INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Accurate Synchronization of Speech and EGG signal using Phase Information

Sunil Kumar. S. B, K. Sreenivasa Rao, Tanumay Mandal

Indian Institute of Technology Kharagpur, India
sunil220552@gmail.com, ksrao@iitkgp.ac.in, tanum.dets@gmail.com

Abstract

Synchronization of speech and corresponding Electroglottographic (EGG) signal is very helpful for speech processing research and development. During simultaneous recording of speech and EGG signals, the speech signal will be delayed by the duration corresponding to the speech wave propagation from the glottis to the microphone relative to the EGG signal. Even in same session of recording, the delay between the speech and the EGG signals is varying due to the natural movement of speaker's head and movement of microphone in case MIC is held by hand. To study and model the information within glottal cycles, precise synchronization of speech and EGG signals is of utmost necessity. In this work, we propose a method for synchronization of speech and EGG signals based on the glottal activity information present in the signals. The performance of the proposed method is demonstrated by estimation of delay between the two signals (speech signals and corresponding EGG signals) and synchronizing these signals by compensating the estimated delay. The CMU-Arctic database consist of simultaneous recording of the speech and the EGG signals is used for the evaluation of the proposed method.

Index Terms: Speech signal, EGG signal, Synchronization.

1. Introduction

Speech processing tasks such as voice conversion, speech synthesis and emotion recognition can benefit from simultaneous recording of speech and EGG signals [1, 2, 3]. The databases such as the CMU Arctic database [4] and Keele University database [5] consist of simultaneous recordings of speech and corresponding EGG signals. The EGG signal is recorded using the EGG device which measures the degree of contact between the vibrating vocal folds during voice production [6]. Microphone is used to record the speech signal of the speaker. During simultaneous recording of speech and EGG signals, the speech signal traverses an additional distance equal to the sum of vocal-tract length (distance between glottis and lips) and the distance between lips and microphone, as compared to the EGG signal. As a result, the speech signal is delayed relative to the EGG signal. The time delay introduced by this phenomenon will be known as glottis-to-microphone delay. Figure 1 shows the time-lag of speech signal with respect to EGG signal. Figures 1(a) and 1(b) show EGG signal and the corresponding speech signal, respectively. Here, the glottis-to-microphone delay can be observed by measuring the time-delay between Glottal Closure Instants (GCIs) determined from EGG signal and the corresponding speech signal. Figures 1(c) and 1(d) show the GCI corresponding to EGG and speech signal, respectively. Positive peaks in first order derivative of EGG (DEGG) signal are used to determine GCIs from EGG signal [7] and Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) [8] method is used to determine GCIs from speech signal. Instants "p" and "q" marked on the time axis of Figure 1(d) correspond to GCI of EGG and speech signal, respectively. From instant "p" and

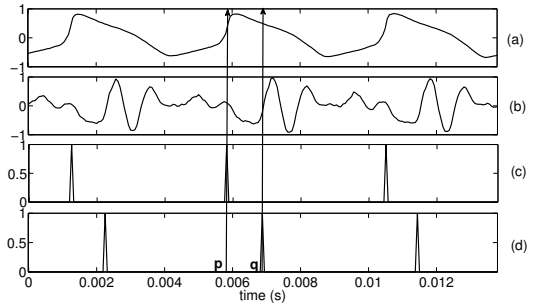

Figure 1: *Illustration of time-lag of speech signal with respect to EGG signal. (a) EGG signal, corresponding (b) speech signal, (c) Glottal Closure Instants (GCIs) determined from EGG signal and (d) GCIs determined from speech signal (p and q are GCIs determined from EGG and speech signals, respectively).*

"q", we can observe that, GCI of speech signal (shown in Figure 1(d)) is delayed in time with respect to GCI of EGG signal (shown in Figure 1(c)).

