Interspeech 2018
2-6 September 2018, Hyderabad

Detection of Glottal Closure Instants from Speech Signals: A Convolutional Neural Network Based Method

Shuai Yang [1], *Zhiyong Wu*[1,3], *Binbin Shen*[2], *Helen Meng*[1,3]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University
[2] Pachira Information Technology (Beijing) Co., Ltd.
[3] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong

yangshua16@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, shenbinbin@pachiratech.com, hmmeng@se.cuhk.edu.hk

Abstract

Most conventional methods to detect glottal closure instants (GCI) are based on signal processing technologies and different GCI candidate selection methods. This paper proposes a classification method to detect glottal closure instants from speech waveforms using convolutional neural network (CNN). The procedure is divided into two successive steps. Firstly, a low-pass filtered signal is computed, whose negative peaks are taken as candidates for GCI placement. Secondly, a CNN-based classification model determines for each peak whether it corresponds to a GCI or not. The method is compared with three existing GCI detection algorithms on two publicly available databases. For the proposed method, the detection accuracy in terms of F1-score is 98.23%. Additional experiment indicates that the model can perform better after trained with the speech data from the speakers who are the same as those in the test set.

Index Terms: glottal closure instants (GCI), pitch mark, convolutional neural network (CNN), classification

1. Introduction

In speech processing, glottal closure instants (GCIs) are referred to the instances of significant excitation of the vocal tract. These particular time events correspond to the moments of high energy in the glottal signal during voiced speech. For speech analysis, closed-phase linear prediction autoregressive analysis techniques have been developed for better estimating the prediction coefficients, which results in a better estimation of the vocal tract resonances [16]. These techniques explicitly require the determination of GCIs. A wide range of applications also implicitly assume that these instants are already known. In concatenative speech synthesis, it is well known that some knowledge of a reference instant is necessary to eliminate concatenation discontinuities [10]. Knowing the GCI location is of particular importance in speech processing. Such information has been put to practical use in applications including prosodic speech modification [1], glottal flow estimation [2], speech synthesis [3][4] and data-driven voice source modelling [5].

Although GCIs can be reliably detected from a simultaneously recorded electroglottograph (EGG) signal (which measures glottal activity directly—see Figure 1.c), it is not always possible or comfortable to use an EGG device during recording [11]. Hence, there is a great interest to detect GCIs directly from the speech signal.

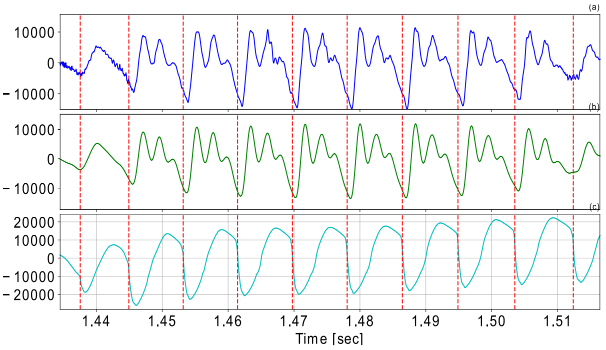

Figure 1: *Example of a speech signal (a), the corresponding lowpass filtered signal (b), EGG signal (c). GCIs are marked by red dashed lines in speech-based and glottal-based signals respectively. Note the delay between speech and EGG signals.*

Various algorithms have been proposed to detect GCIs directly in speech signals. Most conventional methods to detect GCI are based on signal processing technologies and different GCI candidate selection methods. For example, most previous work have been devoted to convert original speech waveform to a new signal in which features related to GCI locations are easily to be identified, followed by identifying a series of GCI candidates from which accurate GCI locations are identified.

Several approaches relying on the Hilbert Envelope (HE) have been proposed in the literature [6][7][8]. The Dynamic Programming Phase Slope Algorithm (DYPSA) [9] estimates GCIs by the identification of peaks in linear prediction residual of speech in a similar way to the HE method. The Speech Event Detection using the Residual Excitation and a Mean-based Signal (SEDREAMS) algorithm was proposed in [10] as a reliable and accurate method for locating both GCIs and GOIs from the speech waveform. The ERT-P3 algorithm applies extremely randomized trees (ERT) trained on relevant features extracted around potential locations of GCIs (peaks in speech waveforms) to classify whether or not a peak corresponds to a true GCI [11]. It was shown that the ERT-P3 algorithm and the technique proposed in [11] clearly outperformed other state-of-the-art methods.

