ISSN: 2572-4193

Xi et al. J Otolaryngol Rhinol 2023, 9:130
DOI: 10.23937/2572-4193.1510130
Volume 9 | Issue 1
Open Access

Journal of
Otolaryngology and Rhinology

ORIGINAL RESEARCH

Combining Artificial Intelligence and Automatic Analysis to Study of the Closure Characteristics of Healthy Human Vocal Cords during Phonation Using Synchronous Electroglottography and Laryngeal High-Speed Videoendoscopy

***Wang Xi*[1]*, Xuan Jiacheng*[2]*, Zhao Caidan*[2]*, Zhuang Peiyun*[3*] *and Huang Lianfen*[3]**

[1]*School of Medicine, Xiamen University, China*

[2]*School of Information Engineering, Xiamen University, China*

[3]*Department of Voice, Zhongshan Hospital, Xiamen University, China*

***Corresponding author:** *Zhuang Peiyun, Department of Voice, Zhongshan Hospital, Xiamen University, 361000, China, Tel: 8613003989899*

Abstract

Objective: To use an artificial intelligence neural network to automatically analyze images from high-speed videoendoscopy and synchronous electroglottographic signals in healthy vocal folds to obtain the accurate glottal opening and closing instants in the glottal cycle. This method will be compared to the traditional electroglottography point derivation method to explore the glottal opening and closing characteristics of the vocal folds and better interpret the opening and closing instances of the electroglottography.

Methods: Images from high speed videoendoscopy (HSV) and signals from an electroglottogram (EGG) were simultaneously collected for 20 subjects when pronouncing the vowel /i/. High-speed videoendoscopy was used to create a periodic one-dimensional waveform diagram of the glottal area during phonation which was used to calibrate the original EGG signal and obtain the glottal closing instant and the glottal opening instant. This data was combined into a time domain change diagram and used as a training model of the neural network. A prediction of the closed point coordinate values for a segment of EGG signals were calculated and compared against the real value given by videoendoscopy. This process trains the network model to predict a wide range of EGG signals.

Results: 1) The positions of the opening and closing instants of the vocal cords identified by the DEGG method were closer to the peaks than those identified by the analysis of the HSV using Glottal Image Explorer software on the same periodic waveform, and the positions predicted by the neural network model were very close to those identified by the calibration method; 2) The data analyzed using the derivative EGG method had an average OQ was 52.27% and an average CQ was 49.52%. In comparison, the data analyzed using the neural network had an average OQ of 34.88% and an average CQ is 66.12%. As such, the EGG analyzed using the neural network was significantly closer to the calibration method, which had an average OQ was 34.60% and an average CQ was 65.42%.

Conclusion: There is a high consistency between the glottal opening and closing instants predicted by the trained neural network and the HSV, which can identify and predict the position of the glottal opening and closing points on EGG. This method can provide a more convenient and accurate method for EGG signal analysis and can be applied to the automatic analysis of vibration patterns of various voice diseases.

Keywords

High-speed videoendoscopy, Electroglottography, Vocal cord vibration, Artificial intelligence

Introduction

The periodic opening and closing of the glottis are fundamental elements of the vocal vibration pattern. Irregularities or inconsistencies with the vibrations of the vocal folds lead to alterations in voice quality

Citation: Xi W, Jiacheng X, Caidan Z, Peiyun Z, Lianfen H (2023) Combining Artificial Intelligence and Automatic Analysis to Study of the Closure Characteristics of Healthy Human Vocal Cords during Phonation Using Synchronous Electroglottography and Laryngeal High-Speed Videoendoscopy. J Otolaryngol Rhinol 9:130. doi.org/10.23937/2572-4193.1510130

Accepted: January 07, 2023: **Published:** January 09, 2023

which can be perceived as voice disorders. Vocal fold vibrations are usually observed using laryngoscopy which is done using either stroboscopic light or more recently high speed videoendoscopy (HSV) [1,2]. Both of these techniques require the use of computer assisted evaluation due to the difficulty to evaluate the vibration pattern of vocal cords though direct human observation.

