ISSN 2347 - 3983
Volume 10. No.9, September 2022
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter011092022.pdf
https://doi.org/10.30534/ijeter/2022/011092022

Study of Generative Adversarial Networks for Acoustic Signal Enhancement: A Review

Shibani Kar[1]
[1]Assistant Professor,
Department of Electronics and Communication Engineering,
Sambalpur University Institute of Information Technology,Jyoti Vihar,Burla,Odisha
skar@suiit.ac.in

Received Date: August 2, 2022 Accepted Date : August 24, 2022 Published Date : September 07, 2022

ABSTRACT

Acoustic signals enhancement is an important research topic. It has many applications like cochlear implants, speech and speaker recognition, hearing aids, mobile phones etc. The signals processed by these system are always susceptible to noises. Hence, algorithms are required to extract clean signal from noisy ones. Nowadays , deep neural network are the most sought after tool for signal enhancement. Generative Adversarial Network(GAN) is also one of the recent approaches applied to signal enhancement domain. More work is performed by GANs in image and video processing. To the best of my knowledge no review work on the usage of GANs for acoustic signal enhancement have been done. This paper is a review on the use of GANs for acoustical signals enhancement where speech signal is used as acoustic signal. The paper provides in a summarized manner about the basic GAN architectures and its limitations, feature sets used as input to GAN, limitations, performance evaluation measures and future directions.

Key words: Generative adversarial network, signal enhancement, acoustic signal. deep neural network, adversarial network.

1. INTRODUCTION

Nowadays due to advancement in machine learning techniques and improvements in deep learning methods, play a key role in acoustic signal enhancement. Acoustic signal examples are sounds, speech, audio that represents variation in acoustic pressure w .r .t time. Acoustic signals are used in a number of applications like mobile communication and telephony, cochlear implants, automatic speech (or speaker) recognizer, hearing aids, design of silent speech interfaces and many more applications that involve the use of audio, speech sound as the primary source of data.

Due to its wide usage as mentioned above, they are more prone to be corrupted by external noises that may affect the accuracy of the data carried by the signal at the reception point. The external noises are mainly environmental noises, sounds generated by birds, humans, animals, sounds generated by industries, office, environment etc. when these signals are processed to remove these interferences , the intelligibility and quality of the signal get severely affected. This provides a distorted output and affect the overall objective of the applications as the data is lost due to these processing methods. The de-noising of acoustic signals without affecting their intelligibility and quality are known as acoustic signal enhancement methods. Many classical methods for acoustic signal enhancement as per the available literature are:

1. Spectral Subtraction Method
2. Subspace Method
3. Wiener filter
4. Neural Networks
5. Auto-encoders
6. Deep Neural Networks-RNN,CNN
7. Minimum mean square error

Recently the researchers are using deep learning algorithms like RNN and CNN for enhancing the quality of acoustic signals. GAN are one such deep learning algorithms that is widely used for processing image signals. Due to their rising popularity for image signal processing, the GAN architectures are used nowadays for acoustic signal enhancement. GAN network are used in adversarial scenarios where a generator network and a discriminator network are two neural networks that are competing against each other trying to fool each other. Using an appropriate loss function for each neural network the overall performance of the GAN network can be enhanced. The paper compares the performance of the various type of GAN architectures used for acoustic signal enhancement where the comparison is done with respect to the type of neural network used in generator and discriminator, the type of loss function used for training and testing, the datasets used for evaluating the performance, types of noises and noise

conditions used, evaluation metrics, type of input features used, application and the gaps in the available literature.

