ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096484

# PITCH MARK DETECTION FROM NOISY SPEECH WAVEFORM USING WAVE-U-NET

*Hyun-Joon Nam*[1], *Hong-June Park*[1]

[1]Department of Electronic and Electrical Engineering,
Pohang University of Science and Technology, Republic of Korea

## ABSTRACT

Pitch mark (PM) is a time point corresponding to the closing time of vocal fold in voiced speech. PMs are useful for real-life speech processing because of their noise immunity. Wave-U-PM, a Wave-U-Net based neural network, is proposed to detect PMs from noisy speech. The ground truth PMs are generated from clean speech by using REAPER; this increases the available speech dataset for training to 100 hours, while the dataset for the electroglottograph (EGG) based PM detection is less than 5 hours. Wave-U-PM has an encoder and two decoders. The first decoder generates a sinusoidal PM waveform, whose positive peak times represent the PMs. The second decoder generates a combined pitch and formant waveform below 1000Hz. Wave-U-PM outperforms previous PM detection works by 11% and 31% for the voiced and the entire speech intervals, respectively, in the identification rate (IDR) at 0 dB SNR. The second decoder enhances IDR by 2.5% for the entire speech interval.

***Index Terms***— glottal closure instance (GCI), epoch extraction, deep learning, raw speech, multi-task learning

## 1. INTRODUCTION

Pitch mark (PM) or glottal closure instant (GCI) is a closing time point of vocal fold in voiced speech. PMs are widely used in speech processing, such as speech synthesis [1], speech enhancement [2], speech dereverberation [3,4], speech emotion recognition [5], and speaker identification [6].

Methods to detect PMs from clean speech can be classified as data-driven or non-data-driven. Data-driven methods [7–10] usually show better performance than non-data-driven methods [11–14] that use heuristic algorithms. For real-life speech processing, PMs should be detected from noisy speech also. Non-data-driven methods to detect PMs from noisy speech include the single frequency filtering (SFF) method [15] and the probabilistic source filter model (PSFM) method [14]. To reduce the effect of noise, the SFF method uses averaging in the frequency domain, and the PSFM method assumes a Gaussian noise model. A data-driven method, Deep Convolutional Neural Network based PM detection (DCNN) [8] trains a deep learning model with noisy speech for babble and white noise.

All the above-mentioned PM detection methods [7–15] use PMs identified from electroglottograph (EGG) data as ground truth; this, however, limits the dataset size for deep learning based PM detection because of the constraint that electrodes should be attached to the throat of a speaker to get the EGG data while recording speech with a microphone. The widely used speech and EGG datasets, the CMU Arctic [16] and the sentence set of APLAWD [17] are around 4 hours and 20 minutes long clean speech with 4 and 10 speakers, respectively. Furthermore, the above-mentioned methods except [9] and [13] are focused on the voiced time interval of speech, so the non-voiced intervals (unvoiced or silent intervals) of speech are excluded in evaluations. However, the input speech of PM detection for speech processing in real-life scenarios contains noise and entire (voiced, unvoiced and silent) intervals of speech because classifying whether voiced or not is still a challenging task [18].

In this work, a deep learning based PM detection from the entire noisy speech is developed by using the clean speech based PMs as ground truth. The clean speech based PMs are detected by running REAPER [12] on the entire clean speech input; this enables long data size and various speakers for deep learning. REAPER is chosen because it marks the best identification rate (IDR) among the non-data-driven methods evaluated on the clean speech input and EGG-based ground truth [10]. PMs from REAPER on clean speech match the EGG data with an average IDR of 94% in [10]. Also, the method detects PMs from the entire clean speech input because it can distinguish voiced or not.

A Wave-U-Net [19] based deep neural network, Wave-U-PM, is trained in a multi-task manner with two target waveforms which are inspired by [8, 9]. The first target waveform is a sinusoidal PM waveform with a positive peak time being a PM; PMs are generated by REAPER from clean speech. The second target waveform is a low-frequency portion of clean speech waveform below 1,000 Hz, and is generated by low-pass filtering the clean speech waveform. The second target waveform includes all of the pitches (F0) and the F1 formants and a part of F2 formants. Wave-U-PM learns PMs from two targets, directly from the first target and indirectly from the

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2022R1A2C2003451).

second. Around 116 h of clean speech dataset were used; they include CMU Arctic [16], APLAWD [17], a part of Librispeech [20], and a part of CHIME [21]. The noise dataset of this work totals about 12 h; it includes CHIME [21], MS-SNSD [22], and a part of NoiseX92 [23]. The proposed model is trained in various noisy environments.

