Detection of Glottal Closure Instants using Deep Dilated Convolutional Neural Networks

Prathosh A. P*, Mohit Goyal* and Varun Srivastava

Abstract—Glottal Closure Instants (GCIs) correspond to the temporal locations of significant excitation to the vocal tract occurring during the production of voiced speech. Detection of GCIs from speech signals is a well-studied problem given its importance in speech processing. Most of the existing approaches for GCI detection adopt a two-stage approach - (i) Transformation of speech signal into a representative signal where GCIs are localized better, (ii) extraction of GCIs using the representative signal obtained in first stage. The former stage is accomplished using signal processing techniques based on the principles of speech production and the latter with heuristic-algorithms such as dynamic programming and peak-picking. These methods are thus task-specific and rely on the methods used for representative signal extraction. However in this paper, we formulate the GCI detection problem from a representation learning perspective where appropriate representation is implicitly learned from the raw-speech data samples. Specifically, GCI detection is cast as a supervised multi-task learning problem which is solved using a deep dilated convolutional neural network jointly optimizing a classification and regression cost. The learning capabilities of the proposed model is demonstrated with several experiments on standard datasets. The results compare well with the state-of-the-art algorithms while performing better in the case of presence of real-world non-stationary noise.

Index Terms—GCI detection, epoch extraction, dilated convolutional neural networks, multi-task learning.

I. INTRODUCTION

A. Background and Previous work

Production of voiced speech is accompanied with sustained oscillations of the vocal folds [1] resulting in a quasi-periodic flow of air-pulses which constitutes the excitation signal to the vocal tract [2]. The instant of significant excitation (within each period) is termed the Epoch which coincides with instant of closure of the glottis [3]. The problem of detecting the precise locations of such Glottal Closure Instants (GCIs) from speech signal has been studied for decades given its importance in several speech processing tasks [10-18]. Most of the successful GCI detectors adopt a two-stage approach - (i) Obtaining a representative signal from the speech signal in which GCIs are better localized and (ii) detecting GCIs from the representative signal using custom-made heuristic algorithms. The former stage is based on the observation that GCIs do not manifest well in the raw speech-signal domain but exhibits a better localization in some other domain (Figure 1 in [9]). Many algorithms rely on the source-filter model for speech production [1] and use signal-processing techniques to estimate a correlate of the source-signal, in which GCIs are better manifested. For instance, [5], [6], [10], [13] choose either linear-prediction residual or glottal flow derivative as the representative signal. Other class of algorithms do not explicitly make any model assumption for speech production rather indirectly use the properties of excitation signal (such as its impulsive nature) and estimate appropriate representations (E.g., zero-frequency filtered signal [7], mean-based signal [8], wavelet-decompositions [14], singularity exponents [11]). During the second stage, aforementioned algorithms employ several heuristics to extract (or refine) the GCIs. These include dynamic programming [5], [6], peak-picking [9], [15] and optimization with regularity constraints [11]. All these methods perform reasonably well albeit they depend largely on their choice of representative signals.

B. Context and Scope

With the recent advances made in the area of data-driven representation learning [16], it is possible to operate directly in the signal space and let the learning machinery obtain the ‘best’ representation given a task and data. This approach has found tremendous success in multiple domains with image, text and audio data [17]. Specifically, convolutional neural networks (CNN) have found their utility in a range of speech processing tasks such as phoneme recognition [18], [19], feature/front-end learning for LVCSR [20], [21], [22], voice-activity detection [23], spoofing detection [24], emotion recognition [25], [26] and speaker identification [27]. The underlying theme in all these methods is to directly operate on the raw speech signal and let the CNN learn the ‘best’ representation for the given task. All these papers demonstrate that features learned from data using CNNs lead to superior performance as compared to hand-crafted feature learning. Motivated by the above observations, in this paper, we approach the problem of end-to-end GCI detection from a deep-learning perspective.

The key difference in the present formulation as compared to the previous works which uses CNNs for various speech processing tasks is that, by-and-large the scope of previous works is classification of an utterance or segment of speech into classes (phonemes, emotions, speakers) while in the proposed work the interest is in detecting a temporal event (GCI) in the signal. This objective is met by formulating the problem in a novel joint classification-regression framework wherein a temporal event (GCI in this case) is simultaneously detected and localized in a frame. Various learning and generalization experiments are performed on multiple datasets comparing
with four state-of-the-art algorithms to demonstrate the model-capacity of the trained network through improved metrics.