Accurate estimation of glottis-to-microphone delay helps in precise synchronization of EGG and speech signals. It is very important to correlate the signals like speech, EGG and stroboscopic to explore the vibrating pattern of the vocal folds in view of various speaking style. In medcial and speech processing applications, the synchronization of these signals helps to detect and catagorize the various types of vocal folds disorders [9, 10]. For example, one of the characteristic feature of Parkinson's disease is the deterioration of utterance and speech [11]. In [12], an analysis was conducted on EGG and speech signals recorded from patients having Parkinson's disease subjected to deep brain stimulation therapy. In this study, impact of deep brain stimulation therapy on utterance and speech of patients having Parkinson's disease was analyzed. In such medical applications, it is necessary to have precise synchronization of EGG and speech signals at the level of glottal cycle. Precise estimation of glottis-to-microphone delay can be used to determine the length of the vocal-tract. The vocal-tract length has been used to normalize the spectral features. Normalized spectral features are used to implement speaker independent applications such as automatic phone recognition and voice conversion systems [13] etc.

In literature, to synchronize EGG and speech signals, a fixed glottis-to-microphone delay has been assumed, and speech and EGG signals were synchronized according to that fixed timing delay. In [14], CMU-Arctic database was used to evaluate the accuracy of epoch extraction method with an assumption of 0.7 ms glottis-to-microphone delay. The glottis-to-microphone delay depends on two factors namely, the length of vocal-tract of the speaker and the distance between speaker's lips and microphone. Length of vocal-tract varies from speaker

http://dx.doi.org/10.21437/Interspeech.2017-1374

to speaker and natural movement of speaker's head results in variation of distance between lips to microphone. Hence, assumption of such a fixed glottis-to-microphone delay results in imprecise synchronization of speech and EGG signals.

In this work, a method has been proposed to estimate the glottis-to-microphone delay based on the glottal activity information present in speech and EGG signals. In this method, the delay between the speech signal and EGG signal is computed continuously frame by frame i.e., the delay is calculated for each and every frame of these two signals. The difficulties in deriving similar glottal activity information from EGG and speech signals are, (i) EGG and speech signals represent different types of physical phenomenon. EGG records the glottal activity by measuring the electrical impedance variation around speaker's neck, whereas, speech signal is the measure of speech pressure variation emanating from glottal activity (ii) The vocal-tract is often configured as a time varying bandpass filter with pass band at a frequency above the frequency of the glottal activity. Hence, vocal-tract may suppress glottal activity information present in the speech signal (iii) Acoustic signal can be easily degraded by ambient noise.

The organization of this paper is as follows: Section 2.1, discusses the proposed method to derive common glottal activity information present in EGG and speech signals. Section 3 introduces a method to estimate glottis-to-microphone delay. The performance of the proposed method is evaluated in Section 4. Finally, Section 5 summarize the contributions of this work.

2. Proposed Method

2.1. Glottis-to-microphone delay estimation

The main objective of the EGG signal is to analyze the glottal activity. Appropriateness of using EGG signal in observation and analysis of glottal activity has long been known from imaging studies [15]. In [16], Murty *et al.,* have shown that the zero frequency filtered (ZFF) speech signal can be used to characterize the glottal activity. Hence, we can conclude that ZFF-speech signals and EGG signals carry the information related to glottal activity. This claim can be verified from Figure 2. Figures 2(a) and 2(b) show the speech signal of voiced speech segment and its corresponding ZFF-speech signal. Figure 2(c) shows the EGG signal. ZFF-speech signal and EGG signal in Figures 2(b) and 2 (c) are observed to be similar, except for the amplitude envelope of signals and time-lag between signals due to glottis-to-microphone delay. Hence, glottis-to-microphone delay can be accurately estimated, if amplitude envelope variations between ZFF-speech signals and EGG signals are compensated. Signals shown in Figures 2(b) and 2(c) have some inherent amplitude envelope variations, but the variations with respect to time are observed to be same in both the signals. The signal variation with respect to time is mainly related to the phase of the signal. Therefore, ZFF-speech and EGG signals may have more similar phase information.

Real signals like ZFF-speech signals and EGG signals can be described in terms of its amplitude envelope and phase. Here, EGG signal $E(n)$ can be represented as

$$E(n) = A_e(n)exp(j\Phi_e(n)) \tag{1}$$

where, $A_e(n)$ and $\Phi_e(n)$ are amplitude envelope and phase of $E(n)$, respectively.