Wave-U-PM was compared to previous PM detection works in various noisy environments with 90 unseen speakers. Using the voiced decision of REAPER and the Levenshtein distance based metric [10], the results of evaluations were presented for both entire (voiced, unvoiced, and silent) speech and voiced speech. This work outperforms previous works in the IDR at 0 dB SNR; this work gives the IDR better than the EGG-based PM detection works by at least 31.2% and 11.1% for entire and voiced speech interval. Also, the addition of the second decoder that is trained on the low-frequency clean speech target waveform enhances IDR by 2.5% at 0 dB SNR for the entire speech.

## 2. PROPOSED DEEP LEARNING METHOD

A deep neural network is used in this work to identify the PM impulse train output from a noisy speech input waveform. Denoiser [24], a Wave-U-Net [19] based deep learning model for speech enhancements, was modified to implement the deep neural network of this work. The Wave-U-Net was adopted in this work because it demonstrates excellent waveform-to-waveform conversion and this work converts a noisy speech waveform to two target waveforms (PM waveform, a low-frequency portion of clean speech waveform).

Denoiser was modified to generate the deep neural network of this work (Wave-U-PM), such that Wave-U-PM has two decoders. Wave-U-PM includes a five-layer CNN encoder, a two-layer unidirectional LSTM (Uni-LSTM), and two five-layer CNN decoders (Fig. 1) with 24 initial hidden channel sizes ($H$), the kernel size ($K$) of 6, the stride ($S$) of 2, the growth rate of 2, no up-sampling, and $2^4H$ LSTM states. H controls the $C_{out}$ of each encoder and decoder layer with growth rate; the $C_{out}$ is doubled from H to $2^4H$ for each encoder, then halved from $2^4H$ for each decoder except the last decoder layer ($Decoder_1$). The $Decoder_1$ has $C_{out}$ of 1 and does not include the last SELU activation function. Two decoders are used for training but only the first decoder is used for inference; 3.6M parameters are used for inference while 4.1M parameters are used for training.

While training, Wave-U-PM takes a noisy speech input waveform ($x$) and generates two output waveforms; one is a sinusoidal PM waveform ($\hat{y}_s$) to represent the PM impulse train and the other is a combined pitch and formant waveform that targets a low-pass filtered clean speech waveform below 1,000 Hz. The pitch and formant waveform helps the neural network to learn PMs. Wave-U-PM is trained to predict two target waveforms ($y_s$, $y_c$) by minimizing the loss function

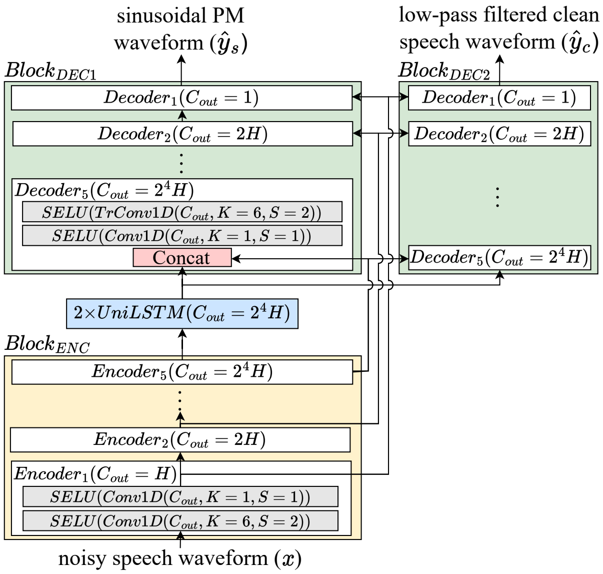
sinusoidal PM waveform ($\hat{y}_s$)
low-pass filtered clean speech waveform ($\hat{y}_c$)
$Block_{DEC1}$
$Block_{DEC2}$
$Decoder_1(C_{out}=1)$
$Decoder_2(C_{out}=2H)$
$Decoder_5(C_{out}=2^4H)$
$SELU(TrConv1D(C_{out}, K=6, S=2))$
$SELU(Conv1D(C_{out}, K=1, S=1))$
Concat
$Decoder_1(C_{out}=1)$
$Decoder_2(C_{out}=2H)$
$Decoder_5(C_{out}=2^4H)$
$2\times UniLSTM(C_{out}=2^4H)$
$Block_{ENC}$
$Encoder_5(C_{out}=2^4H)$
$Encoder_2(C_{out}=2H)$
$Encoder_1(C_{out}=H)$
$SELU(Conv1D(C_{out}, K=1, S=1))$
$SELU(Conv1D(C_{out}, K=6, S=2))$
noisy speech waveform ($x$)

