2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT) | 979-8-3503-9763-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEEICT56924.2023.10157492

CycleGAN based Speech Enhancement Using Time Frequency Masking

Didin Skariah[a,b*], Rajeev Rajan[c] and Joshua Thomas[a]
[a]*Department of Electronics and Communication*, College of Engineering Trivandrum, Thiruvananthapuram, India
[b]*Bharat Sanchar Nigam Limited*, Thiruvananthapuram, India
[c]*Department of Electronics and Communication*, Government Engineering College, Barton Hill, Thiruvananthapuram, India
*e.mail:didinskariah@gmail.com

***Abstract*— The process of separating one or more desired speech signals from a collection of mixed speech signals or noise is known as audio source separation. The proposed work aims to restore clear audio by reducing noise in mixed audio. These technologies, which are now among the most researched in the field of audio signal processing, are essential to the accomplishment of successful noise removal. Initial source separation is accomplished through time frequency filtering and conditional adversarial networks are used to enhance the quality of the separated sound. A pitch tracking method is used to determine the pitch tracks that are present in the mixed speech and binary masks are constructed that correspond to each pitch track and its harmonics. Using conditional adversarial networks built on CycleGAN and working on the spectrogram images, the noise is removed and speech source is separated. Cycle GAN is used to train the masked audio signal spectrograms and the clean ground truth spectrograms. By using the inverse Short-Time Fourier Transform, the reconstructed spectrogram is transformed back into speech signals. The model's effectiveness is evaluated objectively and the performance of the proposed model is comparable to that of the baseline models.**

Keywords*— *audio source separation, speech enhancement, GAN

I. Introduction

For humans, speech is the main form of communication. We frequently encounter speech signal degradation due to noise and other factors when utilizing speech technologies in real-world settings. Such impacts distort the intended speech signal. The research community has become more interested in the critical and difficult topic of audio source separation in recent years. Speech enhancement refers to techniques that aim to lessen these distortions and improve the pleasantness of speech . Speech enhancement is used in a wide range of applications including mobile phones, teleconferencing systems, voice recognition and hearing aids. For all speech processing techniques, including automatic speech recognition (ASR), speech enhancement is the primary pre-processing task to remove the noise.

In recent years, audio source separation has seen the extensive application of Generative Adversarial Networks (GANs). When training is carried out utilizing a lot of data, generative adversarial networks can be utilized to improve speech quality. It has been found that various GAN architectures are very effective in dealing with speech enhancement challenge.

This paper proposes an audio enhancement method that employs CycleGAN to do the transformation using spectrogram images. A pitch detection method is used to first estimatethe pitch tracks that are present in the mixed audio. These pitch recordings are used to create binary masks, which are then time frequency filtered. The masked spectrograms and the ground truth spectrograms are fed to the CycleGAN for training. The refined spectrogram obtained from the CycleGAN is used to reconstruct the clean speech source by adding phase from the mixed audio. One classification of speech enhancement using GAN is the type of input applied to the model. The input to the GAN is a raw noisy audio waveform in [1], [2], [3] while the magnitude spectrogram or Mel-spectrogram by taking the STFT of the raw audio is the input to the GAN in [4], [5], [6]. The magnitude spectrogram does not have any phase information in it. Reconstruction of clean speech from the magnitude spectrogram is possible only if the phase information is added to it. Generally, the phase from the original noisy audio is extracted for this purpose.

Another classification of speech enhancement methods is mapping-based [1], [2], [3], [4] and masking-based methods [5], [6], [7]. In mapping-based method, the algorithm tries to learn a function that relates the noisy input and clean target audio. A mask is computed and applied to the input noisy spectrogram in the masking-based method in order to reduce the noise. GAN training is used to further improve this masked output. When speech enhancement was done using conventional methods which rely mostly on first-order statistics, a limited number of noise conditions are only got enhanced. SEGAN [3] was introduced as the first attempt at speech enhancement using GAN which operates at the waveform level and incorporates various noise conditions. The model is trained to reduce the loss which includes both adversarial losses and L1 losses. The advantages of SEGAN include fast speech denoising because of the presence of an encoder-decoder model with a fully-convolutional structure, the absence of recursive operations in the generator and less number of parameters.