For HSV, a common method of analysis is by tracking the motion of the vocal fold edge [3]. The two most common parameters extracted from the motion tracking is the Open Quotient (OQ) which is the ratio between the time of glottal opening over the duration of the glottal cycle and the Contact Quotient (CQ) which is the is ratio between the time of glottal closure over the duration of the glottal cycle. These quotients provide a way to measure changes in the composition of the glottal cycle when studying the pathogeneses of voice diseases and different vocal patterns [4,5]. These studies have broad applications from understanding vocal variation in human development to clinical studies including Christopher S, et al. which used the CQ value to evaluate the effect of semi-closed vocal tract training on vocal fold vibration and found that the CQ was positively correlated with semi-closed vocal tract training and lead to an increase in voice quality [6]. For studies examining vocal fold vibration and function, it is critical that these vocal fold vibration parameters such as OQ and CQ are measured accurately.

A non-invasive alternative to high-speed laryngography to detect the opening and closing of vocal cords during phonation is electroglottography (EGG). Compared with the semi-invasive acquisition of HSV, the acquisition of EGG is faster, less invasive, and more convenient. However, only the HSV of the glottal area can directly visualize the opening and closing state of the glottis and obtain the exact opening and closing instants of the vibrating vocal folds. Previous studies using EGG have observed that there are maximal accelerations in opposite directions at the opening and closing phases of the glottal cycle, so the first order derivative of the original signal of the EEG signal can be obtained which gives the DEGG. From this, the opening phase can be obtained by the temporal location of the maximum value of the derivative and the closing phase can be obtained by the temporal location of the minimum value of the derivative. Using the maximum acceleration found in DEGG, many scholars have been able to use EGG alone to evaluate parameters of phonation such as OQ and CQ, but this technique has not been fully adopted since it still lacks a certain level of accuracy in comparison to HSV.

Therefore, considering the respective advantages and disadvantages of HSV and EGG, this paper proposes an artificial intelligence system of teacher-student network mode based on cross-mode supervision. Through the teacher-student network model, the HSV visual signal can be used as the supervision signal to calculate the characteristic instants in the time-varying signal of the glottis region, and the synchronous EGG signal can be used as the input to train the teacher-student network model. After the training, using only the EGG signal as input, the system should be able to accurately estimate the characteristic moments of glottal changes from the EGG signal alone, without the need for HSV data as a monitor and reference.

To provide a more convenient and accurate method for the analysis of EGG signals, the synchronous acquisition of EGG signals and high-speed imaging of healthy human vocal cords vibration was performed. These signals were analyzed using an automatic analysis performed by an artificial intelligence neural network. Once trained, it is expected that this neural network will improve the accuracy of vocal cord analysis using EGG which will be used as a minimally invasive measure of the vibrational patterns of phonation.

Subjects and Methods

Study object and collection method

Twenty subjects aged 20-53 years (mean 32 ± 2.5 years) without irregular phonation, respiratory diseases and other laryngeal lesions were recruited. During testing, each subject was observed using an endoscopic high-speed camera and a double electrode plate from the electroglottograph on both sides of the thyroid cartilage to collect images from the laryngeal HSV and synchronous EGG signals respectively. Measurements were collected over the span of 8 seconds during which the subjects pronounced the vowel /i/. The acquisition frequency of the HSV was set to 4K frames per second, and the acquisition frequency of the EGG was 48 KHz. An example of the synchronous measurement is presented in Figure 1.

Research methods

Calibration method: Acquisition of the opening and closing points of the vocal cords on the electroglottogram signal: To calibrate the signal from the electroglottogram, the HSV must first be processed. Initially, the individual frames of the high-speed video were recorded and processed using the Glottal Image Explorer software (GIE). This software created a periodic one-dimensional waveform which measured the glottal area in respect to time when the vocal cord vibrated (Figure 2a). From this waveform, the glottal closing moment was marked as the time in each glottal cycle when the glottal area was equal to zero, and the glottal opening moment was marked as the first non-zero point after the closing moment (Figure 2b). Next, the waveform signals were upscaled to create a one-to-one correspondence between the original EGG signals and the one-dimensional images. This was done using sampling pulses to convert the continuous signals into discrete signals of time and amplitude.