**Fig. 1**. Wave-U-PM model proposed in this work

$$L(\hat{y}, y) = \frac{1}{N}\sum_{n=0}^{N-1}\left[\left|\hat{y}_s(n) - y_s(n)\right| + \left|\hat{y}_c(n) - y_c(n)\right|\right]. \quad (1)$$

Two target waveforms ($y_s$, $y_c$) are used to train Wave-U-PM (Fig. 2). The first target $y_s$ is a sinusoidal waveform based on the ground-truth PMs that were generated by applying the REAPER [12] to clean speech input. The sinusoidal PM waveform $y_s$ has a constant amplitude of 1 with the positive peak time that is synchronized to the ground-truth PMs; the period of the sine waveform is the time interval between two adjacent PMs. PMs are derived from $\hat{y}_s$ by using a peak-picking procedure similar to [9]. The second target $y_c$ is a low-frequency portion of clean speech waveform below 1,000 Hz; it is generated by applying an inverse fast Fourier transform (FFT) to the low-frequency portion of the FFT result of an entire clean speech waveform below 1,000 Hz.

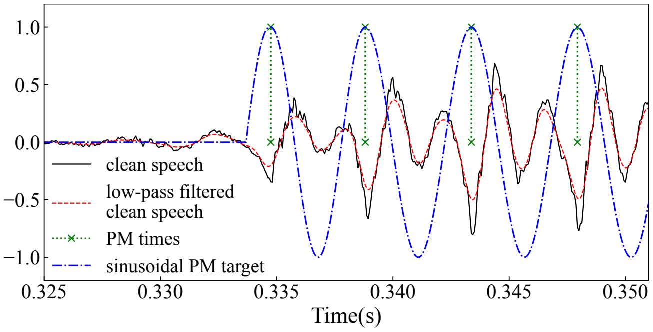

**Fig. 2**. Generated sinusoidal PM target waveform (blue dot-dashed) from ground-truth pitch marks (green dotted) and 1 kHz low-pass filtered clean speech target waveform (red dashed) on clean speech (black solid).

## 3. EXPERIMENTAL RESULTS

### 3.1. Datasets for Training, Validation, and Test

The clean speech and the noise datasets were collected to train and test Wave-U-PM (Table 1). The datasets were sampled at 16 kHz. The clean speech datasets include 352 real speakers to enhance accuracy for unseen speakers, while the number of real speakers does not exceed 40 in most of the previous data-driven methods of PM detection works [7, 8, 10]. The ground-truth PMs were generated by applying to the clean speech dataset. WebRTC VAD algorithm [25] was applied to the REAPER output to remove a few detected PMs in unvoiced and silent time intervals of speech.

**Table 1**. Collected datasets for training, validation and test. The number of speakers is denoted as #SPK.

| **Type** | **Dataset** | **Subset** | **#SPK** | **Time** (hours) |
|---|---|---|---|---|
| | | **Training and Validation datasets** | | |
| Clean | CHIME3 [21] | dt_05, et_05 | 8 | 1.3 |
| | CMU Arctic [16] | BDL, SLT, JMK | 3 | 2.8 |
| | LibriSpeech [20] | train-100 | 251 | 100.6 |
| Noise | CHIME3 [21] | backgrounds | | 8.4 |
| | MS-SNSD [22] | noise_train, noise_valid | | 3.3 |
| | | **Test dataset** | | |
| Clean | APLAWD [17] | sentence set | 10 | 0.4 |
| | LibriSpeech [20] | dev-clean, test-clean | 80 | 10.8 |
| Noise | NoiseX92 [23] | 12 types | | 1.0 |

One-second-long noisy data sets were used for training and validation. The noisy dataset was generated by adding a randomly-chosen combination of a clean speech and noise, both one second long. The data augmentations are applied to the noisy data for training and validation. A random value was chosen for each noisy data with SNR of 0, 5 dB SNR or clean, maximum amplitude of -30, -25, -20, or -17 dBFS, resampling rate in the range from -10% to 10% of 16 kHz, time-shifting in the range from -0.25 s to 0.25 s, and polarity inversion factor of -1 or 1. The random polarity inversion was used to address the polarity mismatch of recorded waveforms. The test dataset is 11.2 h long and recorded from 90 unseen speakers and 12 types of noise, and was tested in several noisy environments of at least 0 dB SNR. The test dataset does not include clean speech or noise used for training.

