GLOTTAL INSTANTS EXTRACTION FROM SPEECH SIGNAL USING GENERATIVE ADVERSARIAL NETWORK

K. T. Deepak[†]*, Pavitra Kulkarni*[⋆]*, U Mudenagudi*[⋆]*, S. R. M. Prasanna*[⋆†]

[†] Electronics and Communication Engineering Department, IIIT Dharwad
[⋆] School of Electronics and Communication Engineering, KLE Technological University
[⋆†] Electrical Engineering Department, IIT Dharwad
deepak@iiitdwd.ac.in, pavitrakulkarni.pvk@gmail.com, uma@kletech.ac.in, prasanna@iitdh.ac.in

ABSTRACT

The Glottal Closure and Opening instants (GCIs and GOIs) form important events in excitation source signal. These instants represent closing and opening events of vocal folds while producing voiced speech signal. Estimation of such instants from speech signal is beneficial and several applications rely on accurate estimation of Closure and Opening instants. In this work, Electroglottographic like (EGG-like) signal is synthesized from speech signal using Generative Adversarial Network (GAN). The Glottal Closure and Opening instants are located using the derivative of EGG-like signal, which is essentially a difference EGG-like signal. The proposed method is evaluated on CMU-Arctic database, as the database consists of simultaneous recordings of speech and EGG signal, respectively. To evaluate the results, the locations obtained from synthesized EGG-like signal are compared with the reference difference EGG signal. The results are evaluated for both seen and unseen conditions. It is shown that the performance of GCI and GOI estimation is comparable to existing state-of-the-art methods.

Index Terms— Glottal closure instants, glottal opening instants, electroglottograph, generative adversarial network

1. INTRODUCTION

Speech signal is produced as a result of exciting vocal tract system with excitation source. The excitation source consists of different types, among which voiced excitation forms the major type of excitation source. Voiced excitation is due to vibration of vocal folds, which essentially is a quasi periodic cyclic oscillations [1]. The oscillations consists of closing and opening phase of vocal chord. However, during such phases, there are time instants at which the excitation source registers the near complete closing and opening of vocal folds called Glottal Closure Instant (GCI) and Glottal Opening Instant (GOI), respectively. The excitation source gets spectrally modified by vocal tract response to produce voiced speech signal [2].

Accurate estimation of GCIs and GOIs from speech signal is of interest to speech community. There are many applications in the literature that involve these instants for processing [3–5]. However, it is difficult to estimate the accurate position of GCIs and GOIs directly from speech signal. There are many methods proposed in the literature for robust estimation of GCIs from speech signal [6–8]. Majority of the proposed methods are signal processing based approaches. Most of the methods require the estimation of fundamental frequency from speech signal [6]. The tuning of parameters become difficult as the characteristics of the speech signal varies and recording scenario changes. For example, the parameters tuned for normal speaking scenario may not give the same performance in case of emotional or singing voice [9, 10].

It is possible to estimate GCIs and GOIs using several invasive methods precisely [11]. However, it is not practically possible to use such invasive techniques for most of the applications. Alternatively, non-invasive techniques like Electroglottography (EGG) are equally reliable for the estimation of GCIs and GOIs. A EGG equipment is needed to record EGG signal and a pair of electrodes need to be strapped around the neck of a speaker. Again it is not always possible to carry EGG equipment and record EGG signal. With the introduction of efficient deep learning algorithms and increased computational performance, it should be possible to synthesize EGG signal from raw speech signal. Generative Adversarial Network (GAN) is one such deep learning technique that can help synthesize the EGG signal [12]. The EGG signal can be derived directly from raw speech signal without the necessity for hand-crafted features using GAN.

In this work, a synthesized EGG-like signal is obtained from trained models of GAN. EGG signal has more attributes than GCIs and GOIs. Since we have not fully characterized the synthesized EGG, we are limiting ourselves to call this as synthesized EGG-like signal. The GCIs and GOIs are estimated from the derivative of synthesized EGG-like signal, as in case of EGG signal [13]. The CMU-Arctic database is used for both training and testing [14]. The proposed work

is compared with state-of-the-art techniques available in the literature.

The rest of this paper is organized as follows: In Section 2, the proposed method for GCI and GOI extraction is discussed. In Section 3, we present the evaluation of the proposed scheme. Finally, the paper is summarized and concluded in Section 4.

