Glottal Instants Extraction From Speech Signal using GAN

A Mini-Project Report

Submitted by

Rohan Jijju	181EC138
Roshan Rangarajan	181EC139
Bhairavi Giridharan	181EC208
Harish Bachu	181EC214

Under the guidance of
Dr. A. V. Narasimhadhan

as apart of the course
Speech and Audio Processing (EC347)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SURATHKAL, MANGALORE - 575025

Contents

1 Abstract

The Glottal Closure Instants (GCIs) and Glottal Opening Instants (GOIs) represent the instants in time when the vocal folds close and open respectively while producing voiced signals. Many applications rely on the accurate estimation of Opening and Closure instants. In the following sections, the Electroglottographic like (EGG-like) signal is synthesized from speech signal using Generative Adversarial Networks (GAN). To identify the Glottal Closure and Opening Instants, we first compute the derivative of the EGG-like signal, which is essentially a difference EGG-like signal. The database we are using 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 obtained from the Speech Enhancement Generative Adversarial Network (SEGAN) are of appreciable accuracy.

2 Introduction

Speech signals are produced by excitations in the vocal tract. The voiced speech signals are mainly of interest to us in locating the opening and closure of the vocal folds which correspond to the GCO and GCI respectively. This is useful in the speech community in many applications.

Most of the earlier techniques used to detect the GCIs and GOIs have relied on fundamental frequency and are largely invasive. In most applications of invasive techniques are not very useful so we present a non invasive approach using electroglottography (EGG) that is equally reliable for the estimation of GCIs and GOIs. A pair of electrodes are strapped to the speaker's neck to obtain a EGG plot.

Since this equipment is not easy to carry around and not always available, we suggest a deep learning approach to obtain the EGG signal using the raw speech signal. In this work, we have used a Generative Adversarial Network proposal: SEGAN, to convert raw speech data to an EGG signal for determining GCIs and GOIs. Original EGG plots contain much more information than just the closing and opening instants so we are calling the plot obtained using the GAN as an EGG-like signal. To identify the GCIs and GOIs, we compute the derivative of the EGG-like signal. The results of the SEGAN model is compared with the performance of an Autoencoder model as well.

3 Literature Survey

There has been a huge amount of research conducted in the fields of speech enhancement and identification of Glottal Closing and Opening Instants. Various techniques and algorithms have been developed and implemented to improve the accuracy and performance.

In [6], the Dynamic Programming Projected Phase-Slope Algorithm has been used to automatically estimate the Glottal Closure Instants from voiced speech. Estimation doesn't require an EGG signal and instead the algorithm uses a phase-slope function along with a projection technique to estimate GCI from the speech signal.

In [1], a fully convolutional neural network is used to map the speech waveform to a target signal from which GCIs are obtained through peak picking. They train the CNN using high quality synthetic speech which has perfect ground truth to overcome problems which occur through training with imperfect EGG signals.

In [4], the GCI detection problem is approached through a supervised multi-task learning approach. This is then solved using a CNN which is optimized by minimizing the classification and regression cost.

In [5], GCIs and GOIs are estimated along with high resolution Voiced/Unvoiced Detection. This is achieved through an algorithm where the structure of the glottal flow derivative is exploited to estimate the two opening and closing instants by using simple time-domain criteria.

In [3], different GCI detection algorithms are compared and analysed. From the metrics calculated for each technique, it was observed that Speech Event Detection using the Residual Excitation And a Mean-based Signal algorithm and Yet Another GCI Algorithm performed best in terms of accuracy and identification rate.

In [7], Generative Adversarial Networks are used to perform speech enhancement. The training is done at the waveform level, and uses 28 speakers and 40 different noise conditions.

In [2], a mean based signal is computed. Further the intervals where speech events should occur are estimated and a precise position is found by locating a discontinuity in the linear prediction residual.

In this paper, a SEGAN is used to generate an EGG-like signal. Further this is used to calculate GCI and GOIs through the SIGMA algorithm.

4 Background

The dataset used is the CMU_ARCTIC dataset. It was created at the Language Technologies Institute at Carnegie Mellon University. It consists of single speaker speech database and have nearly 1150 phoenetically balanced English utterances. It consists of 4 main sets of recordings, BDL, JMK, KED, and SLT. The first three (BDL, JMK and KED) are male speech and SLT is female speech. The speech signals were recorded at 16kHz along with EGG signals. We will be using data from the speakers BDL, JMK, and SLT from the dataset for both training and testing the model.

