INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Glottal Closure Instants Detection from Speech Signal by Deep Features Extracted from Raw Speech and Linear Prediction Residual

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

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

Abstract

Glottal closure instants (GCI) also called as instants of significant excitation occur during abrupt closure of vocal folds is a well-studied problem for its many potential applications in speech processing. Speech signal or its transformed linear prediction residual (LPR) is the most popular signal representations for GCI detection. In this paper, we propose a supervised classification based GCI detection method, in which, we train multiple convolution neural networks to determine the suitable feature representation for efficient GCI detection. Also, we show that the combined model trained with joint acoustic-residual deep features and the model trained with low pass filtered speech significantly increases the detection accuracy. We have manually annotated the speech signal for ground truth GCI using electroglottograph (EGG) as a reference signal. The evaluation results showed that the proposed model trained with very small and less diverse data performs significantly better than the traditional signal processing and most recent data-driven approaches.

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

1. Introduction

The human speech production can be represented as source (glottal source signal)/filter (vocal tract) model [1], where a sequence of impulses (ideally) due to vocal fold vibrations excites the system to produce the voiced speech. The instants of impulse excitation to the system are termed as GCI because during which the vibrating vocal folds close momentarily, results in impulse-like excitation to vocal tract system. The GCI is also called as epochs [2] or instants of significant excitation due to the high signal-to-noise ratio of the signal during this event. Although GCI can be extracted directly from EGG signal recorded from a dedicated Electroglottograph device [3], it is not always feasible or not so easy to record EGG as easily as speech, which we can record directly from any general purpose microphones attached to handheld devices such as smartphones. Also, most of the speech available at various sources do not contain simultaneously recorded EGG hence, there is a great need for extracting GCI directly from speech [4].

The accurately extracted GCI finds applications in many speech related tasks [5] such as fundamental frequency extraction from vocal music [6, 7], prosody modification of speech [8, 9], neutral to target emotion conversion [10, 11], speech synthesis [12], speech segmentation [13], speech enhancement and deverberation [14], voice conversion [15], glottal flow estimation [16], speaker recognition [17], voice source modeling [18], speech pathology [19].

We can broadly (not exhaustively) classify the available GCI detection methods into i) classical signal processing [20, 21, 22, 23, 2, 24, 25, 26] and ii) most recent classification based data driven [4, 27, 28] approaches. Most of the popular signal processing GCI detection methods relies on designing signal processing pipelines to obtain the exemplary signal which emphasizes the locations of GCIs in the speech signal [28]. Further, GCIs from the exemplary signal is obtained from hand-crafted heuristics. Two approaches are popular for exemplary signal extraction: a) source/filter modeling to extract the linear prediction residual whose peaks corresponds to the candidate epochs [20, 21, 22, 23]. b) Other methods which rely on the properties of excitation signal such as impulse nature [2, 24, 25, 26] to obtain the exemplary signal. All aforementioned methods work either on LPR or on speech signal but not on both to detect the GCI locations. Also, the hand-crafted heuristics are specific to extracted exemplary signal, requires manual tuning of many parameters. On the other hand, data-driven approaches automatically learn the required parameters from the data to predict the GCI locations [4, 27, 28] from the speech signal.

Existing signal processing and data-driven methods hinge either on LPR or on the raw speech signal to obtain the GCI locations by relying on the amplitude of the discontinuity of the signal near GCI location which results in missed GCI locations near low excitation regions. Also, existing methods use overlapping frames to detect the GCI locations [28] which requires additional post-processing methods to obtain the genuine GCI location from the multiple frames flagged as GCI frames near the actual GCI location. In this paper, we show an approach to improve the GCI detection accuracy even in the weak excitation regions such as transitions and weakly voiced regions by selecting suitable feature representation around the GCI location which is independent of the amplitude of GCI location. In this work, we treat GCI detection as two class classification problem [4] where each frame of the signal is classified as a GCI or non-GCI frame. We also show that the non-overlapping frames alleviate the need for additional post-processing method to obtain genuine GCI locations. Specifically, we explore the significance of various input representations to significantly improve the performance of GCI detection. We train several convolutional neural network (CNN) models one for each input feature representation to determine their significance of classifying a frame into GCI or non-GCI. In order to reduce the miss rates and false alarms, we train joint acoustic-residual models, combine their posterior probabilities in maximum likelihood sense to significantly improved the GCI identification rate.

