Interspeech 2018
2-6 September 2018, Hyderabad

Glottal Closure Instant Detection from Speech Signal Using Voting Classifier and Recursive Feature Elimination

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 our previous work, we introduced a classification-based method for the automatic detection of glottal closure instants (GCIs) from the speech signal and we showed it was able to perform very well on several test datasets. In this paper, we investigate whether adding more features (voiced/unvoiced, harmonic/noise, spectral etc.) and/or using an ensemble of classifiers such as a voting classifier can further improve GCI detection performance. We show that using additional features leads to a better detection accuracy; best results were obtained when recursive feature elimination was applied on the whole feature set. In addition, a voting classifier is shown to outperform other classifiers and other existing GCI detection algorithms on publicly available databases.

Index Terms: glottal closure instant (GCI), pitch mark, classification, voting classifier, recursive feature elimination

1. Introduction

Glottal closure instants (GCIs) (also called *pitch marks* or *epochs*) refer to peaks in *voiced parts* of the speech signal that correspond 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).

Precise detection of GCIs plays a key role in *pitch-synchronous* speech processing which is used in many speech-technology applications [1, 2, 3, 4]. 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), it is desirable to detect GCIs directly from the speech signal only.

A number of algorithms have been proposed to detect GCIs in the speech signal. 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 [5], YAGA [2], or [6]), wavelet components [7], or multiscale formalism (MMF) [8]. The latter include Hilbert envelope, Frobenius norm, zero-frequency resonator, or SEDREAMS [9]. Dynamic programming is often used to refine the GCI candidates [5, 2]. A universal postprocessing scheme to correct GCI detection errors was also proposed [10].

This research was supported by the Technology Agency of the Czech Republic (TA CR), project No. TH02010307. The access to the MetaCentrum clusters provided under the programme LM2015042 is highly appreciated.

In our previous work [11], we elaborated a classification-based method for the automatic detection of GCIs [12, 13, 14] from the speech signal in which the detection is viewed as a two-class classification problem: whether or not a peak in a speech waveform represents a GCI. We showed it was able to perform very well on several test datasets. The advantage of this method is that once a training dataset is available, classifier parameters are set up automatically without manual tuning.

In this paper, we answer other two research questions: (i) whether adding more features (voiced/unvoiced, harmonic/noise, spectral etc.) can help the classifier perform better; (ii) whether using an ensemble of classifiers or a meta-classifier (such as a voting classifier) can further improve GCI detection performance.

2. Experimental data

2.1. Speech material

The development and testing of the proposed classifiers were performed on clean 16kHz-sampled speech recordings available at our workplace (hereafter referred to as UWB). The recordings were primarily intended for speech synthesis. We used 63 utterances (≈9 minutes of speech) for the development and 20 utterances (≈3 minutes of speech) for testing. The set of utterances was the same as in [15] – 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 development and test datasets. Reference GCIs produced by a human expert (using both speech and EGG signals) were available for each utterance and were synchronized with the corresponding minimum negative sample in the speech signal.

For the purpose of the proposed classification-based GCI detection, speech waveforms were low-pass filtered by a zero-phase Equiripple-designed filter with 0.5 dB ripple in the pass band, 60 dB attenuation in the stop band, and with the cutoff frequency of 800 Hz to reduce the high-frequency structure in the speech signal. 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 [16, 17], 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 purpose of training and testing, the location of each reference GCI was mapped to a corresponding negative peak in the filtered signal. There were 73,205 and 20,338 candidate peaks in the development and test datasets respectively (marked by both ○ and ● in Figure 1), 39,931 and 10,807 of them corresponded to true GCIs (marked by ● only).

10.21437/Interspeech.2018-1147

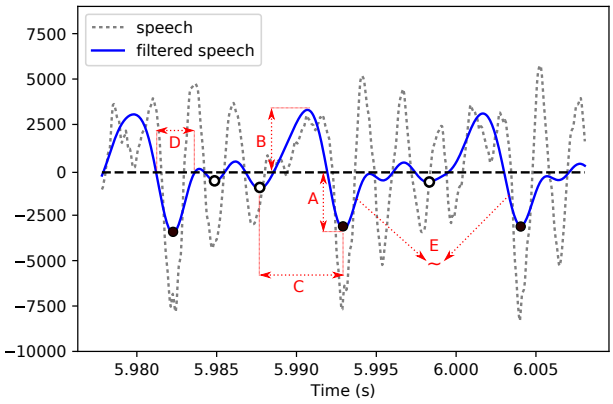

Figure 1: *Illustration of features extraction: amplitude of a negative peak (A, negAmp), amplitude of a positive peak (B, posAmp), difference between two negative peaks (C, timeDiff), width of a negative peak (D, width), correlation between waveforms of two negative peaks (E, corr). GCI candidates are marked by ∘, true GCIs by ●.*