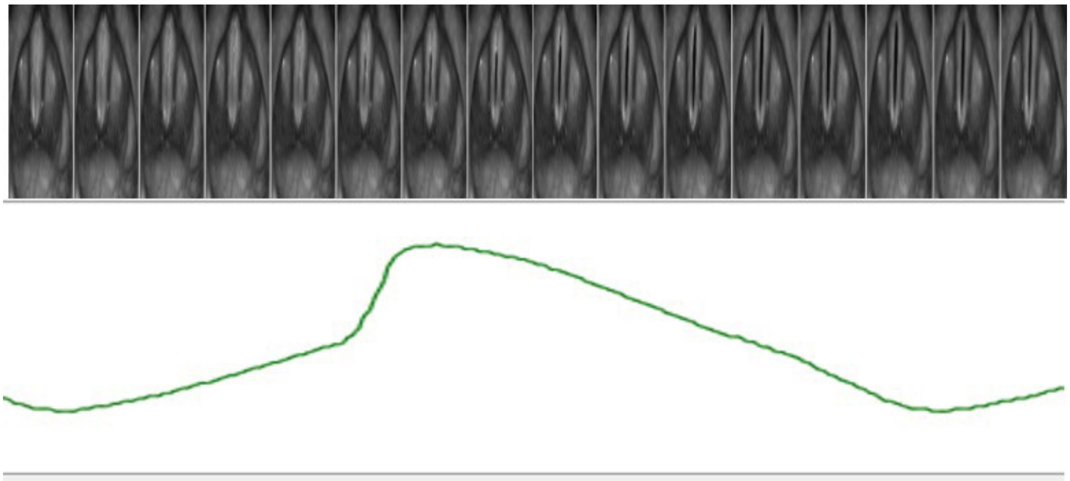

Figure 1: Schematic diagram of synchronously acquired HSV and EGG signals: each frame of the high-speed video are correlated to a point at the same time point on the EGG waveform.

