Non-invasive photoglottography for use in the lab and the field

Eike Suthau[1]*, Peter Birkholz*[2]*, Alexander Mainka*[3] *and Adrian P. Simpson*[4]

[1]Centre of Microtechnical Manufacturing, TU Dresden, 01062 Dresden, Germany
Email: eike.suthau@tu-dresden.de
Web: www.iavt.de
[2]Institute of Acoustics and Speech Communication, TU Dresden, Germany
[3]Division of Phoniatrics and Audiology, ENT-Department, TU Dresden, Germany
[4]Institute of German Linguistics, Friedrich Schiller University Jena, Germany

Abstract

A non-invasive method of photoglottography is described that locates both light source and light sensor on the outside surface of the subject's neck. The study builds on the pioneering work of Honda, Maeda and colleagues, trying to overcome the problems of passing the light signal through the tissue of the neck twice. The specification of the light source and sensor have been improved, as has the optimal placement of the light sensor, now located centrally in the superior thyroid notch.
Index Terms: photoglottography, transillumination

1 Introduction

Various methods have been developed to observe the activity of the vocal folds during speech. Initially, this was done using a mirror attached to the end of a stiff rod inserted into the mouth and illuminated with an external light source. With these early types of laryngoscopy it was only possible to observe laryngeal activity during open vowels, at most with consonants being produced in the pharynx or larynx [1, 2]. Since these early beginnings, direct observation of the glottis during running speech has been made possible using a camera attached to the end of a nasal endoscope inserted through the nose and entering the pharynx through the velopharyngeal port [3, 4]. Despite the possibility of direct observation, the most widespread technique of observing and quantifying activity of the vocal folds, especially in a clinical context, is electroglottography (EGG) [5]. A weak current is passed between electrodes placed on the neck at approximately the height of the vocal folds and changes in impedance reflect changes in vocal fold contact. This method has been successfully used to characterize differences in normal voice quality, but has also been an important diagnostic tool for quantifying and classifying voice pathologies. EGG has certain advantages over direct visual observation of the vocal folds, the most important of which is that it is a completely non-invasive technique and it is also possible to capture vocal fold activity even when direct visual contact is lost due to laryngeal configurations required in particular voice qualities or consonantal articulations.

Photoglottography (PGG) is another indirect method of capturing vocal fold activity. Instead of reflecting changes in vocal fold contact, the magnitude of a light signal passing through the glottis correlates with the size of the glottal opening [6]. Approaches differ as to whether the light sensor or the light source is located above the glottis, but in general one is inserted into the pharynx using a nasal endoscope, the other being attached to the surface of the neck below the glottis, either at the height of the gap between the cricoid and thyroid cartilages, or between two of the upper tracheal cartilages. One major drawback of the majority of PGG setups is the need for introduction of either the light source or the light sensor through the nose. Besides the obvious disadvantages of discomfort and the possible detrimental effects on articulation, the idea of having to have a tube inserted through the nose is often enough to discourage potential subjects from taking part in an experiment in the first place. In addition to these drawbacks, qualified medical assistance is required to introduce the endoscope through the nose. One group of researchers have tried to overcome this problem by placing both light sensor and light source on the surface of the neck [7–10]. The light source, a high-power LED is placed laterally against the neck above the thyroid cartilage, the light sensor located centrally below the cricoid cartilage. However, the small number of publications that have arisen since the initial publications suggest that having a light signal pass twice through the tissues of the neck met with considerable problems. Nevertheless, if successful, a non-invasive method of transilluminating the glottis having both light source and light sensor outside the body has a number of considerable advantages, not least allowing it to be implemented in the field without the need for medical assistance.