2.2. GCI detection measures

GCI detection techniques are usually evaluated by comparing locations of the detected and reference GCIs. The measures typically concern the *reliability* and *accuracy* of the GCI detection algorithms [5]. The former includes the percentage of glottal closures for which exactly one GCI is detected (*identification rate*, IDR), the percentage of glottal closures for which no GCI is detected (*miss rate*, MR), and the percentage of glottal closures for which more than one GCI is detected (*false alarm rate*, FAR). The latter includes the percentage of detections with the identification error $\zeta \leq 0.25$ ms (*accuracy to* ± 0.25 *ms*, A25) and standard deviation of the identification error ζ (*identification accuracy*, IDA).

We also use a more dynamic evaluation measure [18]

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

that combines the reliability and accuracy in a single score and reflects the local T_0 pattern (determined from the reference GCIs). 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 $N_{\zeta > 0.1T_0}$ 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 [18].

3. Features

3.1. Baseline features

The baseline features used in [11] are illustrated in Figure 1. Inspired by [14], the features were associated with negative peaks in the low-pass filtered speech waveforms. Each peak is described by a set of local descriptors reflecting the position and shape of other 3 neighboring peaks [11]. Thus, only 32 features were used in total: the amplitudes of the given negative peak and 6 neighboring (3 prior and 3 subsequent) negative peaks (7 features, denoted as A in Figure 1), 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 (a distance between two zero-crossings) 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).

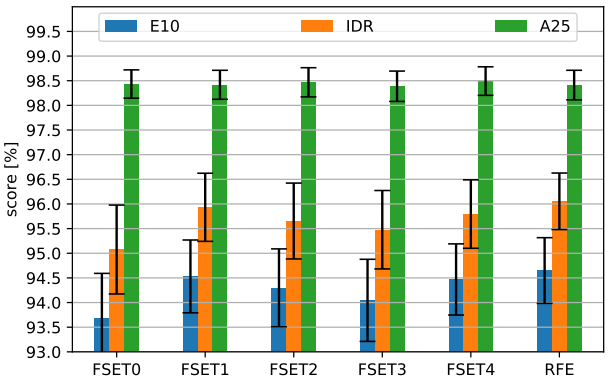

Figure 2: *Comparison of different feature sets with respect to GCI detection performance including 95% confidence intervals.*

3.2. Feature engineering

To answer the first research question of this paper, we defined several feature sets and investigated their influence on the GCI detection performance. We extended the baseline feature set (32 features) described in Section 3.1 (hereinafter referred to as FSET0) with *acoustic features* (zero-crossing rate (ZCR), log energy, harmonic-to-noise ratio (HNR), voiced/unvoiced, peak ratio to 6 neighboring peaks – denoted as FSET1, +10 features), *spectral features* (spectral centroid, spectral bandwidth, and spectral roll-off – FSET2, +3 features), and *mel spectral frequency coefficients* (MFCCs – FSET3, +13 features). All features were calculated from 10ms-long speech segments extracted around every peak candidate.

We also experimented with the full feature set consisting of all 58 features described above (FSET4) and with a feature set designed automatically by a feature selection algorithm. For the feature selection, *recursive feature elimination* (RFE) algorithm was chosen [19]. Starting from the full feature set, the RFE algorithm recursively prunes out the least important features until the desired number of features is reached. The feature importance was assigned by an external estimator (*extremely randomized trees* [20] in our case), and the desired number of features was selected by 10-fold cross-validation technique. The optimal feature set selected by RFE consisted of the following 37 features (feat$\pm n$ means features related to the n-th preceding and n-th succeeding peak):

- **Baseline features** (25 features): negAmp, negAmp$\pm$1, negAmp$\pm$2, negAmp$\pm$3, posAmp$\pm$1, posAmp$\pm$2, posAmp$\pm$3, timeDif$\pm$1, timeDif$\pm$2, timeDif$\pm$3, corr$\pm$1, corr$\pm$2, corr$\pm$3;
- **Acoustic features** (8): voiced/unvoiced, ZCR, logEnergy, HNR, negPeakRatio$\pm$1, negPeakRatio$\pm$3;
- **Spectral features.** (2): specCentroid, specRolloff
- **MFCC-based features** (2): MFCC0, MFCC1.

