Classification-Based Detection of Glottal Closure Instants from Speech Signals

Jindřich Matoušek[1,2], *Daniel Tihelka*[2]

[1]Department of Cybernetics, [2] New Technology for the Information Society (NTIS)
Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Rep.

jmatouse@kky.zcu.cz, dtihelka@ntis.zcu.cz

Abstract

In this paper a classification-based method for the automatic detection of glottal closure instants (GCIs) from the speech signal is proposed. Peaks in the speech waveforms are taken as candidates for GCI placements. A classification framework is used to train a classification model and to classify whether or not a peak corresponds to the GCI. We show that the detection accuracy in terms of $F1$ score is 97.27%. In addition, despite using the speech signal only, the proposed method behaves comparably to a method utilizing the glottal signal. The method is also compared with three existing GCI detection algorithms on publicly available databases.

Index Terms: glottal closure instant (GCI), pitch mark, classification

1. Introduction

Pitch-synchronous methods of speech processing rely on the knowledge of moments of glottal closures. These moments are called *glottal closure instants* (GCIs), *pitch marks* or *epochs*. They can be defined as locations of a speech signal amplitude extreme that corresponds to the moment of glottal closure, a significant excitation of a vocal tract. The distance between two succeeding GCIs then corresponds to one vocal fold vibration cycle and can be represented in the time domain by a local *pitch period* value (T_0) or in the frequency domain by a local *fundamental frequency* value (F_0). Note that GCIs are present only in voiced segments of speech as there is no vocal fold vibration in unvoiced speech segments.

Precise detection of GCIs was reported to be important in many speech-technology applications [1, 2, 3] such as pitch tracking, prosodic speech modification [4, 5], various areas of speech synthesis [4, 6, 7], phonetic segmentation [8], voice conversion and transformation [9, 10], speech enhancement and dereverberation [11], glottal flow estimation [12] and speaker recognition [13], closed-phase linear prediction analysis [14], data-drive voice source modeling [15], and causal-anticausal deconvolution of speech signals [16].

Although GCIs can be reliably detected from a simultaneously recorded electroglottograph (EGG) signal (which measures glottal activity directly; thus, it is not burdened by modifications that happen to a flow of speech in the vocal tract—see Figure 1c), it is not always possible (e.g. in the case of using existing speech recordings) or comfortable to use an EGG device during recording. Hence, there is a great interest to detect GCIs directly from the speech signal.

Various algorithms have been proposed to detect GCIs directly in speech signals. They principally identify GCI candidates from local maxima of various speech representations and/or from discontinuities or changes in signal energy. The former include linear predictive coding (e.g. DYPSA [17], YAGA [2], or [18]), wavelet components [19], or multiscale formalism (MMF) [20]. The latter include Hilbert envelope, Frobenius norm, zero-frequency resonator, or SEDREAMS [21]. Dynamic programming is often used to refine the GCI candidates [17, 2]. A universal postprocessing scheme to correct GCI detection errors was also proposed [22]. A nice overview of the algorithms can be found in [1].

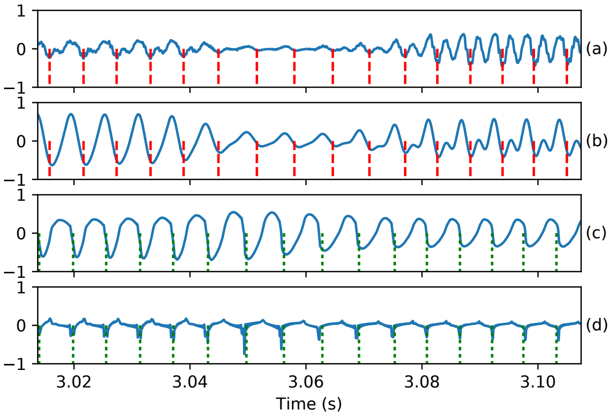

Figure 1: *Example of a speech signal (a), the corresponding low-pass filtered signal (b), EGG signal (c) and its difference (dEGG) signal (d). GCIs are marked by red dashed and green dotted lines in speech-based and glottal-based signals respectively. Note the delay between speech and EGG signals.*

In this paper we present a classification-based method for the automatic detection of GCIs from speech signals. It is based on a classification framework in which a classifier is trained on relevant features extracted around potential locations of GCIs (peaks in speech waveforms) and used to classify whether or not a peak corresponds to a true GCI. Unlike the above mentioned methods which require some manual tuning of their parameters, the proposed method is purely data-based in that the parameters of the classifier are set up automatically based on a training dataset.

