges and Ulrich Eysholdt

Department of Phoniatrics and Pedaudiology
FAU Erlangen, Bohlenplatz 21, 91054 Erlangen
Thomas.Wittenberg@phoni.med.uni-erlangen.de

A digital high-speed camera in combination with a light intensifier and a flexible endoscope is used to record running human speech. Due to the light absorbtion of the flexible endoscope, the recorded image sequences are too dark for human examination and too poor for digital motion analysis. Therefore, two different image enhancement methods are compared to improve the image sequences. Resulting trajectories of motion analysis of vocal cord vibrations during running speech will be shown.

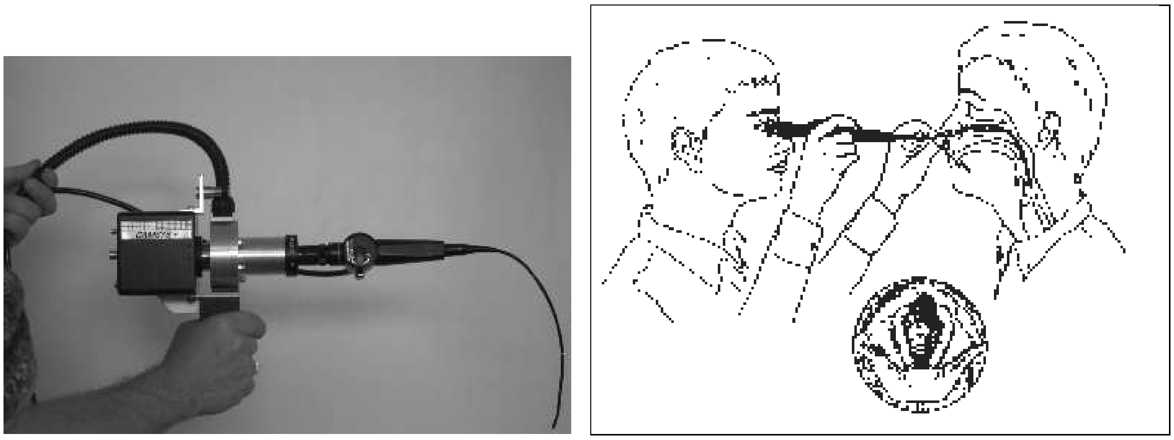
Keywords: light-intensifier, digital high-speed camera, digital image enhancement

1 Motivation

The most common method for the examination and diagnosis of functional voice disorders (dysphonias) is video-stroboscopy. Since the unstressed,human voice has a mean frequency in the range between 150 and 300 Hz, but standard video cameras only record at a maximum speed of 50 half frames per second, the disadvantage of this method is clearly the subsampling of the oscillation of the vocal cords. The examiner registers these oscillations as a so called 'virtual vibration' and therefore cannot judge it properly. With a high-speed camera, however, the vocal cord vibrations can be filmed with a recording rate high enough to observe even asymmetric and aperiodic vibration modes. Furthermore, variations of the amplitude (*jitter*) and period (*shimmer*) can be registered. Since these parameters are symptomatic for hoarse or creaky voices [2], the high-speed glottography in combination with semi-automatic motion analysis of vocal fold vibrations can be a mighty tool for the clinical diagnosis.

2 Recording System and Light-Intensifier

The Department of Phoniatrics in Erlangen has been involved with digital high-speed cameras for about 15 years [6]. For the past three years, a clinically applicable version of a high-speed camera has been in use for the recording of vocal cord vibrations with a grant from the DFG (DFG Ey 15/5-3). This high-speed recording system is based on a commercially available digital high-speed camera (FhG-IIS *CAMSYS+*) with a full resolution of 128^2 pixel and a maximum recording speed of 10 000 fps. The duration of a



(a) (b)

Figure 1: Light-intensifier system connected to the flexible endocope and the head of the highspeed camera containing the 128^2-CCD-Chip (a). Schematic of the insertion of the flexible endoscope through the nose into the larynx (b)[3].

recording depends on the selected recording speed and the available memory extension. Our system has a 16 megabyte circular memory, which can hold a recording with a duration of half a second at maximum recording speed. This time span is sufficient to record the voice onset and about 50 periods of vocal cord vibrations. The circular memory can be extended up to 256 megabyte, yielding a maximum recording span of 4 seconds, if necessary. Furthermore, the high-speed camera is synchronized with two digital data channels, which are used for the simultaneous recording of the audio and electroglottography-signals.

