Improved Time-Frequency Mask-based Speech Enhancement using Convolutional Generative Adversarial Network

Varun Kumar M
Department of ECE
NIT Karnataka
Surathkal, India
171EC251

Deeksha MS
Department of ECE
NIT Karnataka
Surathkal, India
171EC113

Sriram A
Department of EEE
NIT Karnataka
Surathkal, India
171EE144

Praveen Iyer
Department of EEE
NIT Karnataka
Surathkal, India
171EE134

***Abstract*—The task of speech enhancement (SE) to improve the intelligibility of signals aims to suppress noise while enhancing the quality of speech component. This is an active area of research which has slowly drifted from traditional signal processing algorithms to more robust supervised modelling using deep neural networks. Recent literature shows that Time-Frequency (TF) Masking method where model learns the mask explicitly outperforms implicit masking for direct generation of enhanced spectrograms. In this study we compare these methods with deep convolutional neural networks (CNN) and CNN-Generative Adversarial Networks (GANs) for mask estimation. GANs are the state-of-art generative models with adversarial optimization proven to be superior to discriminative CNNs with Maximum-likelihood optimization in an acoustic setting. We show the need for T-F mask with the range of values greater than 1 using linear activation to enhance the speech signal. The experimental results show the superiority of explicit mask-based CNN-GAN to their corresponding non-masking GANs and discriminative models. The performance evaluation in terms of both the predicted mask and the objective metrics, dictates the improvement in the speech quality, while simultaneously reducing the speech distortion observed in the noisy mixture.**

***Index Terms*—Deep Learning, Convolutional Neural Networks, Generative Adversarial Networks, Gammatone Features, Time-Frequency Analysis**

I. Introduction

The speech Enhancement (SE) task has been an active research topic for the Signal Processing Community not just because of the widespread practical use of SE techniques but also because of the vast room for improvement still possible. SE is the suppression or reduction of the additive noise present in the noisy mixture and thus increasing the speech intelligibility [1]. Challenges in SE include generating noise-robust speech features so as to capture the relevant information present in the speech and suppressing the additive noise. Applications of SE are numerous and include Speech and Speaker Recognition tasks as an enhanced speech would be extremely valuable in these applications [2], [3]. Additionally, enhanced speech signals can also come in handy for cochlear implants and hearing aids as it significantly improves speech intelligibility [4].

There have been various methods that have been tried out for this task which include spectral subtraction [5] and Wiener filtering [6]. These methods have not been effective in certain settings like low Signal-to-Noise (SNR) ratio and non-stationary noise conditions. Hence the recent focus has been on the state-of-the-art technique of the mask learning approach. Due to the availability of large amounts of data supervised learning based deep learning has been among the more successful techniques. In this approach, a deep neural network architecture is trained to learn a mapping function between features of the noisy speech and the time-frequency mask.

Since Speech comprises sequential data, a Convolutional Neural Network (CNN) fits the bill perfectly as it takes advantage of location-based data as well as reduces computational complexity by sharing the weights. Furthermore, the importance of predicting a mask instead of directly predicting the Enhanced Speech can be explained by understanding the drawbacks of directly predicting the Enhanced Speech. Non-Masking approaches fail to preserve higher harmonics and have poor noise suppression. Deep learning techniques using CNN rely on maximum likelihood-based optimization function. As an alternative we make use of in this paper, Generative Adversarial Networks (GAN) learns a mapping function through a discriminative process and since CNN models even though have shown promising results have not been the best. Hence a CNN-based GAN is an approach to predict the time-frequency mask is what we explore and eventually propose.

II. Related Work

In the recent years, focus has shifted from STFTs to wavelet transforms due to the availability of variable size windows which improves the time-frequency resolution of speech signals. Anirban et al. [7] used voice speech probability-based wavelet decomposition to perform speech enhancement tasks. In the first stage, Voiced speech probability was calculated using the gaussian mixture models (GMMs). Gain estimators were integrated in the wavelet decomposition stage to reduce

the mean squared error. The proposed method was able to do better than the existing traditional algorithms like spectral subtraction [5] and wiener filtering [6].