II. METHODOLOGY

A. Problem formulation and Data generation

In this work, the problem of GCI detection is formulated using a block-processing approach. It is known that GCI is a temporal event that occurs utmost once in every pitch-period of the voiced speech that could range between 2 milli-seconds (ms) to 20 ms [28]. Based on this observation, we define a detection speech window \( w_d \) as any speech frame of length equal to 2 ms (minimum possible pitch period). Note that, by definition, every detection window can have utmost one GCI within it. Given a \( w_d \), the primary task is to detect whether or not there is GCI within it, which is formulated as a binary classification problem. Having detected a GCI within a \( w_d \), the next step is to localize the GCI which is cast as a regression task, estimating the distance between the onset of \( w_d \) and the location of GCI. Even though each \( w_d \) could be considered as a single data-point, in most of the practical scenarios, a window of 2 ms does not contain a GCI (since the average pitch period is much higher than 2 ms for most speakers). Further, a 2 ms window comprises too less number of samples for meaningful feature learning via a deep CNN. To address these issues, a symmetric context of 5 ms at either sides of the detection window is considered to generate the input frame \( w_i \). Hence in summary, every input data-sample \( w_i \) will be of length 12 ms, with the classification and regression for GCI detection and localization being carried over the 2 ms detection window \( (w_d) \) at the centre of the frame. Note that despite each input frame being 12 ms long, classification and regression takes place over the centre 2 ms window that would contain utmost one GCI. Figure 1 depicts a segment of voiced speech with three input frames marked along with their corresponding \( w_d \) and ground truth GCIs. Finally, multiple input data-frames are generated by an overlapping window method with a shift of one sample between successive input frames. In principle, the shift could be anything less than \( w_d \), however, a shift of sample is a employed to facilitate more data.

B. Dilated CNNs and Multi-task learning

Recently, Deep Dilated Convolutional Neural Networks (DCNNs) are shown to learn useful representations for standard speech processing tasks very well [29][30]. Since, a dilated convolution is mathematically equivalent to regular convolution with a kernel with zeros inserted in between, stacked dilated convolutions allow an exponential increase in context, while still being linear in the number of parameters. The proposed DCNN (Figure 2) comprises of convolutional layers that function as feature extractors, so that a fixed sized representation is obtained for a speech window. Max pooling is used after every convolutional layer with a kernel size and stride of 2, in order to halve the signal size, while doubling the dilation at every convolutional layer to increase the context of the representation learnt at any layer. Scaled Exponential Linear Units (SELU) [31] activation functions are used between convolutional layers in order to normalize the activations. SELU ensures the benefits of explicit normalization (E.g., batch-normalization [32]) while being identical implementation-wise during train and test phase, unlike techniques such as batch normalization.

Since both the classification and regression tasks are to be performed on the same input frame \( w_i \) it is natural to share the features learned across the tasks which is the core idea behind multitask learning [33]. Thus, an input frame \( w_i \) is first fed to the DCNN to extract the feature vector that goes into two identical subnetworks performing classification and regression. A sigmoid activation is used at the output of the classification branch and a Hard-tanh [34] function is used for regression (instead of a linear activation to suppress out-of-window predictions). The activations of the penultimate layer of the feature extractor are reshaped such that the effective receptive field for the classifier/regressor subnetwork corresponds to \( w_d \) in the input-signal space. Note that this is an important step that ensures the restriction of the classification and regression tasks over the centre \( w_d \) within each data-frame of length \( w_i \).

The task is to output the probability of presence of a GCI \( y_c \), as well as its location \( y_r \) (\( 0 \leq y_r < w_d \)) within each detection window. Ideally, for a window with a GCI, \( y_c = 1 \) and \( y_r = d \) (where \( d \) is the distance between the onset of the window and location of the GCI within the window) and for a window without a GCI, \( y_c = 0 \), and \( y_r \) is arbitrary (Refer Fig. 1). Since a single network predicts both the probability of occurrence of the GCI and location of the GCI, the loss

![Figure 1: Depiction of data generation for GCI detection - (a) A voiced speech segment with three pitch periods, (b) Differentiated electroglotograph (dEGG) signal with three ground-truth GCI peaks. Three input data-frames \( (w_i^{1–3}) \) have been marked (dotted arrows) along with their corresponding detection windows \( (w_d^{1–3}) \). It could be seen that \( w_i^1 \) does not have a true GCI within it \( (y_c^1 = 0) \), both \( w_i^2 \) and \( w_i^3 \) have the same ground-truth GCI \( (y_c^2, y_c^3 = 1) \) with different regression values \( (y_r) \).](image1)