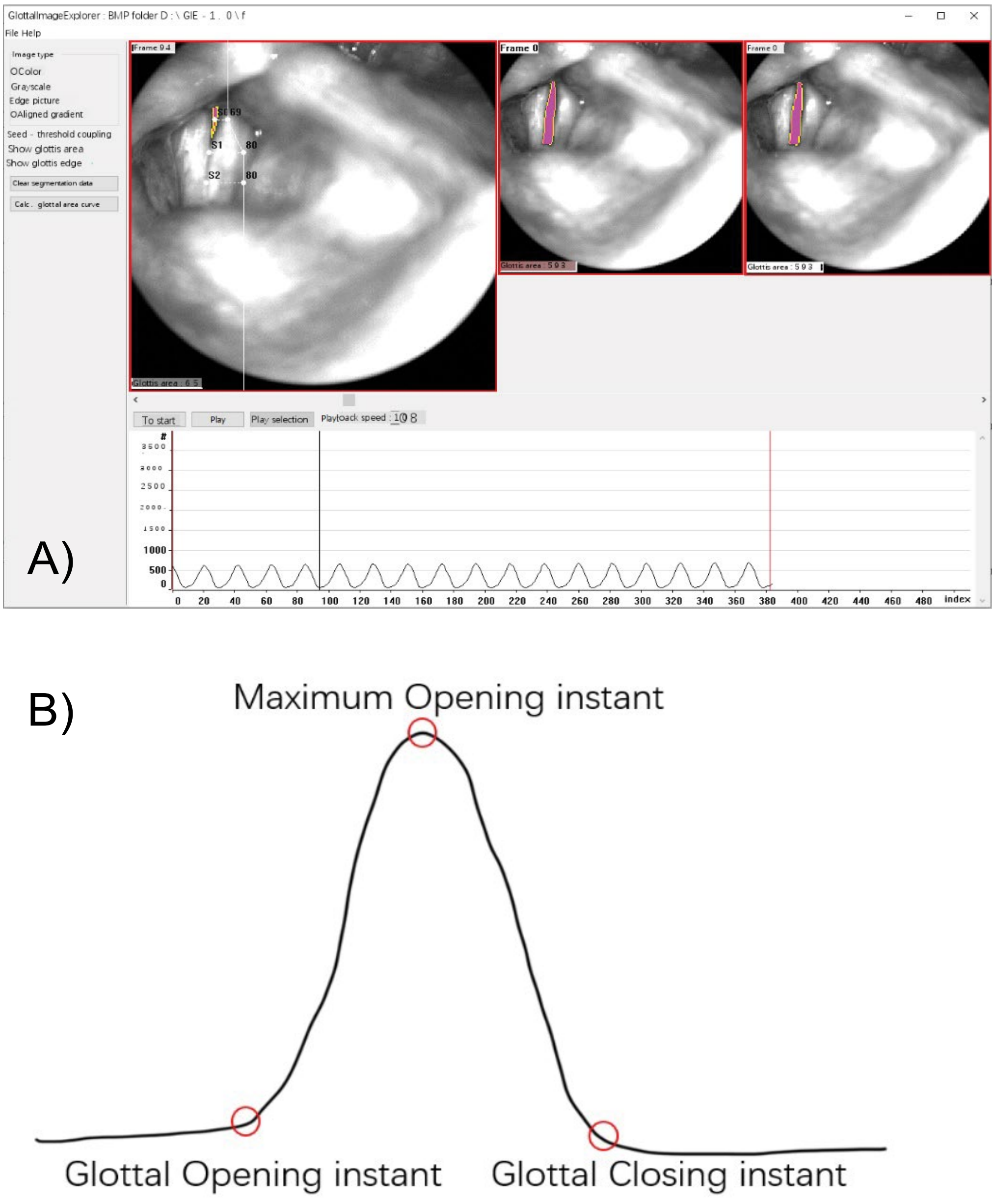

Figure 2: (a) One-dimensional waveform of GIE segmentation of vocal fold vibration as a function of time domain; (b) Schematic representation of the location of key points in a single glottal region change cycle.

Construction of the neural network model

Teacher-student network model: Teacher-student network model is a widely used method in the field of model compression, which can be used to compress deep neural network models into a more simplified model. The teacher network is the original complex neural network which will have the best performance and generalization ability when dealing with the tasks. This teacher network then simplifies the parameters and instructs a more lightweight student network how to achieve the best model effect. This model architecture can exchange information between both single and multi-modal signals, thus providing cross-modal supervised training [7-9].

Punar Jay, et al. detected voice activity information through audio signals. This method is a person voice activity detector model based on video signals and provides cross-modal supervised training. The experimental results show that the trained detector model based on specific person activity of video signals has better performance than the general detector [10]. Ming Zhao, et al. Constructed a human pose recognition model using synchronous RF signals and visual signals as inputs. They extracted accurate pose information from video signals to guide the overall training process of the model. Once the network training is completed, the model only needs to use RF signals as inputs. The activity posture of the human body can be detected in a visible scene. In addition, the human pose recognition model based on RF signals can also accurately detect the 2D pose of the human body in occluded scenes. Therefore, through cross-modal supervised training, we can achieve interactive learning of network models in multi-modal signals. This method can not only provide more feature information for the analysis task, but it also makes the network model more efficient and flexible in the implementation of tasks [11]. In this paper, we attempt to apply this network model architecture to accurately detect glottal closure time only through EGG signal without HSV signal.

The schematic diagram of the teacher-student network model used in this paper is shown in Figure 3. The goal of this Network is to use the original EGG signal as an input to predict the corresponding glottal opening and closing time (Figure 3).