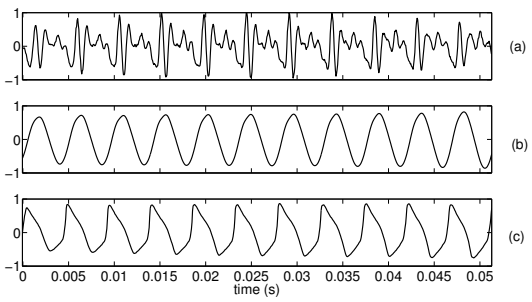

Figure 2: *Illustration of similar glottal activity information present in EGG and speech signals. (a) Acoustic signal, (b) its corresponding zero frequency filtered signal and (c) EGG signal.*

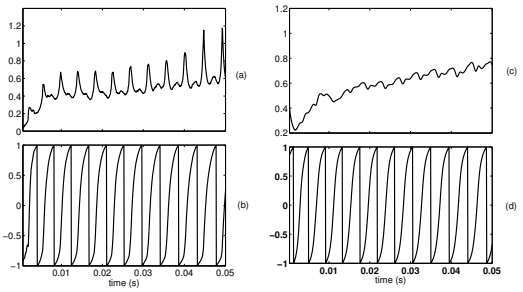

Figure 3: *(a) and (b) Amplitude envelope and phase of EGG signal shown in Fig. 2 (c). (c) and (d) Amplitude envelope and phase of zero frequency filtered speech signal shown in Fig. 2 (a).*

2.2. Extraction of amplitude envelope and phase of ZFF-speech and EGG signals

In this work, Hilbert transformation method is used for extracting amplitude envelope and phase of EGG and ZFF-speech signals.

For any real signal $S(n)$, its analytic signal $S_a(n)$, can be expressed as,

$$S_a(n) = S(n) + jS_h(n) \tag{2}$$

where, $S(n)$ and $S_h(n)$ are Hilbert transformation pair. Amplitude envelope $A(n)$ and phase $\Phi(n)$ can be computed using the following relations.

$$A(n) = \sqrt{S(n)^2 + S_h(n)^2} \tag{3}$$

$$\Phi(n) = tan^{-1}\frac{S_h(n)}{S(n)} \tag{4}$$

In this work, the phase range has been normalized between -1 to +1. Figures 3(a) and 3(b) show amplitude envelope and phase of EGG signal. Figures 3(c) and 3(d) show amplitude envelope and phase of ZFF-speech signal. From Figure 3, it can be concluded that phase of EGG and ZFF-speech signals are similar, except that the phase of ZFF-speech signal lags in time by glottis-to-microphone delay, with respect to phase of EGG signal. Hence, we can express phase of ZFF-speech signal $\Phi_a(n)$, as

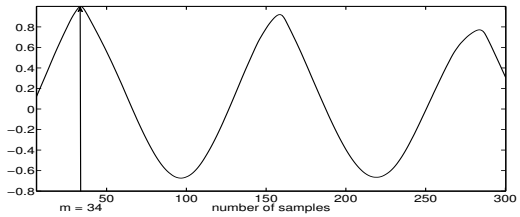

Figure 4: *Cross-correlation between phase of ZFF speech and EGG signals.*

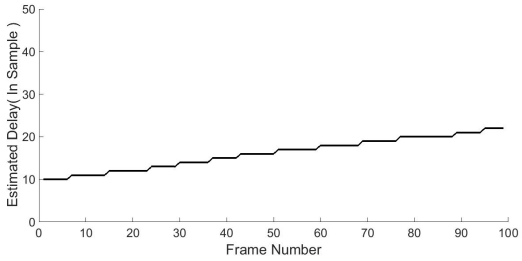

Figure 5: *Estimation of delay between phase of ZFF-speech and EGG signals with frame by frame analysis.*

$$\Phi_a(n) \approx \Phi_e(n-m) \tag{5}$$

where, m is number of samples by which speech signal lags EGG signal due to glottis-to-microphone delay. Hence, glottis-to-microphone delay can be measured by estimating number of samples with which phase of ZFF-speech signal ($\Phi_a(n)$) lags the phase of EGG signal ($\Phi_e(n)$).