2. Review of GAN architectures for Acoustic Signal Enhancement

2.1 Speech Enhancement Generative Adversarial Network(SEGAN)

SEGAN is one of the state of art model for speech enhancement. The method uses adversarial learning for enhancing signals in time domain. The model consist of generator and discriminator network. The generator network is an encoder decoder model . The generator model design is made fully convolution to focus on temporally-close relations in the input signal. The fully convolution layer reduces the number of training parameters and training time. In the encoding stage, strided convolution layers followed by parametric Rectified linear units(PReLUs). The decoder consist of fractional strided transposed convolutions followed by PReLUs. The Generator layer uses Skip connections connecting each encoding layer to respective decoding layer. The generator model uses L1 norm as a loss function to measure the distance between the generator output and clean samples. The Discriminator layer is one-dimensional convolution layer, it uses two input structure as the generator's encoder stage. It has two input channels and uses virtual batch Norm and uses leaky ReLU as activation function. The model uses open source dataset and generalized by using number of noises. The model is trained for 30 speakers and Voice bank Corpus is used to provide clean sentences. For training 28 speakers and for testing 2 speakers are used. Demand Dataset is used for providing 10 types of noise. During training phase 40 noise conditions are created using 4 types of SNRs(0dB,5dB,15dB,20dB). The model is trained with 80 epochs, lr=0.0002 and batch size of 400. The model has worse PESQ measure as compared to wiener model but it has low speech distortion and removes noise more effectively[1].

2.2 Speech Enhancement Conditional Generative Adversarial Networks(SEcGAN)

The model uses conditional GAN network for the enhancement of speech signal. The spectrogram of speech signal is used as input to the GAN model. The generator uses pix to pix algorithm to transform noisy spectrogram to clean spectrogram. The Generator consist of U-NET network and Discriminator consist of Patch GAN network. Pix to Pix algorithm not only maps the noisy speech to enhanced speech but also learns loss function for the training of GAN. . Loss function consist of adversarial loss(G,D) and L1 norm that measure the distance between the output of G and real samples, L2 norm and perceptual losses. In the proposed model, L1 loss is used as it provides less blurred spectrogram at the output and generalize better as compared to perceptual losses. The signal is sampled at 16KHZ and 512 point STFT is obtained using hamming window. Only 256 samples are chosen as input to the model as the spectrogram image given as an input to model is 256x256x1channel to Generator and 256x256x2channel to discriminator. For training the model, stochastic gradient descent algorithm is used along with ADAM optimizer to optimize the weights of the model during training, epochs=10 and batch size is one. The initial weights of network are initialized with normal distribution where mean is 0 and standard deviation is 0.02. The l1 loss is added to the GAN loss(G,D) by using a scaling factor having value set at 100. At the output signal is reconstructed by using inverse STFT. For performance analysis , PESQ and STOI metric are used. The model is compared to MMSE-STSA and DNN_SE(with IRM mask). Only 5 type of noise: Gaussian , Canteen, babble, Aero plane, Market. The model outperforms MMSE-STSA but the performance of pix to pix is comparable to DNN-SE. The pix to pix has better PESQ but suffers from STOI measures. For Airplane noise, MMSE-STSA performance is best[2].

2.3 Spectral Feature Mapping using FSEGAN

The method discusses the use of GAN for speech enhancement to train an automatic speech recognition system to make it more robust to noise. SEGAN uses raw waveform as input and reduces additive noise but in Automatic Speech Recognition GAN need to suppress both additive and reverberation noise. This task is very complex in time domain and frequency domain values improves the performance of ASR. The approach used in this method of speech enhancement using GAN for training automatic speech enhancement system is spectral feature mapping approach where the input to the GAN is log Mel filter bank spectra that requires less computation and more robust to reverberant noise. The GAN with spectral feature mapping approach is known as FSEGAN that performs spectral feature mapping using Pix to Pix algorithm. FSEGAN is fully convolution where the generator consist of encoder and decoder having 7 layers each and skip connections are used. The final layer of decoder uses linear activation and output a single channel features The source of clean speech data is Wall street journal. Large stereo data set of music and ambient signals are sources of additive noises for MTR training. This data set is collected from You tube recordings and real life environment recordings. The time domain data is converted to frequency domain using STFT with a window size of 32ms and hop size of 10ms. The magnitude spectrum is retained and phase spectrum is discarded. Calculate triangular windows for a bank of 128 filters where filter frequency are equally spaced on Mel scale between 125Hz and 7500Hz. After applying transform to magnitude spectrum we take logarithm of output and normalize each frequency bin to have zero mean and unit variance. The performance of the model lags than the tradition multistyle training approach for automatic speech recognizer. The study shows that the FSEGAN model works better than SEGAN for ASR applications[3].