2. Proposed GCI Detection Method

The proposed parallel/multi-column CNN classification based GCI detection is shown in Fig. 1.

http://dx.doi.org/10.21437/Interspeech.2019-1981

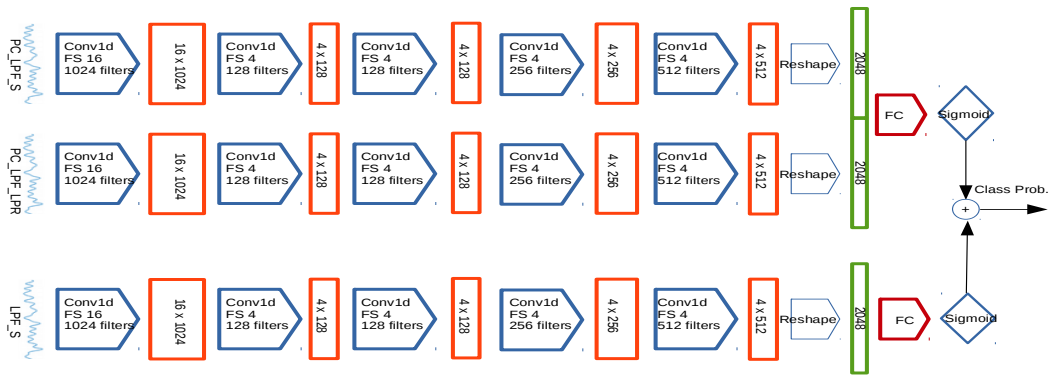

Figure 1: *Proposed parallel/multi-column CNN classification based GCI detection model (FC = fully connected, FS = filter size).*

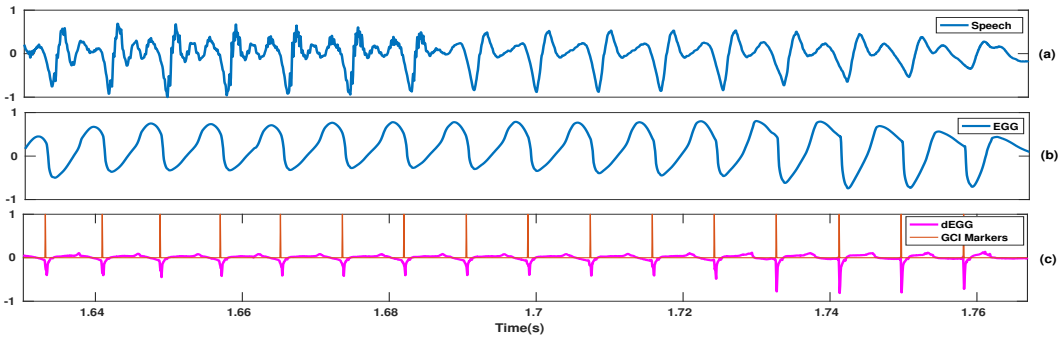

Figure 2: *Illustration of GCI ground truth creation.*

2.1. Dataset

The CMU_ARCTIC dataset [29] consists of the simultaneously recorded EGG and speech from female (SLT) and male (BDL, JMK, and KED) speakers are used as training and testing data. The differenced EGG (dEGG) is taken as a reference signal after compensating delay between EGG and speech signal to mark GCI locations. All speech signals are downsampled to 16KHz, switched to negative polarity prior to annotation. An example illustrating the speech, EGG, and dEGG used to mark the GCI locations is shown in Fig. 2. From Fig. 2(c), we can observe that negative peaks in the dEGG signal are used as a reference to assign GCI markers for the corresponding speech signal.