The selection suggests that the baseline features [11] generally perform well except for the width-based features that were not selected at all. It also seems that especially acoustic features are an appropriate complement to the original set.

The extremely randomized trees (ERT) classifier [20] with the default hyper-parameters (according to the Scikit-learn

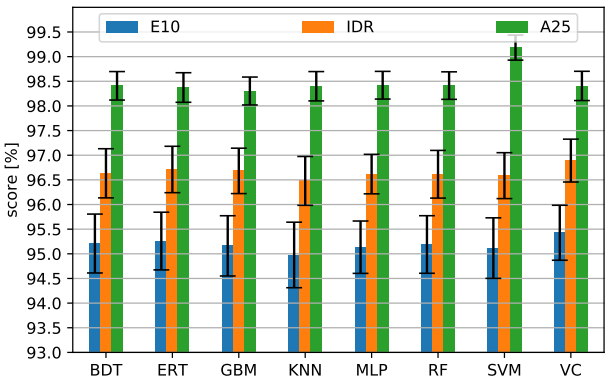

Figure 3: *Comparison of classifiers' GCI detection performance incl. 95% confidence intervals on the validation dataset.*

toolkit [21]) used in [11] was trained and evaluated on the development data described by the different feature sets using a leave-one-out cross-validation strategy on the utterance level. It means that GCIs from each utterance were used once as a validation set while the GCIs from all remaining utterances formed the training set. The comparison of the different feature sets in Figure 2 indicate that the RFE algorithm yields the best results and thus it outperforms the baseline features (FSET0) used in [11].

4. Classifiers

The second research question concerned the classifier used to detect GCIs. We examined a number of "single" classifiers, both *linear* and *non-linear*. Non-linear classifiers clearly outperformed linear classifiers; support vector machines (SVM) with a Gaussian radial basis function (RBF) kernel, multilayer perceptron (MLP), and k-nearest neighbors (KNN) showed the best performance.

We also investigated another class of classifiers called *ensemble models*. This kind of models combines predictions of several single classifiers built with a given learning algorithm (decision trees are typically used). The following classifiers showed the best performance: bagged decision trees (BDT) [22], random forests (RF) [23], extremely randomized trees (ERT) [20]), and gradient boosting machines (GBM) [24]). As can be seen in Figure 3, ensemble classifiers generally outperformed "single" classifiers.

To design the proposed classifiers, the standard procedure was applied: (i) feature scaling/normalization was applied (except for decision-tree based classifiers); (ii) classifier training and extensive hyper-parameter tuning using grid search with 10-fold cross validation on the utterance level was conducted on the development dataset. For the hyper-parameter optimization, E10 measure (1) was used. The RFE-based feature set described in Section 3.2 was utilized. Since the detected GCIs correspond to peaks in the filtered speech signal, they were shifted towards the minimum negative sample in the original speech signal, see (a,b,c,d) in Figure 4.

A comparison of the classifiers is shown Figure 3. If the voting classifier (VC) described further in Section 4.1 is not counted in, the best classifiers were ERT (achieving $E10 = 95.26\%$ and $IDR = 96.71\%$) and SVM ($A25 = 99.18\%$) which is consistent with the results of an extensive study in which 179 classifiers belonging to a wide collection of 17 families were evaluated on 121 datasets from various domains [25].

