INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Glottal Closure Instants Detection from EGG Signal by Classification Approach

Gurunath Reddy M, K. Sreenivasa Rao, Partha Pratim Das

Indian Institute of Technology, Kharagpur, India
mgurunathreddy@sit.iitkgp.ernet.in, ksrao@cse.iitkgp.ac.in, ppd@cse.iitkgp.ernet.in

Abstract

Electroglottography is a non-invasive technique to acquire the vocal folds activity across the larynx called EGG signal. The EGG is a clean signal free from vocal tract resonances, the parameters extracted from such a signal finds many applications in clinical and speech processing technology. In this paper, we propose a classification based approach to detect the significant parameter of the EGG such as glottal closure instant (GCI). We train deep convolutional neural networks (CNN) to predict if a frame of samples contain GCI location. Further, the GCI location within the frame is obtained by exploiting its unique manifestation from its first order derivative. We train several CNN models to determine the suitable input feature representation to efficiently detect the GCI location. Further, we train and evaluate the models on multiple speaker dataset to determine and eliminate any bias towards the speaker. We also show that the GCI identification rate can be improved significantly by the model trained with joint EGG and derivative (dEGG) signal. The deep models are trained with manually annotated GCI markers obtained from dEGG as reference. The objective evaluation measures confirmed that the proposed method is comparable and better than the traditional signal processing GCI detection methods.

Index Terms: GCI, EGG, CNN, Electroglottograph, dEGG

1. Introduction

The human speech production can be approximated by the source-filter model where the glottal source signal excites the vocal tract filter to produce the acoustic speech [1]. The vibration of vocal folds creates the source of excitation to the vocal tract system. The instant during which the vocal folds make maximum contact is the glottal closure instant (GCI) during which the vocal tract is excited with maximum extent [2]. The GCI detection from the speech is a hard problem due to the presence of vocal tract resonances, lip radiations, and other unwanted external noises. Hence, the natural choice for GCI detection is the simultaneously recorded EGG along with the speech which is the correlate of the glottal source signal. The accurately detected GCI can be used in many speech related applications: the inverse of the difference between consecutive GCIs forms the fundamental frequency of the speaker [3] [4] which can be used for speech synthesis [5], speech recognition [6], speaker verification [7], singing tonic identification and so on. The GCI markers can be used for various pitch synchronous analysis of speech [8], prosody modification [9]. Also, GCI parameters extracted from the EGG can be used for studying the pathological condition of vocal folds of a vocal disorder patient non-invasively [10] [11] [12] [13].

We can broadly (not exhaustively) classify the available GCI detection methods into three categories: 1) methods which directly work on EGG by amplitude thresholding and its derivative signal dEGG, and combination of both [14] [15] [16] [17] [18], 2) wavelet transform based approaches [19] [20] [21] [22] [23], and 3) empirical mode decomposition methods [24] [25] [26]. It should be noted that most of the aforementioned methods depend heavily on amplitude thresholding derived either from the pattern observed from the signal or empirically from the dataset used to design the detection technique. Also, most of the methods use handcrafted heuristics and rules to pick the GCI location which further degrades the detection accuracy. Although wavelet transform based methods are quite popular for GCI detection, they require careful task specific supervision to choose the right parameters such as decomposition mother wavelet, signal decomposition levels, the sampling frequency of the signal. Furthermore, multiple signal processing steps after decomposition further limit the generalizability of the method for entirely new data samples. On the other hand, empirical mode decomposition methods are limited by the empirical selection of the required number of intrinsic mode frequencies, time complexity of computing signal modes, combining empirically chosen signal modes whose characteristic features suitable for extracting GCI locations.

The aforementioned limitations of the existing methods motivated us to propose a parametric classification based approach to GCI detection. Our contributions in this paper include 1) we propose a supervised classification based GCI detection from EGG, which was not been explored before to the best of our knowledge. 2) We model GCI detection as a two class classification problem by training deep convolutional neural networks which automatically learns the features and the model parameters from the EGG data. The deep model predicts whether a frame of EGG samples contain GCI location. The target GCI labels for model training are obtained by manual annotation with dEGG as a reference signal. 3) We train multiple CNN models to determine the suitable input feature representation to efficiently detect the GCI locations. 4) Further, we investigate the dependency of the model towards the speaker/gender. 5) We show that the GCI identification rate can be improved significantly by the model trained with joint EGG and its derivative signal. The proposed GCI detection method is evaluated by accuracy and reliability measures and compared with several state-of-the-art GCI detection methods.