2. Classification-based GCI detection

The problem of GCI detection could be viewed as a two-class classification problem: whether or not a peak in a speech waveform represents a GCI. We experimented with many classifiers, and the following ones showed the best performance: *support vector machines* (SVM) with a Gaussian radial basis function (RBF) kernel, *extremely randomized trees* (ERT), *k-nearest neighbors* (KNN), and *multilayer perceptron* (MLP).

This research was supported by the Czech Science Foundation (GA CR), project No. GA16-04420S. The access to the MetaCentrum clusters provided under the programme LM2015042 is highly appreciated.

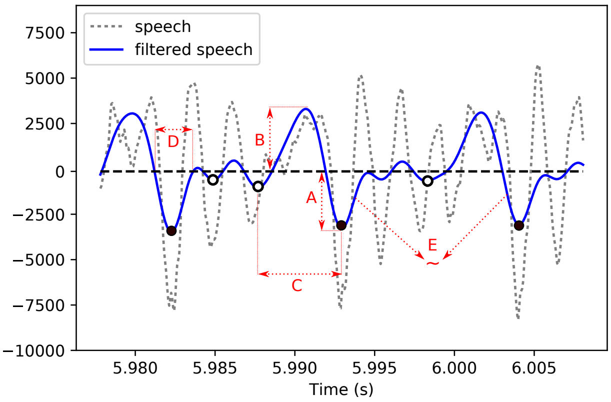

Figure 2: *Illustration of peak-based features extraction: amplitude of a negative peak (A), amplitude of a positive peak (B), difference between two negative peaks (C), width of a negative peak (D), correlation between waveforms of two negative peaks (E). GCI candidates are marked by ○, true GCIs by ●.*

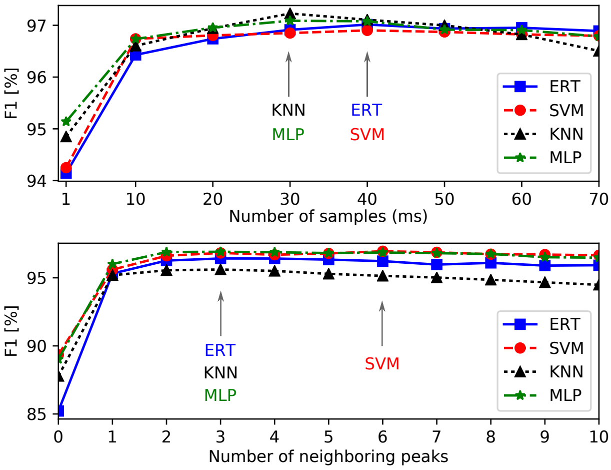

Figure 3: *Feature optimization: the optimal number of waveform samples (top) and neighboring peaks (bottom).*

2.1. Experimental data

The training and testing of the examined classifiers were performed on clean speech data available at our workplace (hereafter referred to as UWB). The speech recordings were primarily intended for speech synthesis. We used 63 utterances (≈9 minutes of speech) for training and 19 utterances (≈3 minutes of speech) for testing. The set of utterances was the same as in [23]; it comprised various Czech (male and female), Slovak (female), German (male), US English (male), and French (female) speakers. All speakers were part of both the training and test datasets. All speech waveforms were sampled at 16 kHz. Reference GCIs produced by a human expert (using both speech an EGG signals) were available for each utterance (51,629 in total).

2.2. Features

Before the features were extracted, speech waveforms were low-pass filtered by a zero-phase filter with cutoff frequency of 700 Hz to reduce the high-frequency structure in the speech signal (see Figure 1b or 2). The signals were then zero-crossed to identify peaks (both of the negative and positive polarity) that are used for feature extraction in further processing. Since the polarity of speech signals was shown to have an important impact on the performance of a GCI detector [24, 25], all speech signals were switched to have the negative polarity, and only the negative peaks were taken as the candidates for the GCI placement. For the purposes of training and testing, the location of each reference GCI was assigned to a corresponding negative peak in the filtered signal (see Figure 1b). There were 66,130 and 18,026 candidate peaks in the training and test datasets respectively, 40,938 and 10,691 of them corresponded to true GCIs.