For the diagnosis of functional voice disorders, a rigid 90^0 laryngoscope is connected to the head of the high speed camera via a C-mount adapter with a 32 mm lens [2]. For the purpose of recording the motion of the vocal cords and the larynx during running speech, a rigid endoscope is not feasible, since it prevents the articulary movement of the tongue and causes phonation problems. For this reason, a flexible pharyngo-laryngoscope (Olympus *ENF-P3*) has been applied, which can be inserted through the nose and moved forward into the larynx. Since the flexible endoscope has a diameter of only 3.6 mm, its glassfiber cables provide a scarce illumination of the larynx. To record the larynx movements under these extreme circumstances, the high speed camera system has been extended by an especially designed light-intensifier with an amplification factor 3, built by the Fraunhofer Institute for Integrated Circuits in Tennenlohe. For illumination, a 250 Watt cold light source is used. Fig. 1a shows the head of the high-speed camera with the microphone on top, the light-intensifier and the pharyngo-laryngoscope.

For the recording process the flexible endoscope is detached from the camera system and is inserted through the nose under direct visual control through the ocular. When the tip of the endoscope has reached its final position in or above the larynx, the ocular can be re-attached to the light-intensifier and the camera for the recording process [7]. Fig. 1(b) shows the insertion of the flexible endoscope through the nose into the larynx.

3 Image Enhancement

Even though the light-intensifier has an amplification factor of 3, the recorded image sequences are still to dark for a meaningful subjective interpretation and an objective analysis using our motion analysis tool [8] (see Fig. 2a). For these reasons different digital image enhancing methods have been applied to the image sequences.
Image enhancement techniques can usually be separated into three different categories:

- Point processing (intensity transformations)
- Spatial filtering (neighborhood operations) and
- Frequency-domain filtering.

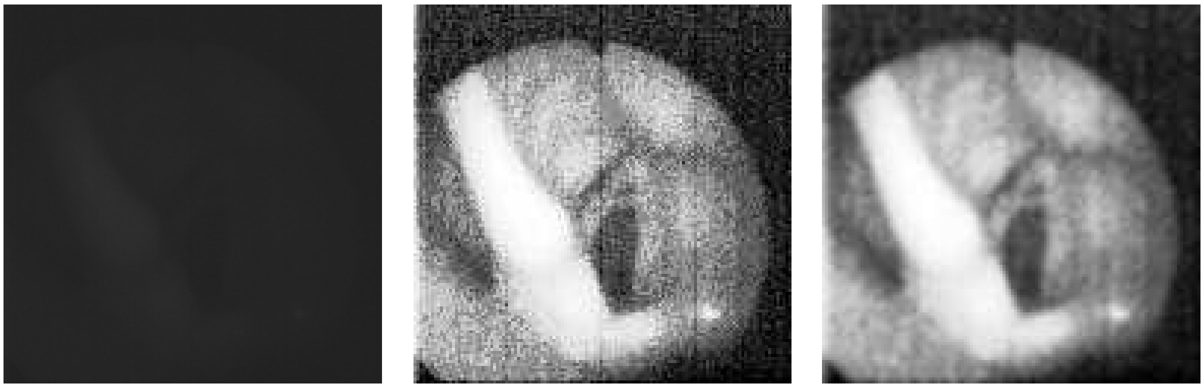
Since the enhancement operation has to be applied as fast as possible to the complete image sequence (16 MB) in order to speed up the examination and diagnosis for both the patient and the examiner, the class of frequency-domain techniques have been neglected for the time being. From the remaining two categories, a typical representative has been selected, applied to the image sequences and judged under both, objective and subjective points of view.