3. Proposed method for glottis-to-microphone delay estimation

In the proposed method, phase of EGG and ZFF-speech signals are used to estimate glottis-to-microphone delay. From equation 5, it is observed that, estimation of glottis-to-microphone delay depends on accurate detection of number of samples by which phase of ZFF-speech signal lags the phase of EGG signal. In this work, lag between speech and EGG signals is estimated as time instant at which cross-correlation between $\Phi_e(n)$ and $\Phi_a(n)$ is maximum. Cross-correlation, $cr(n)$ between $\Phi_e(n)$ and $\Phi_a(n)$ can be computed using the following equation.

$$cr(n) = \sum_k \Phi_e(k)\Phi_a(n-k) \tag{6}$$

The effectiveness of the proposed method is demonstrated using two experiments. In the first experiment, the microphone is kept at a fixed distance of 55 cm from the lips and the electrodes of the Electroglottograph device are placed outside of the thyroid cartilage. The speaker was asked to utter the vowel /a/ continuously at least for 5 seconds with a comfortable loudness and intensity. The recorded signals are sampled at 16 kHz sampling rate with 16 bits per sample resolution. After the simultaneous recording of speech and corresponding EGG signals, the proposed method is applied on these signals. Figure 4 shows the cross-correlation between $\Phi_e(n)$ and $\Phi_a(n)$. EGG and speech signals used for generating the cross-correlation signal shown in Figure 4. From Figure 4 it is observed that maximum cross-correlation occurs at 34th sample of $cr(n)$. Therefore, a lag of 34 samples corresponds to glottis-to-microphone delay of 2.125 ms, which is equivalent to glottis-to-microphone distance of 70.76 cm (assuming speech waves propagate at 330 meter/second). For male speaker, the average vocal-tract length is 17 cm [17]. Hence, in this experiment, approximate glottis-to-microphone distance (sum of vocal-tract length and distance between lips and microphone) is 72 cm, which is close to the estimated distance of 70.76 cm using proposed method. In the second experiment, we recorded the speech and its corresponding EGG signal simultaneously by varying the distance between lips and microphone. In this experiment, we gradually increase the distance between lips and microphone from 5 cm to 30 cm. In this whole process, the subject utters the vowel /a/ at comfortable loudness and intensity. As the distance between lips and microphone gradually increases, the delay between speech and EGG signal is also observed to be increasing uniformly with respect to time. Figure 5 shows the sequence of estimated delay values (in terms of number of samples) at frame level between the phase of ZFF-speech and EGG signals,while microphone is shifting away from the lips. Here, the delay between the signals is estimated at frame level (frame length=100 ms and frame shift=50 ms)using cross correlation. From Figure 5 , it is observed that the estimated delay is varying from 10 samples to 22 samples over 100 frames. This indicate that at the beginning, the glottis to microphone distance is about 0.625 ms and at the end of 100 frames, it is about 1.375 ms. The two delays corresponds to physical distance of 20.625 cm and 45.375 cm respectively for initial and final positions of the microphone. Here, initial and final positions of the microphone corresponds to 5 cm and 30 cm, respectively away from the lips. The appropriate distance of microphone from the glottis at initial and final positions are 22 cm and 47 cm with the assumption of glottis to lips distance is 17 cm for a male speaker. It is observed that the estimated delays (i.e., 20.625 cm and 45.375 cm) are very close to the actual delays (22 cm and 47 cm). From the study, it can be observed that the proposed method can effectively estimate the delays between EGG and speech signals and it is also noted that the proposed method can provide the delay at frame level i.e., delay estimation can be carried out with respect to time whereas in the literature, the delay between EGG and speech signal for CMU-Arctic database is considered as fixed or constant delay for the whole sentence as well as same for all sentences. Therefore, the proposed method will be very useful for the precise synchronization between EGG and speech signals. The precise synchronization (at frame level) may be required for the analysis of source and tract information within a glottal cycle.