5 Methodology

5.1 Models used

5.1.1 Autoencoder Model

An autoencoder model with skip connections was built for EGG construction, with an architecture similar to the Generator of the SEGAN model. The autoencoder takes a speech segment as input, and outputs the constructed EGG signal. Then we used a mean absolute error to find the loss between our constructed EGG signal and the expected one. However, the fault with using errors such as mse and mae is that they might try to map into an average value of EGG signal, and might produce a jittery output as it prioritizes reducing loss instead of creating a realistic looking EGG signal. To overcome this disadvantage, the following model is used.

5.1.2 SEGAN Model

The SEGAN model proposed in [7] consists of a GAN-like model which helps map the input speech signal into an EGG signal. Here the discriminator is used to provide a better differentiation between real and fake EGG signals, and will push the generator to learn more realistic looking feature for the constructed EGG signal.

The generator for the GAN is an autoencoder model with skip connections that maps the preemphasized speech to an output. Preemphasis of speech is achieved by applying the following filter-

$$y[n] = x[n] - \alpha x[n-1] \tag{1}$$

Where α is a constant set to 0.97 here. The discriminator network maps the generator output to a single tanh output in the range [-1,1]. The GAN is modelled after the WGAN, with the exception of the final output node having a tanh activation instead of the standard linear activation. This is to limit the value of the Wasserstein loss.Thus the discriminator loss becomes-

$$loss_D = y_{true} \cdot y_{pred} \tag{2}$$

Where y_{true} is -1 for real EGG samples, and +1 for the generator EGG samples. Thus the network is made to predict +1 for real samples and -1 for generated samples in order to minimize the discriminator loss.

To map the input speech into its corresponding egg signal, instead of learning a random mapping, an L1 normalization term is added to the standard generator loss. Thus the generator loss becomes-

$$loss_G = -\lambda D(G(x)) + ||G(x) - e||_1 \tag{3}$$

Where x is the input speech signal, e is the required egg signal, λ is a hyper-parameter to control effect of the added L1 norm term(set to 0.01 for our model), $G(x)$ is the output of the generator and $D(G(x))$ is the output of the discriminator on the generated sample.

Thus we create a model that should be able to create accurate and realistic looking EGG signals from a corresponding speech signal.

5.2 Performance Analysis

In order to quantitatively analyze the result of the constructed EGG signal, we have used a number of metrics which compare the occurrences of GCI's and GOI's of the reconstructed EGG signal against the original EGG signal. The method used for getting the glottal instants from the EGG signal is the SIGMA algorithm[8]. The SIGMA algorithm is especially robust in the transition at the end of voicing periods, and has comparable or superior results when compared with other glottal extraction algorithms. The flowchart for the SIGMA algorithm is shown in Fig. 1.

The modules of the SIGMA algorithm are as follows-

- Stationary Wavelet Transform: The stationary wavelet transform(SWT) is used to get the multiscale product. The multiscale product is used to convert the EGG signal into glottal singularities. The positive rectified and negative rectified multiscale product is used for detection of GCI's and GOI's respectively.
- Group Delay Function: The group delay function is used to detect peaks from the output of the SWT module. To provide robustness against noise, the energy weighted group delay is used. Then the zero crossings of the energy weighted group delay give the candidates for glottal instants.