![Figure 2: Proposed architecture: The Convolution layer parameters are depicted as "Conv" - Kernel Size/Dilation/Channels. The input is of length \( w_i \), detection and localization happens over the centre \( [0, w_d] \) window.](image2)
function consists of two terms, the classification error which is standard cross entropy loss and the regression error which is the mean squared loss between the actual and the predicted location of the GCI within a window. Mathematically,

$$\mathcal{L} = \frac{w_c}{N} \sum_{j=1}^{N} \left[ t_j \log(y_j^c) + (1 - t_j) \log(1 - y_j^c) \right] + \frac{w_r}{(\sum_{k=1}^{N} c_k)} \sum_{j=1}^{N} (t_j - y_j^c)^2$$

(1)

where $t_j$ is 1 if the $j^{th}$ sample window has a GCI, and is 0 otherwise. $t_j$ is the actual location of the GCI within the $j^{th}$ window, $N$ is the number of datapoints within a training batch. $w_c$ are $w_r$ are the weights of the classification and regression loss respectively. Note that the regression loss is divided by the number of true GCIs in a batch (instead of total number of datapoints) to account for the fact that there are fewer samples for the regression block in every batch compared to the classification block. The ratio of $w_r$ to $w_c$ is set to 1:10 to account for the numerical range of the classification ([0,1]) and regression arms ( [0,32] at a sampling rate of 16 kHz). The output of the classification sigmoid is thresholded at 0.5 and only the detections with $y_c \geq 0.5$ are called the candidate GCIs and considered for further processing.

C. Inference and Clustering

In the present formulation since, (i) every detection window can contain utmost one GCI and (ii) data-frames are generated using an overlapping window approach, a single ground-truth GCI would have been predicted (and regressed) over several successive windows ($w_2^r$ and $w_3^c$ in Fig. 1 and solid dots in Fig. 3 (a)) and hence have to be merged. Given a speech signal and associated candidate GCIs, the first step is to construct a weighted histogram of the candidate GCIs with a bin size of $B$ which is a hyper-parameter of the method. Since every detection is associated with a probability representing the confidence of presence of GCI, the weight for the histogram is taken to be the detection probabilities. Now since GCI is a periodic event, the histogram thus obtained will contain local-groups of contiguous bins of non-zero values (Fig. 3, (b)). The final GCIs are hypothesized to be the means (measure of central tendency) of each local-groups (dotted lines in Fig. 3(b)). By construction, the means of the local-groups tend to be at the locations with large number of high-probability candidate GCIs which is desirable. The bin-size $B$ is set at 5 samples through-out this work albeit there is a trade-off between erroneous detections (higher the $B$ less are the false insertions) and resolution of predicted GCIs (lower the $B$, better is the localization of detected GCIs). The entire procedure for clustering is illustrated in Figure 3.

III. EXPERIMENTS AND RESULTS

A. Experimental details

The standard practice to evaluate GCI detection algorithms is to use datasets containing simultaneous recordings of speech and Electroglottography (EGG) signals. The negative peaks of differentiated EGG (dEGG) signals are considered the ground truth for GCI locations (Figure 1 (b)). In this paper, we use speech signals from the standard CMU Arctic [36] and the APLAWD datasets [37] for our experiments. Standard performance metrics namely, Identification Rate (IDR - % of correct detections, higher the better), Miss Rate (MR - % of missed detections, lower the better), False Alarm Rate (FAR - % of false insertions, lower the better) and Identification Accuracy (IDA - standard deviation of distance between the true and predicted GCIs, lower the better) are employed for evaluation. A detailed description of the performance metrics is omitted here for the lack of space and may be found in Figure 2 of [5] and elsewhere [12], [9]. There are a total of 351226 ground-truth GCIs from 13 speakers in the combined dataset. The models are implemented using PyTorch [38] using the ADAMAX optimizer [39] with standard parameter settings. Models parameters were initialized as per the scheme outlined in [31]. Detection experiments were carried out on clean speech as well as speech corrupted with additive synthetic white noise and real-world babble noise (background multi-speaker chatter obtained from [40]) with SNRs ranging from 0 to 25 dB in steps of 5 dB. For a given SNR, both the training and testing are carried on the corresponding corrupted speech at the same SNR. The results of the proposed algorithm (DCNN) are compared with four state-of-the-art algorithms, namely, Zero-frequency resonator [7], Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) [8], Dynamic Plosion Index (DPI) [9] and Micro-canonical Multi-scale Formalism (MMF) [11], by evaluating all four on the same test dataset as that of DCNN. As mentioned earlier, the bin-size parameter is fixed at 5 (this yields the best performance) for all experiments. The following experiments are carried out for assessing the performance and the generalization abilities of the proposed algorithm - (a) Baseline experiments: Training and testing on CMU and APLAWD datasets with speaker-overlap between the training and test data. For this case, training and testing both the training and test data contains non-overlapping utterances from all speakers, (b) Cross-dataset experiments: Training on
IEEE SIGNAL PROCESSING LETTERS

