Proceedings of the International Conference on Inventive Computation Technologies (ICICT 2023)
IEEE Xplore Part Number: CFP23F70-ART; ISBN: 979-8-3503-9849-6

2023 International Conference on Inventive Computation Technologies (ICICT) | 979-8-3503-9849-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICT57646.2023.10134295

Design of a Real-Time GAN based Speech Recognizer for Consumer Electronics

Pubali Roy
Department of CORI and ECE
PES University
Bengaluru, India
pubaliroy@pesu.pes.edu

Pranav Bidare
Department of CORI and ECE
PES University
Bengaluru, India
pranavrbidare@pesu.pes.edu

Priya Bharadwaj
Department of CORI and ECE
PES University
Bengaluru, India
priyabharadwaj@pesu.pes.edu

Manikandan J
Department of CORI and ECE
PES University
Bengaluru, India
manikandanj@pes.edu

Abstract **– Modern consumer electronics including automotive electronics, televisions, microwave ovens, music systems, refrigerators with speech controlled features and hands-free operation have spearheaded research in designing smart electronic devices for consumers. Real-time speech recognizer is the main module for these systems and a lot of research is in progress with the design of real-time speech recognizers with a quicker recognition time being considered as one of the challenges. Generative Adversarial Networks (GAN) are mainly used with two dimensional signals such as image for applications such as recognition, synthesis, translation etc. In this paper, an attempt is made to design and evaluate a real-time GAN based pattern recognizer for one-dimensional speech signal. In order to achieve this, the one-dimensional speech signal is first converted into a two dimensional spectrogram and fed to the GAN model for recognition. The proposed speech recognizer yielded a maximum recognition accuracy of 100% with a recognition time of 49.10ms per word. The proposed work can be easily employed to design various smart consumer electronics.**

Keywords – Consumer Electronics, Generative Adversarial Networks, Speech Recognition.

I. INTRODUCTION

Consumer electronics are becoming smarter with the advent of machine learning algorithms for speech recognition. Design and implementation of a voice recognition chat bot for customer assistance is proposed in [1]. Design of a system to identify driver based on voiceprint and acoustic sensing is proposed in [2]. Design of a real-time emotion recognition and analysis system for smart home assistants is proposed in [3]. Design of real-time speech based systems for consumer electronics using various machine learning algorithms are in huge demand with several consumer electronic companies working in these areas.

Generative Adversarial Networks (GAN) have become a rapidly changing and exciting field capable of generating realistic examples across a range of problem domains dealing with two dimensional input signals, mostly images. Use of GAN for image enhancement with specular highlight removal is proposed in [4], GAN for face recognition is proposed in [5], GAN for multiple object tracking in UAV videos is proposed in [6]. GAN is extensively used for image to image translation tasks too such as translating photos of day to night or summer to winter and many more transformations. Translation of visible images into infrared domain using GAN is proposed in [7] and transformation of human faces into animated characters using GAN is proposed in [8]. Use of GAN for face age synthesis capable of translating an input face into an aging face is reported in [9].

A review on research in progress towards using GAN for speech processing, speech synthesis, speech enhancement, speech conversion in automatic speech recognition is reported in [10], with most of the papers focusing on data augmentation to enhance the performance of the speech recognizer. Use of GAN to generate audio samples to improve readability of words to recuperate from dyslexia is reported in [11], use of GAN for data augmentation generating newer samples thus enlarging the size and diversity of training data for robust speech recognition is proposed in [12], GAN based augmentation for gender classification using convolutional neural network (CNN) classifier is proposed in [13], use of GAN to address the practical issue of visual imperfections for audio visual speech enhancement is proposed in [14] and many more such applications of GAN are reported in literature.

In this paper, design and evaluation of a real-time GAN based speech recognizer for consumer electronics is proposed wherein the one dimensional input speech signal is first converted into a two dimensional spectrogram and later fed to the GAN model for recognition.