10.21437/Interspeech.2018-1281

This paper proposes a classification method to detect glottal closure instants from speech waveforms using convolutional neural network (CNN). The procedure is divided into two successive steps. Firstly, a low-pass filtered signal is computed, whose negative peaks are taken as candidates for GCIs placement. Secondly, a CNN-based classification model classifies each peak into two categories, deciding whether or not a peak corresponds to a GCI.

Unlike the above conventional algorithms which require some manual tuning of parameters (such as window length), the proposed method is similar to ERT-P3 algorithm that it is purely data-driven as the parameters of the classifier are set up automatically based on a training database.

The paper is structured as follows. The proposed method is fully described in Section 2. In Section 3, we present our experiments and results obtained on the CMU ACRTIC database [12]. The proposed method is compared with DYSPA, SEDREAMS and ERT-P3 algorithms according to their GCI detection performance. Finally, Section 4 concludes the paper.

2. Proposed Method

The problem of GCI detection could be viewed as a two-class classification problem in deciding whether a GCI candidate (peaks in speech waveforms) represents a reference GCI [13].

2.1. GCI candidates

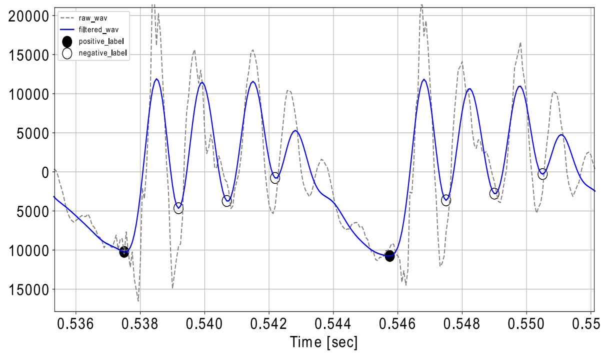

Figure 2: *Illustration of GCI candidates' selection: raw signal (dashed line), low-pass filtered signal (solid line), GCI candidates (circle), true GCIs (solid points)*

Since glottal wave makes up the low-frequency content of the raw speech signal, high-frequency content helps little on GCI detection. Speech waveforms were low-pass filtered by a 6th-order low-pass Butterworth filter with a cutoff frequency of 700 Hz to reduce the high-frequency structure in the speech signal. The low-pass filtered signal needs to be move ahead about 1.5ms (depends on signal period and cutoff frequency) to fixed the time delay caused by low-pass filtering. Negative peaks in the low-pass filtered signal were taken as the candidates for the GCI placement. Details on selecting GCI candidates are illustrated in Figure 2.

2.2. Features

In [11], best sample-based classifier could not outperform the best peak-based classifier. In this work, CNN based method is adopted to deliver better classification accuracy for GCI on the raw waveform samples. Raw waveform samples in a window surrounding the negative peak (GCI candidates) were taken as features in our proposed method. For the speech waveform sampled at 16 kHz, if the window length is 30ms (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.

2.3. Classifier

Different from the multilayer perceptron (MLP) classifier which directly classify candidates on waveform samples in the window [11], a 9-layer Convolutional Neural Network (CNN) model whose structure is described as Table 1 was used in our proposed method to extract different high-level features from the whole window, followed by a multilayer fully connected network to make the decision based on the extracted features. Cross entropy with L2 norm is used as the loss function.

Table 1: *Detailed Specifications of the Proposed Network for GCI classification*

Layers	**Output shape (channels * time dimension)**	**padding**	**Kernel Size / Strides**
Output	2*1		
Fully Connected	64*1		
Max_Pool_9	512*1	same	1*3 / 2
Conv_9	512*2	same	1*3 / 1
Max_Pool_8	512*2	same	1*3 / 2
Conv_8	512*4	same	1*3 / 1
Max_Pool_7	512*4	same	1*3 / 2
Conv_7	512*8	same	1*3 / 1
Max_Pool_6	512*8	same	1*3 / 2
Conv_6	512*16	same	1*3 / 1
Max_Pool_5	256*16	same	1*3 / 2
Conv_5	256*31	same	1*3 / 1
Max_Pool_4	128*31	same	1*3 / 2
Conv_4	128*61	same	1*3 / 1
Max_Pool_3	64*61	same	1*3 / 2
Conv_3	64*121	same	1*3 / 1
Max_Pool_2	32*121	same	1*3 / 2
Conv_2	32*241	same	1*5 / 1
Max_Pool_1	16*241	same	1*3 / 2
Conv_1	16*481	same	1*7 / 1
Input	1*481		

The number of CNN layers depends on time dimension of the input vector (481). In this work, the number of layers is 9. With this number of layer, the proposed structure can exactly down sample on time dimension from 481 to 1. A schematic view of the resulting network is depicted in Figure 3.