Minimum mean squared error estimator is a well-known speech enhancement technique. Tahmina et al. [8] proposed an improvised version of minimum mean square error (MMSE) noise estimator. Binary search was integrated with a first-in-first-out MMSE noise reduction algorithm, the noise spectral minima was computed using binary search which was fast and efficient. The proposed algorithm was tested on real time data and showed better performance compared to other MMSE algorithms . With advancement in deep learning over the last few years, Aaron et al. [9] investigated the use of deep learning methods for MMSE with the objective of producing intelligible enhanced speech quality. A residual long short-term memory cell (ResLSTM) was utilized to estimate the priori SNR. The results showed that the performance of MMSE has significantly increased in terms of quality and intelligibility scores using deep learning.

Numerous techniques using neural networks have achieved good denoising performance over the last few years. One such approach was tried by Yupeng et al. [10] for speech enhancement. A convolutional neural network architecture along with skip connections was proposed to learn the residual error between noisy and clean speech. Finally, the learnt residue error was subtracted from noisy speech to give clean speech. Recently progressive learning has shown capability to improve speech enhancement using deep networks [11] and LSTMs [12]. Progressive learning is a supervised speech enhancement algorithm that directly does not map noisy and clean but it does the process in different stages. Andong et al. [13] proposed a progressive learning convolutional recurrent neural network to reduce the complexity of the progressing learning. Results showed that the proposed algorithm consistently performed better than the progressive learning using deep networks and LSTMs by reducing the number of parameters drastically.

Generative adversarial networks (GANs) [14] are the state-of-the-art generative models within the deep learning framework. Santiago et al. [15] proposed time domain speech enhancement using GANs called the SEGAN. The generator network is a CNN-Autoencoder with skip connections (between encoder and decoder) that takes noisy signals as input and compresses it in the time domain and tries to reconstruct the clean signal. Training is done using the min-max loss function. Results show that the SEGAN performs better than MMSE, and other deep learning models tuned for speech enhancement. Motivated by the performance of GANs in image processing, Daniel et al. [16] explored the potential of conditional GANs (cGANs) for speech enhancement. The generator tries to enhance the noisy spectrogram using a Pix2Pix framework that uses L1 regularization. Metrics generated from the proposed model showed that the cGANs outperformed the MMSE algorithms and results are comparable with the deep neural networks.

Recently Time-Frequency based masking for speech enhancement has outperformed traditional techniques. Neil et al. [17] propose a CNN based GAN for inherent mask estimation. GAN takes advantage of the adversarial learning over maximum likelihood estimation. Training is done using min-max loss of GAN along with a RMSE loss for the generator. The paper also shows the need for time-frequency based masking for speech enhancement. The results were compared to traditional methods, deep learning networks, and other GAN architectures. Comparison shows that using Time-Frequency masking along with GAN improves speech quality and speech intelligibility while reducing the speech distortion observed in the noisy mixture.

III. Methodology

A. *Explicit mask-based SE*

Explicit mask prediction has shown prominence in improving the quality and intelligibility of speech signals [18] The models here are trained to output the mask values which are then multiplied to the input signal spectrogram for noise removal. Designing a supervised task and training a model sufficient to solve the problem will result in improved time-frequency representation of signals.

1) ***CNN for mask-based SE****:* A deep convolutional neural network is trained to predict the mask using the noisy log-spectrum as the input to minimise the mean-squared error between enhanced-input and clean log-spectrums [20]. The output of the last layer is treated as the TF-mask, which is multiplied with the input signal before calculation error and backpropagation step.