II. PROPOSED GAN BASED SPEECH RECOGNIZER

The block diagram of proposed GAN based real-time speech recognizer for consumer electronics is shown in Fig. 1. The microphone captures input speech signal and is fed to end-point detection block to identify whether the input signal is a speech signal or just ambient noise. Once a valid speech signal is detected using threshold energy in a frame, the endpoint detection block crops the valid speech signal from the actual input as shown in Fig. 2 and Fig. 3. The

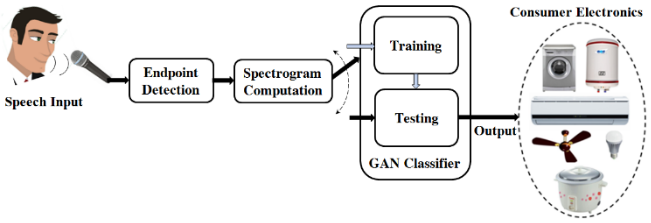


Fig. 1. Proposed real-time speech recognizer for consumer electronics

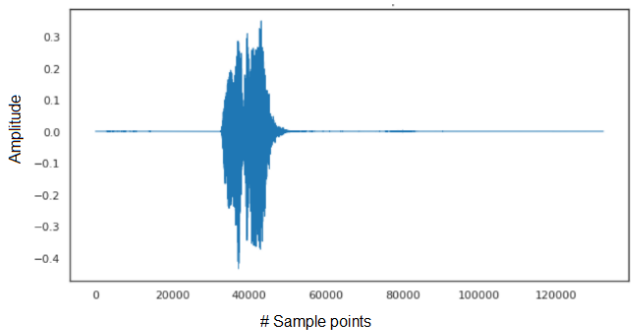


Fig. 2. Input speech signal (Uttered Digit 0) to endpoint detection block

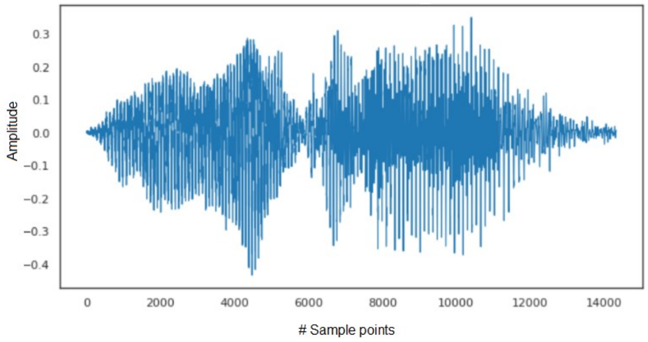


Fig. 3. Cropped speech signal output from endpoint detection block

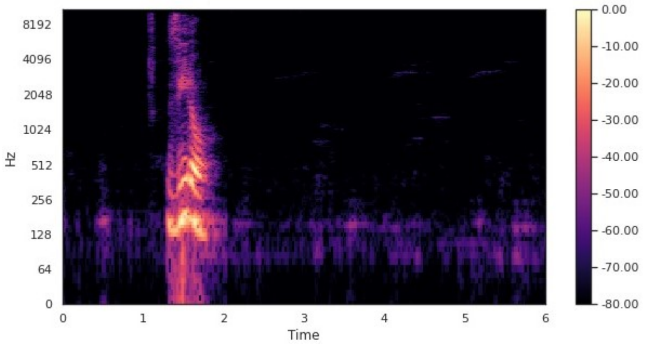


Fig. 4. Spectrogram of the input speech signal (Digit 0)