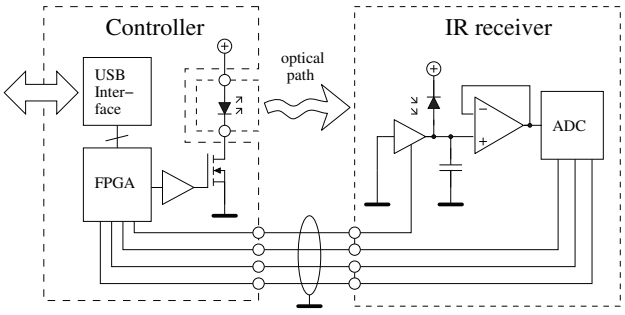

Figure 1: *Simplified block schematic of the PGG setup*

In the present paper we describe a technique that builds on the pioneering work of Honda, Maeda and colleagues. It has initially been designed for the study of the production mechanisms behind epiphenomenal and 'real' ejectives in three languages, German, English and Georgian, and in two locations [11]. The PGG method needs to be non-invasive, so as not to discourage subjects from already small speaker populations (e.g. Georgian speakers in Jena) taking part, and the method needs to be transportable so that it can be used in the field (Suffolk English speakers). Finally, as already mentioned above, this method also does not require qualified medical assistance. The technique we describe makes innovations in the light source and sensor used, but also, perhaps most surprisingly, in the optimal placement of the light sensor, centrally in the superior thyroid notch.

ISBN 978-3-8007-4275-2

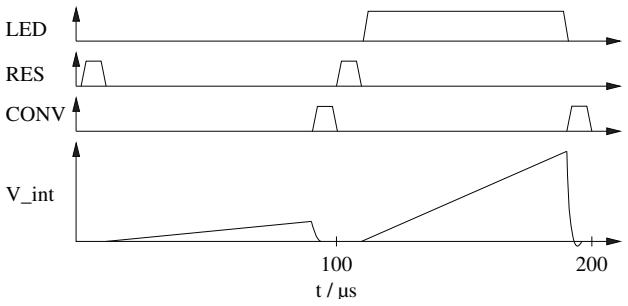

Figure 2: *Timing of LED, integrator reset and ADC conversion*

2 Method

2.1 Measurement electronics

The main challenge in the design of a PGG system is the high attenuation along the optical path. Near infrared light is least attenuated by the human skin [12]. Nevertheless, our experimental results show that light is attenuated by approximately 70 dB for a wavelength of 850 nm. This motivates both a high-power infrared (IR) emitter and a highly sensitive IR receiver.

As shown in Figure 1, the measurement setup consists of three printed circuit boards (PCBs), the controller, the IR emitter and the IR receiver. A sampling frequency of 5 kHz was chosen to provide adequate time domain resolution.

2.1.1 IR emitter and controller board

Nine surface-mounted Osram SFH4235 emitters are placed on a 25 mm by 52 mm aluminum core printed circuit board (PCB), arranged as identical strings of three diodes. As shown in Figure 3, highly thermally conductive epoxy adhesive secures a heat sink and a 50 mm diameter fan on the LED PCB, thus offering low thermal resistance from the IR emitter to the environment. This setup delivers approximately 5 W of optical power into the speaker's skin without causing excessive heating of the LEDs and subsequent discomfort for the subject.

A combination of field programmable gate array (FPGA) and industry standard USB FIFO IC is used for high accuracy, synchronized gate drive, IR receiver control and buffered high speed data transmission to a host computer. The device is controlled using a LabView program on the host computer which is used to visualize and store the recorded data.

2.1.2 IR receiver

For maximum sensitivity an Osram BPW34FA photo diode has been selected for its 7 mm^2 area, low dark current, built-in daylight filter and sufficient dynamic properties.

Transimpedance amplifiers (TIAs) are well suited to the acquisition of rapidly changing signals but introduce significant noise and have been used by previous PGG solutions. Compared to the previous PGG solution by Honda et al., we found that an integrating current to voltage converter as shown in Figure 1 is more suitable for PGG applications. For a sampling frequency of 5 kHz, a TIA-based circuit of the same bandwidth was found to have five times higher signal noise than the integrating circuit.

The timing of the integrating TIA is shown in Figure 2. After discharging the integrating capacitor using a tristate logic buffer, the photo current is integrated for the desired time of exposure. In order to cancel out ambient light, photo diode dark current, operational amplifier offset voltage and ADC offset sampling is performed at 10 kHz alternating the IR emitter between ON and OFF state.

Maximum transimpedance is achieved for a minimal integrating capacitor. It was found that a discrete capacitor can in fact be omitted completely, thus using the photo diode's parasitic junction capacitance as the integrating capacitor. Since the diode's junction capacitance is a function of junction voltage, it introduces a full-scale linearity error of about $\pm 1\%$ which is compensated for during post-processing.