3.1 Intensity Transformation

Intensity transforms can be described as point operations $s = T(r)$, which transform a grayscale value g_k of the original image into a new grayscale value h_l, and comply to the following two criterions [4]:

- $T(r)$ is monotonically increasing in the interval $0 \leq r \leq 255$, which preserves the order in the grayscale, and
- $T \leq T(r) \leq 255$ for $0 \leq r \leq 255$, which guarantees a consistent mapping in the allowed color range.

. This kind of transformation is typically characterized by a probability function P, which describes the mapping from the grayscale distribution of the old image to the new distribution. A typical representative for such a transformation is the histogram linearization method, which has a uniform distribution as characteristic probability function. Intensity transforms such as histogram equalization change the pixel values only in a subjective way for a viewer. It does not change the relative contents of an image and therefore does not affect the adjacent motion analysis operation except for a change in the threshold levels. For this reason and for the extreme pixel noise created by the light intensifier (see Fig. 2b), we despeckle the equalized image using a 3×3-medianfilter, yielding a subjective smooth image, which can also be tracked using our semi automatic motion analysis tool [8] (see Fig. 2c). From the objective point of view, the combination of histogram linearization and median filtering has enhanced the laryngeal image sequence, so it can be tracked automatically, but from subjective point of view, the images seem very crude and have an extreme contrast.



(a) (b) (c)

Figure 2: First frame of original image sequence, recorded with a highspeed camera with a recording speed of 2000 fps and a resolution of 128×64 pixel using a light intensifier and a flexible endoscope (a), after histogram equalization (b), and after smoothing with a 3×3 median filter (c).

3.2 Spatial filtering

A very good way to combine both criterions, filtering and dynamic contrast modification, is the use of spatial filtering methods. Lee [5] proposed an algorithm for image enhancement, which can be expressed as

$$f'(x,y) = \mu + \alpha\, a(x,y) + \beta\,[f(x,y) - a(x,y)] \tag{1}$$

where $f(x,y)$ and $f'(x,y)$ are the original and processed images and $a(x,y)$ is the arithmetic mean of brightness of a $(n \times n)$ window that is centered on the pixel position (x,y). The parameters α, β and μ are real numbers. If $\beta > 1$, the image $f'(x,y)$ will be sharpened like using a high-pass filter. If $0 \leq \beta < 1$, the processed image $f'(x,y)$ will be smoothed as if passed through a low-pass filter. In the extreme case with $\beta = 0$, the image $f'(x,y)$ will be equal to the mean image $a(x,y)$ [5]. The optimal parametrization is usually task dependent. We have used $\alpha = .35$ and $\beta = 1.0$

Deng, Cahill and Tobin [1] have modified Lee's algorithm under the hypothesis, that the human visual system can be compared to a logarithmic image processing (LIP) system. Their new algorithm can be expressed as:

$$log(F'(x,y)) = \alpha log(A(x,y)) + \beta\,[log(F(x,y)) - log(A(x,y))] \quad \text{with} \tag{2}$$

$$log(A(x,y)) = \frac{1}{n \times m} \sum_{k=x-n/2}^{x+n/2} \sum_{l=y-m/2}^{l+m/2} log(F(k,l)) \tag{3}$$

and $F(x,y) = M - f(x,y)$ with M being the number of total colors used (for an 8-bit image, $M = 256$). The parameters α and β are real numbers and modify the dynamic range and the sharpness of the resulting image, respectively. The only restriction is made on β, which must be within the range of $[0, M]$ [1]. Figure 3 shows the image of 2a which has been enhanced, using Dengs et al. modified algorithm of Lee's method with a variation of α and β.