Training data set: For each of the 20 subjects, video frames were extracted to train the neural network. In total, 640,000 (4K*8s*20) video frames and 7.68 million (64,000*12) EGG signals were used for training the neural network.

Approximate glottis opening and closing moment obtained by DEGG: In accordance with the methods, the EGG signal at the same time were derived to obtain the DEGG change graph of the signal across time. The maximum DEGG value (glottis closing moment) and the minimum DEGG value (glottis opening moment) were obtained. An example of the DEGG graph is shown in Figure 4.

Comparison

The instantaneous opening and closing time of the vocal folds were obtained under the three methods (calibration, model training output, DEGG). The OQ and CQ parameters of each method were compared to find the accuracy of each method.

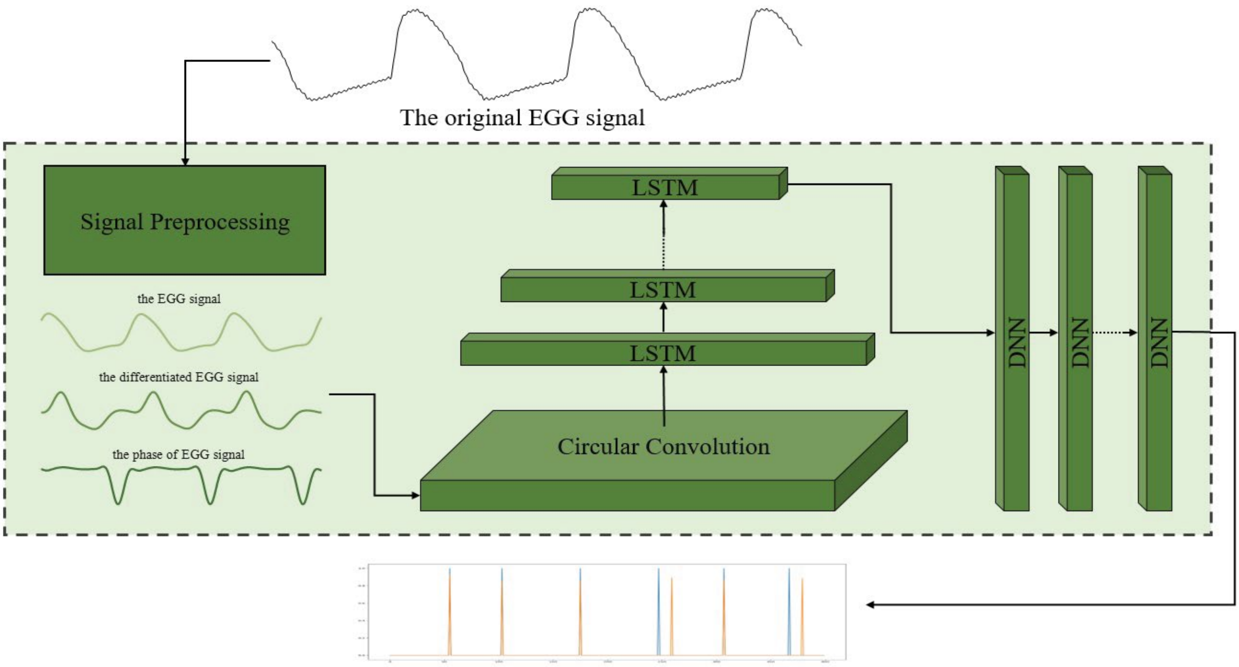

Figure 3: Schematic diagram of input and output of neural network.