2.2. Feature Representation

To leverage the advantages of both raw speech and LPR, we train CNN models on the following input representations: 1) Low pass filtered speech (**LPF_S**). The signal is low pass filtered since high frequencies do not contribute to the low-frequency glottal signal. 2) The lowpass filtered LPR (**LPF_LPR**) with LP order 12. LPR is a correlate of glottal source signal obtained after removing vocal tract resonances [30] contains high-frequency noise hence, it is low pass filtered. 3) The positive clipped low-pass filtered LPR (**PC_LPF_LPR**) and 4) The positive clipped low-pass filtered speech (**PC_LPF_S**). Signals are positive clipped since all speech signals are switched to negative polarity and the GCI information is mostly predominant at the negative portion of the signal. The low-pass filter used is a sixth order zero phase Butterworth filter with 1 KHz cut-off frequency. The training data is created by non-overlapping speech frames with 16 samples (1msec) at 16 KHz sampling rate. Each frame is labeled with magnitude 0 or 1 which represents the presence or absence of the GCI location within that frame. The 16 samples around the glottal closure (we call it as GCI frame) is used as a feature for GCI detection shown in Fig. 3 with box plot on respective input signal representations. It should be noted that the non-overlapping GCI frame shown in Fig. 3 captures the overall shape, slope, and amplitude features. Reduces multiple frames being assigned with very high probabilities which requires post-processing to eliminate the spurious GCI frames. In an unreported experiment, we found that the overlapping frames (which captures the whole negative GCI peak lobe) result in multiple frames being assigned with very high-class probabilities. Also, the classifier predictions bias towards the amplitude of the negative peaks in the signal results in classifying frames with all secondary (spurious) peaks as GCI frames and also results in assigning weak or very low probabilities in the low voiced and transition regions of the speech signal.

2.3. Classification Model

The proposed multi-column/parallel deep CNN GCI classification model is shown in Fig. 1. It consists of three CNN networks. Each network is trained with different input representation discussed in subsection 2.2 with input dimension of 16 samples around the GCI location as a GCI frame shown in Fig. 3. Each CNN network consists of five convolutional layers. 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 output probability for each frame. In our model,