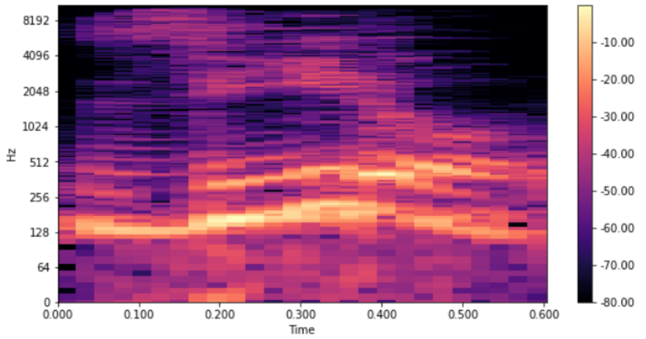


Fig. 5. Spectrogram of the cropped speech signal (Digit 0)

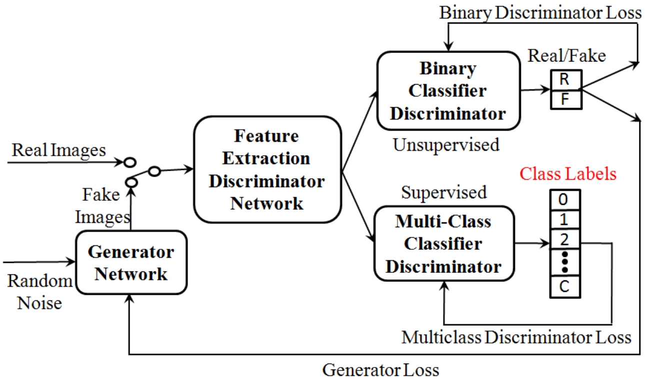

Binary Discriminator Loss
Real/Fake
R
F
Binary Classifier Discriminator
Unsupervised
Real Images
Fake Images
Feature Extraction Discriminator Network
Supervised
Class Labels
Multi-Class Classifier Discriminator
0
1
2
C
Generator Network
Random Noise
Multiclass Discriminator Loss
Generator Loss

Fig. 6. Block diagram of GAN Model for proposed speech recognizer.

spectrogram of actual input speech signal is shown in Fig. 4, wherein it can be observed that the valid speech segments have a threshold greater than −20dB. The spectrogram of cropped speech signal from endpoint detection block is shown in Fig. 5. The spectrogram image of speech signal is then resized to 128×128 pixels and fed to the GAN architecture for training and classification, as shown in Fig. 6. The GAN architecture comprises of Generator and Discriminator, with the Generator using random inputs to generate fake images similar to real images. Details of the generator network are given in Table I. It can be observed from Table I that a random vector of 100 values are first fed to a Dense layer to generate 262144 values, which are reshaped to an image of size 16×16 pixels with 1024 layers. The dense layer output is then fed to a series of four (convolution + relu) layers to generate a 128×128 pixel image with 3 layers (RGB), which are similar to the real images. The Discriminator module of GAN model discriminates between fake and real images. Features are also extracted from the images of all the classes by the discriminator and it gets trained on class labels. Details of the multi class discriminator network are also given in Table I. The spectrogram image of size 128×128 with 3 layers is fed through a series of four (convolution + relu) layers with the image size reduced by half after each layer, but the number of convolution filters are increased with the fourth layer output having 8×8 pixels with 512 layers. This is in turn flattened into a one dimensional data of size 32768 and the number of outputs is equal to the number of classes to be recognized. After training, the test images are fed to feature extraction block of discriminator network, followed by multi-class classifier discriminator to predict the class label for the test input. Based on the word recognized, the corresponding consumer electronics will carry out the necessary function.

TABLE I. GAN ARCHITECTURE FOR PROPOSED SPEECH RECOGNIZER

Generator		Discriminator	
Layer	Output Size	Layer	Output Size
Input	1×100	Input	128×128×3
Dense	1×262144 Reshaped to 16×16×1024	Conv + Pool	64×64×32
		Conv + Pool	32×32×64
		Conv + Pool	16×16×128
Conv + Pool	32×32×512	Conv + Pool	8×8×512
Conv + Pool	64×64×128	Flatten	1×32768
Conv + Pool	64×64×64	Dense	1×32768
Conv + Pool	128×128×3	Output	1×10