2. GCI Detection from EGG

2.1. Dataset

We have created ground truth GCI for training CNN model from CMU_ARCTIC dataset [27]. The CMU_ARCTIC dataset consists of simultaneously recorded speech and EGG for female (SLT) and male (BDL, JMK) speakers. The negative peaks in the differenced EGG (dEGG) are taken as a reference to mark GCI locations. All EGG signals are downsampled to 16 kHz

http://dx.doi.org/10.21437/Interspeech.2020-1189

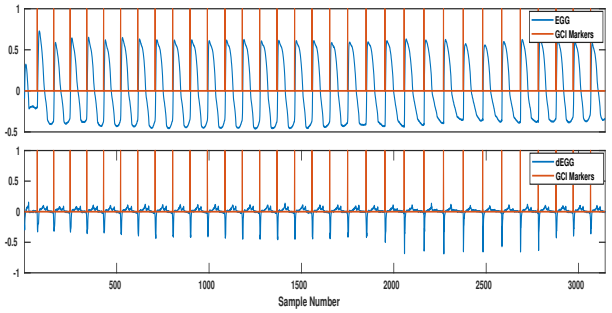

Figure 1: *Illustration of EGG, dEGG and GCI markers.*

prior to annotation. A sample data from CMU_ARCTIC dataset illustrating EGG and dEGG used to mark the GCI locations is shown in Fig1. From Fig. 1, we can observe that the negative peaks of the dEGG signal are used to mark the reference GCI locations. It should be noted that even though GCI manifests as a negative peak in the dEGG signal, it is not trivial to extract the GCI locations by thresholding due to varying strength of excitation, double peaks, abrupt vocal folds vibration and so on [28, 29, 30, 31].

2.2. Feature Representation

Most of the popular GCI detection methods depend on EGG or its derivative signal to obtain the GCI locations. In this work, we explore both EGG and dEGG to determine the importance of both representations to reliably identify the GCI locations. The slow moving vocal tract structure corrupts the EGG signal recorded from the EGG device by modulating low frequency baseline oscillations. Hence, initially, we remove the baseline oscillations by high pass filtering the EGG signal with 20 Hz cutoff frequency [32]. The high pass filter is a minimum order finite impulse response (FIR) implementation of *Matlab*® with delay compensation. The input feature vector to the CNN model is the non-overlapping frames of 16 samples. Each frame is labeled with binary 1 or 0 which represents the presence or absence of GCI location within the frame as a target label for the CNN model. We call the frame containing GCI location as GCI frame and the frame without GCI as non-GCI frame. The specific reasoning for choosing non-overlapping frames, feature vector selection around GCI location, and the decision for selecting 16 samples (1 ms) feature vector frames for classification are discussed elaborately in the following subsections.

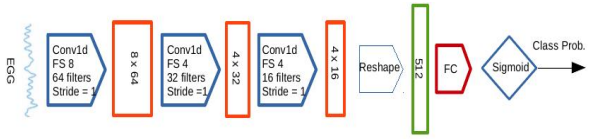

Figure 2: *Proposed CNN classification based GCI detection model (FS = filter size, FC = fully connected).*

2.3. GCI Classification Model

The proposed deep CNN GCI classification model is shown in Fig. 2. The CNN model consists of three CNN layers. The input to the CNN model is the frames of non-overlapping samples of EGG or dEGG or combination of both. Each convolution layer is followed by batch normalization. The d-dimensional deep feature vector from the last convolution layer is connected densely to the sigmoid activation function to predict the class probability for each frame. The GCI is manifested as a negative peak in the dEGG and also as a sudden change of slope in EGG signal hence, we drop max pooling layers in the proposed model to avoid the model bias towards the maximum amplitude of EGG and dEGG near GCI. In order to avoid the model being over-fitted to training data, a dropout layer with a dropout probability of 0.25 is added after each batch normalization layer. The network is trained to minimize the binary cross entropy loss between the target label y and the predicted label $\hat{y}$

$$L(y, \hat{y}) = \sum_{i=1}^{2} (-y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)) \quad (1)$$

The loss function is optimized by ADAM optimizer with a learning rate of 0.0001. The model is trained for 1000 epochs with batch size of 2048 frames randomly selected from the training set for each iteration. The index of the minimum value of the negative derivative of the samples of the predicted frame with its two neighbors of EGG signal is hypothesized as GCI location.