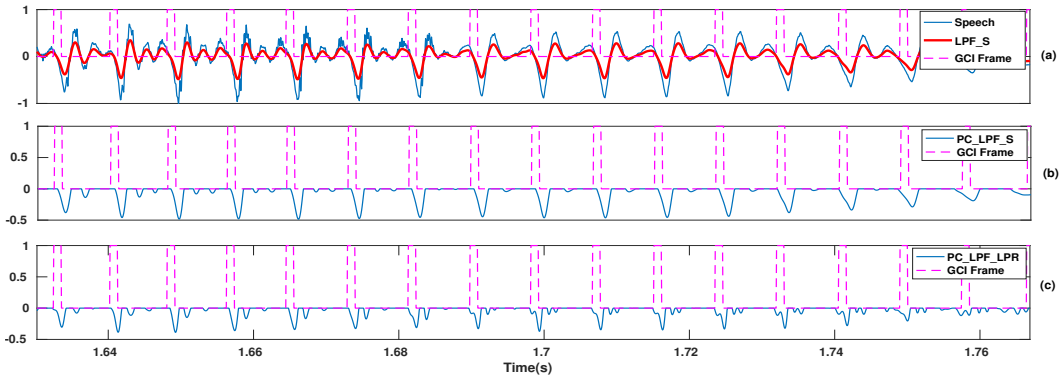

Figure 3: *Feature Representation.*

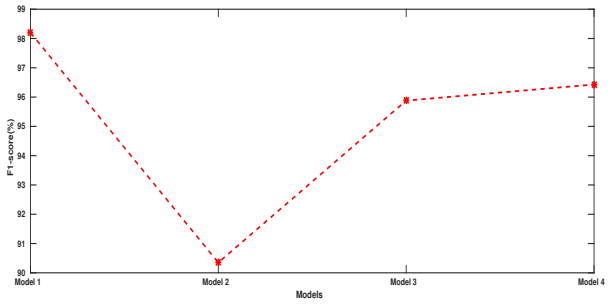

Figure 4: *F1-scores of Model 1, Model 2, Model 3, Model 4.*

we dropped max-pooling layers because we want to capture the variations in the GCI region which are due to quasi-stationary nature of vocal folds vibration (also, due to low dimension input feature, models trained without max-pooling layers gave better results than one with max-pooling layers). 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 learning rate 0.0001. The model is trained for 30 epochs with the batch size 2048 randomly selected from the training set.

2.4. Experiments

Initially, we have trained a single column CNN model for each feature representation separately discussed in subsection 2.2 to evaluate the significance of input representation in classifying the GCI frame efficiently. In an unreported experiment, we trained the models on male speaker and test on female speaker and vice-versa resulted in poor results. Hence, we trained the CNN models with a mixture of JMK (male) and SLT (female) datasets. The models are tested on BDL (male) dataset. We denote the models trained with inputs **LPF_S**, **LPF_LPR**, **PC_LPF_S** and **PC_LPF_LPR** as **Model 1**, **Model 2**, **Model 3** and **Model 4** respectively. Fig. 4 shows the F1-score (F1-score is the harmonic average of precision and recall, higher the score better the model performance) for each model. From Fig. 4, we can observe that the models trained with **LPF_S**, **PC_LPF_S** and **PC_LPF_LPR** i.e., **Model 1**, **Model 3** and **Model 4** achieves high F1-score than model trained with **LPF_LPR** i.e., **Model 2**. Further investigation into predicted probabilities of **Model 1**, **Model 3** and **Model 4** revealed that **Model 3** predicts strong GCI frames i.e., frames with strong impulsive GCI with very high probabilities, also, assigns little high probability to the non-GCI frames where the secondary excitations or the peaks whose amplitudes are comparable to the GCI peaks (shown in Fig. 5, marked with ellipses). Note that using **Model 3** alone for GCI classification requires careful thresholding of predicted probability scores to reduce the false alarms. Where as **Model 4** assigns very low probability scores in transition and low/weak voiced regions (shown in Fig. 5, marked with boxes) which results in high miss rates. It should be also noted that **Model 4** assigns very low probability scores for frames with spurious or secondary excitations which significantly reduces miss rates. From the predicted posterior class probabilities of **Model 3** and **Model 4**, we can conclude that **Model 3** trained with **PC_LPF_S** significantly reduces miss rates in the regions of weakly voiced and transitions regions where the excitation strength is very low. Where as **Model 4** considerably reduces false alarms due to secondary or spurious peaks. Fig. 5 shows the predicted posterior class probabilities for **Model 3** and **Model 4**. From Fig. 5, we can observe that **Model 3** assigns little high probabilities in the secondary excitation regions marked with ellipses, **Model 4** assigns very low probabilities marked with boxes in the low excitation regions. We train a joint acoustic-residual model to reap the benefits of both the models shown in Fig. 1. In this joint model, we concatenate d-dimensional feature vector from the last convolution layer of **Model 3** and **Model 4** to train densely connected sigmoid activation function to predict the class probabilities. Since the model (**Model 1**) trained with features from low pass filtered speech signal **LPF_S** also gave good results, we combine the posterior probabilities of joint acoustic-residual model and model trained with **LPF_S** i.e, **Model 1** in maximum likelihood sense to predict the final class probability as

$$P = \prod_{i=1}^{2} (p(i) + \epsilon) \quad (2)$$

where $p(1)$ and $p(2)$ are the posterior probabilities of joint acoustic-residual model and **Model 1**, ϵ is a small value to prevent numerical underflow. The frames which achieves class probability greater than or equal to 0.1 are classified as GCI frames. The location of maximum negative peak in the classified GCI frame is considered as glottal closure instant.