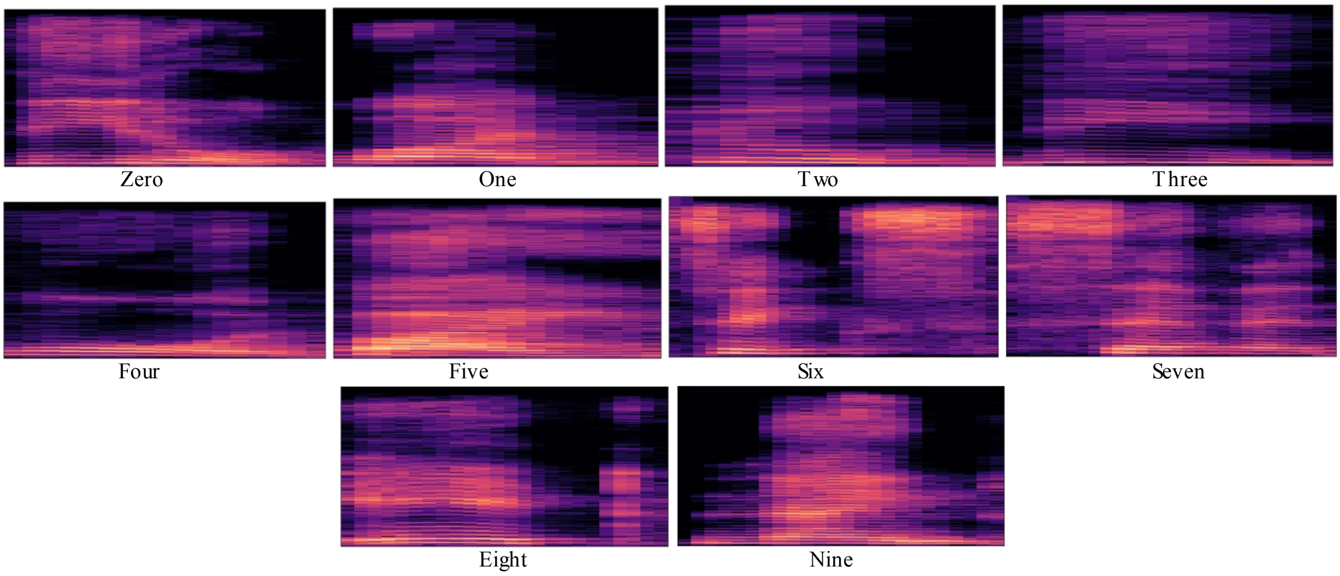


Fig. 7. Spectrogram of digits to be recognized by the proposed GAN based speech recognition model

Most of the consumer electronics have limited functionalities and recognition of isolated words based on the consumer electronic device should suffice. Hence the proposed model is designed to recognize real time digits uttered as an input speech, thus making it a 10-class pattern recognition problem. The spectrogram of all the ten digits is shown in Fig. 7.

III. Experimental Results

The proposed GAN based real-time speech recognizer model is designed and evaluated using Python scripts and resources available for MNIST handwritten digit classification problem in [15]. A Python package called Librosa is used to retrieve information for audio and music analysis. The function librosa.effects.trim( ) performs end point detection trimming the leading and trailing silence from an audio signal based on the top threshold set, below which everything is considered as silence. −20dB is set as the top threshold for proposed work. The function librosa.stft( ) computes discrete Fourier Transform over short overlapping windows of 93ms using 2048 point DFT, which is nothing but the spectrogram of the trimmed speech input signal.

In order to assess the performance of proposed GAN based real-time speech recognizer, for isolated digits, 15 utterances of each digit from three speakers (one male S1 and two female S2, S3) totaling to 450 samples were collected.