By employing an integrating amplifier the presented PGG setup offers a significantly better resolution and signal-to-noise ratio than the solution presented by Honda and colleagues, and is further improved upon by placing the 14-bit successive-approximation register analog-to-digital converter (SAR ADC) immediately next to the buffer amplifier on the sensor PCB. Furthermore, our design has the advantage of passing only digital signals to the controller, thus completely avoiding analog line drivers and noise pick-up via the transmission line. The overall size of the sensor PCB is 11.5 mm by 19.5 mm.

2.1.3 Sample data post-processing

In addition to the correction of the sensor signal's linearity, signal noise can be reduced through sample data post-processing. The sensor board's 14-bit SAR ADC has a random error of approximately 3 LSB. Since the ambient light signal changes very slowly when compared to the PGG sampling rate, the observed signal is free of 50 Hz mains hum.

2.2 Light source and sensor placement

Optimal placement of the light sensor to achieve the best possible signal turned out to be one of the most challenging problems. Although lateral placement on the neck above the thyroid cartilage seems the most natural choice, we found extreme variation in signal strength both between speakers, presumably due to individual differences in tissue density and thickness, as well as, more disturbingly, in repeated recordings of the same speaker. Such variation casts serious doubts on reproducibility, as even data from the same speaker uttering the same expression on two separate occasions may vary wildly. In the end, the most reliable and reproducible signals were collected when the light sensor was placed centrally, directly in the superior thyroid notch (see Figure 4). This position has a number of advantages. It represents the shortest possible distance to the glottis, but perhaps more importantly, it is also a tightly defined location, that can be easily found again for the same speaker and also replicated across different subjects in the same sample.

3 Results

The system we are describing in this study is still in the initial stages of development, so at present we are only in a position to provide qualitative observations of some of the short and long term patterns in the transillumination signals. These have been elicited during vowel sequences

ISBN 978-3-8007-4275-2

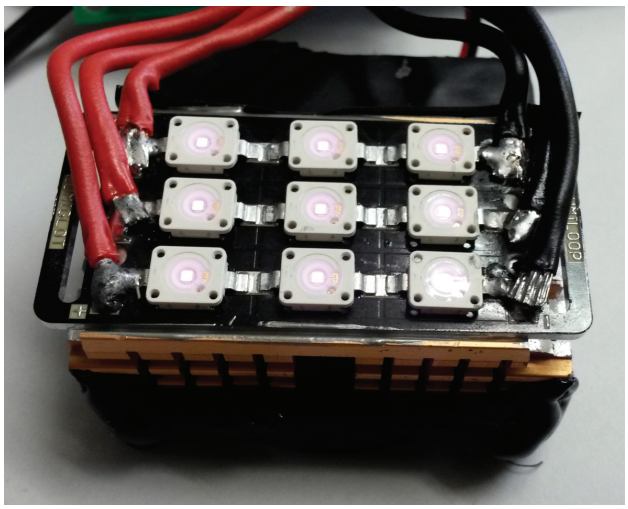

Figure 3: *Infrared LED PCB with heat sink and fan. Electrical contacts were sealed before application of the device on the skin of the neck.*

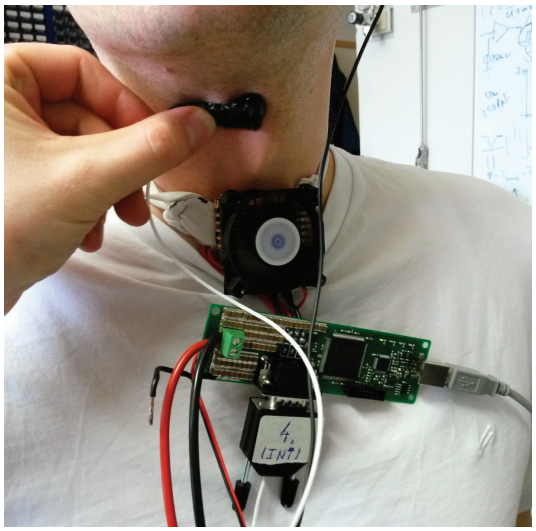

Figure 4: *Placement of LED array below the larynx and light sensor above the larynx in the superior thyroid notch.*

[i e a o u ø y], as well as in simple disyllables [VsV] and [V?V] produced by the first two authors. In a first attempt to examine the relationship between the transillumination signal and glottal area, high-speed videos of vocal fold activity were also recorded from a further subject producing breathy and modally phonated vowels.