We set up our network with an architecture similar to encoder-decoder models. The encoding stage consists of 4 layers of 64, 128, 256 and 512 filter depth. Each filter has a kernel size of 4 with stride length as 2. This encoder learns the spatial downsampling while extracting the low-dimensional features which is structured as a bottleneck with 100 units. The decoder is the transpose of encoder as it aims to project the learnt features to original space. It consists of filters of size 512, 256, 128, and 64 in each layer with kernel size 4 and stride length of 2. This model learns details from every level compressing and processing the data with a series of convolutions and deconvolutions.

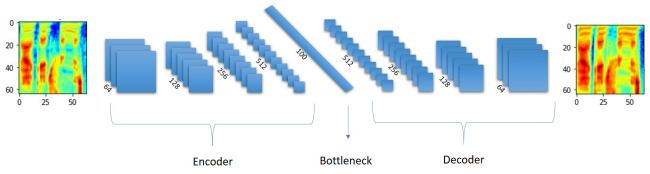


Fig. 1: Encoder-Decoder architecture of CNN. The same architecture is used for G network of CNN-GAN

2) ***CNN-GAN for mask-based SE****:* Being a discriminative model, CNN alone is prone to instability as it cannot predict the samples outside the discontinuous probability distribution. This results in more distortions in enhanced samples in such

cases. Moreover, these discriminative models generally fail to learn optimum parameters [14]. Therefore, a generative model with adversarial learning is preferred as they can generate new samples from the training distribution. We experiment the efficiency of using CNN-GAN models for speech enhancement and establish its superiority over discriminative models.

The GAN consists of two networks: Generator (G) and Discriminator (D). The G is trained to learn the TF mask and the discriminator tries to distinguish between enhanced and clean spectrograms. In the training process, these two networks compete with each other until the Nash equilibrium is obtained, wherein the G produces almost perfect-fake samples indistinguishable by the D. In addition to the adversarial loss, the network G is improved by minimising the MSE error between the enhanced and clean log-spectrum [19]

$$min_D V(D) = -E_{x\sim}[log(D(x))] \\ -E_{y\sim\mathcal{Y}}[log(1 - D(G(y)))]$$

$$min_G V(G) = -E_{y\sim\mathcal{Y}}[log(D(G(y))] \\ +\frac{1}{2}E_{x\sim,y\sim\mathcal{Y}}[log(x) - log(G(y))]^2$$

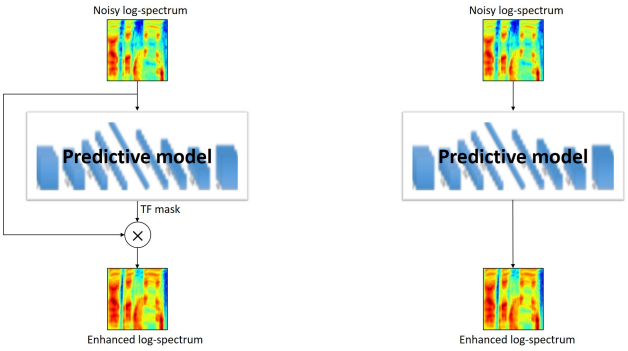
We model G with similar architecture as CNN, while D has six layers of filter size of [64, 128, 256, 512, 64, 1] dimension with kernel size 4 and stride length of 2.

B. Implicit (non-masking) SE

In contrast to explicit mask-based models, the implicit mask prediction (commonly known as non-masking method, directly predicts the spectrogram, hence eliminating the step of multiplication as shown in Fig[2]. This method is more prone to model uncertainty witnessed when the bias gets added as noise to all-zero sample. Therefore, this strategy fails to obtain accurate enhancement of the speech signals [20] [21]. We experiment the ability of both CNN and CNN-GAN to predict spectrograms directly, while learning the mask inherently. We employ the same architecture as described in the previous section as we intend to establish the superiority of masking methods but the output of CNN and G network is log-spectrum which is directly passed into the D network and backpropagation.