We evaluate the proposed GCI detection using Identification rate (IDR): measures the percentage of GCI detected exactly one per glottal cycle. False alarm rate (FAR): the percentage of glottal cycles for which more than one GCI is detected. Miss rate (MR): the percentage of glottal cycles for which no GCI is detected. Identification accuracy (IDA): the standard deviation of the timing error between the detected and the corresponding reference GCI [23].

2.4. Experiments

In this section, we describe several experiments which are conducted to investigate the significance of feature vector selection around the GCI location as GCI frame, model bias to the speaker/gender of the data used for training the model, and the importance of combining features to improve the GCI classification accuracy. All models are trained with SLT dataset with train, test, validation split of 80%, 10%, and 10% respectively unless otherwise explicitly mentioned.

2.4.1. Non-overlapping frames for classification

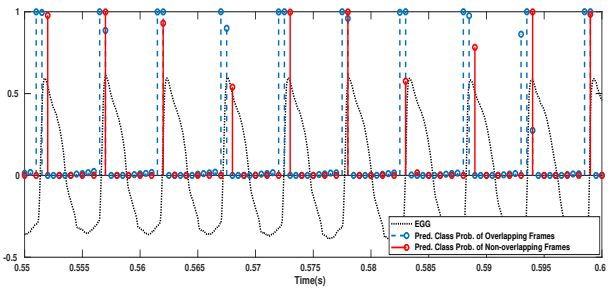

Figure 3: *Posterior class prediction probabilities for overlapping and non-overlapping EGG frames.*

The proposed CNN model discussed in 2.3 is trained with frames of 16 samples dimension to predict the class (GCI or non-GCI frame) probability. We trained CNN models to infer whether overlapping or non-overlapping frames are suitable for GCI classification. We trained independent models for

overlapping (16 samples frame width with 8 samples overlap) and non-overlapping (16 samples non-overlapping) frames. We found that multiple frames around the GCI location are predicted with high class probabilities for overlapping frames. This is due to overlapping frames around the GCI location shares relatively common information about GCI, results in multiple frames being predicted as GCI frames which requires additional post-processing techniques to reliably detected GCI location within the predicted frames around the GCI, which not only increases computational complexity but also requires additional handcrafted rules. On the other hand, the model trained with non-overlapping frames confidently assigns high probability scores only for GCI frames and very low or negligible probability scores for non-GCI frames. The predicted class probabilities of the models trained on overlapping and non-overlapping frames of EGG is shown in Fig. 3. From Fig 3, we can observe that multiple frames around GCI location are assigned with high posterior class probabilities by the model trained with overlapping frames whereas only GCI frames predicted with high class probability by the model trained with non-overlapping frames.

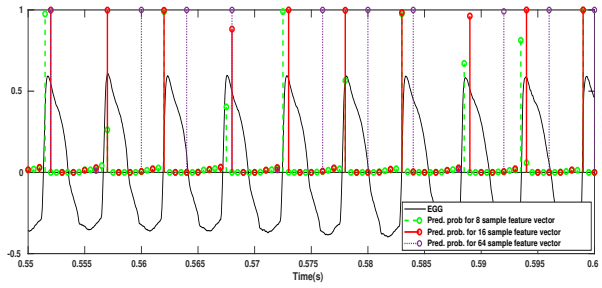

Figure 4: *Class prediction probabilities for varying feature vector dimensions.*

2.4.2. Frame size

We trained several models to confirm the significance of frame size/length for the unambiguous classification of the input feature vector. We trained models with non-overlapping frame lengths of 4, 8, 16, 32, 64, and 128 samples. We found that most of the times model shows very low confidence in predicting the GCI frame with high probability for low dimensional frames of 4 and 8 samples. This is due to the low dimensional feature vector captures incomplete GCI related features or it fails to accommodate the characteristic features of GCI within the low dimensional vector results in GCI frames predicted with weak probabilities. On the other hand, models trained with high dimensional frames: 32, 64, and 128 predicts frames with very high confidence but results in many false alarms i.e., most of the non-GCI frames are being classified as GCI frames. We found that 16 samples frame size is the optimal frame to unambiguously predict the GCI frame with high confidence. In summary, the dimension of the feature vector should be lesser than the minimum pitch period of the glottal source signal but large enough to capture the GCI information around the GCI location. The 16 samples non-overlapping frame size corresponds to 1 millisecond (ms) or 1000Hz at 16KHz sampling rate i.e., we can classify GCI frames unambiguously with vocal frequencies up to 1000Hz above which all non-GCI frames will also get classified as GCI frames since each frame covers more than one pitch period. Also, the fundamental frequency of the male or female speech rarely crosses 1000Hz hence, the choice of 1ms