Details about the computation time required by proposed real-time speech recognizer for each digit uttered are reported in Table II. Google Colab is used to evaluate the proposed model. The results reported in Table II for each digit is an average for ten runs with a unique sample used for each run. It can be observed from Table II that the overall recognition time is 49.10ms with endpoint detection consuming 1.84ms, computation of spectrogram consuming 2.42ms and GAN classification consuming 44.8ms.

Different combinations of dataset were considered to assess the performance of proposed speech recognizer and the results are reported in Table III using the samples collected. It is observed from Table III that proposed recognizer yielded recognition accuracies ranging between 90.00−100.0%.

Table II. Computation Time For Proposed Recognizer

Digit	Endpoint Detection	Spectrogram Computation	GAN Classification	Total
0	1.77ms	2.45ms	45.2ms	49.42ms
1	1.86ms	2.30ms	43.6ms	47.76ms
2	1.69ms	2.48ms	45.1ms	49.27ms
3	1.88ms	2.52ms	45.4ms	49.80ms
4	1.96ms	2.56ms	44.5ms	49.02ms
5	1.90ms	2.29ms	44.7ms	48.89ms
6	1.87ms	2.36ms	45.6ms	49.83ms
7	1.79ms	2.44ms	44.9ms	49.13ms
8	1.83ms	2.49ms	45.1ms	49.42ms
9	1.88ms	2.34ms	44.2ms	48.42ms
Avg.	**1.84ms**	**2.42ms**	**44.8ms**	**49.10ms**

Table III. Performance Evaluation Of Proposed Recognizer

Train (#Samples)	Test (#Samples)	Accuracy	Epochs
S1 (100)	S1 (50)	100.0%	50
S2 (100)	S2 (50)	90.00%	50
S3 (100)	S3 (50)	92.00%	50
S1(100) + S2(100)	S1(50), S2 (50)	95.00%	50
S2(100) + S3(100)	S2 (50), S3 (50)	91.00%	50
S1(100) + S3(100)	S1 (50), S3 (50)	96.00%	50
S1 + S2 + S3 (150)	S1 (50)	100.0%	200
S1 + S2 + S3 (150)	S2 (50)	94.00%	200
S1 + S2 + S3 (150)	S3 (50)	98.00%	200
S1 + S2 + S3 (150)	S1+S2+S3 (150)	97.33%	200

It was also observed during training that the model gave NaN error after 50 epochs and the number of epochs could be increased on increasing the number of training samples.

Details about the variation in training, validation and testing accuracies for different epochs on using a combination of S1+S2+S3 samples with 150 samples each for training, validation and testing is shown in Fig. 8. The confusion matrix for the model at 200 epochs is given in

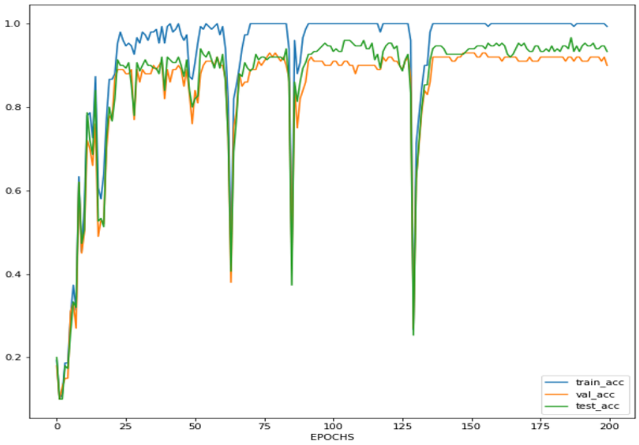
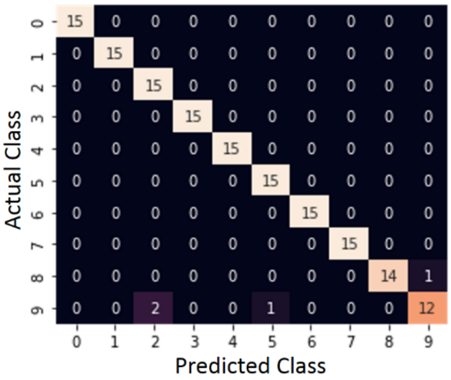