Following sequence of steps summarizes the estimation procedure of glottis-to-microphone delay:

1. From the speech signal $a(n)$, obtain ZFF-speech signal $Z_a(n)$ using zero frequency resonator [14].
2. From $Z_a(n)$, derive the phase of ZFF-speech signal $\Phi_a(n)$ using equations 2 and 4.
3. Derive the phase of EGG signal $\Phi_e(n)$ using equations 2 and 4.
4. Compute cross-correlation $cr(n)$ between $\Phi_e(n)$ and $\Phi_a(n)$ using equation 6.
5. The desired glottis-to-microphone delay is estimated as the number of samples by which speech signal lags by EGG signal which is given by, an instant where $cr(n)$ is maximum.

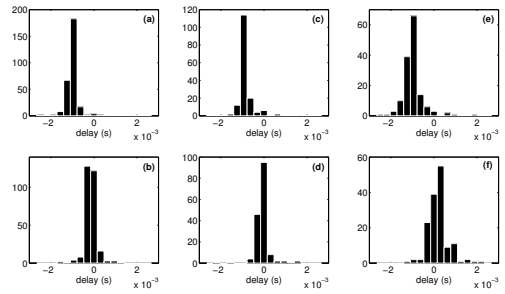

Figure 6: *Histogram of the delays between the GCIs detected from EGG signal and corresponding speech signal for speaker (a) M1 before synchronization, (b) M1 after synchronization (c) M2 before synchronization, (d) M2 after synchronization, (e) F1 before synchronization and (f) F1 after synchronization.*

4. Evaluation of the proposed method to synchronize EGG and speech signals

The proposed method for synchronization of EGG and speech signals is evaluated using CMU-Arctic database [4]. The CMU-Arctic database consists of simultaneous recording of EGG and speech signals spoken by one female (F1) and two male (M1 and M2) speakers. In literature, EGG and speech signals are synchronized by assuming a constant glottis-to-microphone delay. In [14], Murty *et al.,* synchronized EGG and speech signals of CMU-Arctic database by assuming a constant glottis-to-microphone delay of 0.7 msec. In this work, the speech and EGG signals of CMU-Arctic database are divided into frames of 100 msec with a frame shift of 50 msec and for each frame, the glottis-to-microphone delay has been estimated. The mean of the estimated glottis-to-microphone delays of the entire CMU-Arctic database is found to be 0.713 msec which is comparable to assumed glottis-to-microphone delay of 0.7 ms in [14]. Table 1 shows the mean and standard deviation of glottis-to-microphone delay before and after synchronization of all speech and EGG signals of speakers M1, M2 and F1. A total of 1114 sentences from each speaker were considered in this evaluation. In Table 1 columns 2 and 4 indicate the estimated average delay, columns 3 and 5 introduced the STD (Standard Deviations) before and after synchronization of speech and EGG signals. It is observed that the average delay after synchronization is not exactly zero, because we have considered 100 ms speech and EGG segments for estimating the delay as well as correcting the delay. The considered segment consists of 10 to 25 glottal cycles (pitch periods). During correction, we have incorporated the delay correction uniformly for all the glottal cycles present in 100 ms segment. Since, the GCIs detection methods (for speech and EGG) are different, the delay associated to individual glottal cycles may fluctuate. Hence, after performing the delay correction also there exists a delay between the GCIs of speech and EGG signals.

For further evaluation of the proposed synchronization method, the delay between the GCIs detected from speech and the corresponding EGG signals are used. Histogram and statistical measures of the delay before and after synchronization of EGG and speech signals are analysed to evaluate the proposed method. Positive peaks in DEGG signal are used to determine GCIs from the EGG signal. DYPSA method is used to determine GCIs from the speech signal. Signals are synchronized by estimating and compensating the glottis-to-microphone delay

Table 1: *Mean glottis-to-microphone delay estimated using proposed method, for speakers of CMU-Arctic database (Mean Delay – MD, Standard deviation – STD)*

Speaker	Before Synchronization		After Synchronization	
	MD (ms)	STD (ms)	MD (ms)	STD (ms)
M1	1.1	0.29	0.37	0.13
M2	0.99	0.32	0.27	0.12
F1	0.94	0.27	0.25	0.11

using proposed method. Figures 6(a) and 6(b) show histograms of delays, before and after synchronization of EGG and speech signals corresponding to a sentence spoken by a speaker M1. Histogram shows the delays corresponding to 288 GCIs. Average of delays before and after synchronization are -1.1 ms and -0.15 ms, respectively. This reduction in average delay from -1.1 ms to -0.15 ms is due to ability of the proposed method to synchronize speech and EGG signals. Standard deviation of delays before and after synchronization are 0.27 ms and 0.15 ms, respectively.