2.4 Domain Adversarial Training using GAN

The paper discusses noise adaptive speech enhancement system that applies domain adversarial training approach to a basic GAN model. The model consist of encoder -decoder and a discriminator network where the noisy speech is given input to encoder , encoder generates noise invariant features that are given input to the decoder that produces enhanced signal as

output. The discriminator act as a classifier to compare the decoder output with real samples and provides feedback to the encoder to produce accurate noise invariant features to correct the output of the decoder. The signal carries both stationary and non-stationary noises. TIMIT sentences are used as clean utterances where these noises are added to give input to the enhancement model. DAT approach helps to create a system that adapts itself to real time noises. It helps to remove the problem of noise type mismatch that occurs when the system fails to identify the noise type in real world scenarios. This causes degradation of the system performance. Hence, the system need to adapt itself to a variety of noises in real time scenarios and give sustained performance. The model uses BLSTM layer in encoder and decoder and unidirectional LSTM with discriminator. Stationary noises used are pink noise, soft and strong wind noise, car and engine noise. Non-stationary noises used are baby cry, babble noise and cafeteria noise. The input to the model is STFT log power spectra and the output is reconstructed using inverse Fourier transform and overlap and save method. The training data is 500 utterances corrupted with 5 stationary noise types using 6 SNR levels(-5dB to 20dB). For testing 192 sentences are combined with non stationary noises. The BLSTM layer uses 512 nodes as input for both encoder and decoder whereas 257 linear nodes are used at output of decoder for spectrogram estimation. The discriminator layer uses 1024 nodes and fully connected layer with 6 nodes followed by softmax layer to predict the noise type. For training the model Adam optimizer is used with learning rate 0.0001 (SE model) and 0.0005 for discriminator model. For the entire operation of model the hyper parameter value is set below 0.1. The model uses PESQ, SSNR and STOI for measurement of its performance. The result of the DAT method is PESQ(19%), STOI(27%), SSNR(39.3%)[4].

2.5 Speech Reconstruction using GAN

The paper proposes the implementation of speaker dependent end to end model for voiced speech generation based on GAN for pathological applications and design of silent speech interfaces. Many patients suffering from aphonia either produce whispered speech or monotone speech. The speech is highly non audible and need to be converted to voiced sound. The SEGAN model is used in the work for generating voiced speech. The results are compared with RNN and baseline methods. The SEGAN model consist of Generator and discriminator where learning rate of Generator is 0.0001 and for discriminator is 0.0004. ADAM optimizer is used for training. The dataset used is CMUarctic corpus. The results shows the comparison of pitch counters. The model doesn't provide the quality and intelligibility of generated speech[5].

2.6 Adversarial Feature Mapping

The paper uses adversarial feature mapping approach along with discriminator network to enhance the noisy speech samples to clean speech samples by minimizing feature mapping loss function and discriminator loss function. The method is tested with Seone-Aware Adversarial Feature Masking approach to optimize the seone classification loss. Chime-3 data set is used. The input to the network is log Mel filter bank and delta features. The proposed model is used for ASR applications. For feature mapping the model consist of LSTM-RNN model and feed forward DNN is used as discriminator. The word error rate is used as performance measure. The WER rate for AFM real noisy data and feature mapping method is 16.95% and 5.27%. The seone-aware AFM word error rate is 9.85%[6].

2.7 Deep Complex Convolution Recurrent GAN(DCCRGAN)

The model uses deep complex convolution recurrent GAN for speech enhancement. The performance measure metric used are PESQ,STOI,CBAK,CSIG. The dataset used is Voice bank corpus and Demand dataset. The model consist of generator and discriminator. The generator uses encode decoder model that takes noisy waveform as input. The waveform is converted to short time Fourier transform(STFT) using a convolution layer. The LSTM layer is used between encoder and decoder to extract long term information. The encoder decoder network uses complex convolution layer, batch normalization,PReLU layer and skip connections[7].

2.8 Generalized Speech Enhancement using GAN

The proposed GAN model enhances the signal not only with additive and reverberant noise but also signals with speech distortions like clipping, chunk elimination, frequency band removal. A GAN model with acoustic loss and two step training is implemented in the paper. The VCTK corpus is used for training. For training 80 speakers and 14 speakers are used for testing. The model uses multilayer perceptron with PReLu units . The evaluation is performed using Mel Cepstral distortion, F0 room mean square error, voiced/unvoiced frame prediction. The model is compared to SEGAN baseline, SEGAN-Acoustic and SEGANPTACO and performs better than them[8].