In ERT-P3 algorithm, a set of local descriptors reflecting the position and shape of other 2P neighboring peaks are used as GCI candidate features [11]. We think that this kind of artificial designed features limit the performance of the classifier. In our proposed method, CNN extracts features from raw wave samples, and feeds these high-level features to the following fully connected network for final classification.

3. Experiments and Results

In this section, the proposed method was evaluated in the GCI classification task and the GCI detection task by comparing with other algorithms.

3.1. Speech material

The evaluation of GCI detection methods relies on the ground-truth obtained from EGG recordings. Electroglottography (EGG), also known as electro-laryn-gography, is a non-intrusive technique for measuring the impedance between the vocal folds. The EGG signal is obtained by passing a weak electrical current between a pair of electrodes placed in contact with the skin on both sides of the larynx. This was done in this work by HQTX program of Speech Filing System (SFS) [14] for each database.

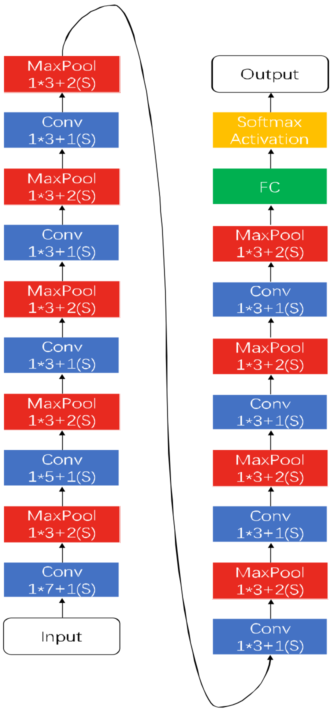

Figure 3: *Network with all the bells and whistles.*

The methods are compared on 6 large corpora containing contemporaneous EGG recordings whose description is summarized in Table 2, where BDL, SLT, KED and JMK database are CMU ARCTIC databases and can be obtained through [12].

Table 2: *Description of Speech Databases*

Database	Number of speakers	Number of utterances	Approximate duration
BDL	1	1132	54 min
SLT	1	1132	54 min
KED	1	452	20 min
JMK	1	1132	55 min
MIX2	2	1200	57 min
MIX4	4	1700	81 min

To prevent CNN-based model from learning speaker-specific features on a one-speaker database, which may cause the network to have poor generalization ability on different speaker datasets, we made a two-speaker database named MIX2 for network training. MIX2 database consists of 800 utterances of BDL database and 400 utterances of JMK database. In the preliminary experiment, we trained model on BDL, SLT, KED and JMK respectively, and then tested the trained model on the other three database. Results indicated the models trained on BDL and JMK got higher classification accuracy. Hence, we construct the MIX2 database by mixing the utterances from BDL and JMK databases.

Furthermore, to verify the effect of speaker characteristics on the classification accuracy, a four-speaker database named MIX4 is further constructed, which consists of 800 utterances of BDL, 400 utterances of JMK, 400 utterances of SLT and 100 utterances of KED database.

3.2. Performance measures

Two kinds of methods are used in our work to measure the performance of the models. One focuses on GCI classification task, and the other focuses on GCI detection task.

3.2.1. GCI classification

Table 3: *Binary Confusion Matrix*

	Positive Class	**Negative Class**
Predict Positive	True Positive (TP)	False Negative (FN)
Predict Negative	False Positive (FP)	True Negative (TN)

GCI classification task is to classify whether or not each GCI candidate is a true GCI. A confusion matrix, depicted in Table 3, is used to show results of binary classification. For this task, three classification accuracy measures are used to measure the performance of the classifier:

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-score = 2*(Precision * Recall) / (Precision + Recall)

3.2.2. GCI detection

GCI detection task is to estimate the location of GCI in a speech signal. The most common way to assess the performance of GCI detection techniques is to compare the estimates with the reference locations extracted from EGG signals. For the task, we here make use of the performance measure defined in [15]. The first three measures describe how reliable the algorithm is in identifying GCIs:

- Identification Rate (IDR): the proportion of larynx cycles for which exactly one GCI is detected.
- Miss Rate (MR): the proportion of larynx cycles for which no GCI is detected.
- False Alarm Rate (FAR): the proportion of larynx cycles for which more than one GCI is detected.

and two indicators characterizing the timing error probability density:

- Identification Accuracy (IDA): the standard deviation of the distribution.
- Accuracy to ± 0.25ms (A25): the proportion of detections for which the timing error is smaller than this bound.