CMU and testing on APLAWD and vice-versa, (c) Cross-speaker experiments: Training and testing on non-overlapping set of speakers, (d) Cross-SNR experiments: Training on a given SNR and testing on other SNRs. While experiment (a) evaluates the detection capabilities of the proposed method (hence this is used for comparison), the rest of experiments will validate its generalization capabilities. In all experiments a random subset of 10% of data is considered for training and rest for testing. Since 90% of the data is used for testing, no cross-validation is employed.

B. Results and Discussion

Table I describes the results of five algorithms on the CMU ARCTIC and APLAWD datasets on clean speech. It can be seen that DCNN has the highest detection rate (99.3 % and 95.5 %) on both the datasets. The IDA of the DCNN method is also the least for both the datasets. It was observed that all the algorithms degrade on the APLAWD dataset which might be because it contains ten speakers of varied pitch frequencies. A similar trend is observed even with noisy speech as can be observed in Figure 4. It is seen that DCNN and ZFR are very robust to noise with the IDR being above 90% even at 0 dB babble noise. Further, the IDA of the DCNN algorithm is consistently lowest for all cases considered.

Table II describes the outcomes on the experiments (b), (c) and (d) that are designed to ascertain the generalization abilities of the proposed algorithm across different speaker, dataset and SNR settings. For the case of cross-dataset studies (Noe that CMU has three speakers while APLAWD has ten), the IDR reduces by 2-3% albeit IDA remains intact. To further confirm the robustness to pitch-period, we trained a model on only male speakers and tested on the female speaker of the CMU and vice-versa. It was the observed that there was no or slight reduction in IDR and IDA, compared to the models that were trained on all speakers. In the third experiment, a model that was trained on 0 dB Babble noise was tested on speech with different SNRs and it was observed that there is no change in the model performance. It is noteworthy that in all these experiments, the model parameters (detection probability threshold: 0.5 and bin-size: 5) were kept the same in spite of which the models generalize well across datasets, recording settings, speakers and noise-levels. However, in practical settings, one can tweak the model hyper-parameters to best suit the data. Given that the proposed method operates on raw speech and fully data-driven, the aforementioned performance is significant. All the aforementioned results confirms the learning abilities of the proposed architecture. The performance of DCNN might be ascribed to the following factors: (i) problem formulation that enforces presence of multiple detections per GCI leading to robustness, (ii) use of shared feature representations across two tasks, (iii) use of DCNN for effective feature learning and (iv) use of weighted histogram for clustering.

IV. Conclusion

In this paper, a data-driven method for GCI detection from raw-speech using a multi-task supervised learning approach is proposed. A dilated CNN is employed for learning representations along with a weighted-histogram based method for inference and clustering. Several experiments were conducted to compare the performance with state-of-the-art algorithms on datasets comprising multiple speakers to demonstrate the efficacy of the proposed approach. Given the representation and generalization abilities of the proposed approach, we believe that a similar methodology could be adopted for detecting multiple landmarks occurring in speech and general time-series data, which could provide directions for future work.
REFERENCES

[1] G. Fant, Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations. Walter de Gruyter, 1971, vol. 2.

[2] B. H. Story, “An overview of the physiology, physics and modeling of the sound source for vowels,” Acoustical Science and Technology, vol. 23, no. 4, pp. 185–206, 2002.

[3] T. V. Ananthapadmanabha and B. Yegnarayanan, “Epoc extraction from linear prediction residual for identification of closed glottis interval,” IEEE Trans. Acoust., Speech, Signal Process., vol. 27, no. 7, pp. 309–319, Aug.1979.

[4] ——, “Epoc extraction of voiced speech,” IEEE Trans. Acoust., Speech, Signal Process., vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[5] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, “Estimation of glottal opening and closing instants in voiced speech using the YAGA algorithm,” IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no.1, pp. 34–43, Jan.2007.