3.1 Vowel quality

Figure 5 illustrates some of the longer term patterns observable in the transillumination signals. Displayed are two tokens of the vowel sequence [a e i o u ɛ ø y] produced by the second author at similar speech rates and with the same voice quality. Here we can observe slower changes in the vertical offset of the signal possibly related to differences in vertical larynx displacement depending on vowel category, but we can also see that the higher vertical signal offset seems to be related to whether a vowel is rounded or not, with [o u ø y] exhibting higher values than the unrounded vowels. There is also a slow change of the signal offset within each vowel, which we assume to be related

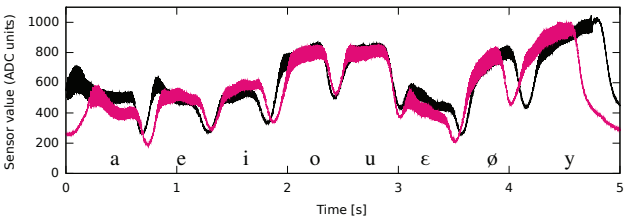

Figure 5: *The same vowel sequence produced twice by the second author at approximately the same speech rate.*

to a change of pitch with a possibly correlated change in larynx position. What is also clearly visible between each of the voiced vocalic stretches is a dip corresponding to the glottal closure which routinely occurs prior to the onset of voicing of vowels at syllable onset in German.

3.2 Consonant articulation

Figure 6 shows two tokens of the voiceless sibilant [s] in three different vowel contexts produced by the first (top) and second (bottom) author. In both tokens the positive vertical displacement of the transillumination signal corresponds to the vocal fold abduction for the voiceless fricative.

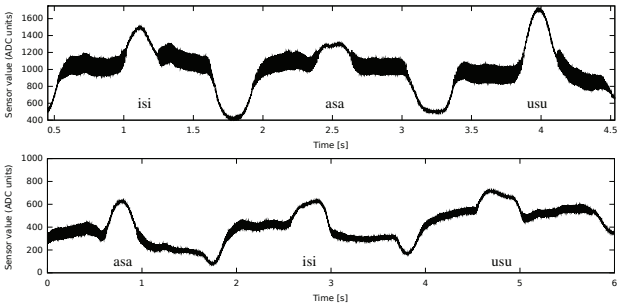

Figure 6: *Tokens of the voiceless sibilant* [s] *produced in the context of front, back and open vowels by the first (top) and second (bottom) author. Transcriptions are positioned under the stretch corresponding to vocal fold abduction.*

3.3 Correlation of PGG signal with glottal area

Previous studies using traditional (invasive) PGG showed that the PGG signal is highly and positively correlated with the glottal area [13]. To test the correlation of the glottal area and the signals obtained with the PGG system, we recorded endoscopic high-speed videos of the vocal folds along with the PGG signal of one trained speech therapist producing the vowel /ɛ/ both in breathy and modal phonation. High-speed videos were recorded at a framerate of 4000 frames/s with the HRES-Endocam (Wolf, Germany; product number: 5562) using a 90° rigid endoscope and proprietary software. The photodetector of the PGG system was held in place at the superior thyroid notch by the subject. The subject sustained the vowel in both voice quality conditions for at least two seconds with flat intonation at a convenient pitch.

From the high-speed videos, the glottal area waveform was extracted using the software *GlottalImageExplorer* [14], which performs glottal area segmentation using seeded region growing based on the method by [15]