2.9 Multistage Enhancement using GAN

The paper discusses the use of multistage enhancement using multiple generators where each generator refines the output received from previous generator. Two new SEGAN framework i.e. iSEGAN and deepSEGAN are used where iSEGAN share the mapping of all enhancement stages and independent mapping is used in deepSEGAN. The model is similar to SEGAN. Dataset used is Voice Bank Corpus and Demand dataset for noise. Raw time domain waveform is the input to the model. The method performs better than baseline SEGAN method[9].

2.10 Improvements in Conditional GAN model

The paper discusses about the need of improvement in conditional GAN(cGAN) for speech enhancement in terms of achieving stability of cGAN model during training. This is achieved by three methods. First the usage of instance normalization in the discriminator model. Second use of Gammatone filter bank in first layer of both generator and discriminator to enhancement speech intelligibility. Third use of pre-emphasis filter in generator by using a trainable

convolution layer of filter length 2 and stride 1. the layer is initialized with weights and trained together with cGAN. The model is similar to SEGAN. The performance metrics and dataset are similar to SEGAN[1]. The proposed cGAN model is compared to LSTM-IRM model and outperforms the model[10].

2.11 Visual Speech Enhancement GAN(VSEGAN)

The model takes both visual and speech data as input to provide enhanced speech output. It uses audio-visual feature fusion strategy. The generator is a multilayer feature fusion convolution network that follows a encoder decoder scheme. The generator consist of audio and video encoder. The performance of the model is enhanced by including visual information[11].

2.12 Dynamic Attention recursive GAN

The paper proposes use of dynamic attention recursive GAN for noise reduction in time-frequency domain. The model uses recursive learning protocol for multistage training of generator, dynamic attention mechanism to control the feature distribution of noise reduction network and Griffin Lim algorithm for reconstructing the signal using phase information. The model uses Voice Bank Corpus and Demand dataset for training and testing the model. The performance of model is better than previous GAN models[12].

2.13 Metric GAN

The proposed method implements a cost function called metric GAN to optimize the objective metrics by connecting the metric with a discriminator. The Voice Bank-Demand Database is used for testing and training and shows better PESQ score. The method is compared with SEGAN,MMSE-GAN, SERGAN, BLSTM, MetricGAN, HiFiGAN and performance is better[13].

2.14 Time domain GAN with Mask Learning

The paper proposes the use of speech enhancement for ASR applications. Most of the enhancement models use neural network that uses feature mapping or mask learning. In the proposed method, the time domain feature mapping and mask learning are integrated and given as an input to GAN model. The waveform is converted two magnitude spectrums using convolution 1-D layers that map the waveforms to spectrogram in complex domain. These speech and noise spectrograms are used to compute the speech mark loss. TIMIT dataset is used and the method outperforms DNN based speech enhancement and SEGAN[14].

2.15 Convolution Neural Network based GAN

The paper discusses the use of Time-Frequency masking based CNN_GAN model for speech enhancement. The model uses Voice bank Corpus and Demand dataset for training and testing. Performance measuring tools are CSIG,PESQ,STOI,CBAK. The model performs better than SEGAN, Pix to Pix L1, Pix to Pix L2, Wiener, CNN models for speech enhancement[15].

2.16 μ-Law SGAN

The paper discusses the use of μ-Law spectrum generative adversarial network for speech enhancement. The model outperforms the available methods. It uses SSNR parametric method that along with other methods-CSIG,CBAK,PESQ,ESTOI that is not used in other methods[16].

2.17 Phase Sensitive GAN(PSMGAN)

The paper proposes a phase sensitive masking based single channel speech enhancement using conditional generative adversarial network. The model uses phase information for speech intelligibility improvement at low SNRs and shows state of art performance in terms of CSIG, CBAK, PESQ, STOI parameters[17].

3. CONCLUSION

The review work shows that the GAN based architectures shows state of art performances. They give better performance for both Time domain input and Time Frequency domain input. The models are tested with a variety of datasets such as Voice bank corpus, TIMIT, Noisex, Demand , ChiMe3, ATR Japanese. The models also uses T-F mask and Phase mask for improving the performances of GAN architectures. The concept of domain adversarial training, multistage generator, usage of discriminator to compute metric losses also helps to enhance the results.