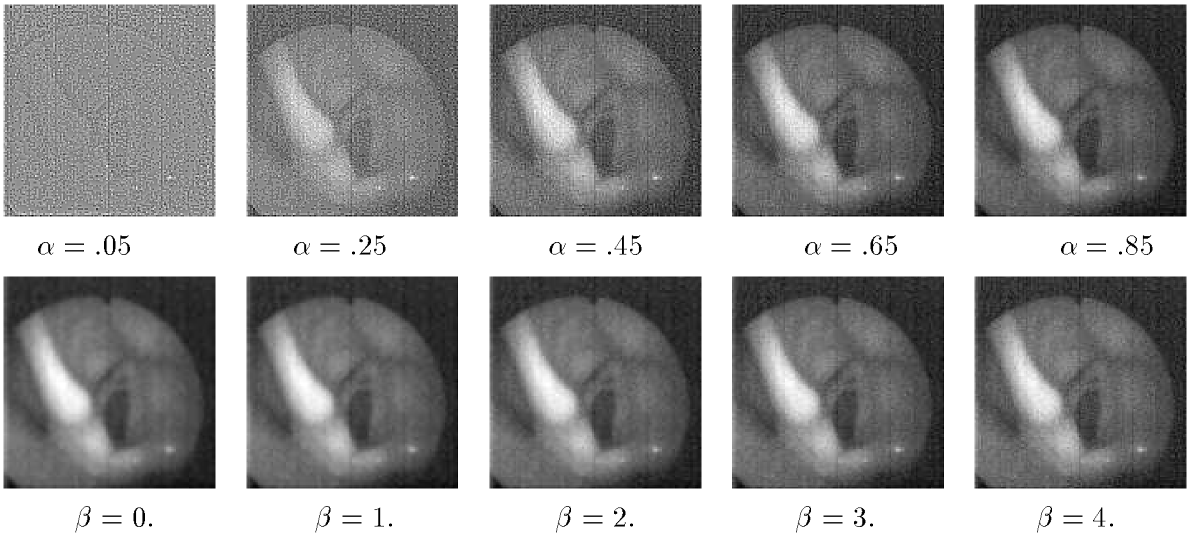


Figure 3: First frame of laryngeal image sequence of (Fig.2a), enhanced with Deng's modified LIP algorithm, varying the values for α and β. Top row $\beta = 1.$, bottom row $\alpha = .35$

4 Results and Conclusion

A high-speed recording of the utterance of the vowel combination /A/–/E/–/I/–/O/–/U/ has been made using the light intensifier and a flexible endoscope. Due to the restriction of the camera memory, only the last four vowels have been recorded. The recording has been enhanced, using both the histogram linearization method with the successive median filtering and Deng's modified spatial LIP filter algorithm. The resulting image sequences have been tracked using our semi-automatic motion analysis tool for vocal fold vibrations [8]. Objectively, both methods seem equal, yielding similar tracking results (see Fig. 4a.) with neglectable differences. Subjectively, the spatial filtering method yields better results for the visual examination of functional voice disorders using light intensified high-speed recordings of running speech.

References

[1] G. Deng, L. W. Cahill, and G. R. Tobin. The Study of Logarithmitic Image Processing Model and its Application to Image Enhancement. *IEEE Transactions on Image Processing*, 4(2), April 1995.

[2] U. Eysholdt, M. Tigges, T. Wittenberg, and U. Pröschel. Direct Evaluation of High–Speed Recordings of Vocal Fold Vibrations. *Folia Phoniatrica et Logopaedica*, pages 163–170, 1996.

[3] H. Ganz. *Lehrbuch der HNO-Heilkunde*. de Gruyer Verlag, 1990.