### 3.2. Evaluation and Results

To demonstrate the performance of Wave-U-PM, a clean speech waveform is mixed with a loud babble noise with 0 dB SNR to generate the input waveform (Fig. 3(a)). The babble noise was taken from the test dataset of MS-SNSD [22]. Wave-U-PM generates a combined pitch and formant waveform (Fig. 3(b)), and PMs that are similar to those from ground truth data; the maximum time difference of PMs was 0.25 ms between the Wave-U-PM output and ground truth.

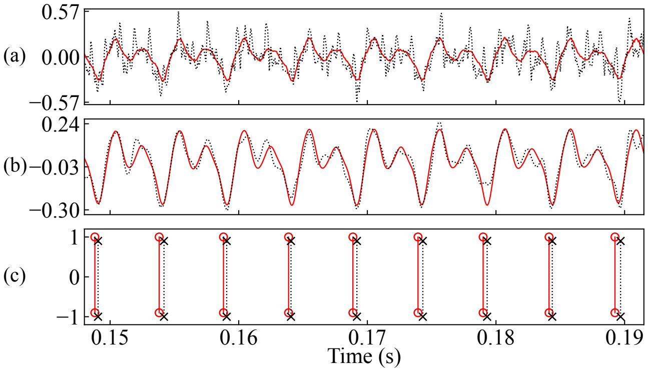

**Fig. 3**. Comparison of Wave-U-PM input/output (black dotted) and ground truth (red solid), (a) input waveforms, (b) combined pitch and formant waveform, (c) PMs derived from sinusoidal PM waveform.

Wave-U-PM was compared to the previously published PM detection methods [12–14] in the identification rate (IDR) and the dynamic evaluation measure (E10) [10] for the 11.2-h test dataset. IDR declares that a predicted PM is correct, only if the predicted PM is located within the range $[t2 - 0.5(t2 - t1), t2 + 0.5(t3 - t2)]$ for a ground-truth PM sequence of $\{t1, t2, t3\}$. E10 declares a predicted PM is correct, only if the predicted PM is located within the range $[t2 - 0.05(t3 - t1), t2 + 0.05(t3 - t1)]$ for the ground-truth PM sequence of $\{t1, t2, t3\}$. Then, the IDR and E10 are computed based on the Levenshtein distance [10], which enables more accurate evaluation on the missing and false-alarms of predicted PMs. The false-alarms of predicted PMs are located in the unvoiced and silent time intervals of speech, and are crucial to both IDR and E10 in the entire speech evaluation.

Among the previously published PM detection methods, REAPER [12], GEFBA [13], and PSFM [14] were chosen for comparison with the Wave-U-PM. REAPER and GEFBA were chosen because they detect PMs from the entire speech including the unvoiced speech interval. PSFM was chosen because it has good noise immunity, although it detects PMs from voiced speech only. For the voiced speech evaluation in each method, false-alarm PMs during unvoiced speech intervals are excluded by using the voiced/unvoiced decision criterion that is generated by REAPER and WebRTC VAD with clean speech.

Wave-U-PM outperforms others at least 11.1% and 31.2% in the IDR for the voiced and the entire noisy speech, respectively, at 0 dB SNR with the test dataset (Table 2, Fig. 4). Also, Wave-U-PM shows the minimum relative differences from clean IDR to 0 dB IDR by 12.9% for both voiced and the entire speech with test dataset. The second minimum relative differences from clean IDR to 0 dB IDR are 13.6% (PSFM$_V$) and 44.1% (GEFBA) for each voiced and entire speech.

An ablation study of the proposed Wave-U-PM with Uni-LSTM indicates that the LSTM contributes the most by en-

**Table 2**. Comparison of IDR and E10 using the test dataset for voiced only ($_V$) and entire (voiced+unvoiced+silence) speech.