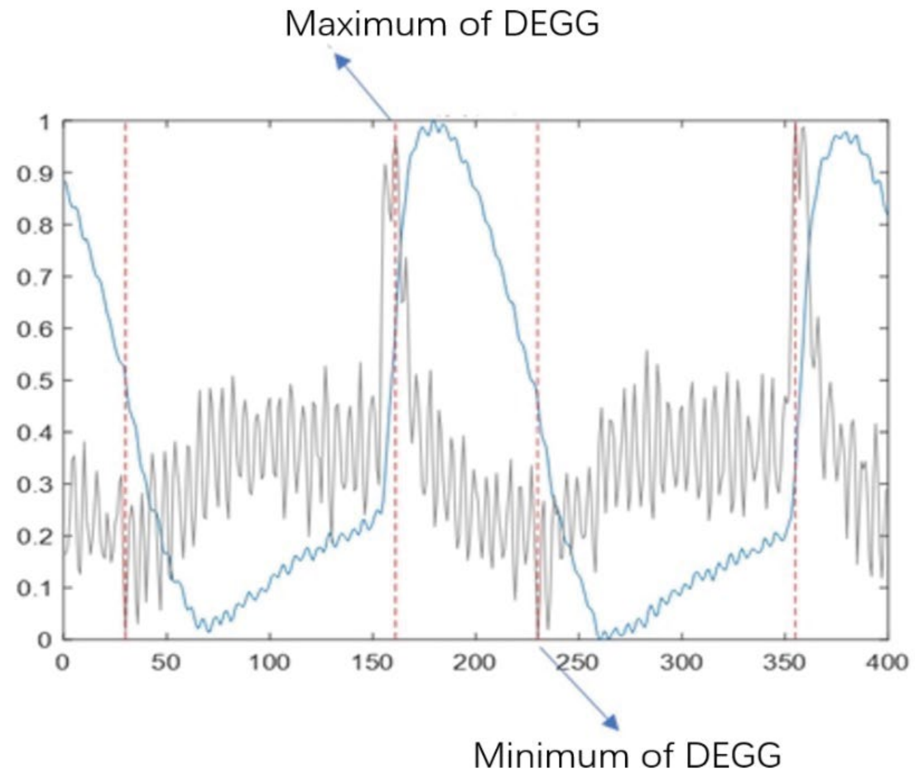

Figure 4: An example of the EGG and DEGG change vs. time graph which shows the glottal opening point (derivative maximum) and glottal closing point (derivative minimum). The DEGG signal is seen in gray and the EGG signal is seen in blue.

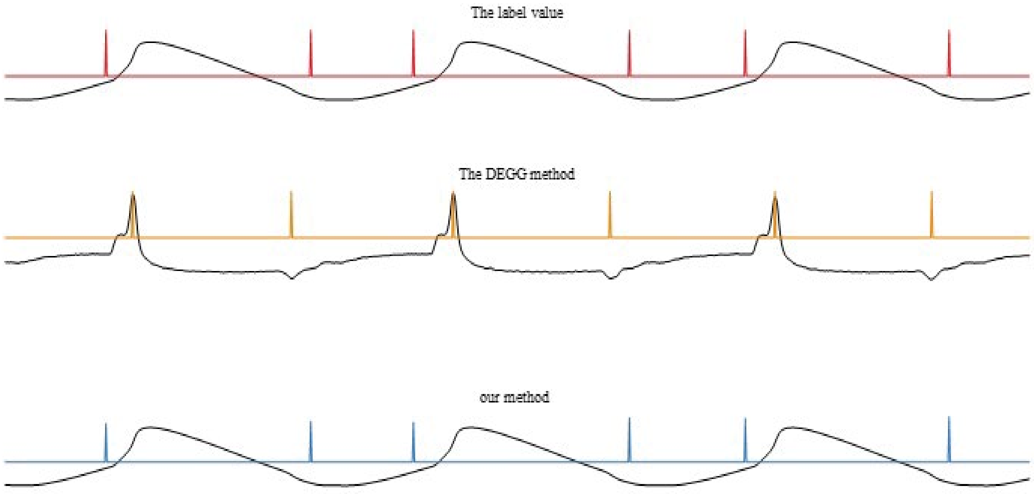

Figure 5: Is a characteristic point diagram of a normal subject by three methods (Red = "GIE analysis of HSV"; Yellow = "DEGG Analysis of EGG data"; Blue = "neural network analysis of EGG data").