[6] M. P. Thomas, J. Gudnason, and P. A. Naylor, “Estimation of glottal opening and closing instants in voiced speech using the YAGA algorithm,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 82–91, Jan. 2012.

[7] K. S. R. Murty and B. Yegnarayanan, “Epoc extraction from speech signals,” IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[8] T. Drugman and T. Dutoit, “Glottal closure and opening instant detection from speech signals,” inProc. Interspeech Conf., 2009.

[9] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, “Epoc 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, Dec. 2013.

[10] A. P. Prathosh, P. Sujith, A. G. Ramakrishnan, and P. K. Ghosh, “Cumulative impulse strength for epoc extraction,” IEEE Signal Processing Letters, vol. 23, no. 4, pp. 421–428, 2016.

[11] 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 (TASLP), vol. 22, no. 12, pp. 1941–1950, 2014.

[12] 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, Lang. Process., vol. 20, no. 3, pp. 994–1006, Mar. 2012.

[13] A. I. Kourtouvelis, G. P. Kafentzis, N. D. Gaubitch, and R. Heusdens, “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 (TASLP), vol. 24, no. 2, pp. 316–329, 2016.

[14] C. D’ Alessandro and N. Sturmel, “Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude,” Sadhana, vol. 36, no. 5, pp. 601–622, 2011.

[15] 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.

[16] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,”IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[17] J. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, Deep learning.MIT press Cambridge, 2016, vol. 1.

[18] D. Palaz, R. Collobert, and M. Magimai-Doss, “Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks,” inProc. Interspeech Conf., 2013.

[19] N. Zeghidour, U. Nicolas, K. Jasonas, S. Thomas, S. Gabriel, and D. Emmanuel, “Learning filterbanks from raw speech for phone recognition,” inIEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[20] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, “Learning the speech front-end with raw waveform CLDNNs,” inProc. Interspeech Conf., 2015.

[21] Z. Tuske, P. Golik, R. Schlueter, and H. Ney, “Acoustic modeling with deep neural networks using raw time signal for LVCSR,” inProc. Interspeech Conf., 2014.

[22] O. Abdel-Hamid, A.-f. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,”IEEE/ACM Transactions on audio, speech, and language processing, vol. 22, no. 10, pp. 1533–1545, 2014.

[23] R. Zazo, T. N. Sainath, G. Simko, and C. Parada, “Feature learning with raw-waveform CLDNNs for voice activity detection,” inProc. Interspeech Conf., 2016.

[24] H. Dinkel, N. Chen, Y. Qian, and K. Yu, “End-to-end spoofing detection with raw waveform cldnn,” inIEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 4860–4864.

[25] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolau, B. Schuller, and S. Zafeiriou, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” inIEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5200–5204.

[26] S. Zhang, S. Zhang, T. Huang, and W. Gao, “Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching,”IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576–1590, 2017.

[27] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, “Towards directly modeling raw speech signal for speaker verification using cnns,” inIEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[28] I. Titze, Principles of Voice Production, 1994, p. 188.

[29] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,”arXiv preprint arXiv:1511.07126v3, 2015.

[30] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499v2, 2016.

[31] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, “Self-normalizing neural networks,” inAdvances in Neural Information Processing Systems, 2017, pp. 972–981.

[32] J. Joffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, 2015.

[33] R. Caruana, “Multitask learning,” inLearning to learn. Springer, 1998, pp. 95–133.

[34] C. Gulcehre, M. Moczulski, M. Denil, and Y. Bengio, “Noisy activation functions,” inInternational Conference on Machine Learning, 2016, pp. 3059–3068.

[35] D. G. Childers and A. K. Krishnamurthy, “A critical review of electroglottography,”CRC Crit. Rev. Bioeng., vol. 12, pp. 131–164, 1985.

[36] J. Kominek and A. W. Black, “The CMU arctic speech databases,” inFifth ISCA Workshop on Speech Synthesis, 2004.

[37] G. Lindsey, A. Breen, and S. Nevard, “SPAR’s achivable actual-word databases,”Univ. College London, London, Tech. Rep., 1987. [Online]. Available:https://www.comspse.ec.ie.ac.uk/~sap/resources/aplawdw/

[38] P. C. Team, “Pytorch: Tensors and dynamic neural networks in python with strong GPU acceleration,” 2017.

[39] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,”arXiv preprint arXiv:1502.03167, 2015.

[40] NoiseX-92. [Online]. Available: www.speech.cs.cmu.edu/comp.speech/SectionI/Data/noisex.html