| Methods | IDR (%) | | | E10 (%) | | |
|---|---|---|---|---|---|---|
| | clean | 5 dB | 0 dB | clean | 5 dB | 0 dB |
| REAPER$_V$ | **99.5** | 69.6 | 50.4 | **99.0** | 65.7 | 46.9 |
| GEFBA$_V$ | 85.5 | 53.9 | 38.1 | 46.6 | 29.0 | 20.5 |
| PSFM$_V$ | 85.9 | 76.8 | 72.3 | 67.6 | 37.8 | 28.0 |
| Wave-U-PM$_V$ | 96.3 | **90.8** | **83.4** | 93.2 | **87.0** | **78.9** |
| REAPER | **99.0** | 68.6 | 49.6 | **98.5** | 64.5 | 45.6 |
| GEFBA | 73.6 | 45.3 | 29.5 | 34.1 | 19.6 | 11.1 |
| Wave-U-PM | 93.7 | **88.4** | **80.8** | 90.6 | **84.4** | **76.0** |

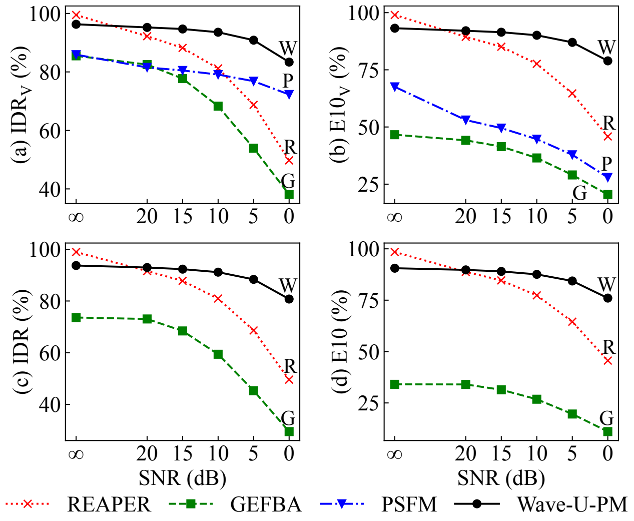

**Fig. 4**. Comparison of Wave-U-PM with others in IDR and E10 versus SNR for the 11.2-h test dataset. (a) IDR for voiced speech (b) E10 for voiced speech (c) IDR for entire speech (d). ∞ denotes that there is no additional noise for clean speech input.

**Table 3**. Ablation study of Wave-U-PM on the test dataset. The effective number of parameters is denoted as #P*.

| Wave-U-PM | #P* | IDR (%) | | E10 (%) | |
|---|---|---|---|---|---|
| | | clean | 0 dB | clean | 0 dB |
| **Proposed** | 3.6 M | 93.7 | 80.8 | 90.6 | 76.0 |
| w/o Causality (Bi-LSTM) | 7.5 M | 94.9 | 84.1 | 92.0 | 80.2 |
| w/o LSTM | 1.3 M | 91.7 | 58.9 | 86.9 | 52.0 |
| w/o Second decoder | 3.6 M | 93.8 | 78.3 | 90.7 | 73.4 |

hancing IDR 21.9% at 0 dB SNR (Table 3). It is assumed that the narrow input receptive field of 9.8 ms without LSTM causes the large degradation of IDR at 0 dB SNR. Although the bidirectional LSTM enhances IDR by 3.3% at 0 dB SNR over the Uni-LSTM, it cannot be used for streaming operation because of its non-causality. The second decoder with the low-pass filtered clean speech target enhances IDR by 2.5%; the second decoder does not increase the number of parameters of the inference model because it is used for training not for inference.

## 4. CONCLUSIONS

The voiced speech waveform is characterized by large periodic peaks called pitches. A PM represents a negative peak time point of the voiced speech waveform; it corresponds to the GCI of the vocal fold. Since PMs are used by the brain to time-synchronize different frequency components sent by the cochlea, they can be used as input features of various signal-processing tasks. Pitches are strong to noise in speech waveforms, because of their large amplitudes. To use this strong noise immunity of PMs, a deep learning model (Wave-U-PM) is proposed to detect PMs from noisy speech. Wave-U-PM is extended from a speech enhancement deep neural network, Denoiser [24] a Wave-U-Net [19]. By using two decoders, Wave-U-PM generates two output waveforms from a noisy speech input waveform; one is a sinusoidal PM waveform that has a constant amplitude of 1 with the positive peak times synchronized to the ground-truth PMs, and the other is a pitch and formant waveform that matches a low-pass filtered clean speech waveform below 1,000 Hz. The model uses 3.6M weights for inference; the input frame width is 9.8 ms, and the input step size is 2 ms at a 16 kS/sec sampling rate. REAPER is one of the best available non-data-driven PM detection methods [12]; REAPER distinguishes the voiced and unvoiced time intervals, also. Almost all the published PM detection methods use the EGG data as the ground truth; this limits the data size available for deep learning because two electrodes should be attached to the throat to get the EGG data. To increase the data size for deep learning, REAPER is used to get the target waveforms from clean speech instead of the EGG data. Clean speech and noise waveforms, around 100 and 10 hours each, are used to train Wave-U-PM. The training dataset randomly combines the clean speech from 260 speakers and the noise dataset with data augmentation. The test dataset uses clean speech data from unseen speakers. Wave-U-PM demonstrates a robust performance for noisy speech input waveforms with SNR down to 0 dB; Wave-U-PM outperforms the conventional PM detection methods in IDR by 11% and 31% for each voiced and entire speech on the 11.2-h-long test dataset. The ablation study revealed that the second decoder using the low-pass filtered clean speech waveform as target improves IDR by 2.5% at 0 dB SNR, while the second decoder does not increase the number of parameters for inference because it is used for training not for inference. From the robust PM detection ability of Wave-U-PM for unseen speakers in noisy environments, it is expected that Wave-U-PM can be used for other speech processing jobs such as noise suppression or automatic speech recognition.