Fig. 8. Performance evaluation of proposed model over different epochs



	0	1	2	3	4	5	6	7	8	9
0	15	0	0	0	0	0	0	0	0	0
1	0	15	0	0	0	0	0	0	0	0
2	0	0	15	0	0	0	0	0	0	0
3	0	0	0	15	0	0	0	0	0	0
4	0	0	0	0	15	0	0	0	0	0
5	0	0	0	0	0	15	0	0	0	0
6	0	0	0	0	0	0	15	0	0	0
7	0	0	0	0	0	0	0	15	0	0
8	0	0	0	0	0	0	0	0	14	1
9	0	0	2	0	0	1	0	0	0	12

Fig. 9. Confusion matrix for the proposed model at 200 epochs

Fig. 9 for a recognition accuracy of 97.33%. It could be observed from Fig. 8 that the training accuracy is 100% beyond 140 epochs, whereas validation and testing accuracy were around 95–98%.

Details about recognition accuracy and other metrics such as precision, recall and F1 score computed for each class from the confusion matrix given in Fig. 9 for proposed recognizer is given in Table IV. Precision represents the quality of positive prediction made by the model, recall quantifies the number of correct predictions made out of all positive predictions that could have been made. F1 score combines the precision and recall metrics into a single metric and is defined as the harmonic mean of precision and recall. The equations employed in proposed work to compute the metrics are as follows:

$$Recognition\ Accuracy = \frac{\#\ Correct\_Predictions}{\#\ Total\_Predictions} \tag{1}$$

$$Precision = \frac{\#\ True\_Positives}{\#\ True\_Positives + \#\ False\_Positives} \tag{2}$$

$$Recall = \frac{\#\ True\_Positives}{\#\ True\_Positives + \#\ False\_Negatives} \tag{3}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

TABLE IV. PERFORMANCE METRICS FOR PROPOSED SPEECH RECOGNIZER

Digit	Accuracy	Precision	Recall	F1 Score
0	100.0%	1.00	1.00	1.00
1	100.0%	1.00	1.00	1.00
2	100.0%	0.88	1.00	0.94
3	100.0%	1.00	1.00	1.00
4	100.0%	1.00	1.00	1.00
5	100.0%	0.94	1.00	0.97
6	100.0%	1.00	1.00	1.00
7	100.0%	1.00	1.00	1.00
8	93.33%	1.00	0.93	0.97
9	80.00%	0.92	0.80	0.86
Avg.	**97.33%**	**0.974**	**0.973**	**0.974**

In order to assess the performance of proposed real-time speech recognizer using GAN, results obtained from proposed work are compared with the results reported in literature in Table V for isolated digit recognition problem using various other machine learning algorithms. It can be observed from Table V that the performance of proposed work in terms of recognition accuracy and recognition time is on par and better in few cases with the results reported in the literature.

TABLE V. COMPARISON WITH RESULTS REPORTED IN LITERATURE

Ref	Algorithm	#Testing	Accuracy	Recog. Time
[16]	CNN	500	99.60%	276.6 ms
[17]	MFCC+SVM	100	99.00%	27.05 ms
[17]	MFCC+RVM	100	98.00%	25.66 ms
[18]	MFCC+SVM (Phoneme)	100	95.00%	671.9 ms
Prop. work	GAN	50	100.0%	49.10 ms
Prop. work	GAN	100	97.33%	49.10 ms

IV. CONCLUSION

Generative Adversarial Networks (GAN) are mostly used for image and video processing. In this paper, design and evaluation of a real-time speech recognition system using Generative Adversarial Networks (GAN) is proposed for isolated digit recognition, focusing on applications related to consumer electronics having limited number of words to be recognized. The proposed model could recognize isolated digits in 49.10ms with a maximum accuracy of 100%. The proposed speech recognizer can be easily implemented on various electronic devices enabling them to act as speech controlled devices.