3.3. Compared methods

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

- The ERT-P3 algorithm [11].
- The Dynamic Programming Phase Slope Algorithm

(DYPSA) [9].

- The Speech Event Detection using the Residual Excitation and a Mean-based Signal (SEDREAMS) algorithm [10].

3.4. Results

For performance of ERT-P3, DYPSA and SEDREAMS, we refer to [11].

3.4.1. *Experiment 1*

The MIX2 database was first divided into training set (800 utterances), validation set (100 utterances) and test set (300 utterances), and then used to train CNN-based model. Table 4 shows the classification accuracy of methods on MIX2 test set, SLT set and KED set. Table 5 shows the detection results of methods evaluated on MIX2 test set, SLT set and KED set.

Table 4: *Classification accuracy of models trained on MIX2 training set and tested on MIX2 test set, SLT set and KED set*

Database	**Method**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
MIX2	CNN	97.26	99.22	98.23
SLT	CNN	**95.16**	98.48	96.79
	ERT-P3	95.07	**99.69**	**97.33**
KED	CNN	**93.07**	96.15	94.21
	ERT-P3	92.73	**96.96**	**94.79**

Table 5: *Detection accuracy of models trained on MIX2 training set and tested on MIX2 test set, SLT set and KED set*

Data set	**Method**	**IDR (%)**	**MR (%)**	**FAR (%)**	**IDA (ms)**	**A25 (%)**
MIX2	CNN	92.33	7.01	0.67	0.04	95.65
SLT	CNN	94.87	4.51	**0.62**	**0.03**	**99.46**
	ERT-P3	**95.18**	1.35	3.47	0.15	95.08
	DYPSA	91.50	2.80	5.70	0.30	81.23
	SEDREAMS	92.96	**1.15**	5.89	0.19	89.09
KED	CNN	91.51	6.87	**1.62**	**0.02**	**96.98**
	ERT-P3	**91.88**	2.94	5.18	0.27	88.02
	DYPSA	89.01	4.62	6.37	0.48	83.70
	SEDREAMS	89.54	**1.16**	9.30	0.56	78.46

The results in Table 4 and Table 5 show that the proposed method performs near or better on SLT and KED set than those state-of-the-art algorithms, especially with respect to the False Alarm Rate (FAR), Identification Accuracy (IDA) and Accuracy to ± 0.25ms (A25) metrics. We attribute this to the superior capabilities of CNN in extracting more representative features from raw speech signals. Taking Miss Rate (MR) into consideration, we review the larynx cycles for which no GCI is detected, most of these cycles are located on the border between voiced and silent sections. Speech signal decays sharply and affects the extraction of features.

3.4.2. *Experiment 2*

The MIX4 database was divided into training set (1300 utterances), validation set (200 utterances) and test set (500 utterances), and was used to train CNN-based model. The rest utterances of SLT (732 utterances) and KED (352 utterances) make up the SLT and KED test set for measuring performance on these two databases. Table 6 shows model classification accuracy on MIX4 test set, SLT test set and KED test set. Table 7 shows the detection results of models.

Table 6: *Classification accuracy of models trained on MIX4 training set and tested on MIX4, SLT, KED test sets*

Database	**Method**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
MIX4	CNN	97.26	98.44	97.85
SLT	CNN	**95.80**	**99.71**	**97.72**
	ERT-P3	95.07	99.69	97.33
KED	CNN	**93.40**	**97.12**	**95.23**
	ERT-P3	92.73	96.96	94.79

Table 7: *GCI detection evaluation of models trained on MIX4 training set and tested on MIX4, SLT, KED test sets*