Later FSEGAN [4], which performs GAN enhancement in the frequency domain employs a spectral feature mapping approach similar to pix2pix. According to the FSEGAN authors, frequency-domain GAN enhancement will be more efficient than time-domain GAN enhancement. FSEGAN showed enhanced ASR performance over the earlier time-domain method and the word error rate is reduced compared to the traditional multi-style training system. The similarities between SEGAN and FSEGAN architecture are both use skip connections, both are deterministic models by removing latent codes from the generator and batch normalization is excluded in both architectures. Various speech enhancement methods and audio separation methods can be seen in [8] to [16]. MetricGAN [10] authors take the idea of increasing human auditory perception by considering the STOI, PESQ, etc. as objective functions to optimize compared to previous methods where the objective function based on adversarial loss and simple L1 or L2 loss is used that does not reflect human auditory perception. Speech Enhancement using Forked Generative Adversarial Networks with Spectral Subtraction [14] employs a forked GAN structure to derive both speech and noise details from the log power spectra (LPS) of noisy speech waveforms.

In Section II, the proposed method is briefly described. Part III provides network architecture and Part IV gives an explanation of the performance evaluation. Section V provides an analysis of the findings and Section VI serves as the paper's conclusion.

II. PROPOSED METHOD

The suggested technique estimates the pitch from the mixed audio signal using a pitch-tracking algorithm. The approximated pitch tracks are later used to create binary masks. Finally, using binary masks, refined spectrograms are produced from the mixed spectrograms. The conditional generative adversarial network is used in image-to-image translation on the masked spectrogram to produce improved spectra. The clean audio is retrieved back from the clean spectrogram using the phase extracted from the mixed speech signal.

The noisy spectrogram is generated from the mixed audio signal. The Duan-Han technique [17] utilizes multi-pitch estimation and streaming to estimate individual pitch tracks. It employs the maximum-likelihood method to determine the pitch tacks. The process, which is based on the maximum likelihood approach to multiple fundamental frequency estimation, uses the power spectrum of a time frame as the observation. Peaks and non-peak regions are included in power spectrum models of the observed data. The total likelihood function is then calculated as the product of the likelihoods determined in both regions.

Finding candidate F0 s in the mixture is aided by the peak region likelihood. To obtain the masked spectrograms, the mask is applied to the input spectrogram. Information about the phases is saved for reconstruction in the end. The original noisy spectrogram and binary mask are shown in Fig.1.

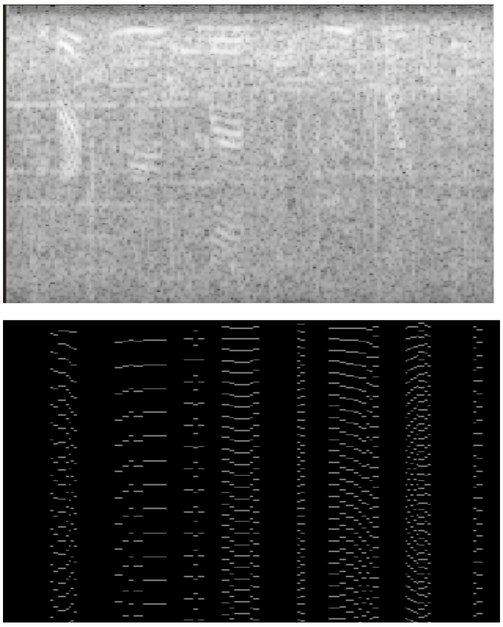

Fig. 1. Noisy speech spectrogram and binary mask

CycleGAN[18], which was originally used in image applications, is used here for speech enhancement. Its main components are Generator and Discriminator. Two generator models are used in this architecture: Generator-A for generating images for the first domain and Generator-B for generating images for the second domain. A discriminator model for every generator exists. Using generated images from Generator-A and real images from Domain- A, the first discriminator model (Discriminator-A) determines if the images are real or fake. Using generated images from Generator-B and real images from Domain-B, the second discriminator model (Discriminator-B) determines whether the images are real or fake. The block diagram of proposed method is shown in Fig. 2.