2. ELECTROGLOTTOGRAPH-LIKE SIGNAL SYNTHESIS

Electroglottography is a non-invasive device used to measure the vocal folds contact area (VFCA) [15]. Although this may not be an accurate measurement, however it is sufficient for many applications including speech pathology [16]. The EGG device consists of two electrodes strapped across the neck of the speaker near the larynx. Due to vibration of vocal folds vertically, the contact area between them varies during oscillatory cycles. The oscillatory pattern is essentially made of near complete closure (contact between vocal folds) and near complete opening (no contact between vocal folds) cycles. The electrical impedance between electrodes is lowest when VFCA is maximum and largest when VFCA is minimum, respectively. Recording of impedance waveform as a function of time is called EGG signal. There is sufficient correlation between EGG signal and the actual VFCA measurement during the production of voiced speech signal [17]. The usual practice is to record both EGG and speech signal for analysis purpose. There is a correspondence between EGG and speech signal in terms of closing phase, opening phase, GCIs, GOIs and instantaneous fundamental frequency. However, as stated earlier it is difficult to carry EGG device for recording EGG signal. It is therefore beneficial to synthesize EGG waveform directly from speech signal.

In this work, an attempt is made to relate these two signals using Generative Adversarial Network (GAN). The Generative model is trained to synthesize EGG-like signal from the corresponding speech signal as input.

2.1. Speech Enhancement Generative Adversarial Network for EGG-like Signal Synthesis

GANs are promising deep learning networks that can learn the mapping from one distribution to another. It was first proposed in [12] for face and digit synthesis in the form of images. A modified network was proposed for speech enhancement and it is named as Speech Enhancement Generative Adversarial Network (SEGAN) [18]. The objective is to map the noisy speech signal distribution to a learned clean speech distribution and thereby synthesize the clean speech signal. Motivated by this, the proposed work enables the Generative network to synthesize EGG-like signal from speech signal.

Let speech samples x belong to to some prior distribution $\mathcal{X}$. The objective of SEGAN in this work is to learn the mapping of samples x to samples e from another distribution $\mathcal{E}$ that belong to EGG signal. The SEGAN is a combination of two networks where one is discriminator called (D) and the other network is called generator (G). Both networks acts adversarial to each other, where G tries to outwit D by faking the training data, while D tries to figure out the real against the fake samples generated by G. The network is having double feedback using which the weights are adjusted through back-propagation algorithm. The proposed EGG synthesis can be defined as the speech signal x which is fed as input to network G for obtaining the synthesized EGG-like signal e. The relationship between the speech and synthesized EGG-like signal is given by the following relationship.

$$e = G(x) \tag{1}$$

The G network is fully a convolutional layer, the advantage of using such a network is that it reduces the number of parameters and thereby reducing the training time. The proposed work is an end to end system, where the network is trained using raw audio files. The training is unsupervised as there are no labels involved in it.

2.2. Glottal Closure and Opening Instants Estimation from Synthesized EGG Signal

The VFCA consists of two oscillatory phases, *viz.,* closing and opening during voiced speech production. However, the instants at which the near complete closing and opening of vocal folds play a major role in several applications. These instants are called Glottal Closure and Opening Instants. It is crucial to estimate the accurate time instants at which these events occur. One such reliable measurement can be obtained using non-invasive EGG signal. The proposed work attempts to estimate these locations using synthesized EGG-like signal as mentioned in Section 2.1.

The Figure 1 illustrates the synthesis of EGG-like signal. Figure 1(a) shows a segment of voiced speech signal from a female speaker taken from CMU-Arctic database [14]. Figure 1(b) shows the EGG waveform recorded simultaneously. It can be noticed that the waveform essentially is a measurement of closing and opening phase. However, it can be noticed that GCIs and GOIs are not evident from the EGG signal. Figure 1(c) shows the difference EGG signal, where the red arrow marks indicate the GCI locations and green arrow marks indicate GOI locations, respectively. The Figure1(d) is the synthesized EGG-like signal generated by the generator network for the input speech signal. It can be noticed that the signal is similar to EGG waveform. Therefore, the GCI and GOI locations can be obtained by taking the difference EGG-like signal as shown in Figure 1(e). Note that there is a close correspondence between difference EGG and synthesized EGG-like signal in terms of GCI and GOI locations.