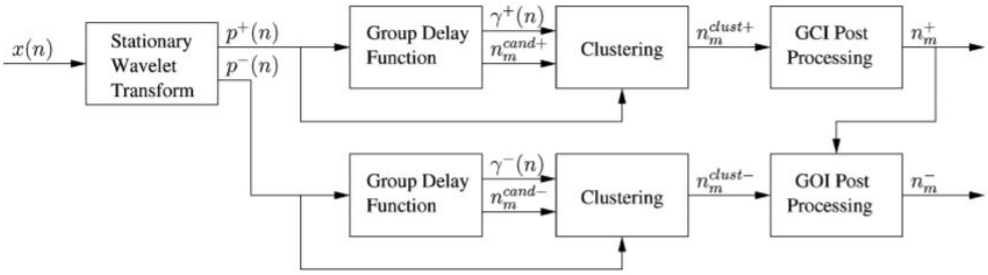

Figure 1: Flowchart of the SIGMA model

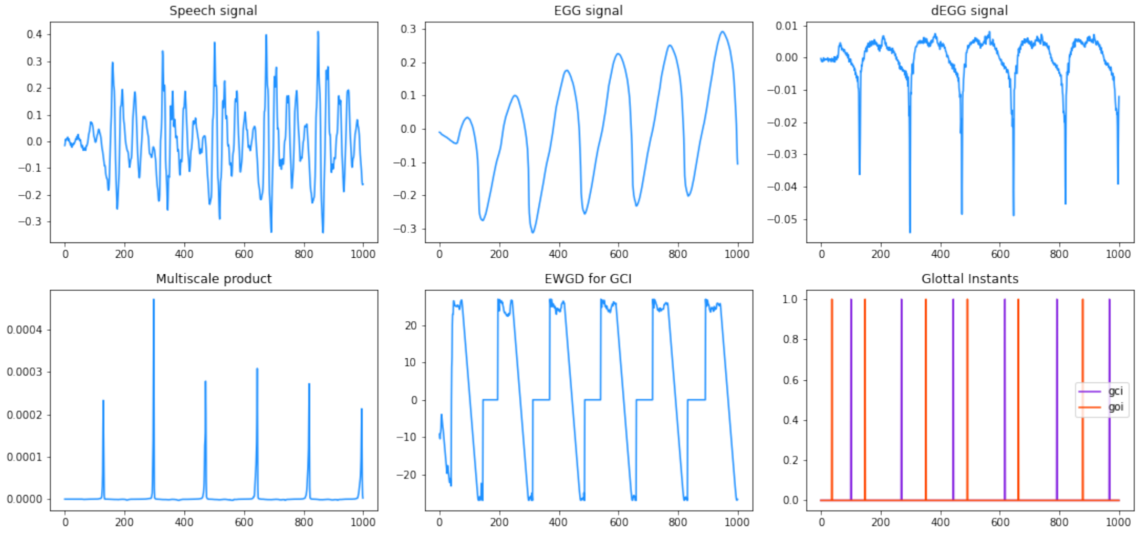

Figure 2: Results at different stages of the SIGMA algorithm on an input EGG signal.

- Clustering: To get rid of false detections, feature vectors are constructed for each glottal instant candidate, and a Gaussian Mixture Model is used to select true glottal instants from erroneous detections.
- GCI Post Processing: To get rid of erroneous candidates caused by swallowing or other disturbances in between utterances, GCI candidates that are isolated from other GCI candidates are removed.
- GOI Post Processing: Since GOI is harder to detect and more prone for false detection that GCI, GCI locations are used to refine the output of the GOI candidates. A single GOI candidate corresponding to each GCI candidates is picked, and if none is found GOI candidate is inserted for the GCI candidate.

The results of various stages in the SIGMA process is shown in Fig. 2.

5.3 Performance Metrics

The performance of the models in generating the EGG signal is determined by their locations of GCI and GOI instants given by the SIGMA algorithm. For each input signal, we have the ground truth EGG signal, from which the correct location of Glottal instants can be extracted. Using each model, we generate a predicted EGG signal, and extract glottal instants from the predicted signal. From the ground truth and predicted signal glottal instants, we can analyse the performance, using the following metrics:

- Identification Rate (IDR): This refers to the proportion of glottal cycles for which only 1 Glottal Instant is detected. Higher Identification rate implies a better prediction.
- Miss Rate (MR): This refers to the proportion of glottal cycles for which no Glottal instant is detected. Lower miss rate is better.
- False Alarm Rate (FAR): This refers to the proportion of glottal cycles for which more than one glottal instant is detected. A lower value is better.
- Identification Accuracy (IDA): This metric tries to find a timing error between the location of the ground truth Glottal Instant, and the predicted Glottal Instant. The lower this value, the better.