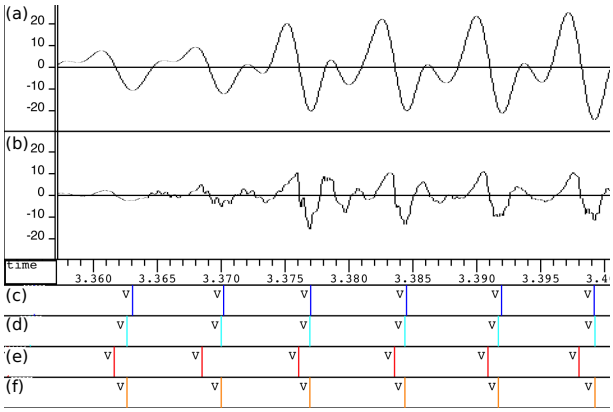

Figure 4: *Illustration of the shift between the detected and final GCI locations in filtered (a) and original speech (b) signals: GCIs detected by a classification-based approach in the filtered signal (c) and shifted to a minimum negative sample in the speech signal (d), and GCIs detected by the SEDREAMS method (e) in the speech signal and shifted to the minimum negative sample in the speech signal (f).*

Table 1: *Top 5 combinations of different learning algorithms within a voting classifier on the validation dataset.*

Voting classifier	**Opt. weights**	**E10** (%)
BDT+GDM+KNN+SVM	(1, 3, 1, 3)	95.42
RF+GDM+KNN+MLP+SVM	(1, 2, 1, 1, 2)	95.42
BDT+GDM+KNN+MLP+SVM	(2, 3, 2, 1, 2)	95.42
RF+GDM+SVM	(1, 4, 4)	95.39
ERT+RF+GDM+MLP+SVM	(2, 1, 3, 2, 3)	**95.43**

4.1. Voting classifier

A special case is a *voting classifier* which can be seen as a "meta-classifier" in that it combines conceptually different machine learning algorithms and uses a majority of votes to make a final prediction. Table 1 shows top five combinations (all of them yielded the same accuracy $E10 = 95.39\%$ when weights of all learning algorithms were set to 1). We further took these combinations and optimized their performance by weighting the contribution of each learning algorithm within the range $[1, 4]$. The best results were achieved for the combination of ERT (with weight 2), RF (1), GBM (3), MLP (2), and SVM (3). This voting classifier is further denoted as VC and is also compared to other classifiers in Figure 3. It can be seen that VC outperforms other classifiers in terms of IDR and E10 measures. As for the A25 measure, SVM behaved best.

5. Comparison with other methods

We compared the proposed voting classifier with three existing state-of-the-art GCI detection methods:

- *Speech Event Detection using the Residual Excitation And a Mean-based Signal* (SEDREAMS) [9] (available in the COVAREP repository [26, 27], 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) [8] (available in [28]);
- *Dynamic Programming Phase Slope Algorithm* (DYPSA)

Table 2: *Summary of the performance of the GCI detection algorithms for the four datasets.*

Dataset	Method	**IDR** (%)	**MR** (%)	**FAR** (%)	**IDA** (ms)	**A25** (%)	**E10** (%)
UWB	VC	**96.66**	**2.25**	**1.09**	**0.22**	98.73	**95.56**
	ERT2017	95.45	3.08	1.47	0.23	**98.85**	94.47
	SEDREAMS	93.14	3.99	2.87	0.29	98.09	91.70
	MMF	85.09	11.42	3.48	0.47	97.86	83.57
	DYPSA	89.62	6.26	4.12	0.37	98.07	88.22
BDL	VC	**93.86**	**2.26**	3.89	0.45	95.61	**90.17**
	ERT2017	93.39	2.99	**3.62**	0.52	95.59	89.72
	SEDREAMS	91.82	3.02	5.16	**0.44**	**97.37**	89.45
	MMF	89.49	4.53	5.98	0.57	96.23	86.67
	DYPSA	88.95	4.32	6.73	0.56	96.81	86.29
SLT	VC	**96.21**	**0.58**	3.21	0.20	98.57	94.81
	ERT2017	**96.21**	0.71	3.07	0.20	98.59	**94.84**
	SEDREAMS	94.67	1.12	4.21	**0.18**	**99.61**	94.34
	MMF	92.48	5.24	**2.28**	0.41	98.89	91.65
	DYPSA	93.23	2.88	3.89	0.31	99.39	92.73
KED	VC	**95.85**	**1.42**	2.73	0.25	99.57	**95.48**
	ERT2017	95.38	1.93	2.69	**0.24**	**99.61**	95.05
	SEDREAMS	92.31	6.03	**1.66**	0.29	99.04	91.77
	MMF	90.24	7.04	2.72	0.37	98.79	89.45
	DYPSA	90.29	7.05	2.66	0.31	99.16	89.71

[5] available in the VOICEBOX toolbox [29].