C. Dataset

We use the dataset released by Valentini et. al. [22] in our experiments. It consists of 30 speakers from the Voice Bank corpus [23] mismatched conditions. The training set contains 28 English speakers and the test set contains 2 English speakers, with around 400 sentences each, both for the clean and noisy set. All the sentences are sampled at 48 kHz. The training set explores 40 different noisy conditions with 10 types of noise and 4 SNR each (15, 10, 5, and 0 dB). The test set comprises 20 different noisy conditions with 5 types of noise and 4 SNR each (17.5, 12.5, 7.5, and 2.5 dB). The noise samples are taken from Demand database [24]. The train and test set contains 11572 and 824 utterances, respectively.



(a) Explicit mask-based (b) Implicit non-masking

Fig. 2: Flowchart of implemented algorithms

D. Preprocessing

The input signals are downsampled from 48kHz to 16kHz and applied a pre-emphasis of 0.95. 64-channel gammatone features are extracted using 20ms Hamming window with 50% (10 ms) overlap with adjacent frames. The log-spectrum is computed from the gammatone features. We experiment the effect of normalisation (log-spectrum values restricted between 0-1) and standardisation i.e., mean normalisation on the model performance, and establish that normalisation is more preferable. Each audio spectrum is further subdivided into 64*64 log-patches. These patches are created using a sliding window with 50% overlap during training and no overlap for test set.

E. Training setup

In our initial exploration, we observed that some of the pixels in the clean log-spectrum have a higher value than the corresponding noisy log-spectrum. This demands the mask value to be higher than 1 at certain places. Hence, we apply linear activation, i.e., effectively no activation function in the last layer of the model as opposed to sigmoid in [17] [25] Each convolution, except the last layer, is followed by batch normalisation and LeakyReLU [26] activation to avoid overfitting while resulting in high-quality TF representations. In all experiments, we train the model for 25 epochs with Adam optimizer [27] for CNN and G network, and SGD [28] for D network. Adaptive optimizer such as Adam is proven to go well in the initial stages, however, it flatlines due to learning rate decay [30]. This prevents the GAN from reaching the equilibrium, in which case SGD can work better [29]. We use a learning rate of 0.0002, using an effective batch size of 200. Out of 11572 training utterances, 11000 random utterances are used for training the network and remaining 572 utterances are used for validation. The epoch showing the least MSE on the validation set is applied on the test set.

IV. Results

The predicted gammatone spectrum for different patches is shown in figure 3. The T-F mask predicted by CNN-GAN preserves finer structures and crucial harmonics. The quality

of the enhanced speech is computed using various objective measures. The Composite measure for Signal Distortion (CSIG) predicts the Mean Opinion Score of the signal (MOS) distortion (from -0.5 to 4.5). The Composite measure for Background interferences (CBAK) and the Overall Composite measure (COVL) (from 1 to 5) predicts the extent of background interferences in the speech and the overall effect, respectively. Perceptual Evaluation of Speech Quality (PESQ) (from -0.5 to 4.5) is a wideband version recommended in ITU-T P.862.2. These metrics are calculated using the implementation given in [31]. Moreover, the Short-Time Objective Intelligibility (STOI) that records the improvement in speech intelligibility [32] is also computed.

Metric	**Noisy**	**CNN**		**CNN-GAN**	
		Implicit	Explicit	Implicit	Explicit
CSIG	3.35	2.84	2.92	3.18	**3.50**
CBAK	2.44	2.26	2.43	2.64	**2.84**
COVL	2.63	2.08	2.35	2.59	**2.90**
PESQ	1.97	1.40	1.83	2.05	**2.34**
STOI	0.91	0.85	0.86	0.89	**0.91**