Data set	**Method**	**IDR (%)**	**MR (%)**	**FAR (%)**	**IDA (ms)**	**A25 (%)**
MIX4	CNN	94.67	4.86	0.47	0.03	97.38
SLT	CNN	**97.51**	2.27	**0.22**	**0.03**	**99.47**
	ERT-P3	95.18	1.35	3.47	0.15	95.08
	DYPSA	91.50	2.80	5.70	0.30	81.23
	SEDREAMS	92.96	**1.15**	5.89	0.19	89.09
KED	CNN	**94.61**	5.12	**0.26**	**0.02**	**98.31**
	ERT-P3	91.88	2.94	5.18	0.27	88.02
	DYPSA	89.01	4.62	6.37	0.48	83.70
	SEDREAMS	89.54	**1.16**	9.30	0.56	78.46

The results in Table 6 and Table 7 show that the proposed method perform better on both SLT and KED databases, if the training set includes utterances from speakers in the test set.

By checking locations of missed GCIs, it was found that most missed GCIs were located at the boundary between silence and voiced phoneme. Features of these boundary candidates might not be extracted as easily as candidates in voiced phoneme segments, therefore, missing rate was higher than what we expected.

4. Conclusions

A CNN based model is proposed to detect glottal closure instants (GCIs) from speech waveform. Experiments show that the proposed method performed very well on several test databases and got near or better performance compared to state-of-the-art methods. Model trained on multi-speaker database especially including speech utterances from the speakers that are in the test database would perform better. I n our future work, we would like to concentrate more on those boundary candidates and investigate method that can further reduce the Miss Rate (MR) of the proposed method.

5. Acknowledgement

This work is supported by National Natural Science Foundation of China-Research Grant Council of Hong Kong (NSFC-RGC) joint research fund (61531166002, N_CUHK404/15), National Natural Science Foundation of China (61433018, 61375027) and National Social Science Foundation of China (13&ZD189).

6. References

[1] E. Moulines, F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Communication*, vol. 9, no. 5–6, pp. 453–467, Dec. 1990.

[2] D. Y. Wong, J. D. Markel, J. A. H. Gray, “Least squares glottal inverse filtering from the acoustic speech waveform,” *IEEE Trans.Acoust., Speech, Signal Process.*, vol. 27, no. 4, pp. 350–355, Aug. 1979.

[3] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 21–29, 2001.

[4] T. Drugman, G. Wilfart, T. Dutoit, “A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis,” in *Proc. Interspeech Conference*, 2009.

[5] M. R. P. Thomas, J. Gudnason, P. A. Naylor, “Data-driven voice source waveform modelling,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.

[6] T. Ananthapadmanabha, B. Yegnanarayana, “Epoch extraction from linear prediction residual for identification of closed glottis interval,” *IEEE Trans. Acoust. Speech Signal Proc.*, vol. 27, pp. 309–319, 1979.

[7] Y. M. Cheng, D. O’Shaughnessy, “Automatic and reliable estimation of glottal closure instant and period,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 1805–1815, Dec. 1989.

[8] K. S. Rao, S. R. M. Prasanna, B. Yegnanarayana, “Determination of instants of significant excitation in speech using Hilbert envelope and group delay function,” *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 762–765, 2007.

[9] P. A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, “Estimation of glottal closure instants in voiced speech using the DYPSA algorithm,” *IEEE Trans. Speech Audio Process.*, vol. 15, no. 1, pp. 34–43, 2007.

[10] T. Drugman, T. Dutoit, “Glottal closure and opening instant detection from speech signals,” in *Proc. Interspeech Conference*, 2009.

[11] J. Matoušek, D. Tihelka, “Classification-Based Detection of Glottal Closure Instants from Speech Signals,” *Interspeech Conference*, 2017:3053-3057.

[12] A. W. Blac, “CMU ARCTIC speech synthesis databases,” Internet: http://festvox.org/cmu_arctic/ , Dec. 25, 2017 [Mar. 21, 2018].

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

[14] M. Huckvale, “UCL Speech Filing System,” Internet: http://www.phon.ucl.ac.uk/resource/sfs,
Dec. 4, 2017 [Mar. 21, 2018].

[15] P. A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, “Estimation of glottal closure instants in voiced speech using the DYPSA algorithm,” *IEEE Trans. Speech Audio Process.*, vol. 15, no. 1, pp. 34–43, 2007.

[16] A. Krishnamurthy, D. Childers, “Two-channel speech analysis”, *IEEE trans. on Acoustics, Speech and Signal Processing*, 34:4, pp. 730-743, 1986.