III. NETWORK ARCHITECTURE

The CycleGAN model is shown in Fig. 3. The discriminator model is in the position to determine if a generated or actual image is fake or real by accepting it as input. A PatchGAN model is used to implement the discriminator.

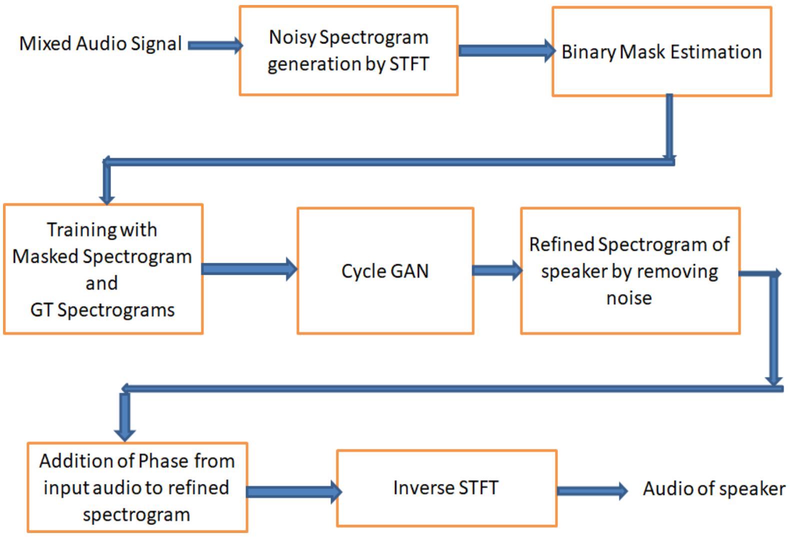
Mixed Audio Signal
Noisy Spectrogram generation by STFT
Binary Mask Estimation
Training with Masked Spectrogram and GT Spectrograms
Cycle GAN
Refined Spectrogram of speaker by removing noise
Addition of Phase from input audio to refined spectrogram
Inverse STFT
Audio of speaker

Fig. 2. Block Diagram of the proposed method

A square or one-channel feature map of predictions by the PatchGAN discriminator model can be done in place of a single value, as is the case with conventional discriminator models. The 70×70 refers to the effective receptive field of the model on the input, not the actual geometry of the output feature map. The architecture of discriminator is shown in Fig. 4.

CycleGAN generator employs a series of residual network (ResNet) convolutional blocks to transform the image, a number of upsampling convolutional blocks to produce the output image, and a sequence of downsampling convolutional blocks to encode the input image. The architecture of generator is shown in Fig. 5.

Cycle GAN is employed to transfer an image's characteristics to another. It can be done by taking an image input and turning it into a reconstructed image using the generator G. Finally, using a generator F, this process is reversed to get original image back. The mean squared error loss between the original and recreated image is then calculated. The intuition that both mappings should be reversals of one another and that both mappings should be bijections is reinforced by the measure of cycle consistency loss.

IV. PERFORMANCE EVALUATION

A. *Data sets*

We used the Voice Bank Corpus dataset [19]. It consists of 28 speakers - 14 male and 14 female of the same accent region (England). Different noises from noise dataset such as DEMAND can be added to the clean speech data set to create a training dataset as a pair. Different noises include voice babble, factory noise, environmental noises etc. We also used NOIZEUS dataset which contains IEEE sentences corrupted by different real-world noises.

B. *Evaluation Metrics*

We used Perceptual Evaluation of Speech Quality (PESQ) metrics for the performance evaluation. The PESQ Algorithm is designed to predict subjective opinion scores of a degraded audio sample. PESQ returns a score from 4.5 to -0.5, with higher scores indicating better quality. PESQ is designed to analyze specific parameters of audio, including time warping, variable delays, transcoding, and noise.

Another objective measure is Segmental SNR (SSNR). Instead of analyzing the entire signal, the segmental signal-to-noise ratio (SSNR) determines the average of the SNR values of brief segments typically 10 to 15 ms.