TABLE I: Metrics obtained for normalised data

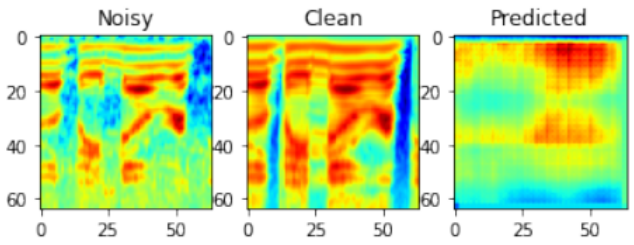
Table I shows the computed metrics for CNN and CNN-GAN architectures whose masks are computed implicitly and explicitly for normalised data. The computed metrics for CNN-GAN (both implicit and explicit) are observed to be better than that computed using the results of CNN architecture. The adversarial characteristics developed between the generator and discriminator in CNN-GAN, gives a significant improvement (in terms of both the predicted mask and objective metrics) over CNN, that lacks the adversarial characteristics in its objective function. The CNN (MSE optimization) only reduces the numerical error between the enhanced and the clean speech, that may not necessarily lead to perceptually optimal enhanced speech.

Table II shows a comparison between the metrics computed when normalised and standardised data is fed to the CNN model. It is clear from the objective scores that the model was able to learn better when the log-Gammatone spectrograms were normalised. Table II also shows a comparison between the objective scores of CNN-GAN when it was trained with Adam optimizer and SGD optimizer. On using the Adam optimizer the generator part of the CNN-GAN model did not converge and consequently, the predicted masks were poor.

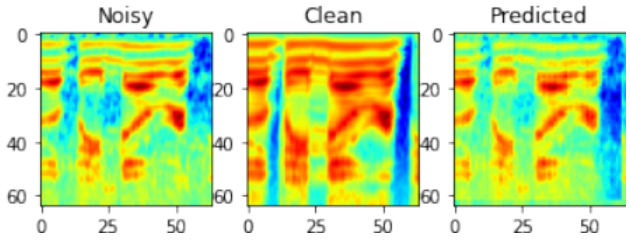
Metric	**Noisy**	**CNN**		**CNN-GAN**	
		Normalised	Standardised	Adam	SGD
CSIG	3.35	**2.84**	2.24	1.01	**3.50**
CBAK	2.44	**2.26**	2.17	1.93	**2.84**
COVL	2.63	**2.08**	1.95	1.06	**2.90**
PESQ	1.97	**1.40**	1.80	1.46	**2.34**
STOI	0.91	**0.85**	0.80	0.71	**0.91**

TABLE II: Comparison between CNN (norm vs std) and CNN-GAN (Adam vs SGD optimizer)

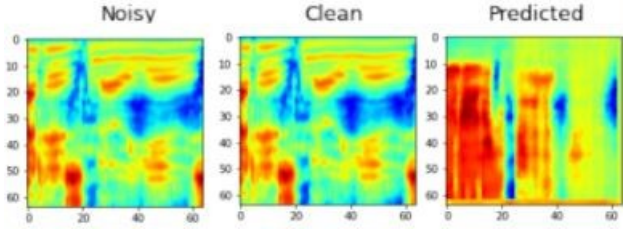
Table III shows the computed metrics for CNN and CNN-GAN models with and without activation. It is clearly observed



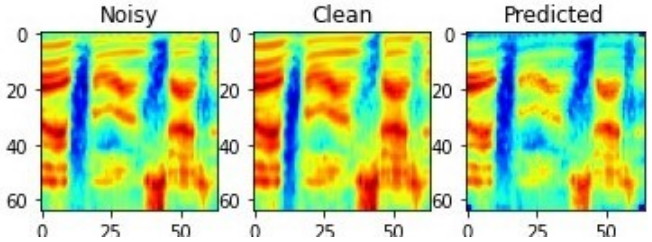
(a) Implicit CNN



(b) Explicit CNN



(c) Implicit GAN



(d) Explicit GAN

Fig. 3: Comparison of predicted spectrograms with different models

Metric	**Noisy**	**CNN**		**CNN-GAN**	
		Linear	Sigmoid	Linear	Sigmoid
CSIG	3.35	**2.84**	2.06	**3.34**	2.43
CBAK	2.44	**2.26**	2.16	**2.66**	1.88
COVL	2.63	**2.08**	1.80	**2.65**	1.84
PESQ	1.97	1.40	**1.69**	**2.02**	1.37
STOI	0.91	**0.85**	0.81	**0.91**	0.75

TABLE III: Effect of activation for CNN and CNN-GAN

that the computed metrics for a model without activation is greater than its counterpart. The removal of sigmoid activation helps the model learn better because it upscales clean pixels and effectively learning a mask closer to the ground truth.