Results

Representative graphs of the EGG signal during vocal fold vibration by the 3 methods in this study are shown in Figure 5.

The degree of glottal opening demarcated by the time domain image of the glottal area captured via High-speed videoendoscopy is shown in red in Figure 5; Characteristic glottal opening and glottal closing values were obtained by the DEGG method are shown in yellow in Figure 5. Finally, after training the OQ and CQ values from the neural network are shown in blue in Figure 5.

When comparing the positions of the feature points in the same domain from the different measuring techniques, it is found that the opening moment and the closing moment of the vocal cords identified by the signals from the traditional DEGG methods are closer to the wave peak on the same periodic waveform than those obtained by the calibration curve analyzed with GIE. This is seen due to the open-close interval in the glottal cycle obtained by the DEGG method being smaller than that obtained by the calibration method. In

Table 1: Comparison of Open Quotient (OQ) and Close Quotient (CQ) of vocal fold vibration in subjects obtained by DEGG, HSV, and EGG using the neural network.

Method	DEGG method	Calibration method	Neural network output
OQ (%)	52.27	34.60	34.88
CQ (%)	49.52	65.42	66.12

addition, when a segment of the original EGG signal was input into the trained neural network, the location of the opening and closing instances were consistent with the calibration points.

Obtain the mean value of vocal fold vibration parameters OQ\CQ (Table 1) through the feature points on vocal fold EGG signals obtained by different methods:

The average OQ and CQ calculated by the DEGG collected from the healthy subjects in this study were 52.27% and 49.52% respectively, the average OQ and CQ calculated using the GIE analysis of the HSV was 34.6% and 65.42% respectively, and the average OQ and CQ calculated using the trained neural network was 34.88% and 66.12% respectively.

Discussion

In this study, the reference/calibration methodology was based on the acquisition of High-speed videoendoscopy in healthy subjects when pronouncing the vowel /i/. Images of the glottal area were extracted and analyzed using the GIE software to segment and produce a periodic change map. When the sampling frequency is high enough, this method has been shown to be the most accurate means of measuring the CQ and OQ of focal fold tissue since it is able to accurately measure the full area of the glottis over the oscillation cycle. The sampling frequency of laryngeal high-speed photography used in this study is 4K f/s which is well above the fundamental frequency of human phonation. This meant that each phonation cycle can be accurately reconstructed using the images which results in the acquisition of accurate characteristic points. In this study, the average CQ and OQ obtained after calibration were 65.41% and 34.60%, respectively. These average CQ and OQ values are consistent with previous studies which confirm the accuracy of the calibration curve.

To acquire the standard accuracy of the CQ and OQ obtained by EGG, DEGG analysis was performed on the EGG signals. The first-order derivation of the signal was used to find the maxima and minima which were mapped on to the original EGG signal. Using this technique, the average CQ and OQ were 49.52% and 52.27%, respectively. In comparison to the calibration curve, this value is significantly different in comparison to the calibration curve. The alternative method to analyze the EGG signals in this study was using the neural network to predict the CQ and OQ based on previous comparisons between synchronous EGG and HSV data. After training, the neural network measured the average CQ as 66.12% and the average OQ as 34.88%. These CQ and OQ values correlate strongly with those calculated by the calibration method. Based on these average CQ and OQ values from the two EGG analysis methods, It can be considered that the predicted vocal cord vibration based on the neural network is closer to the actual vibration of the vocal cords than the vibration predicted by DEGG. This can be seen in Figure 5 where there were significant differences in the characteristics of the vocal fold oscillation between the traditional DEGG point diagram and the neural network point diagram. The differences in the point fields can in part be explained by the fact that the DEGG method is more dependent on a highly stable and simple EGG signal. This is caused by the because there are many factors including interglottal mucus and irregular vibration of vocal cords which affect the relationship between the DEGG and glottal closure [12-17]. In comparison, the neural network is better able to identify and predict the position of the opening and closing of the glottis because it is better able to filter the irregularities in the EGG signal from the HSV training.