The GCIs and GOIs are obtained using Singularity detection In EGG with Multiscale Analysis (SIGMA) algorithm

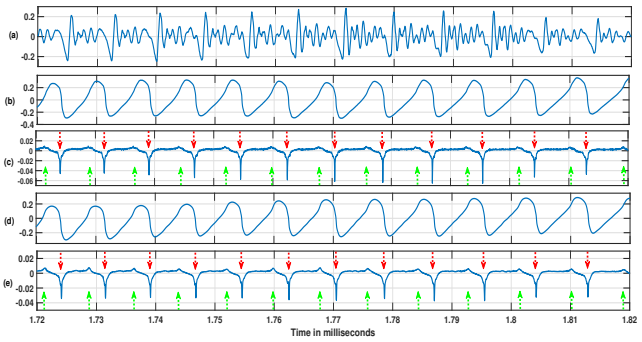

Fig. 1: Illustration of synthesized EGG-like signal. (a) speech signal taken from vowel region, (b) Electroglottograph signal, (c) difference EGG signal, (d) synthesized EGG-like signal using SEGAN, (e) difference synthesized EGG-like signal. Red arrow indicates the locations of Glottal Closure Instants and green arrow indicates the locations of Glottal Opening Instants.

from difference EGG [13]. The GCI and GOI locations obtained from difference EGG signal using SIGMA is considered as reference locations. The procedure is repeated for synthesized EGG-like signal and compared with reference locations for evaluation.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The training and testing of SEGAN is done using CMU-ARCTIC database. The database consists of 2 channel recordings of speech and EGG signal, respectively. Both channels are recorded at a sampling rate of 32 kHz. In this work, a total of 4 (3 Male and 1 Female) speakers data is considered for the experiments. Each speaker's data consists of 1150 utterances.

Two different trained models are considered for experimental purpose. The seen condition is trained and tested using the same speaker's data. However, in unseen case both trained and testing files are exclusive to each other. In order to train the models in both cases 80 epochs are used, while learning rate is set as 0.0002 with a batch size of 100. The generator network consists of 22 one dimensional convolutional layers that has kernel size of 31 and stride size is set as 2.

The Figures 2 and 3 shows the plots of wavegram for a 10 ms segments taken from original EGG signal and synthesized EGG-like signal, respectively. The wavegram is a visual representation of EGG signal and difference EGG signal. This essentially captures the variations with reference to each glottal cycle as shown in the graph (ii) which is similar to spectrogram plots [19]. Subsequently, (iii) and (iv) shows the F0 contour and amplitude variation in decibels. It can be observed that both EGG and EGG-like signal demonstrates similar pattern in all 3 plots. However, this requires a detailed analysis for a sufficiently larger data set. It will be of interest to extend this similarity study to normal, emotional, singing, and pathological conditions.

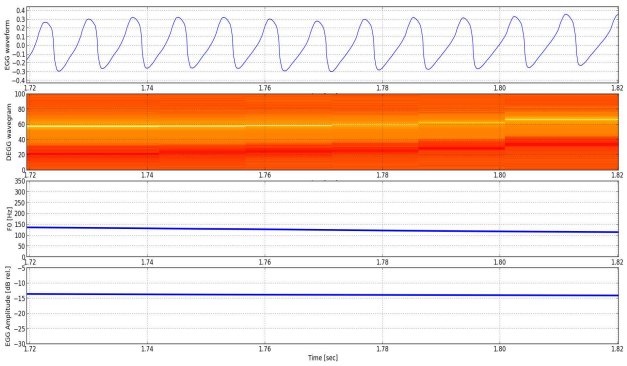

Fig. 2: 10 milliseconds of (i) EGG signal from voiced region, (ii) Wavegram plot of EGG signal, (iii) F0 contour, and (iv) EGG signal amplitude in decibels.

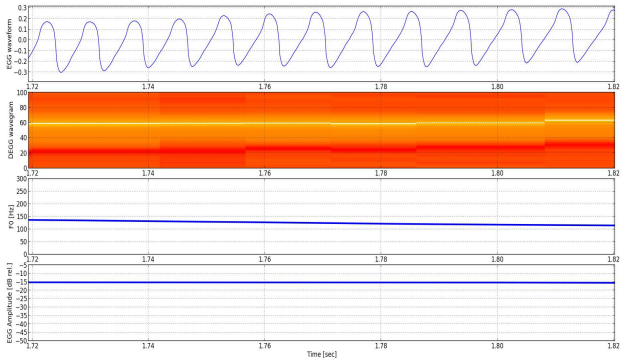

Fig. 3: 10 milliseconds of (i) Synthesized EGG-like signal from voiced region, (ii) Wavegram plot of EGG-like signal, (iii) F0 contour, and (iv) EGG-like signal amplitude in decibels.