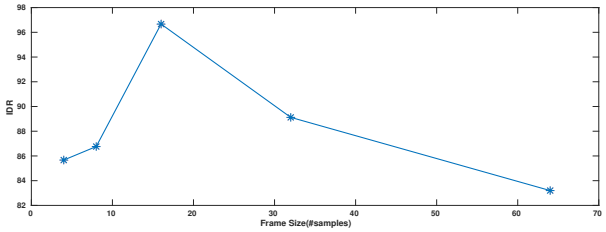

Figure 5: *Identification rate for varying frame sizes.*

frame size is sufficient to reliably classify the input frames as GCI or non-GCI frames. The class prediction probabilities for 8, 16, and 64 sample frame size feature vectors are shown in Fig. 4. From Fig. 4, we can observe that low dimensional feature vectors exhibit low prediction probability for GCI frames. High dimensional feature vectors result in mostly false alarms whereas the right feature vector predicts the GCI frames with high confidence. The GCI identification rates for various frame sizes is shown in Fig. 5. From Fig. 5, we can observe that the 16 samples (1ms) frame size gives the highest identification rate compared to other frame length choices.

2.4.3. GCI frame selection

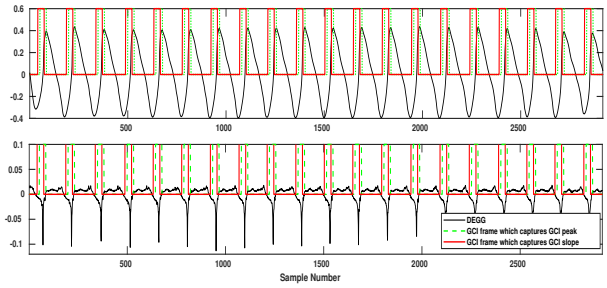

Figure 6: *GCI frame selection around GCI location.*

Through experiments, we found that not-all GCI frames with 1ms frame size around the GCI location are suitable for GCI classification. We found that the feature vector around the GCI location which captures slope of the GCI location, in other words, the feature vector which contains the samples of EGG between the negative peak and positive peak as shown in upper plot of Fig. 6 in solid rectangular boxes and the first half of the dEGG as shown in the lower plot of Fig. 6 in the solid rectangular box is suitable for reliably predict the GCI frames. Any other feature vector which captures the peaks/amplitudes of the GCI locations either in EGG or dEGG (an example is shown in Fig. 6 as dashed rectangular plots for both EGG and dEGG) biases the model to the GCI amplitudes resulting in weakly predicted probabilities for GCI frames of low voiced and transition regions where the GCI strength is significantly low. The GCI identification rate for various frame sizes and GCI frame feature vectors around the GCI location is shown in Table 1. The GCI frame around the GCI location is selected by shifting the ground truth GCI from its original location. A 16 sample GCI shift and 8 samples frame size in Table 1 indicates that the ground truth GCI is shifted by 16 samples and the 8 samples vector GCI frame is extracted by slicing the samples left to the new GCI location of EGG. From Table 1, we can observe that the 16 sample frame size around the GCI location which captures the slope of the GCI is the significant feature vector for

identifying GCI location with significantly high identification rate.

Table 1: *GCI Identification rates for various GCI frames around the GCI location.*

GCI shift\frame size	**8**	**16**	**32**	**64**
0	86.76	96.67	85.12	68.43
16	82.16	86.65	83.80	62.87
32	78.76	65.89	82.44	50.16

2.4.4. Speaker dependency of the model

In this subsection, we train and evaluate the GCI classification CNN models to find the speaker/gender dependency of the models. The CMU_ARCTIC dataset consists of EGG signals of female (SLT) and Male (BDL and JMK) speakers. To determine the speaker's dependency on the models, we train separate models for each speaker and test on the remaining speakers. The average GCI identification rate for each speaker is shown in Table 2. From Table 2, we can observe that the models trained on one speaker and tested on another speaker irrespective of gender shows variations in identification rate. The model trained on female speaker and tested on male speakers do not show much deviation from the identification rate of the test set of the same female speaker. The model trained a male speaker tested on the other male speaker showed negligible performance deviation whereas the same model tested on female speaker showed significant performance deviation. From the evaluation results shown in Table 2, we can infer that models trained with the cross gender data generalize to unknown speakers of any gender.