[4] R. Gonzales and P. Wintz. *Digital Image Processing*. Addison Wesley

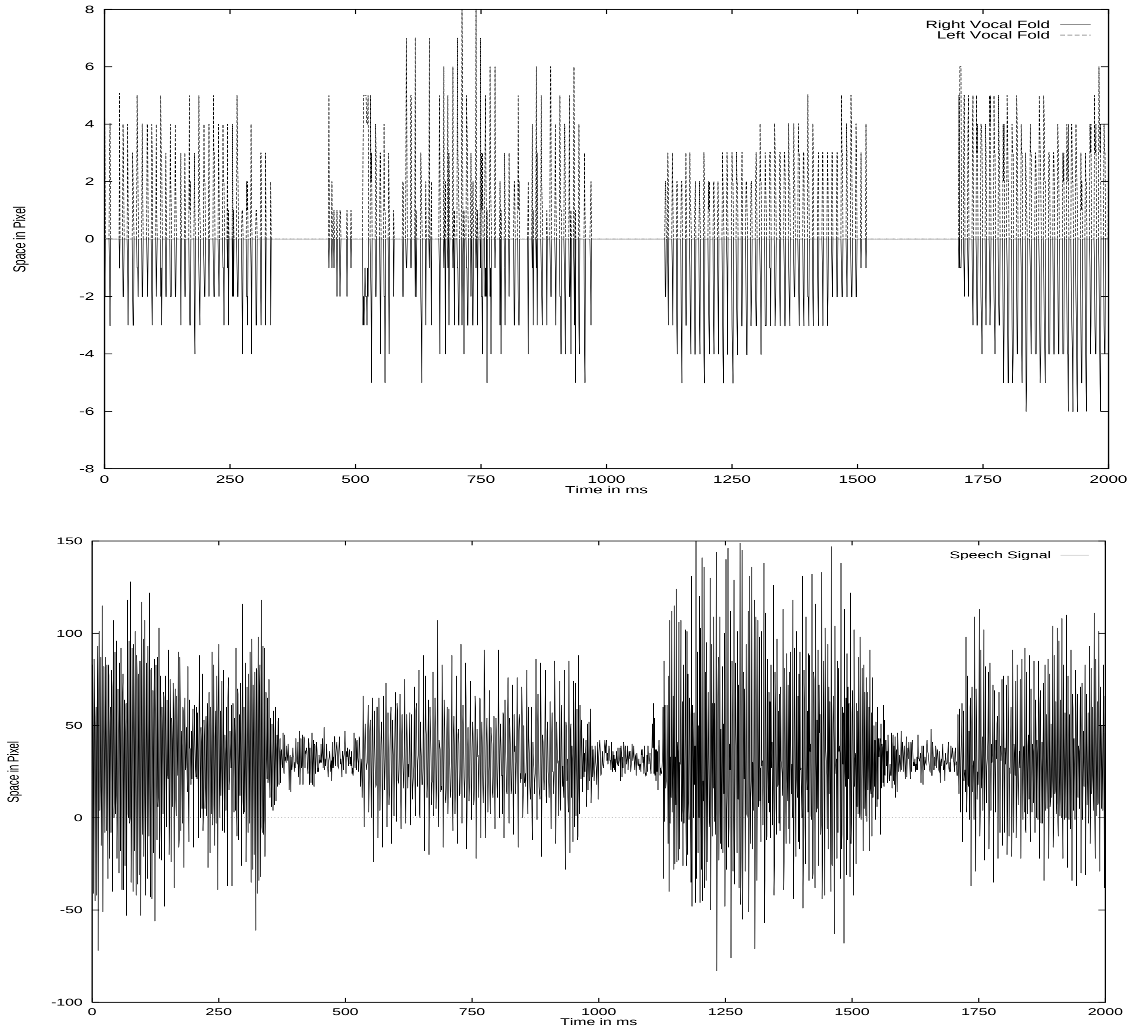


Figure 4: Trajectories of the right and left vocal cord, and coresponding audio signal during the phonation of the utterance of the vowel-combination /E/–/I/–/O/–/U/.

[5] J. S. Lee. Digital Image Enhancement and Noise Filtering by Use of Local Statistics. *PAMI*, (2):165–168, March 1980.

[6] M. Moser and G. Kittel. Digitale Hochgeschwindigkeitsglottographie. In U. Eysholdt, editor, *Differentialdiagnostische Möglichkeiten bei Dysphonien*, pages 27–31, 1990.

[7] M. Tigges. Endoskopische Untersuchungstechnik bei der Hochgeschwindigkeitsglottographie. In *Fortschritte der Untersuchungsmethoden der menschlichen Stimme*, pages 24–26. Verlag Abt. Phoniatrie, Göttingen, 1996.

[8] T. Wittenberg, M. Moser, M. Tigges, and U. Eysholdt. Recording, proccessing and analysis of digital highspeed sequences in glottography. *Machine Vision and Applications(8)*, pages 399–404, 1995.