We used two kinds of features. Perhaps the most intuitive way of describing characteristics of a given peak is to simply use (hanning-windowed) waveform samples in a window surrounding the peak. For the window length of 30 ms ($S = 30$), 481 samples (one sample representing the current peak plus 240 samples to the left and 240 samples to the right) were taken as features.

Alternatively, features inspired by [26] describing the given peak by a set of local descriptors reflecting the position and shape of other $2P$ neighboring peaks were used. For $P = 3$, such peak-based features comprise the amplitudes of the given negative peak and 6 neighboring (3 prior and 3 subsequent) negative peaks (7 features, denoted as A in Figure 2), amplitudes of 6 neighboring positive peaks (6, B), the time difference between the given negative peak and each of the neighboring negative peaks (6, C), the width of the given negative peak and each of the neighboring negative peaks (7, D), the correlation of the waveform around the given negative peak and the waveforms around each of the neighboring negative peaks (6, E). Hence, for $P = 3$, only 32 features were used in total.

2.3. Classifier design

To design the proposed classifiers, the *Scikit-learn* toolkit [27] was employed. The design consisted of the following steps:

1. *Feature optimization.* For each classifier with the default parameter setting (according to the Scikit-learn toolkit), optimal number of features (the number of samples and the number of peaks surrounding the given peak) were found on the training dataset using 10-fold cross validation. The optimal numbers are shown in Figure 3.
2. *Parameter tuning & model selection.* For each classifier and the features selected in the previous step, an extensive parameter tuning using grid search over relevant values of classifier parameters with 10-fold cross validation was conducted on the training dataset. The results of this step are shown in Figure 4.
3. *The best classifier selection.* Based on the results from the previous step, the best classifier for each kind of features was selected—KNN-S30 for sample-based and ERT-P3 for peak-based features.

2.4. Final evaluation on UWB test dataset

The final evaluation of the proposed classifiers was carried out on the UWB test dataset. In addition to the best classifiers using waveform samples (KNN-S30) and peak-based features (ERT-P3) respectively, we also used these classifiers with a combination of both kinds of features (KNN-P3S30 and ERT-P3S30) for the comparison. The results of the comparison and their statistical significance are shown in Table 1 and 2, respectively.

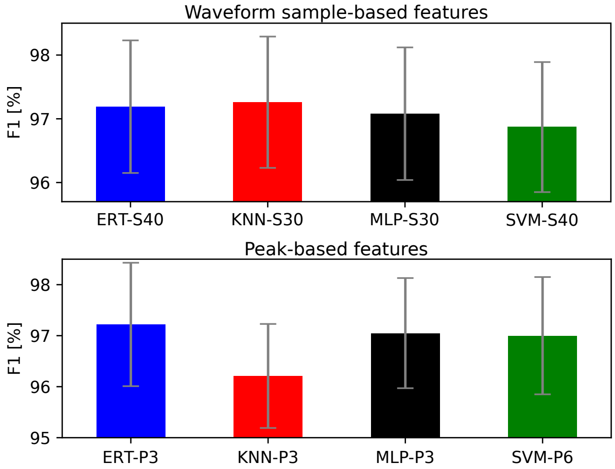

Figure 4: *Comparison of classifiers' performance on the cross-validation dataset for sample-based (top) and peak-based features (bottom) in terms of $F1$ and 95% confidence intervals. CLF-Sn and CLF-Pm denote the classifier type (CLF), the number of samples corresponding to a window of length n ms (Sn), and the number of peaks prior and subsequent to a given peak (Pm).*

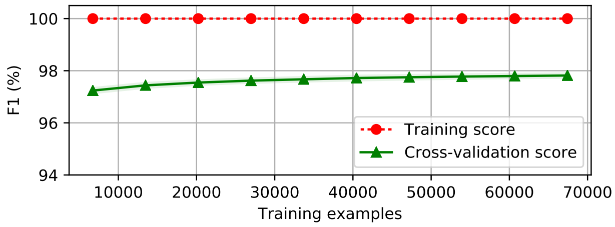

Figure 5: *Learning curves of the ERT-P3 classifier.*

As can be seen, despite relatively similar performances of all classifiers, given the large number of testing examples the ERT-P3 classifier performs significantly better than the others. Considering this finding and the low number of 32 features, ERT-P3 was chosen as the best classifier for our task of GCI detection and its performance was compared with other existing GCI detection algorithms further in Section 3. The learning curve of the ERT-P3 classifier in Figure 5 suggests that there is still some room for improvement if more training data and/or other features would be available.