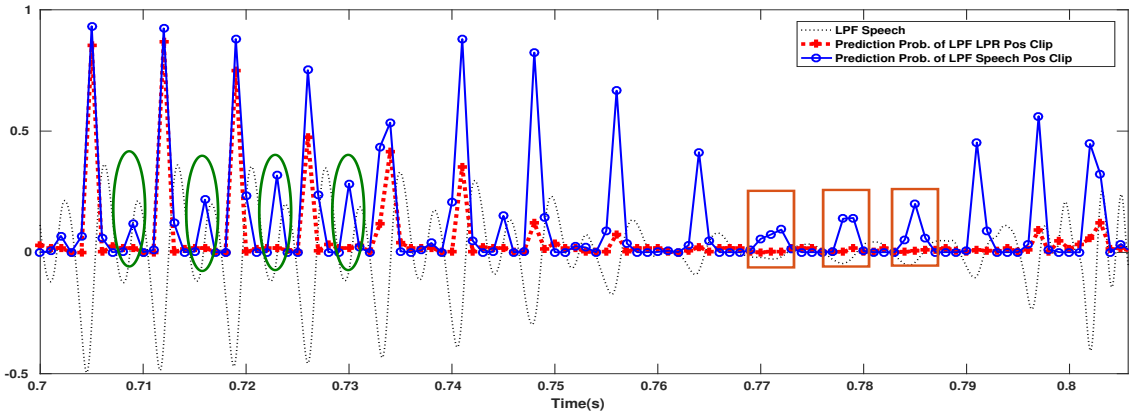

Figure 5: *Posterior probabilities of Model 3 and Model 4 (Model 3 = Predicted Prob. of LPF Speech Pos Clip, Model 4 = LPF LPR Pos Clip).*

Table 1: *Comparison of proposed method on various datasets and other methods.*

Dataset	Method	IDR(%)	MR(%)	FAR(%)	IDA(ms)
BDL	**Proposed**	96.42	1.42	2.16	0.21
	VC	93.86	2.26	3.89	0.45
	ERT-P3	91.96	2.98	5.06	0.41
	SEDREAMS	91.82	3.02	5.16	0.44
	MMF	89.49	4.53	5.98	0.57
	DYPSA	89.95	4.32	6.73	0.56
KED	**Proposed**	97.32	1.55	1.13	0.32
	VC	95.85	1.42	2.73	0.25
	ERT-P3	91.88	2.96	5.18	0.27
	SEDREAMS	92.31	6.03	1.66	0.29
	MMF	90.24	7.04	2.72	0.37
	DYPSA	90.29	7.05	2.66	0.31

3. Evaluation and Results

The proposed classification based GCI method is assessed with reliability and accuracy measures given in [31]. 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. We compared our proposed method with recent state-of-the-art classification based data-driven approach VC [27], ERT P3 [4] and popular unsupervised signal processing based methods: SEDREAMS [24], MMF [26], and DYPSA [20]. For comparison with other methods, the final model (combined joint acoustic-residual and **Model 1**) shown in Fig 1 is trained with GCI labeled utterances from JMK and SLT discussed in 2.2, tested on full dataset of BDL and KED which consists of 1131 and 452 utterances. The GCI detection evaluation results on BDL and KED datasets compared with other methods is shown Table 1. The evaluation results for VC [27], ERT P3 [4], SEDREAMS [24], MMF [26], and DYPSA [20] are obtained from [27] and [4]. From Table 1, we can observe that the **IDA** of the proposed method is significantly better than the state-of-the-art classification based method **VC** and other signal processing methods. It should be noted that the proposed CNN model is trained with less diverse (only two speakers) data and very low dimension raw features from the input data, whereas **VC** is trained with relatively large and diverse data, carefully selected handcrafted features after recursive elimination, classification based on voting classifiers. We can also observe that on average **MR** and **FAR** are significantly less compared to other methods. This can be attributed to the joint acoustic-residual model which significantly eliminates the miss rate and false alarm rates combined with the model trained with low pass filtered speech resulted in a high identification rate.