C. *Experimental Set-up*

Spectrograms are generated using short time fourier transform with a segment size of 256, fft size of 512 and hamming window is used. Masked Spectrograms are obtained after pitch estimation. Image pairs of the masked spectrograms and the spectrogram of the corresponding clean audio (ground truth) are given as inputs to the Cycle GAN model.

For Cycle GAN training, masked spectrogram and ground truth spectrogram pairings of about 10,000 spectrogram images are employed. In order to test and assess performance, 1000 samples are used.

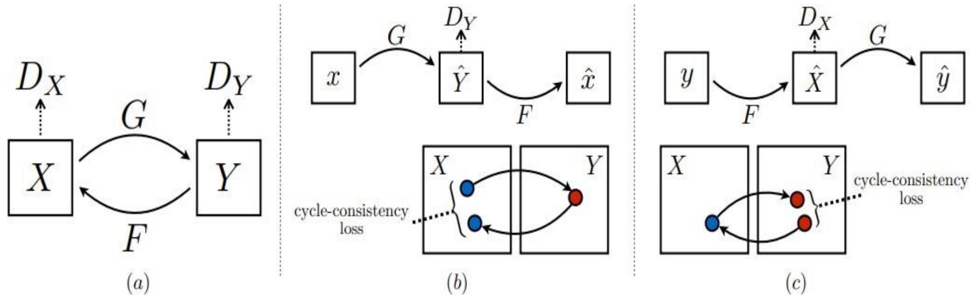

Fig. 3. Cycle GAN[18], G,F – Generators ; Dx,Dy – Discriminators

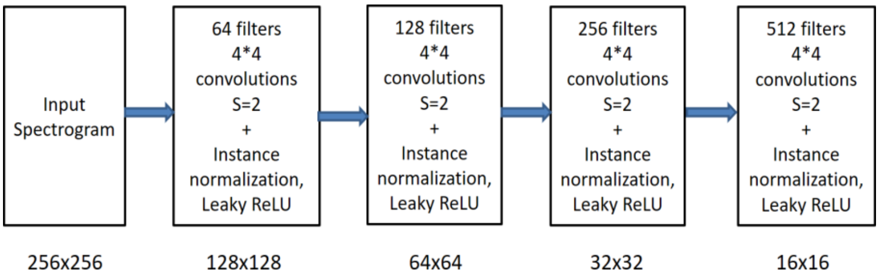
Input Spectrogram
64 filters 4*4 convolutions S=2 + Instance normalization, Leaky ReLU
128 filters 4*4 convolutions S=2 + Instance normalization, Leaky ReLU
256 filters 4*4 convolutions S=2 + Instance normalization, Leaky ReLU
512 filters 4*4 convolutions S=2 + Instance normalization, Leaky ReLU
256x256
128x128
64x64
32x32
16x16

Fig. 4. Cycle GAN[18]- Architecture of Discriminator, s= stride

100 epochs with a learning rate of 0.0002 and a batch size of 1 is used for training. Total number of parameters for each generator is 11.378 M and for each discriminator is 2.765 M. The baseline model is SEGAN [3] which is having an encoder-decoder architecture and raw waveform as the input. It is a mapping based deterministic model having skip connections. With 90 epochs and a batch size of 400, a PESQ of 2.16 and SSNR of 7.73 are observed with voice bank corpus data set.

V. RESULTS AND ANALYSIS

The result analysis can be done in two ways. One is the comparison of spectrograms. Figure 6 displays the outcomes for the CycleGAN reconstructed image. It is evident from the newly created spectrogram image that spectrum aspects of a near-real reconstruction have been accomplished. After adding phase information from the initial mixed audio, inverse STFT is used. The clarity of speech after removal of noise is found good and promising.

The other one is the objective evaluation measure PESQ which was developed to simulate subjective evaluations of voice quality that are frequently used in telecommunications. SSNR gives the average of the SNR values of brief segments in the audio. PESQ and SSNR scores of the proposed method and reference model is shown in TABLE I. The perceptual quality of the results is found promising and satisfactory.