3. Comparison with other methods

In the previous section, the proposed method was evaluated in a standard classification-manner, i.e., how good the classifier is *both in classifying peaks that correspond to true GCIs and, at the same time, in classifying peaks that do not represent GCIs*. Now, however, we will look at the comparison of the GCI detection with some other available detection algorithms.

3.1. Performance measures

The most common way to assess the performance of GCI detection techniques is to compare *locations of the detected and reference GCIs*. The widely used measures, proposed in [17], concern the *reliability*:

- *Identification Rate* (IDR): the percentage of glottal closures for which exactly one GCI is detected;

Table 1: *Final evaluation of the classifiers on the UWB test dataset in terms of recall (R), precision (P), and $F1$ score.*

Classifier	R (%)	P (%)	$F1$ (%)
ERT-P3	96.46	**98.09**	**97.27**
KNN-S30	96.45	97.75	97.10
ERT-P3S30	**96.74**	97.65	97.20
KNN-P3S30	96.68	97.80	97.23

Table 2: *Statistical significance according to McNemar's test [28]. The symbols "≫" and ">" mean that the row classifier is significantly better at the significance level $\alpha = 0.01$ and $\alpha = 0.05$ respectively than the column classifier. The symbol "=" means that the respective classifiers perform the same.*

Classifier	ERT-P3	KNN-S30	ERT-P3S30	KNN-P3S30
ERT-P3	=	≫	≫	≫
KNN-S30	≪	=	<	<
ERT-P3S30	≪	>	=	=
KNN-P3S30	≪	>	=	=

- *Miss Rate* (MR): the percentage of glottal closures for which no GCI is detected;
- *False Alarm Rate* (FAR): the percentage of glottal closures for which more than one GCI is detected;

and the *accuracy* of the algorithms:

- *Accuracy to* ± 0.25 *ms* (A25): the percentage of detections for which the identification error $\zeta \leq 0.25$ ms (the timing error between the detected and the corresponding reference GCI);
- *Identification Accuracy* (IDA): standard deviation of the identification error ζ.

A more dynamic evaluation measure

$$E10 = \frac{N_R - N_{\zeta > 0.1 T_0} - N_M - N_{FA}}{N_R} \tag{1}$$

that combines the reliability and accuracy in a single score and reflects the local T_0 pattern (determined from the reference GCIs) was also defined [29]. N_R stands for the number of reference GCIs, N_M is the number of missing GCIs (corresponding to MR), N_{FA} is the number of false GCIs (corresponding to FAR), and $N_{\zeta > 0.1 T_0}$ is the number of GCIs with the identification error ζ greater than 10% of the local pitch period T_0. For the alignment between the detected and reference GCIs, dynamic programming was employed [29].

3.2. Compared methods

We compared the proposed classification-based GCI detection method with three existing state-of-the-art methods:

- *Speech Event Detection using the Residual Excitation And a Mean-based Signal* (SEDREAMS) [21] (available in the COVAREP repository [30, 31], v1.4.1), shown in [1] to provide the best of performances compared to other methods;
- fast GCI detection based on *Microcanonical Multiscale Formalism* (MMF) [20] (available in [32]);
- *Dynamic Programming Phase Slope Algorithm* (DYPSA) [17] available in the VOICEBOX toolbox [33].

Table 3: *Summary of the performance of the GCI detection algorithms for the four datasets.*

Dataset	**Method**	**IDR** (%)	**MR** (%)	**FAR** (%)	**IDA** (ms)	**A25** (%)	**E10** (%)
UWB	MPA	97.06	0.66	2.28	0.21	84.65	97.03
	ERT-P3	**95.87**	**1.99**	**2.14**	0.29	81.06	**95.93**
	SEDREAMS	91.80	3.54	4.66	**0.24**	**81.51**	91.87
	MMF	83.47	11.42	5.11	0.42	80.72	84.80
	DYPSA	87.40	4.86	7.74	0.40	80.60	87.27
BDL	ERT-P3	**91.96**	2.98	**5.06**	**0.41**	88.41	**91.78**
	SEDREAMS	90.98	**2.35**	6.67	0.54	**91.23**	90.57
	MMF	87.82	5.84	6.34	0.61	90.36	87.77
	DYPSA	86.98	7.59	5.43	0.65	91.16	86.69
SLT	ERT-P3	**95.18**	1.35	**3.47**	**0.15**	**95.08**	**95.07**
	SEDREAMS	92.96	**1.15**	5.89	0.19	89.09	92.61
	MMF	91.16	5.33	3.51	0.37	77.53	91.32
	DYPSA	91.50	2.80	5.70	0.30	81.23	91.24
KED	ERT-P3	**91.88**	2.94	**5.18**	**0.27**	**88.02**	**91.69**
	SEDREAMS	89.54	**1.16**	9.30	0.56	78.46	88.61
	MMF	89.11	4.61	6.28	0.57	83.52	88.92
	DYPSA	89.01	4.62	6.37	0.48	83.70	88.81