4. FUTURE SCOPE

The performance of GAN architectures can be evaluated by using perceptual weights to train the model. Better convolution architectures can be used to enhance the performance of the model. L1 and L2 norm along with perceptual losses can be used to enhance the system performance.

Ref. No	Year	Architecture	Dataset	Features Used	Training and Testing Data	Comparison model	Performance measure	Application	Result	Limitations
[1]	2017	Generative Adversarial network consist of fully connected, end to end convolution network	Demand dataset for Noise, Voice bank corpus for clean sentences	Raw Waveform	Training set: 28 speakers, 40 different noise conditions Testing Set:2 speakersand 20 noise conditions Noise Type: 10 Noise Condition:(0dB,5 dB,10dB, 15dB)	Wiener model	PESQ, COVL, CSIG, CBAK, SSNR	Hearing Aid, Mobile communications, cochlear implants	1.PESQ:Model fails to improve the perceptual quality of speech. 2.CSIG:Model has reduced speech distortion and removes noise more effectively (CBAK,SSNR) 3. Implementation of speaker independent model	1.No evaluation of intelligibility metric like STOI analysis 2.Performance is Compared with one method only 3.Only 10 noise for training and 5 noise for testing isused. 4.Model fails to improve PESQ score of speech signal. 5. Model can be compared with the performance when features other than time domain are given as an input. Future Scope: 1. Better convolution model can be used 2. Perceptual Weights can be used for training the model. 3. Model can be compared with more speech enhancement methods in terms of PESQ and STOI 4. More noise types can be used.
[2]	2017	Conditional Generative Adversarial Networks using Pix to Pix algorithm	TIMIT, RSR2015 Noise Types 1. Babble 2.Cantine 3.Market 4.Airplane 5.White Gaussian Noise	Spectrogram of signal	Five noise specific front ends, one noise general front end trained on all types of noise.	STSA-MMSE, NS-DNN-IRM, NS-Pix to Pix, NG-DNN-IRM, NG Pix to Pix	PESQ STOI	Hearing aids, speech recognition, mobile communication, automatic speaker verification	1.SEcGAN outperforms MMSE-STSA and shows comparable performance to DNN_SE. 2. Pix to Pix has better PESQ metric 3. Pix to Pix suffers for STOI metric. 4. For Airplane noise, MMSE-STSA is good.	1. Limited noise set is used for training the model. 2. Model need to be compared with other models. 3. Model need to be speaker independent and unseen noises must be used for testing the performance of the model. Future Scope: 1. Testing on unseen noises 2. Noise set must be increased 3. better convolution models must be used for training and testing. 4. Use for other features and time domain data as input.
[3]	2018	Spectral Feature Mapping approach using GAN (FSEGAN)	Wall Street Journal dataset for clean speech and You Tube recordings for noise data	Log Mel filter bank spectra	Additive and reverberant noise	SEGAN, Multi Style Training (MTR)	Word Error rate	Automatic Speech recognition	1. FSEGAN performs better than SEGAN for ASR applications 2. Traditional MTR style performs better than FSEGAN	1.PESQ and STOI are not measured 2. The model is not compared with other models or methods. Future scope 1. Method can be tested with other feature inputs.
[4]	2019	Domain Adversarial Approach using GAN	TIMIT Dataset For clean utterances Noises 1.Stationary (car, pink, wind noise, engine noise) 2. Non stationary -baby cry, cafeteria, babble	STFT Log Power Spectra	Stationary noise (Car noise, engine noise, soft and strong wind noise, pink noise) Non-stationary noise (baby cry, cafeteria, babble)	Baseline methods	PESQ (19%) SSNR (39.3%) STOI (27%)	Real time applications	Model performs better than baseline models using few noise data	The model need to be compared with other models.