TABLE I

OBJECTIVE EVALUATION METRICS OF PROPOSED AND REFERENCE MODEL

Models	PESQ	SSNR
Cycle GAN based SE	1.78	6.83
SEGAN [3]	2.16	7.73

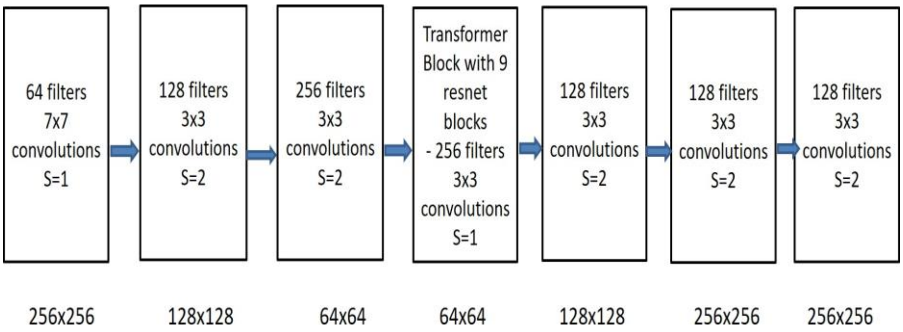

Fig. 5. Cycle GAN [18]- Architecture of Generator, s= stride

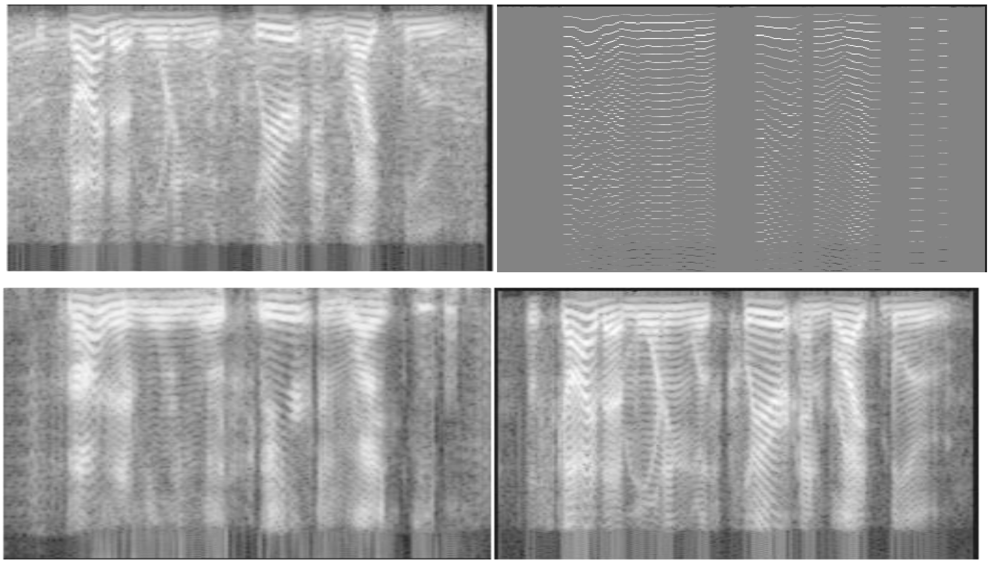

Fig. 6. (a) Spectrogram of noisy speech (b) Masked Spectrogram (c) Spectrogram of Reconstructed speech (d) Ground truth spectrogram.

VI. CONCLUSION

GAN is initially utilized in image processing, but later it is discovered to be beneficial in speech applications. Cycle GAN is used here for speech enhancement for which masked spectrogram is given as the input. The coupled networks of GAN are quite powerful for enhancing speech. The conditional GAN improves the separated speech's quality. The time-frequency masking function is often created using GAN frameworks, but in our approach, masked spectrogram is refined using GAN.

Both approaches validate GAN's planned source separation attempts. Even though the results are comparable, the proposed method using Cycle GAN training is straightforward, reliable and easily convergent because audio samples are processed as images (spectrograms). The key benefit of this approach is that the processing is carried out in the TF domain, which facilitates the masking process and easy as well as generalized learning process. Mel frequency cepstral coefficients (MFCC), Mel spectrograms and magnitude spectrograms are features in the TF domain. In contemporary neural network architecture, acoustic characteristics such as magnitude and mel spectrograms are preferred over MFCC because they are better suited for deep learning techniques. This proposed method, which is a magnitude spectrogram-based feature improvement model, performs well in the field of speech enhancement arena. The application of this work can be further extended to other audio applications such as musical audio track separation and improving the accuracy of speech to text conversion.