The GAN network learns the data distribution rather than the signal characteristics. One of the drawbacks of such modeling technique is that it averages the synthesized signal, while detailed characteristics might be lost. For example, there are chances of missing double or triple peaks of GCIs and GOIs in difference EGG and therefore we limit to call this synthesized signal as EGG-like signal.

The Table 1 shows the GCI and GOI estimation performance using SEGAN EGG for seen and unseen conditions. In order to evaluate the performance, the GCI and GOI locations obtained from original EGG signal is considered as

Table 1: Performance evaluation of GCI and GOI estimation for seen and unseen conditions. Note: IDR, MR, and FAR are expressed in $\%$ while IDA in ms.

model	IDR	MR	FAR	IDA
Seen GCI	97.52	1.41	1.83	0.10
Seen GOI	97.09	1.98	0.43	0.34
Unseen GCI	96.99	2.44	0.57	0.20
Unseen GOI	96.68	2.49	0.84	0.46

Table 2: Performance evaluation of GCI estimation. Note: IDR, MR, and FAR are expressed in $\%$ while IDA in ms.

Method	IDR	MR	FAR	IDA
HE	95.24	2.61	2.15	0.65
DYPSA	97.06	1.49	1.45	0.42
SEDREAMS	98.79	0.45	0.76	0.34
YAGA	98.66	0.37	0.96	0.33
ZFF	95.19	3.13	1.68	0.44
SEGAN(seen)	97.52	1.41	1.83	0.10
SEGAN(unseen)	96.99	2.44	0.57	0.20

ground truth. The reference locations are obtained by passing difference EGG signal as input to SIGMA algorithm. Similarly, the GCI and GOI locations are obtained by passing difference EGG-like signal through SIGMA algorithm. The performance estimation of GCI and GOI are evaluated in terms of identification rate (IDR), miss rate (MR), false alarm rate (FAR), and identification accuracy (IDA). The details of these parameters are explained in [6]. The reported results are computed by taking average performance of all speech files in testing.

The CMU-Arctic database consists of 4 speakers data set *viz.,* BDL, JMK, SLT, and KED. Amonst which BDL, JMK, and KED are male speakers, while SLT is a female speaker data set. In case of seen condition, 80% of speech files from all speakers are used for training the GAN network, while 20% is used for testing the performance. However, in case of unseen conditions the BDL data set is used for training the model, while other speaker's data is used for testing. Note that the test consists of cross gender in case of unseen condition. It can be noticed that the performance of both seen and unseen conditions are similar in terms of IDR, however IDA is relatively better in case of seen condition compared to unseen condition.

The Table 2 shows the comparison between different state-of-the-art methods for GCI estimation. The performance is compared with Hilbert Envelope, DYPSA, SEDREAMS, YAGA, and ZFF [20–24]. Note that the results of other methods were taken from [6]. Most recently a method is proposed to estimate the GCI locations from raw speech signal using Deep Dilated Convolutional Neural Networks [25]. The method makes use of dilated convolutional neural networks (DCNN) for feature extraction, while 2 fully connected networks are used for regression and classification. This method localizes on GCI locations directly, unlike deriving EGG signal from speech signal as proposed in the current approach. However, the DCNN based approach is not compared with the current approach. It can be observed from Table 2 that the performance of the proposed method is comparable in terms of IDR. However, the proposed method is better in terms of identification accuracy compared to other methods.

Table 3: Performance evaluation of GOI estimation. Note: IDR, MR, and FAR are expressed in $\%$ while IDA in ms.

Method	IDR	MR	FAR	IDA
MBS+LPres	98.23	1.36	0.55	0.98
I. DYPSA	95.00	1.90	3.09	1.09
SEGAN (seen)	97.09	1.98	0.43	0.34
SEGAN (unseen)	96.68	2.49	0.84	0.46

Table 3 shows the GOI estimation performance compared to other methods. Compared to GCI, very few existing methods extract GOI in the literature. GOI extraction is difficult compared GCI as it is less pronounced and has relatively low peaks. In case of Mean Based Signal (MBS) approach, the GOI extraction is dependent on GCI locations [22]. However, the GOI estimation is independent to GCI locations in the proposed method. It can be observed that the proposed method is comparable in its performance in terms of identifying the GOI locations. It can be noticed that the proposed method has slight advantage in terms of identification accuracy compared to other methods.