ISBN 978-3-8007-4275-2

and then resampled from 4000 Hz to 44100 Hz. The PGG waveforms were resampled from the native PGG sampling rate of 5000 Hz to the same 44100 Hz and then high-pass filtered with a zero-phase digital filter to remove the slowly varying component in the waveform and to preserve the shape of the glottal pulses. This was done using a 4th order Butterworth filter with a cutoff frequency of 90 Hz that was applied in forward and reverse directions on the PGG signals. From the area waveform, a segment of 50 ms was extracted from the middle of the vowel. Because the laryngoscopic video and the PGG were not strictly synchronized during simultaneous recording, the 50 ms segment was cross-correlated with the PGG signal to detect the corresponding PGG segment by means of the maximum of the cross-correlation function. Figure 7 shows the resulting normalized glottal area waveforms extracted from the high-speed videos (gray curves) and the corresponding PGG waveforms (black curves) for modal phonation in a) and breathy phonation in b). We observe a very good correspondence between gottal area and PGG waveforms for breathy phonation (Pearsons correlation coefficient $r = 0.992$) and a slightly worse correspondence for modal phonation ($r = 0.915$). A less optimal placement of the photodetector may be one of the reasons for the poorer correlation between glottal area and PGG curves in modal phonation. However, complete glottal closure is not reflected with such an abrupt change in the PGG curve as it is in the glottal curve since light continues to pass through the tissue of the closed glottis, giving rise to a larger discrepancy between the curves at the point of maximum glottal adduction.

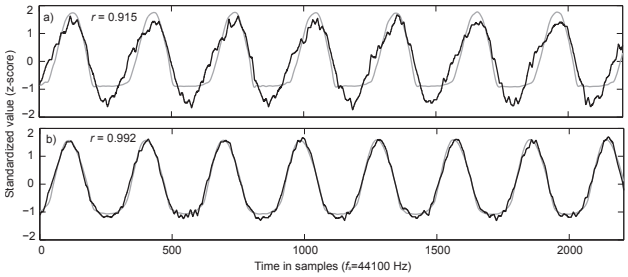

Figure 7: *Glottal area waveforms (gray) and corresponding high-pass filtered PGG waveforms (black) in a 50 ms interval procudcing /ɛ/ in modal (a) and breathy (b) voice quality.*

4 Discussion

We have presented a detailed description and specification for a non-invasive technique of photoglottography together with initial observations on the transillumination signals produced by the system. Our approach has built on the foundations laid by Honda, Maeda and colleagues [7–10]. The method locates both the light source and the light sensor outside the body on the surface of the neck. The optimal placement of the light sensor, in our case, above the vocal folds, proved to be in the superior thyroid notch. This position has a number of advantages. It is easy to find on a speaker, meaning that relocation on the same speaker should be reliable. From an articulatory point of view, it is also a relatively tightly defined spot, keeping differences in placement across a group of subjects to a minimum.

As mentioned in Sec. 2, the light has to pass the skin of the neck twice and is therefore strongly attenuated. Depending on the skin, the signal-to-noise ratio may for some subjects become so low that the PGG signal is barely interpretable. So far, we tested the system with four subjects and obtained very good signals (similar to Figure 7) for three of them, while it was very noisy for the fourth person, thus deserving further investigation.

The system is still in the early stages of development, and already a number of improvements are planned. In particular, sensor sensitivity is presently limited by the operational amplifier and the analog-to-digital converter. As explained above, maximum achievable transimpedance and linearity are currently limited by the photo diode's junction capacitance and its susceptibility to bias voltage variation. An integrating transimpedance amplifier based on an operational amplifier with purely capacitive feedback, higher transimpedance, and lower noise will therefore be developed to overcome these disadvantages. Additionally, operating the ADC in oversampling mode for further noise reduction will be investigated. If the sensor's sensitivity could be increased by a factor of 20 to 30 without evoking more noise, the full scale of the ADC would be used and illumination by a single LED would suffice for measurements with a better signal to noise ratio than the ones presented in this paper. Consequently, the instrument's size could be reduced significantly, a fan be avoided, and the entire device be designed as a bus-powered USB device.

Moreover, the LEDs are at present attached to a stiff base. A flexible mounting is being considered that will make better contact with the surface of the neck. Alternatively, a more compact arrangement of the LEDs could work equally well with a rigid PCB.

The current PGG system is capable of recording the speaker's voice using the measurement computer's sound card for reference and further analysis. Due to hardware and software limitations sound signal and PGG signal cannot be synchronized easily to a higher precision than a few milliseconds. Additionally, the sound card's sampling rate may not be chosen arbitrarily. For these reasons some manual post-processing is required for synchronization. Future PGG systems should therefore feature a dedicated microphone input for truly synchronous sound voice recordings at identical sampling rates.