4. Summary and Conclusions

We proposed deep CNN classification based GCI detection method. Initially, we trained CNN models one for each input feature representation to determine its significance in classifying a frame into GCI and non-GCI frames. The F1-scores of the trained models revealed that the models trained with low-pass filtered speech, low-pass filtered positive clipped linear prediction residual and speech are the most significant input representations for GCI detection. In order to reduce the miss rates and false alarms, we trained a joint acoustic-residual model. The combined posterior probabilities of joint acoustic-residual and model trained with low-pass filtered speech significantly improved the GCI identification rate. In the future, we would like to explore the noise robustness of the model for various environmental noises, emotional speech, telephone and other modes of speech such as conversation, extempore, and shouted speech.

5. References

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

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

[3] 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.

[4] J. Matoušek and D. Tihelka, "Classification-based detection of glottal closure instants from speech signals," *Proc. Interspeech 2017*, pp. 3053–3057, 2017.

[5] P. Rengaswamy, G. Reddy, K. S. Rao, and P. Dasgupta, "A robust non-parametric and filtering based approach for glottal closure instant detection." in *INTERSPEECH*, 2016, pp. 1795–1799.

[6] 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.

[7] 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.

[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] M. G. Reddy and K. S. Rao, "Neutral to joyous happy emotion conversion," in *2017 14th IEEE India Council International Conference (INDICON)*. IEEE, 2017, pp. 1–6.

[11] D. Govind, S. M. Prasanna, and B. Yegnanarayana, "Neutral to target emotion conversion using source and suprasegmental information," in *Twelfth annual conference of the international speech communication association*, 2011.

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

[13] J. Matoušek and J. Romportl, "Automatic pitch-synchronous phonetic segmentation," 2008.

[14] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal averaging-method for enhancement of reverberant speech," in *Digital Signal Processing, 2007 15th International Conference on*. IEEE, 2007, pp. 607–610.

[15] Z. Hanzlíček and J. Matoušek, "F0 transformation within the voice conversion framework," 2007.

[16] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[17] 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.

[18] M. R. Thomas, J. Gudnason, and P. A. Naylor, "Data-driven voice soruce waveform modelling," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3965–3968.

[19] M. G. Reddy, M. Tanumay, and K. S. Rao, "Glottal closure instants detection from pathological acoustic speech signal using deep learning," *Machine Learning for Health (ML4H) Workshop at NeurIPS 2018*, 2018.

[20] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," 2007.

[21] M. R. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[22] A. Prathosh, P. Sujith, A. Ramakrishnan, and P. K. Ghosh, "Cumulative impulse strength for epoch extraction," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 424–428, 2016.

[23] A. I. Koutrouvelis, G. P. Kafentzis, N. D. Gaubitch, and R. Heusdens, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 316–328, 2016.

[24] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[25] C. d'Alessandro and N. Sturmel, "Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude," *Sadhana*, vol. 36, no. 5, pp. 601–622, 2011.

[26] V. Khanagha, K. Daoudi, and H. M. Yahia, "Detection of glottal closure instants based on the microcanonical multiscale formalism," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1941–1950, 2014.

[27] J. Matoušek and D. Tihelka, "Glottal closure instant detection from speech signal using voting classifier and recursive feature elimination," *Proc. Interspeech 2018*, pp. 2112–2116, 2018.

[28] M. Goyal, V. Srivastava *et al.*, "Detection of glottal closure instants using deep dilated convolutional neural networks," *arXiv preprint arXiv:1804.10147*, 2018.

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

[30] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[31] 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.