[5]	2018	Voiced speech Generation using GAN	CMU Arctic Corpus	Time domain wavefo rm	Whisper ed sound to voiced sound	RNN, SEGA N, Baseli ne metho ds	Pitch counters using histogram graphs	Silent Speech Interface Design, Pathological applications	Model performs better	1. The model need to compute the PESQ and STOI for the generated speech 2. No intelligibility and quality enhancement of speech is done here.
[6]	2018	Adversarial Feature Mapping	Chime-3	Log Mel Filter bank	Far field noise data	Feed forwar d DNN, Advers arial feature mappi ng, Seone aware AFM	Word Error rate	Automatic speech and speaker recognition, Mobile speech communicatio n, Hearing aids, Cochlear Implants	AFM word error rate 16.95%, 5.27% over real noisy data and feature mapping baseline. SA-AFM 9.85% WER rate	1. No intelligibility analysis and speech quality analysis. 2. Computation of PESQ and STOI for the enhanced data.
[7]	2022	Deep Complex Convolutional Recurrent GAN	Voice Bank Corpus, Demand Dataset	Time domain input, STFT	Demand Dataset for noise	SEGA N, MMS E-GA N, Metric GAN	PESQ, CSIG, CBAK, COVL	Automatic speech Recognition, Hearing Aids	DCCRGAN works better than state of art GAN.	1. Need to be compared with other feature set.
[8]	2019	GAN model with acoustic regression loss and two step adversarial training	VCTK corpus	Time domain	Speech distortio ns- Clipping , Chunk eliminati on, Frequen cy eliminati on	1. SEGA N-Bas eline 2.SEG AN-A CO 3.SEG AN-P TA	Mel cepstral distortion, F0 toot mean square, Voiced/un voiced frame prediction	Cochlear implants, Hearing aids, Communicati on systems	Better performance	1. Need to compare with other feature set 2. Intelligibility and speech quality need to be determined.
[9]	2020	Multistage GAN	Voice Bank Corpus	Time domain	Demand Dataset	1. SEGA N-Bas eline 2. i-SEG AN 3.deep SEGA N	PESQ, CBAK, CSIG, COVL	ASR, Cochlear implants, Hearing Aids	Better performance	1. need to compare with more models 2.Frequency domain features as input to iSEGAN and deepSEGAN to test the performance.
[10]	2020	Conditional GAN	Voice bank Corpus	Time Domai n	Demand Dataset	1. LSTM -IRM 2.SEG AN-B aseline	STOI, PESQ, CD LLR segSNR	ASR, Cochlear implants, Hearing Aids	Better Performance than IRM-LSTM	1. the model need to be compared with other available speech enhancement model.

[11]	2022	VSEGAN	GRID dataset for Video recordings and TCD-TIMIT for audio recordings	Magnitude spectrum(T-F)	12 types of noise-Car, room, instrument, engine, train,talker,air brake, water street, mic-noise, ring bell, music	1. Looking to listen model 2.Online visual Augmented model 3.AV(SE)^{2}model	PESQ STOI	Speech recognition, Speech-coding	Better performance	1. use better convolution structure 2.include weights for audio and visual modalities fusion blocks
[12]	2020	Dynamic Attention Recursive GAN	Voice Bank corpus	Magnitude spectrum (STFT)	Demand Database	1. SEGAN 2. SERGAN 3. GSEGAN 4.MMSE-GAN	PESQ, CBAK. COVL, CSIG	Automatic Speech recognition, Hearing Assistive Devices	Better performance	1.No intelligibility analysis using STOI is conducted and compared with other models.
[13]	2021	Metric GAN	Voice Bank Corpus	Magnitude Spectrogram	Demand Database	SEGAN-GAN, MMSE-STSA	PESQ, CBAK. COVL, CSIG	Automatic Speech recognition, Hearing Assistive Devices	Better performance	
[14]	2020	GAN	TIMIT	Time domain waveform	Noise-100	DNN-SE SEGAN	PESQ ,STOI	Automatic Speech recognition	Better performance	Model need to be compared with more models of speech enhancement
[15]	2018	Convolutional Neural Network GAN	Voice Bank Corpus	T-F mask	Demand Database	Pix-ti-Pix l1, Pix-to-Pix-l2	CSIG, CBAK, COVL,PESQ, STOI	Automatic Speech recognition, Hearing Assistive Devices	Better performance than Pix 2 Pix and SEGAN	
[16]	2021	μ-Law SGAN	Voice Bank Corpus	T-F mask	Demand dataset	SEGAN DSEGAN WaveNet	CSIG, CBAK, PESQ, ESTOI SSNR	Automatic Speech Recognition	Better performance	The model need to be compared with different datasets.