We used the implementations available online; no modifications of the algorithms were made. Since all three algorithms estimate GCIs also during unvoiced segments, authors recommend to filter the detected GCIs by the output of a separate voiced/unvoiced detector. We applied an F_0 contour estimated by the RAPT algorithm [34] as implemented in the Wavesurfer tool [35] for this purpose (we removed GCIs with the undefined F_0 value). As each of the algorithms places GCIs slightly differently, the locations of GCIs were shifted towards the neighboring negative peak of the corresponding filtered speech signal. The same filtering and shifting were applied also on the output of the proposed ERT-P3 classifier.

3.3. Test datasets

Firstly, the evaluation was carried out on the UWB test dataset (≈3 minutes of speech) described in Section 2.1. GCIs produced by a human expert were used as reference GCIs. Since contemporaneous EGG recordings were also available, the Multi-Phase Algorithm (MPA) [29] that detects GCIs by thresholding the dEGG signal was included in the comparison as an upper bound for GCI detection.

Secondly, two voices, a US male (BDL) and a US female (SLT) from the CMU ARCTIC databases intended for unit selection speech synthesis [36, 37] were used as a test material. Each voice consists of 1132 phonetically balanced utterances of a total duration ≈54 minutes per voice. Additionally, KED TIMIT database [37] comprising 453 phonetically balanced utterances (≈20 min.) of a US male speaker was also used for testing. All these datasets comprise clean speech. Since there are no hand-crafted GCIs available for these datasets, GCIs detected from contemporaneous EGG recordings by the MPA algorithm [29] were used as the reference GCIs. Original speech and EGG signals were downsampled to 16 kHz. Note that MPA also synchronizes GCIs with speech signal to compensate for the delay between the speech and EGG signals (see Figure 1). No correction of the automatically obtained reference GCIs were made. It is important to mention that no speaker from these datasets was part of the training dataset used to train the proposed classification-based GCI detection method.

3.4. Results

The results in Table 3 show that the proposed classification-based method (ERT-P3) consistently outperforms other methods in terms of *reliability*, especially with respect to the identification (IDR) and false alarm (FAR) rates, for all tested datasets. It also gives the highest score that combines reliability and dynamic detection accuracy (E10).

As for the *accuracy* itself, ERT-P3 performed very well with the highest identification accuracy (IDA) for all datasets except for the UWB dataset. Together with the SEDREAMS algorithm it also yielded the smallest number of timing errors higher than 0.25 ms (A25). It is obvious that the proposed classification-based approach to GCI detection from speech signals performs very well and mostly outperforms existing state-of-the-art methods on the four test datasets. In addition, its performance is not much worse in comparison with the MPA algorithm which utilizes glottal (EGG) signal for the detection of GCIs.

4. Conclusions

A classification-based method was proposed to detect GCIs from speech signals. Being a data-based method, the only requirement is a set of reference GCIs to train the classifier. No manual tuning of parameters is required—classifier parameters are set up automatically during the training process. We showed that the proposed method performed very well on several test datasets and outperformed other state-of-the-art methods in terms of detection reliability and mostly also in terms of accuracy. This was also true for datasets whose speakers were not included in the training data[1].

In our future work, we would like to investigate more closely whether more training data from more speakers could further increase the performance of the proposed method. We also plan to incorporate some other features (e.g. pitch-based, voiced/unvoiced or harmonic/noise related) to the currently used peak-based feature set. Robustness of the proposed method to noisy signals will also be investigated.

[1] Data relevant to the described experiments are available online [38].

5. References

[1] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, mar 2012.