Figures 6 (c) and 6 (d) show histograms of delays (corresponding to 162 GCIs) before and after synchronization of EGG and speech signals for a sentence spoken by speaker M2, respectively. Figures 6 (e) and 6 (f) show histograms of delays (corresponding to 147 GCIs) before and after synchronization of EGG and speech signals for a sentence spoken by speaker F1, respectively.

5. Conclusion

In this work, a method to synchronize speech and EGG signals has been proposed. The proposed method is based on similar glottal activity information present in speech and EGG signals. The phase of ZFF-speech and EGG signals are used to derive similar glottal activity information from speech and EGG signals, respectively. Derived glottal activity information from speech and EGG signals has been used to synchronize speech and EGG signals. The method is computationally very simple and accurate. At present, the glottis-to-microphone delay is considered as constant through out the database. The proposed method will estimate this delay for every frame of 100 ms with a frame shift of 50 ms. It will provide flexible and accurate synchronization between speech and EGG signals. Further, the proposed method can be used to synchronize speech and EGG signals at fine levels such as for each glottal cycle.

6. Acknowledgements

The present work is carried out under the project entitled "Accurate analysis of vocal folds activity for speech and biomedical applications (AVS)" sponsored by Ministry of Human Resource Development (MHRD), Govt. of India.

7. References

[1] K. S. Rao, "Voice conversion by mapping the speaker-specific features using pitch synchronous approach," *Computer Speech and Language*, vol. 24, pp. 474–494, 2010.

[2] B. Yegnanarayana and S. Gangashetty, "Epoch-based analysis of speech signals," *Sadhana*, vol. 36, no. 5, pp. 651–697, 2011.

[3] M. E. Ayadia, M. S. Kamelb, and F. Karrayb, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572–587, 2010.

[4] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *ISCA Speech Synth. Workshop*, 2004, pp. 223–224.

[5] F. Plante, "A pitch extraction reference database," *Children*, vol. 8, no. 12, pp. 30–50, 1995.

[6] D. G. Childers and J. Larar, "Electroglottography for laryngeal function assessment and speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. 31, no. 12, pp. 807–817, 1984.

[7] D. Childers, D. Hicks, G. Moore, and Y. Alsaka, "A model for vocal fold vibratory motion, contact area, and the electroglottogram," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1309–1320, 1986.

[8] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[9] A. Yamauchi, H. Yokonishi, H. Imagawa, K.-I. Sakakibara, T. Nito, N. Tayama, and T. Yamasoba, "Quantification of vocal fold vibration in various laryngeal disorders using high-speed digital imaging," *Journal of Voice*, vol. 30, no. 2, pp. 205–214, 2016.

[10] H. I. Turkmen, M. E. Karsligil, and I. Kocak, "Classification of laryngeal disorders based on shape and vascular defects of vocal folds," *Computers in Biology and Medicine*, vol. 62, pp. 76–85, 2015.

[11] L. Naranjoa, C. J. Prez, Y. Campos-Roca, and J. Martna, "Addressing voice recording replications for parkinsons disease detection," *Expert Systems With Applications*, vol. 46, pp. 286–292, 2016.

[12] M. A. Mate, I. Cobeta, F. J. Jimnez-Jimnez, and R. Figueiras, "Digital voice analysis in patients with advanced parkinsons disease undergoing deep brain stimulation therapy," *Journal of Voice*, vol. 26, no. 4, pp. 496–501, 2012.

[13] C. L. Huang, C. Hori, H. Kashioka, and B. Ma, "Joint analysis of vocal tract length and temporal information for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7432–7436.

[14] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *EEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[15] D. Teaney and A. Fourcin, "The electrolaryngography as a clinical tool for the observation and analysis of vocal fold vibration," *The Voice Foundation*, 1980.

[16] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469–472, 2009.

[17] L. R. Rabiner and B. Juang, *Fundamentals of speech recognition*. Englewood Cliffs: PTR Prentice Hall, 1993.