Table 2: *GCI identification rates for evaluating speaker dependency.*

Dataset	**Female(SLT)**	**Male(BDL)**	**Male(JMK)**
Female(SLT)	96.67	94.03	96.44
Male(BDL)	90.47	95.99	95.76
Male(JMK)	89.76	91.50	95.79

2.4.5. Combining features

In the previous subsection, we inferred that the models trained with cross-gender data improve the GCI identification rate. In this subsection, we discover the significance of combining EGG and dEGG (whose negative peaks indicate the GCI locations). We train a classification model by concatenating 16 dimensional feature vectors of EGG and dEGG with GCI frames extracted around GCI as discussed in 2.4.3 with the combined cross-gender datasets SLT and BDL. Fig. 7 shows the GCI identification rate for models trained with EGG, dEGG, and combination of both features. From Fig. 7, we can observe that the model trained with the combined EGG and dEGG (EGG-dEGG) features outperforms the models trained with individual features. The improved identification rate can be attributed to dEGG features that act as complementary to EGG in reliably classifying GCI frames.

2.4.6. Evaluation

We compare the proposed method with popular GCI detection techniques: zero frequency filtering (ZFF) [33], SIGMA [23],

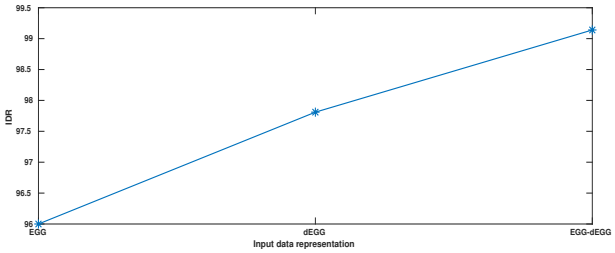

Figure 7: *Identification rate for models trained with EGG, dEGG and combined EGG-dEGG features.*

TXGEN [34] and HQTx [34]. Table 3 shows the comparison of proposed method with other techniques. The CNN classification model used for comparing with other methods is the one trained with combined EGG and dEGG features discussed in 2.4.5 with SLT, BDL as the training set and JMK as the test set. From Table 3, we can observe that the identification rate of the proposed method is comparable and better than the other popular GCI detection techniques. It is also observed that the GCI miss rate of the proposed method is significantly low compared to other methods. Also, we can observe that the identification accuracy of the proposed method is better than the most recent methods such as ZFF and SIGMA.

Table 3: *Comparison of proposed method with other GCI detection techniques.*

Method	**IDR**	**MR**	**FAR**	**IDA**(ms)
ZFF	96.12	3.76	0.12	0.90
SIGMA	97.05	2.77	0.17	0.50
HQTx	96.81	2.10	1.09	0.04
TXGEN	94.33	5.20	0.47	0.12
Proposed	97.94	1.26	2.07	0.30

3. Summary and Conclusions

In this paper, we proposed a classification based GCI detection from EGG signals. We trained multiple CNN models to determine the suitable input feature representation to efficiently detect the GCI locations. Further, we trained and evaluated the GCI detection models on multiple speaker datasets to determine and eliminate any bias towards the speaker/gender. We showed that the GCI identification rate can be improved significantly by combining EGG and its derivative signal. The proposed method is compared with several state-of-the-art GCI detection methods. As a future work, the proposed method can be extended to glottal opening instant (GOI) detection, voice/non-voice classification of EGG frames, and also for other significant parameter detection from EGG signal.

4. References

[1] R. Lawrence, *Fundamentals of speech recognition*. Pearson Education India, 2008.

[2] D. Childers, D. Hicks, G. Moore, L. Eskenazi, and A. Lalwani, "Electroglottography and vocal fold physiology," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 245–254, 1990.

[3] G. Reddy and K. S. Rao, "Enhanced harmonic content and vocal note based predominant melody extraction from vocal polyphonic music signals." in *INTERSPEECH*, 2016, pp. 3309–3313.

[4] M. G. Reddy and K. S. Rao, "Predominant melody extraction from vocal polyphonic music signal by combined spectro-temporal method," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 455–459.

[5] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," 2007.

[6] S. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

[7] Z. Ćirović, M. Milosavljević, and Z. Banjac, "Multimodal speaker verification based on electroglottograph signal and glottal activity detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, p. 65, 2010.