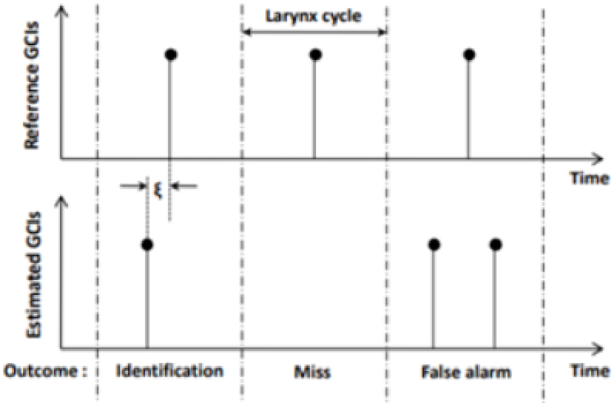

Figure 3: Naylor Metrics

6 Model Framework

6.1 Autoencoder model

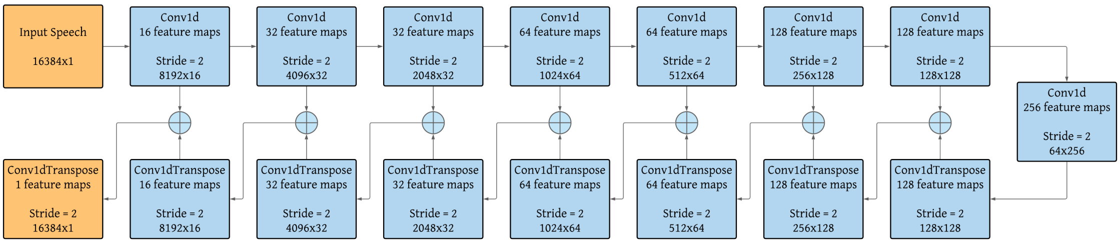

Figure 4: Autoencoder flowchart

The autoencoder model we have used is an exact copy of the SEGAN generator network, without the latent noise sample concatenation at the central encoded layer. It's encoding layers have a depth of 10 layers, each layer downsampling the input signal, using a filter width of 31, with a stride of 2 and 'same' zero padding. The decoding layers are of similair dimensions as the encoding layers, to facilitate skip connections from earlier layers.

The input speech signal is given to an autoencoder model with skip connections to stop excessive loss of information due to vanishing gradients, and to facilitate learning from lower level features. The encoding layers are implemented by using 1 dimensional strided convolutions and the decoding layers are implemented using 1 dimensional fractionally strided Convolutions layers. All hidden layers are given a "PReLU" activation function, and the final layer is given a tanh activation function.

The model is evaluated using a Mean Absolute Loss error, and optimized using an Adams Optimizer with a learning rate of 0.01

6.2 SEGAN model

The input speech preemphasized signal is given to a generator, which tries to map it into the corresponding EGG signal. It is aided by the discriminator, which helps make the output of the generator look more EGG-like by discriminating between real and fake egg signals.

The generator model is shown in Fig. 6. It consists of an autoencoder style network followed by some convolutional layers. The autoencoder layers have a 5 layer deep encoding and decoding parts respectively, and are implemented by using Conv1d and Conv1dTranspose layers. Maxpooling and upsampling has been replaced by using a stride of 2 and 'same' zero padding in the convolutional layers. Layers in the encoding

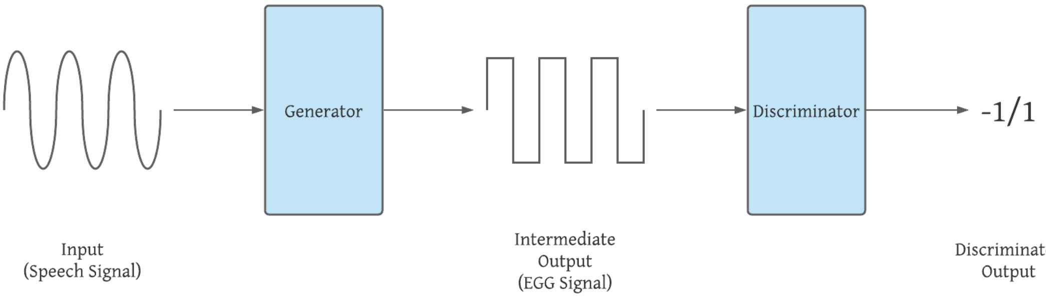

Figure 5: SEGAN flowchart

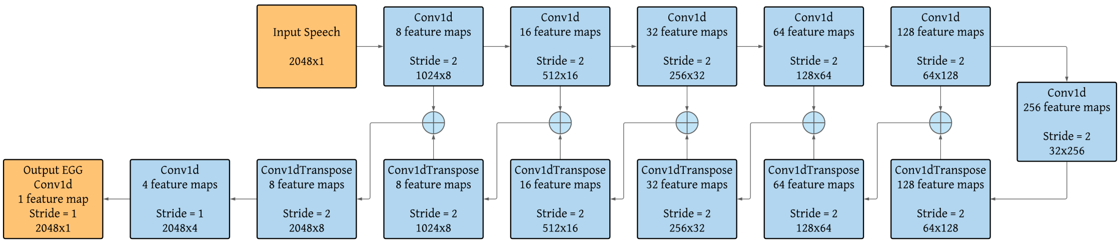

Figure 6: Generator Model

part are concatenated with layers in the decoder using skip connections to avoid too much loss of information while encoding. Each Conv1d and Conv1dTranspose layer except the last one is followed by a batch normalization layer and a PReLU activation function. The last convolutional layer, which is used to create the output, is a normal convolutional layer followed by a "tanh" activation function. Each convolutional layer has a kernel size of 32 and is padded to maintain size(except for the stride).