[2] M. R. P. 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, jan 2012.

[3] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech and Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[4] 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.

[5] T. Ewender and B. Pfister, "Accurate pitch marking for prosodic modification of speech segments," in *INTERSPEECH*, Makuhari, Japan, 2010, pp. 178–181.

[6] T. Dutoit and B. Gosselin, "On the use of a hybrid harmonic/stochastic model for TTS synthesis-by-concatenation," *Speech Communication*, vol. 19, no. 2, pp. 119–143, 1996.

[7] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *INTERSPEECH*, Brighton, Great Britain, 2009, pp. 1779–1782.

[8] J. Matoušek and J. Romportl, "Automatic pitch-synchronous phonetic segmentation," in *INTERSPEECH*, Brisbane, Australia, 2008, pp. 1626–1629.

[9] Z. Hanzlíček and J. Matoušek, "F0 transformation within the voice conversion framework," in *INTERSPEECH*, Antwerp, Belgium, 2007, pp. 1961–1964.

[10] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *INTERSPEECH*, Brisbane, Australia, 2008, pp. 1829–1832.

[11] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal averaging method for enhancement of reverberant speech," in *International Conference on Digital Signal Processing*, Cardiff, Great Britain, 2007, pp. 607–610.

[12] 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, aug 1979.

[13] S. R. 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.

[14] A. Krishnamurthy and D. Childers, "Two-channel speech analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 730–743, 1986.

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

[16] B. Bozkurt and T. Dutoit, "Mixed-phase speech modeling and formant estimation, using differential phase spectrums," in *ISCA Tutorial and Research Workshop VOQUAL03*, Geneva, Switzerland, 2003, pp. 21–24.

[17] 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.

[18] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 12, pp. 2471–2480, 2013.

[19] V. N. Tuan and C. D'Alessandro, "Robust glottal closure detection using the wavelet transform," in *EUROSPEECH*, Budapest, Hungary, 1999, pp. 2805–2808.

[20] 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*, vol. 22, no. 12, pp. 1941–1950, 2014.

[21] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *INTERSPEECH*, Brighton, Great Britain, 2009, pp. 2891–2894.

[22] P. Sujith, A. P. Prathosh, R. A. G., and P. K. Ghosh, "An error correction scheme for GCI detection algorithms using pitch smoothness criterion," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 3284–3288.

[23] M. Legát, J. Matoušek, and D. Tihelka, "On the detection of pitch marks using a robust multi-phase algorithm," *Speech Communication*, vol. 53, no. 4, pp. 552–566, 2011.

[24] M. Legát, D. Tihelka, and J. Matoušek, "Pitch marks at peaks or valleys?" in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007, vol. 4629, pp. 502–507.

[25] T. Drugman, "Residual excitation skewness for automatic speech polarity detection," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 387–390, 2013.

[26] E. Barnard, R. A. Cole, M. P. Vea, and F. A. Alleva, "Pitch detection with a neural-net classifier," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 298–307, 1991.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. M. B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perror, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] S. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, vol. 328, pp. 317–328, 1997.

[29] M. Legát, J. Matoušek, and D. Tihelka, "A robust multi-phase pitch-mark detection algorithm," in *INTERSPEECH*, Antwerp, Belgium, 2007, pp. 1641–1644.

[30] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - A collaborative voice analysis repository for speech technologies," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Florence, Italy, 2014, pp. 960–964.

[31] "A Cooperative voice analysis repository for speech technologies." [Online]. Available: https://github.com/covarep/covarep

[32] "Matlab codes for Glottal Closure Instants (GCI) detection." [Online]. Available: https://geostat.bordeaux.inria.fr/index.php/downloads.html

[33] "VOICEBOX: Speech Processing Toolbox for MATLAB." [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[34] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science, 1995, ch. 14, pp. 495–518.

[35] K. Sjölander and J. Beskow, "Wavesurfer – an open source speech tool," in *INTERSPEECH*, vol. 4, Beijing, China, 2000, pp. 464–467.

[36] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Speech Synthesis Workshop*, Pittsburgh, USA, 2004, pp. 223–224.

[37] "FestVox Speech Synthesis Databases." [Online]. Available: http://festvox.org/dbs/index.html

[38] "Data used for classification-based glottal closure instants (GCI) detection." [Online]. Available: https://github.com/ARTIC-TTS-experiments/2017_Interspeech