4. SUMMARY AND CONCLUSIONS

The proposed work is an attempt to synthesize EGG signal from speech signal using Generative Adversarial Network. We stop short of calling the synthesized signal as EGG signal, since all the attributes of EGG are not verified. The study is limited to extract GCIs and GOIs from synthesized EGG-like signal. It is observed that the performance is comparable to other state-of-the-art methods in the literature in terms of identification rate. However, the estimation accuracy is comparatively better. Only clean speech signal were considered for the evaluation. Hence, the robustness of the proposed work needs further exploration. Also, it is of interest to evaluate the performance of the proposed work in varying pitch scenario like emotional speech, singing voice, and pathological conditions. The proposed work seems to be promising and other attributes of EGG needs to be evaluated.

5. REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *An Introduction to Digital Speech Processing (Foundations and Trends in Signal Processing)*, Now Publishers Inc, 2007.

[2] Mark Tatham and Katherine Morton, *A Guide to Speech Production and Perception*, Edinburgh University Press, Edinburgh, 2011.

[3] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded condition," *IEEE Trans. Audio, Speech and Lan-*

uage Processing, vol. 19, no. 8, pp. 2552–2565, May 2011.

[4] B. Yegnanarayana, S. R. M. Prasanna, J. M. Zachariah, and S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed text speaker verification system," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 4, pp. 575 – 582, July 2005.

[5] K. T. Deepak, Biswajit Dev Sarma, and S. R. Mahadeva Prasanna, "Foreground speech segmentation using zero frequency filtered signal," in *Interspeech*, September 2012.

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

[7] Andreas I. Koutrouvelis, George P. Kafentzis, Nikolay D. Gaubitch, and Richard Heusdens, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 2, pp. 316–328, 2016.

[8] M. V. Achuth Rao and Prasanta Kumar Ghosh, "PSFM - A probabilistic source filter model for noise robust glottal closure instant detection," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 26, no. 9, pp. 1645–1657, 2018.

[9] S. R. M. Prasanna and D. Govind, "Analysis of excitation source information in emotional speech," in *Interspeech*, September 2010.

[10] K. T. Deepak and S. R. M. Prasanna, "Epoch extraction using zero band filtering from speech signal," *Circuits, Systems, and Signal Processing*, vol. 34, pp. 2309–2333, December 2014.

[11] Hans Larsson, *Methods for measurement of vocal fold vibration and viscoelasticity*, Ph.D. thesis, Karolinska Universitetssjukhuset, 2009.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NIPS)*, p. 2672–2680., 2014.

[13] M. R. P. 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, pp. 1557–1566, 2009.

[14] John Kominek and Alan W. Black, "The cmu arctic speech databases.," in *SSW*, 2004.

[15] A. S. Krishnamurthy and D. G. Childers, "Two-channel speech analys," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 730–743, August 1986.

[16] Vit Hampala, Maxime Garcia, Jan G. Svec, Ronald C. Scherer, and Christian T.Herbst, "Relationship between the electroglottographic signal and vocal fold contact area," *Journal of Voice*, vol. 30, no. 2, pp. 161–171, 2016.

[17] Martin Rothenberg, "A multichannel electroglottograph," *Journal of Voice*, vol. 6, no. 1, pp. 36–43, 1992.

[18] Pascual, Santiago, B. Antonio, and S. Joan, "Segan:speech enhancement generative adversarial network," in *Interspeech*, August 2017.

[19] Christian T. Herbstb, W. Tecumseh S. Fitch, and Jan G. Švec, "Electroglottographic wavegrams: A technique for visualizing vocal fold dynamics noninvasively," *J. Acoust. Soc. Am.*, vol. 128, no. 5, pp. 3070–3078, 2010.

[20] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *IEEE Signal Process. Letters*, vol. 14, pp. 762–765, October 2007.

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

[22] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Interspeech*, 2009, pp. 2891–2894.

[23] M. R. P. Thomas, J. Gudnanson, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 20, pp. 82–91, June 2012.

[24] K. Sri Rama Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 16, pp. 1602–1613, November 2008.

[25] Prathosh A. P., Mohit Goyal, and Varun Srivastava, "Detection of glottal closure instants using deep dilated convolutional neural networks," *IEEE Signal Processing Letters*, vol. abs/1804.10147, 2018.