The discriminator model is shown in Fig. 7. It take the output of the generator, and passes it through a ReLU activated convolutional layer. The layer is then batch normalized and flattened, after which it is passed to a fully connected ReLU activated layer. It is finally connected to a single node which is activated with a tanh activation.

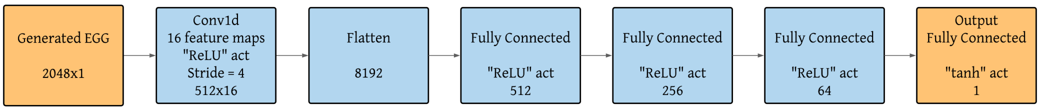

Figure 7: Discriminator model

7 Results and Discussions

7.1 Autoencoder Model

The Autoencoder model was trained on a dataset of 3000 speech signals sampled at 16KHz trimmed to a window size of 16384 samples (1 second). The model was trained for 150 epochs with a learning rate of 0.01. The real GCI and GOI results for Autoencoder are shown in in Fig. 8 and Fig. 9.

7.2 SEGAN

The SEGAN model was trained for a window size of 2048(0.125 ms). To reduce noise in the generator output, the generator output is passed through a butterworth filter for smoothening. The results of the smoothened SEGAN generated EGG run on the sigma algorithm for 2000 samples is shown in Fig. 10. The SEGAN network is able to create realistic looking EGG signal with corresponding glottal instants.