[17]	2022		Phase sensitive mask-conditi on GAN	Voice Bank Corpus, TIMIT Corpus, ATR Japanese	STFT-magnitude	Demand, Noisex ChMe-3	SEGAN, Deep Feature Loss, Metric GAN, AECNN,	CSIG, CBAK, PESQ, STOI	Automatic Speech Recognition	Better performance	The model is tested with a variety of datasets. The model shows good performance at low SNR. Improves speech intelligibility at low SNR.

REFERENCES

[1] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Interspeech 2017*, Aug. 2017, pp. 3642–3646. doi: 10.21437/Interspeech.2017-1428.

[2] D. Michelsanti and Z.-H. Tan, "Conditional Generative Adversarial Networks for Speech Enhancement and Noise-Robust Speaker Verification," in *Interspeech 2017*, Aug. 2017, pp. 2008–2012. doi: 10.21437/Interspeech.2017-1620.

[3] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5024–5028. doi: 10.1109/ICASSP.2018.8462581.

[4] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise Adaptive Speech Enhancement Using Domain Adversarial Training," in *Interspeech 2019*, Sep. 2019, pp. 3148–3152. doi: 10.21437/Interspeech.2019-1519.

[5] S. Pascual, A. Bonafonte, J. Serrà, and J. A. González López, "Whispered-to-voiced Alaryngeal Speech Conversion with Generative Adversarial Networks," in *IberSPEECH 2018*, Nov. 2018, pp. 117–121. doi: 10.21437/IberSPEECH.2018-25.

[6] Z. Meng, J. Li, Y. Gong, and B.-H. (Fred) Juang, "Adversarial Feature-Mapping for Speech Enhancement," in *Interspeech 2018*, Sep. 2018, pp. 3259–3263. doi: 10.21437/Interspeech.2018-2461.

[7] H. Huang, R. Wu, J. Huang, J. Lin, and J. Yin, "DCCRGAN: Deep Complex Convolution Recurrent Generator Adversarial Network for Speech Enhancement," Feb. 2022, pp. 30–35. doi: 10.1109/ISEEIE55684.2022.00013.

[8] S. Pascual, J. Serrà, and A. Bonafonte, "Towards Generalized Speech Enhancement with Generative Adversarial Networks," in *Interspeech 2019*, Sep. 2019, pp. 1791–1795. doi: 10.21437/Interspeech.2019-2688.

[9] H. Phan *et al.*, "Improving GANs for Speech Enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, 2020, doi: 10.1109/LSP.2020.3025020.

[10] D. Baby, "iSEGAN: Improved Speech Enhancement Generative Adversarial Networks." arXiv, Feb. 20, 2020. doi: 10.48550/arXiv.2002.08796.

[11] X. Xu *et al.*, "VSEGAN: Visual Speech Enhancement Generative Adversarial Network," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 7308–7311. doi: 10.1109/ICASSP43922.2022.9747187.

[12] A. Li, C. Zheng, R. Peng, C. Fan, and X. Li, "Dynamic Attention Based Generative Adversarial Network with Phase Post-Processing for Speech Enhancement." arXiv, Jun. 12, 2020. Accessed: Aug. 24, 2022. [Online]. Available: http://arxiv.org/abs/2006.07530

[13] S.-W. Fu *et al.*, "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement," in *Interspeech 2021*, Aug. 2021, pp. 201–205. doi: 10.21437/Interspeech.2021-599.

[14] J. Lin, S. Niu, A. J. van Wijngaarden, J. L. McClendon, M. C. Smith, and K.-C. Wang, "Improved Speech Enhancement Using a Time-Domain GAN with Mask Learning," in *Interspeech 2020*, Oct. 2020, pp. 3286–3290. doi: 10.21437/Interspeech.2020-1946.

[15] N. Shah, H. A. Patil, and M. H. Soni, "Time-Frequency Mask-based Speech Enhancement using Convolutional Generative Adversarial Network," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov. 2018, pp. 1246–1251. doi: 10.23919/APSIPA.2018.8659692.

[16] H. Li, Y. Xu, D. Ke, and K. Su, "μ-law SGAN for generating spectra with more details in speech enhancement," *Neural Networks*, vol. 136, pp. 17–27, Apr. 2021, doi: 10.1016/j.neunet.2020.12.017.

[17] "Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network - ScienceDirect." https://www.sciencedirect.com/science/article/abs/pii/S0885230821000759 (accessed Aug. 28, 2022).