Paragrah on the clinical potential for this neural network (Ex. Non-invasive measure of OQ, CQ, and other vibration parameters), provide some limitations that it may have (Ex. Since this neural network analysis is based on a dataset, abnormal vocal fold vibration which is not represented in the dataset would not be accurately analyzed using this neural network), and what potential future studies will be used to address these limitations (Ex. Use this neural network with patients who have a wide variety of abnormal phonation to better train the NN).

Conclusion

The cross-modal teacher-student network model we designed extracts the important moments of glottal movement in a quasi-periodic way by inputting EGG signals, and calculates the vocal fold vibration parameters, which provides accurate data for further analysis of the opening and closing of vocal fold vibration and is expected to be applied to the follow-up study of EGG on pathological vocal fold vibration characteristics.

Statement of Equal Authors' Contribution

Zhuang Peiyun and Huang Lianfen participated in the design of this study. Xuan Jiacheng and Zhao Caidan carried out the study and collected important background information. Wang Xi drafted the manuscript and performed the statistical analysis. All authors read and approved the final manuscript.

References

1. Kendall Katherine A (2009) High-Speed laryngeal imaging compared with videostroboscopy in healthy subjects [J]. Arch Otolaryngol Head Neck Surg 135: 274-281.
2. Deliyski DD, Petrushev PP, Bonilha HS, et al. (2008) Clinical implementation of laryngeal high-speed videoendoscopy: Challenges and evolution. Folia Phoniatr Logop 60: 33-44.
3. Lohscheller J, Svec JG, Döllinger M (2013) Vocal fold vibration amplitude, open quotient, speed quotient and their variability along glottal length: Kymographic data from normal subjects. Logoped Phoniatr Vocol 38: 182-192.
4. Mandal T, Rao K S (2018) Robust Detection of Glottal Activity Using Unwrapped Phase Electroglottographic Signal. ICASSP 5584-5589.
5. Winkler R, Walter S (2006) EGG open quotient in aging voices-changes with increasing chronological age and its perception. Logoped Phoniatr Vocol 31: 51-56.
6. Gaskill CS, Quinney DM (2012) The effect of resonance tubes on glottal contact quotient with and without task instruction: A comparison of trained and untrained voices. J Voice 26: e79-e93.
7. Hinton G, Vinyals O, Dean J (2015) Distilling the Knowledge in a Neural Network. Computer Science 14: 38-39.
8. Zhou G, Fan Y, Cui R, Bian W, Zhu X, et al. (2017) Rocket Launching: A Universal and Efficient Framework for Training Well-performing Light Net.
9. Gupta S, Hoffman J, Malik J (2016) "Cross Modal Distillation for Supervision Transfer," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2827-2836.
10. Chakravarty P, Tuytelaars T (2016) Cross-Modal Supervision for Learning Active Speaker Detection in Video. ECCV 285-301.
11. Zhao M (2018) "Through-Wall Human Pose Estimation Using Radio Signals," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 7356-7365.
12. Mendelsohn AH, Zhang Z, Luegmair G, Orestes M, Berke GS (2015) Preliminary Study of the Open Quotient in an Ex Vivo Perfused Human Larynx. JAMA Otolaryngol Head Neck Surg 141: 751-756.
13. Henrich N, Alessandro C, Doval B, Castellengo M (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. J Acoust Soc Am 115: 1321-1332.
14. Baken RJ (1992) Electroglottography. J Voice 6: 98-110.
15. Rothenberg M, Mahshie JJ (1988) Monitoring vocal fold abduction through vocal fold contact area. J Speech Hear Res 31: 338-351.
16. Hosokawa K, Yoshida M, Yoshii T, Takenaka Y, Hashimoto M, et al. (2012) Effectiveness of the computed analysis of electroglottographic signals in muscle tension dysphonia. Folia Phoniatr Logop 64: 145-150.
17. Verdolini K, Druker DG, Palmer PM, Samawi H (1998) Laryngeal adduction in resonant voice. J Voice 12: 315-327.