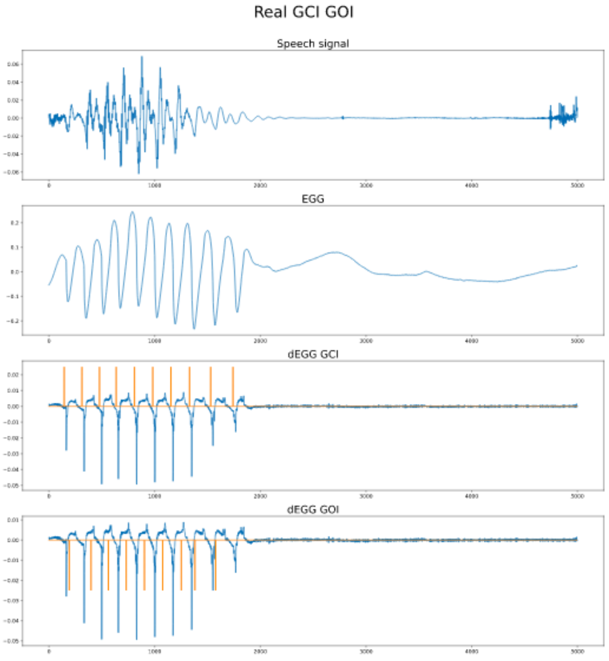

Figure 8: Real GCI GOI positions for Autoencoder

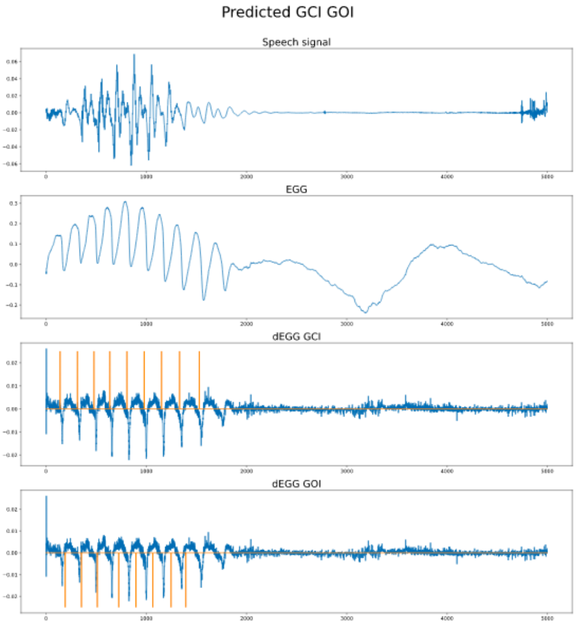

Figure 9: Predicted GCI GOI positions for Autoencoder

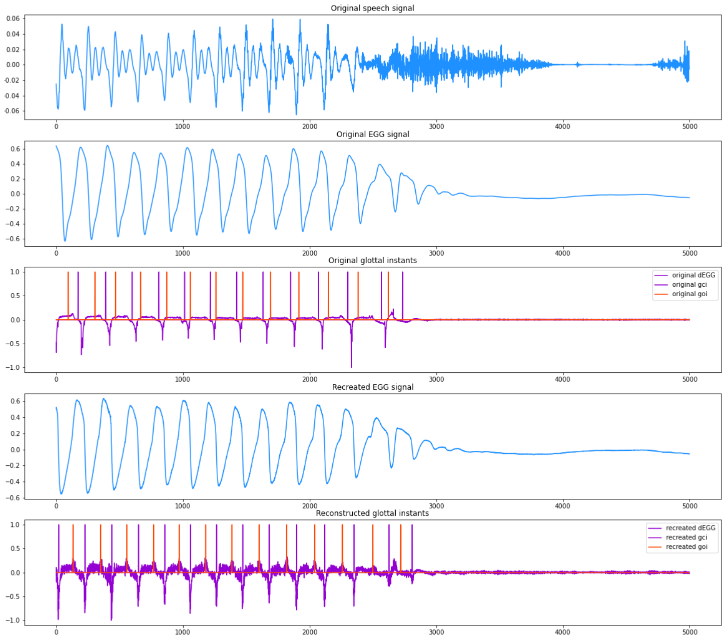

Figure 10: SEGAN Real and Predicted GCI and GOI

7.3 Quantitative Analysis

As mentioned before, we use Naylor metrics for quantitative analysis. The results of quantitative analysis on the Autoencoder and SEGAN model for GCI are shown in table 1 and for GOI are shown in table 2.

Table 1: GCI Quantitative analysis

Quantitative Metrics	**Model**	
	Autoencoder	***SEGAN***
IDR	0.8997	0.9378
MR	0.0865	0.0350
FAR	0.0137	0.0404
IA	0.0013	0.0004

Table 2: GOI Quantitative analysis

Quantitative Metrics	**Model**	
	Autoencoder	***SEGAN***
IDR	0.9297	0.9221
MR	0.0641	0.0542
FAR	0.0061	0.0235
IA	0.0003	0.0004

8 Conclusion

The GAN is able to synthesize the EGG signal with a higher degree of accuracy, and does not show drastic variations at lower frequencies, unlike the Autoencoder.

As a result, the GAN has a higher Identification Rate than the Autoencoder. This makes it a more reliable tool for accurate identification of Glottal Opening and Closing Instants.

9 Future works

The future works include the following:

- We could further try making our model robust enough to work on singing and emotional speech.

- Another improvement that can be implemented is the ability of the EGG-like signal to pick up other attributes as well apart from the GCIs and GOIs alone. This may give more accurate results.

References

[1] Luc Ardaillon and Axel Roebel. “GCI Detection from Raw Speech Using a Fully-Convolutional Network”. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 6739–6743. DOI: 10.1109/ICASSP40776.2020.9053089.

[2] Thomas Drugman and Thierry Dutoit. *Glottal Closure and Opening Instant Detection from Speech Signals*. 2019. arXiv: 2001.00841 [cs.SD].

[3] Thomas Drugman et al. “Detection of Glottal Closure Instants From Speech Signals: A Quantitative Review”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.3 (2012), pp. 994–1006. DOI: 10.1109/TASL.2011.2170835.

[4] Mohit Goyal, Varun Srivastava, and A P Prathosh. “Detection of Glottal Closure Instants from Raw Speech Using Convolutional Neural Networks”. In: Sept. 2019, pp. 1591–1595. DOI: 10.21437/Interspeech.2019-2587.

[5] Andreas I. Koutrouvelis et al. “A Fast Method for High-Resolution Voiced/Unvoiced Detection and Glottal Closure/Opening Instant Estimation of Speech”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.2 (2016), pp. 316–328. DOI: 10.1109/TASLP.2015.2506263.

[6] Patrick A. Naylor et al. “Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.1 (2007), pp. 34–43. DOI: 10.1109/TASL.2006.876878.

[7] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. *SEGAN: Speech Enhancement Generative Adversarial Network*. 2017. arXiv: 1703.09452 [cs.LG].

[8] Mark R. P. Thomas and Patrick A. Naylor. “The SIGMA Algorithm: A Glottal Activity Detector for Electroglottographic Signals”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.8 (2009), pp. 1557–1566. DOI: 10.1109/TASL.2009.2022430.