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 filtering the detected GCIs by the output of a separate voiced/unvoiced detector. Unlike [11], the Robust Epoch And Pitch EstimatoR (REAPER) [30] was applied in this work. There is no need to apply such a postprocessing on GCIs detected by the proposed classification-based approach since the voiced/unvoiced pattern was included directly in the feature set (see Section 3.2). To be consistent with the proposed classification-based approach, the detected GCIs were shifted towards the neighboring minimum negative sample in the original *non-filtered* signal[1], see (b,e,f) in Figure 4.

5.1. Test datasets

Firstly, the evaluation was carried out on the UWB test dataset (≈3 minutes of speech) described in Section 2. GCIs produced by a human expert were used as reference GCIs.

Secondly, two voices, a US male (BDL) and a US female (SLT) from the CMU ARCTIC databases intended for unit selection speech synthesis [31, 32] 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 [32] 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 handcrafted GCIs available for these datasets, GCIs detected from contemporaneous EGG recordings by the Multi-Phase Algorithm (MPA) [18] (again shifted towards the neighboring minimum negative sample in the speech signal) were used as the reference GCIs[2]. Original speech and EGG signals were downsampled to 16 kHz. It is important to mention that no speaker from these datasets was part of the training dataset used to train the proposed classifiers.

[1]Note that the shift was made slightly differently than in [11] where it was made towards the neighboring negative peak in the corresponded *filtered* signal.

[2]The reference GCIs and other data relevant to the described experiments are available online [33].

5.2. Results

The results in Table 2 confirm that the proposed voting classifier (VC) working on the feature set designed by the RFE technique generally outperforms the ERT classifier working on the baseline feature set (FSET0) [11] (denoted as ERT2017 in Table 2).

It is also evident that the proposed classification-based approach (and especially the voting classifier) also outperforms other methods for all datasets with respect to most detection measures, especially in terms of the identification rate (IDR), miss rate (MR), and dynamic detection accuracy (E10). Together with the SEDREAMS algorithm it also yielded the smallest number of timing errors higher than 0.25 ms (A25) and the smallest standard deviation of the timing error (IDA).

6. Conclusions

In this paper, two research questions regarding the improvement of the classification-based method to detect GCIs from the speech signal proposed in [11] were answered. Firstly, we showed that the performance of classification-based GCI detection can be further improved by employing features selected from an extended set of features using the recursive feature elimination technique. Secondly, a combination of single classifiers (such as extremely randomized trees, random forests, gradient boosting machines, multilayered perceptron, and support vector machines with RBF kernel) into a voting classifier outperformed each of the single classifiers. The resulting voting classifier with the RFE-based feature set performed very well in comparison with other state-of-the-art methods on several test datasets.

In our future work, we plan to investigate whether a deep learning algorithm could further increase the performance of the proposed classification-based GCI detection method. Robustness of the proposed method to noisy signals and/or to emotional or expressive speech will also be researched.

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

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

[11] J. Matoušek and D. Tihelka, "Classification-based detection of glottal closure instants from speech signals," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3053–3057.

[12] I. S. Howard and M. A. Huckvale, "Speech fundamental period estimation using a trainable pattern classifier," in *SPEECH'88: 7th FASE Symposium*, Edinburgh, UK, 1988.

[13] J. R. Walliker and I. S. Howard, "Real-time portable multi-layer perceptron voice fundamental-period extractor for hearing aids and cochlear implants," *Speech Communication*, vol. 9, no. 1, pp. 63–72, 1990.

[14] 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.

[15] 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.

[16] 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.

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

[18] M. Legát, J. Matoušek, and D. Tihelka, "A robust multi-phase pitch-mark detection algorithm," in *INTERSPEECH*, vol. 1, Antwerp, Belgium, 2007, pp. 1641–1644.

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1/3, pp. 389–422, 2002.

[20] P. Geurts and D. E. L. Wehenkel, "Extremely Randomized Trees," *Machine Leaming*, vol. 63, no. 1, pp. 3–42, 2006.

[21] 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.

[22] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[23] ——, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[25] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.

[26] 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.

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

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

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

[30] "REAPER: Robust Epoch And Pitch EstimatoR." [Online]. Available: https://github.com/google/REAPER

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

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

[33] "Data used for classification-based glottal closure instant detection." [Online]. Available: https://github.com/ARTIC-TTS-experiments/2018_Interspeech