Table IV compares the results of a variety of models which include Pix2Pix model with L2 and L1 regularization, SEGAN [21], Wiener [21], CNN, CNN-GAN [17] and the Modified CNN-GAN which we propose in this paper.

Paper	CSIG	CBAK	COVL	PESQ	STOI
Noisy	3.35	2.44	2.63	1.97	0.91
Pix2Pix-L2	2.81	2.57	2.42	2.15	0.88
Pix2Pix-L1	1.98	1.63	1.54	1.29	0.74
SEGAN	3.48	2.94	2.8	2.16	0.93
Wiener	3.23	2.68	2.67	2.22	-
CNN	1.64	1.72	1.31	1.12	0.62
CNN-GAN	3.55	2.95	2.92	2.34	0.93
MCNN-GAN (Proposed)	**3.50**	**2.84**	**2.90**	**2.34**	**0.91**

TABLE IV: Metric Comparison for Various Models

Figure 4 summarizes and compares the results of the implemented models. It is clear that the CNN-GAN which learns the mask explicitly has a higher score compared to all other models.

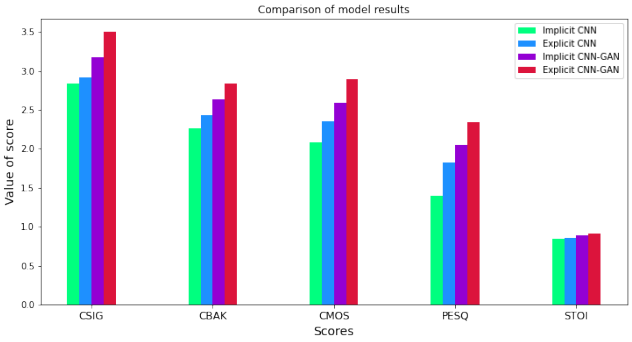


Fig. 4: Comparison of model results

V. Conclusion

Removing the Activation function from the last layer for both our Explicit CNN and Explicit GAN which predicts the mask helped to have a flexible mask which can take up values in the desired ranges without restriction hence making our model a better pragmatic choice. Also the experimentation of using both Standardised data and Normalised data proved that using Normalised data yields better results as shown in table 2. Furthermore, we established that the usage of masks (Explicit models) yields significantly better results than the implicit models which directly predict the spectrogram. Also, the usage of Adam optimizer for the Generator in GAN along with switching to SGD optimizer for Discriminator helped with the training of GAN.

VI. Future Work

Usage of an RNN-based-GAN might be of some help because of the inherent memory cell based characteristic of RNNs. This is especially helpful because speech is a temporal form of data. Also, c-GAN (Conditional GAN) could have been a good try. c-GAN is a type of GAN that involves the conditional generation of images by a generator model. Convergence also tends to be faster for c-GAN which generally is a huge task for GANs.

References

[1] P. C. Loizou, "Speech enhancement: Theory and practice," 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[2] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 22, no. 4, pp. 826–835, 2014.

[3] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. W. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), Liberec, Czech Republic, 2015, pp. 91–99.

[4] D. Wang and J. H. Hansen, "Speech enhancement based on harmonic estimation combined with MMSE to improve speech intelligibility for cochlear implant recipients," in INTERSPEECH, Stockholm, Sweden, 2017, pp. 186–190.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE/ACM TASLP, vol. 27, no. 2, pp. 113–120, 1979.