REFERENCES

[1] M Kathirvelu et al., "Voice Recognition Chat bot for Consumer Product Applications," *Int. Conf. on Data Sci. and Inform. Sys.*, Hassan, India, 29-30 July 2022, pp. 1-5.

[2] KH Chan and CM Chao, "DriverID: Driver Identity System Based on Voiceprint and Acoustic Sensing," *Int. Conf. on Consumer Electronics - Taiwan*, Taipei, Taiwan, 06-08 July 2022, pp. 45-46.

[3] R Chatterjee et al., "Real-Time Speech Emotion Analysis for Smart Home Assistants," in *IEEE Tran. on Cons. Electr.*, Volume 67, no. 1, pp. 68-76, Feb. 2021.

[4] Guangwei Hu et al., "Mask-guided cycle-GAN for specular highlight removal," *Patt. Recog. Letters*, Volume 161, September 2022, pp. 108-114, ISSN 0167-8655.

[5] Samik Banerjee and Sukhendu Das, "LR-GAN for degraded Face Recognition," *Patt. Recog. Letters*, Volume 116, December 2018, pp. 246-253.

[6] Hongyang Yu et al., “Conditional GAN based individual and global motion fusion for multiple object tracking in UAV videos,” *Patt. Recog. Letters*, Vol. 131, 2020, pp. 219-226.

[7] Mehmet and Sedat Ozer, “InfraGAN: A GAN architecture to transfer visible images to infrared domain,” *Patt. Recog. Letters*, Vol. 155, March 2022,pp. 69-76.

[8] S. Re, J. Li, Y. Li and J. Mao, "Improved GAN Model for Image Animation," *Int. Conf. on Information Systems and Computer Aided Education*, Dalian, China, 23-25 September 2022, pp. 838-842.

[9] Chenglong Shi et al., “CAN-GAN: Conditioned-attention normalized GAN for face age synthesis,” *Patt. Recog. Letters*, Volume 138, October 2020, pp. 520-526.

[10] Aamir Wali et al., “Generative adversarial networks for speech processing: A review,” *Computer Speech & Language*, Vol. 72, March 2022, 101308.

[11] Atkar G and Jayaraju P., “Speech synthesis using generative adversarial network for improving readability of Hindi words to recuperate from dyslexia,” *Neural Comput Appl.*, February 2021, Vol. 33, Issue : 15, pp. 9353-9362.

[12] Yanmin et al., “Data augmentation using generative adversarial network for robust speech recognition,” *Speech Commn.*, Vol. 114, November 2019, pp. 1-9.

[13] H Bořil and S Horn, “GAN-Based Augmentation for Gender Classification from Speech Spectrograms,” *Int. Conf. on Electr., Comp. and Energy Techn.*, Prague, 20-22 July 2022, pp. 1-6.

[14] Mandar Gogate, Kia Dashtipur and Amir Hussain, “Towards robust realtime audiovisual speech enhancement,” Cornell University, Dec 2021, pp.1-12.

[15] Sreenivas B, “Semi-supervised GAN”, https://github.com/bnsreenu/python_for_microscopists/tree/master/259_semi_supervised_GAN, March 2022.

[16] Pavan et al., “Design of a Real-Time Speech Recognition System using CNN for Consumer Electronics,” *Zooming Innovation in Consumer Technologies Conf.*, Novi Sad, Serbia, 26-27 May 2020, pp. 5-10.

[17] Shruthi et al., “Design and Evaluation of a Real-Time Speech Recognition System,” *Int. Conf. on Advances in Comp., Commun. and Informatics*, Bangalore, India, 19-22 September 2018, pp. 425-430.

[18] Karan et al., “Design of a phoneme based voice controlled home automation system,” *Int. Conf. on Cons. Electr. Asia*, Bengaluru, India, 05-07 October 2017, pp. 31-35.