## 5. REFERENCES

[1] Y. Cui, X. Wang, L. He, and F. K. Soong, "A New Glottal Neural Vocoder for Speech Synthesis," in *Proc. Interspeech*, 2018, pp. 2017–2021.

[2] K. T. Deepak and S. R. M. Prasanna, "Foreground Speech Segmentation and Enhancement Using Glottal Closure Instants and Mel Cepstral Coefficients," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1205–1219, 2016.

[3] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," in *2004 12th European Signal Processing Conference*, 2004, pp. 809–812.

[4] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal Averagingmethod for Enhancement of Reverberant Speech," in *2007 15th International Conference on Digital Signal Processing*, 2007, pp. 607–610.

[5] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Comparison of Glottal Closure Instants Detection Algorithms for Emotional Speech," in *ICASSP*. IEEE, 2020, pp. 7379–7383.

[6] A. G. Ramakrishnan, B. Abhiram, and S. R. M. Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL469–EL475, 2015.

[7] S. Yang, Z. Wu, B. Shen, and H. Meng, "Detection of Glottal Closure Instants from Speech Signals: A Convolutional Neural Network Based Method," in *Proc. Interspeech*, 2018, pp. 317–321.

[8] M. Goyal and V. Srivastava, "Detection of Glottal Closure Instants from Raw Speech Using Convolutional Neural Networks," in *Proc. Interspeech*, 2019, pp. 1591–1595.

[9] L. Ardaillon and A. Roebel, "GCI Detection from Raw Speech Using a Fully-Convolutional Network," in *ICASSP*. IEEE, 2020, pp. 6739–6743.

[10] J. Matoušek and D. Tihelka, "A Comparison of Convolutional Neural Networks for Glottal Closure Instant Detection from Raw Speech," in *ICASSP*. IEEE, 2021, pp. 6938–6942.

[11] P. A. Naylor, A. Kounoudes, et al., "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.

[12] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR," https://github.com/google/REAPER, 2015.

[13] A. I. Koutrouvelis et al., "A Fast Method for High-Resolution Voiced/Unvoiced Detection and Glottal Closure/Opening Instant Estimation of Speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 316–328, 2016.

[14] A. R. MV and P. K. Ghosh, "PSFM—A Probabilistic Source Filter Model for Noise Robust Glottal Closure Instant Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1645–1657, 2018.

[15] G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Detection of Glottal Closure Instants in Degraded Speech Using Single Frequency Filtering Analysis," in *Proc. Interspeech*, 2018, pp. 2300–2304.

[16] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Workshop on Speech Synthesis (SSW 5)*, 2004, pp. 223–224.

[17] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," *University College London," Technical Report*, 1987.

[18] X. Tan and X. L. Zhang, "Speech Enhancement Aided End-To-End Multi-Task Learning for Voice Activity Detection," in *ICASSP*. IEEE, 2021, pp. 6823–6827.

[19] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," *arXiv preprint arXiv:1806.03185*, 2018.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[21] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *ASRU*, 2015, pp. 504–511.

[22] Reddy, C. K., et al., "A Scalable Noisy Speech Dataset and Online Subjective Test Framework," in *Proc. Interspeech*, 2019, pp. 1816–1820.

[23] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[24] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. Interspeech*, 2020, pp. 3291–3295.

[25] Google, "WebRTC," https://webrtc.org, 2011.