[8] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[9] T. Ewender and B. Pfister, "Accurate pitch marking for prosodic modification of speech segments," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[10] H. Banjara, V. Mungutwar, D. Singh, and A. Gupta, "Objective and subjective evaluation of larynx in smokers and nonsmokers: a comparative study," *Indian journal of otolaryngology and head & neck surgery*, vol. 66, no. 1, pp. 99–109, 2014.

[11] K. Hosokawa, M. Ogawa, M. Hashimoto, and H. Inohara, "Statistical analysis of the reliability of acoustic and electroglottographic perturbation parameters for the detection of vocal roughness," *Journal of Voice*, vol. 28, no. 2, pp. 263–e9, 2014.

[12] B. Yamout, Z. Al-Zaghal, I. El-Dahouk, S. Farhat, A. Sibai, and A.-L. H. Hamdan, "Mean contact quotient using electroglottography in patients with multiple sclerosis," *Journal of Voice*, vol. 27, no. 4, pp. 506–511, 2013.

[13] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted speech," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3050–3061, 2013.

[14] S. N. Awan and J. A. Awan, "The effect of gender on measures of electroglottographic contact quotient," *Journal of Voice*, vol. 27, no. 4, pp. 433–440, 2013.

[15] D. G. Childers and A. K. Krishnamurthy, "A critical review of electroglottography." *Critical reviews in biomedical engineering*, vol. 12, no. 2, pp. 131–161, 1985.

[16] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1321–1332, 2004.

[17] M. Rothenberg and J. J. Mahshie, "Monitoring vocal fold abduction through vocal fold contact area," *Journal of Speech, Language, and Hearing Research*, vol. 31, no. 3, pp. 338–351, 1988.

[18] D. M. Howard, "Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers," *Journal of Voice*, vol. 9, no. 2, pp. 163–172, 1995.

[19] A. Bouzid and N. Ellouze, "Local regularity analysis at glottal opening and closure instants in electroglottogram signal using wavelet transform modulus maxima," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[20] ——, "Multiscale product of electroglottogram signal for glottal closure and opening instant detection," in *Computational Engineering in Systems Applications, IMACS Multiconference on*, vol. 1. IEEE, 2006, pp. 106–109.

[21] A. Bouzid and N. Elouze, "Electroglottographic measures based on gci and goi detection using multiscale product," *International Journal of Computers Communications & Control*, vol. 3, no. 1, pp. 21–32, 2008.

[22] A. Bouzid and N. Ellouze, "Voice source parameter measurement based on multi-scale analysis of electroglottographic signal," *Speech Communication*, vol. 51, no. 9, pp. 782–792, 2009.

[23] M. R. Thomas and P. A. Naylor, "The sigma algorithm: A glottal activity detector for electroglottographic signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1557–1566, 2009.

[24] R. Sharma, K. Ramesh, and S. Prasanna, "Analysis of electroglottograph signal using ensemble empirical mode decomposition," in *India Conference (INDICON), 2014 Annual IEEE*. IEEE, 2014, pp. 1–6.

[25] P. S. Deshpande and M. S. Manikandan, "Effective glottal instant detection and electroglottographic parameter extraction for automated voice pathology assessment," *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, pp. 398–408, 2018.

[26] G. J. Lal, E. Gopalakrishnan, and D. Govind, "Accurate estimation of glottal closure instants and glottal opening instants from electroglottographic signal using variational mode decomposition," *Circuits, Systems, and Signal Processing*, vol. 37, no. 2, pp. 810–830, 2018.

[27] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[28] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A quantitative comparison of glottal closure instant estimation algorithmson a large variety of singing sounds," 2013.

[29] T. Mandal, K. S. Rao, and S. K. Gupta, "Classification of disorders in vocal folds using electroglottographic signal." in *Interspeech*, 2018, pp. 3002–3006.

[30] M. Reddy, T. Mandal, and K. S. Rao, "Glottal closure instants detection from pathological acoustic speech signal using deep learning," in *Proc. Mach. Learn. Health Workshop*, 2018.

[31] G. Reddy, K. S. Rao, and P. P. Das, "Glottal closure instants detection from speech signal by deep features extracted from raw speech and linear prediction residual." in *INTERSPEECH*, 2019, pp. 156–160.

[32] C. T. Herbst and J. C. Dunn, "Fundamental frequency estimation of low-quality electroglottographic signals," *Journal of Voice*, 2018.

[33] V. K. Mittal, B. Yegnanarayana, and P. Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1932–1941, 2014.

[34] M. Huckvale, "Speech filing system: Tools for speech," *University College London, Tech. Rep*, 2004.