[6] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," IEEE/ACM TASLP, vol. 26, no. 3, pp. 197–210, 1978.

[7] Bhowmick, Anirban, and Mahesh Chandra. "Speech enhancement using voiced speech probability based wavelet decomposition." Computers and Electrical Engineering 62 (2017): 706-718.

[8] Gouhar, Tahmina, Nabih Jaber, and Pallavi Kuntumalla. "Speech enhancement using new iterative minimum statistics approach." 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE, 2017.

[9] Nicolson, Aaron, and Kuldip K. Paliwal. "Deep learning for minimum mean-square error approaches to speech enhancement." Speech Communication 111 (2019): 44-55.

[10] Shi, Yupeng, Weicong Rong, and Nengheng Zheng. "Speech enhancement using convolutional neural network with skip connections." 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2018.

[11] Gao T, Du J, Dai L-R, Lee C-H. SNR-based progressive learning of deep neural network for speech enhancement. In: INTERSPEECH. p. 3713–7.

[12] Gao T, Du J, Dai L-R, Lee C-H. Densely connected progressive learning for LSTMbased speech enhancement. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 5054–8.

[13] Li, Andong, et al. "Speech enhancement using progressive learning-based convolutional recurrent neural network." Applied Acoustics 166 (2020): 107347.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in NIPS, Montral,Canada, 2014, pp. 2672–2680.

[15] Pascual, Santiago, Joan Serra, and Antonio Bonafonte. "Time-domain speech enhancement using generative adversarial networks." Speech Communication 114 (2019): 10-21.

[16] Michelsanti, Daniel, and Zheng-Hua Tan. "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification." arXiv preprint arXiv:1709.01703 (2017).

[17] M. H. Soni, N. Shah and H. A. Patil, "Time-Frequency Masking-Based Speech Enhancement Using Generative Adversarial Network," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5039-5043, doi: 10.1109/ICASSP.2018.8462068.

[18] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," IEEE/ACM TASLP, vol. 22, no. 12, pp. 1849–1858, 2014.

[19] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros,"Context encoders: Feature learning by inpainting," in CVPR, Nevada, USA, 2016, pp. 2536–2544.

[20] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in IEEE ICASSP, Brisbane, Australia, 2015, pp. 4390–4394.

[21] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech Enhancement Generative Adversarial Network," in INTERSPEECH, Stockholm, Sweden

[22] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating CNN-based speech enhancement methods for noise-robust

Textto- Speech," http://dx.doi.org/10.7488/ds/1356, Available Online; Last accessed 17-January-2018

[23] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in International Conference on Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (OCOCOSDA/CASLRE), Gurgaon, India, 2013, pp. 1–4.

[24] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings," The Journal of the Acoustical Society of America (JASA), vol. 133, no. 5, pp. 3591–3591, 2013.

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

[26] AL Maas, AY Hannun, AY Ng. Proc, "Rectifier nonlinearities improve neural network acoustic models" ICML 30 (1), 3, 4497, 2013.

[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in IEEE ICLR, San Diego, USA, 2015, pp. 1–15

[28] Bottou, Léon (2004), "Stochastic Learning", Advanced Lectures on Machine Learning, LNAI, 3176, Springer, pp. 146–168, ISBN 978-3-540-23122-6

[29] Kontonis, Vasilis, Sihan Liu and Christos Tzamos. "Convergence and Sample Complexity of SGD in GANs." ArXiv abs/2012.00732 (2020)

[30] Wilson, A., R. Roelofs, Mitchell Stern, Nathan Srebro and B. Recht. "The Marginal Value of Adaptive Gradient Methods in Machine Learning." NIPS (2017).

[31] P. C. Loizou, "Speech enhancement: Theory and practice," 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013

[32] H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short time objective intelligibility measure for time-frequency weighted noisy speech," in IEEE ICASSP, Texas, USA, 2010, pp. 4214–4217