REFERENCES

[1] D. Baby and S. Verhulst, ”Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty,” in 2019 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc., 2019

[2] H. Phan et al., ”Improving GANs for speech enhancement,” IEEE SignalProcess. Lett., vol. 27, 2020.

[3] S. Pascual, A. Bonafonte, and J. Serra, ”SEGAN: Speech enhancement generative adversarial network,” in Proc. Annu. Conf. Int. Speech Communication Association Interspeech 2017.

[4] C. Donahue, B. Li, and R. Prabhavalkar, ”Exploring speech enhancement with generative adversarial networks for robust speech recognition” in 2018 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2018

[5] M. H. Soni, N. Shah, and H. A. Patil, ”Time-Frequency maskingbased speech enhancement using generative adversarial network,” in 2018IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc., 2018

[6] A. Pandey and D. Wang, ”On adversarial training and loss functions for speech enhancement,” in 2018 IEEE Int. Conference Acoustics, Speech and Signal Processing Proc., 2018.

[7] Lin, J., Niu, S., Wijngaarden, A.J.v., McClendon, J.L., Smith, M.C., Wang, K.-C. (2020) Improved Speech Enhancement Using a Time- Domain GAN with Mask Learning. Proc. Interspeech 2020, 3286-3290, doi: 10.21437/Interspeech.2020-1946

[8] Kim HY, Yoon JW, Cheon SJ, Kang WH, Kim NS. A multi-resolution approach to gan-based speech enhancement. Applied Sciences. 2021 Jan13;11(2):721.

[9] Su J, Jin Z, Finkelstein A. HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. arXiv preprint arXiv:2006.05694. 2020 Jun 10.

[10] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, “MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” arXiv preprint arXiv:1905.04874, 2019

[11] Pascual, S.; Serra, J.; Bonafonte, A. Time-domain speech enhance- ment using generative adversarial networks. ”Speech communication”,1 Novembre 2019, vol. 114, p. 10-21.

[12] Adiga N, Pantazis Y, Tsiaras V, Stylianou Y. Speech Enhancement for Noise-Robust Speech Synthesis Using Wasserstein GAN. InINTER- SPEECH 2019 Sep (pp. 1821-1825).

[13] Li H, Fu SW, Tsao Y, Yamagishi J. iMetricGAN: Intelligibility enhance- ment for speech-in-noise using generative adversarial network-basedmetric learning. arXiv preprint arXiv:2004.00932.

[14] Lin J, Niu S, Wei Z, Lan X, Wijngaarden AJ, Smith MC, Wang KC. Speech enhancement using forked generative adversarial networks with spectral subtraction. Proceedings of Interspeech 2019. 2019 Sep.

[15] Abdulatif S, Armanious K, Guirguis K, Sajeev JT, Yang B. Aegan: Time- frequency speech denoising via generative adversarial networks. In2020 28th European Signal Processing Conference (EUSIPCO) 2021 Jan 18 (pp. 451-455). IEEE.

[16] Joseph S, Rajan R. Cycle GAN-Based Audio Source Separation Using Time–Frequency Masking. Circuits, Systems, and Signal Processing. 2023 Feb;42(2):1163-80.

[17] Zhiyao Duan, Bryan Pardo, and Changshui Zhang. Multiple funda- mental frequency estimation by modeling spectral peaks and non-peak regions. IEEE Transactions on Audio, Speech, and Language Processing, 18(8):2121–2133, 2010.

[18] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired imageto-image translation using cycle-consistent adversarial networks. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2242–2251, 2017.

[19] C. Veaux, J. Yamagishi, and S. King, ”The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” in 2013 Int. Conf. Oriental COCOSDA 2013 Conf. Asian Spoken Language Research and Evaluation, 2013.