As can be seen in Figure 4 the light sensor is being held in position by the investigator. A collar is being designed that will keep the sensor in its position in the superior thyroid notch even when the thyroid cartilage itself moves up and down, as it does during the course of a natural utterance.

High-speed endoscopic video recordings of a single subject producing modal and breathy-voiced vowels have revealed a high correlation between directly observed glottal area changes and the PGG signal. It will also be interesting to see how the transillumination signals compare with those produced by an electroglottographic system.

The signals presented here include the low frequency DC component, not least because the fourth author is interested in studying glottal adduction during the production of epiphenomenal and 'real' ejectives in English, German and Georgian, and it is the DC component which is of prime interest. However, from our discussion of the short-term, higher frequency patterns associated with vocal fold vibration during different phonation types, it is clear that it would be appropriate to remove the DC component with high-pass filtering.

ISBN 978-3-8007-4275-2

References

[1] J. N. Czermak, *Der Kehlkopfspiegel und seine Verwertung für Physiologie und Medizin*. Leipzig: Wilhelm Engelmann, 1860.

[2] M. Garcia, "On the invention of the laryngoscope," *Transactions of the International Medical Congress of London*, vol. 2, p. 3, 1881.

[3] M. Sawashima and H. Hirose, "New laryngoscopic technique by use of fiber optics," *Journal of the Acoustical Society of America*, vol. 43, p. 168, 1968.

[4] H. Hirose, "Investigating the physiology of laryngeal structures," in *The handbook of phonetic sciences* (W. J. Hardcastle, J. Laver, and F. E. Gibbon, eds.), pp. 130–152, Chichester: Wiley-Blackwell, 2 ed., 2013.

[5] A. J. Fourcin, "Laryngographic assessment of phonatory function," in *ASHA report 11: Proc. of the Conference on the Assessment of Vocal Pathology* (C. L. Ludlow and M. O. Hart, eds.), (Rockville, MD), pp. 116–127, The American Speech-Language-Hearing Association, 1981.

[6] P. Hoole, "Investigation of the devoicing gesture," in *Coarticulation: theory, data and techniques* (W. J. Hardcastle and N. Hewlett, eds.), pp. 294–299, Cambridge: Cambridge University Press, 1999.

[7] K. Honda and S. Maeda, "Glottal-opening and airflow pattern during production of voicelss fricatives: a new noninvasive instrumentation," *Journal of the Acoustical Society of America*, vol. 123, p. 5, 2008.

[8] M. Yeou, K. Honda, and S. Maeda, "Laryngeal adjustments in the production of consonant clusters and geminates in Moroccan Arabic," in *Proceedings of the 8th International Seminar on Speech Production*, pp. 249–252, 2008.

[9] K. Honda and S. Maeda, "Non-invasive photoelectroglottography method and device," Oct. 7 2010. US Patent App. 12/664,562.

[10] J. Vaissière, K. Honda, A. Amelot, S. Maeda, and L. Crevier-Buchman, "Multisensor platform for speech physiology research in a phonetics laboratory," *The Journal of the Phonetic Society of Japan,*, vol. 14, no. 2, pp. 65–78, 2010.

[11] A. P. Simpson, "Ejectives in English and German – linguistic, sociophonetic, interactional, epiphenomenal?," in *Advances in sociophonetics* (C. Celata and S. Calamai, eds.), pp. 187–202, Amsterdam: Benjamins, 2014.

[12] A. N. Bashkatov, E. A. Genina, V. I. Kochubey, and V. V. Tuchin, "Optical properties of human skin, subcutaneous and mucous tissues in the wavelength range from 400 to 2000nm," *Journal of Physics D: Applied Physics*, vol. 38, no. 15, p. 2543, 2005.

[13] W. Habermann, J. Jiang, E. Lin, and D. G. Hanson, "Correlation between glottal area and photoglottographic signal in normal subjects," *Acta Oto-Laryngologica*, 2000.

[14] P. Birkholz, "*GlottalImageExplorer* – An open source tool for glottis segmentation in endoscopic high-speed videos of the vocal folds," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2016* (O. Jokisch, ed.), (Dresden), TUDPress.

[15] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Döllinger, "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos," *Medical image analysis*, vol. 11, no. 4, pp. 400–413, 2007.

ISBN 978-3-8007-4275-2