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

Neil Shah, Hemant A. Patil and Meet H. Soni
Speech Research Lab,
Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India
E-mail: {neil_shah,hemant_patil,meet_soni}@daiict.ac.in

***Abstract*—Speech Enhancement (SE) system deals with improving the perceptual quality and preserving the speech intelligibility of the noisy mixture. The Time-Frequency (T-F) masking-based SE using the supervised learning algorithm, such as a Deep Neural Network (DNN), has outperformed the traditional SE techniques. However, the notable difference observed between the oracle mask and the predicted mask, motivates us to explore different deep learning architectures. In this paper, we propose to use a Convolutional Neural Network (CNN)-based Generative Adversarial Network (GAN) for inherent mask estimation. GAN takes an advantage of the adversarial optimization, an alternative to the other Maximum Likelihood (ML) optimization-based architectures. We also show the need for supervised T-F mask estimation for effective noise suppression. Experimental results demonstrate that the proposed T-F mask-based SE significantly outperforms the recently proposed end-to-end SEGAN and a GAN-based Pix2Pix architecture. The performance evaluation in terms of both the predicted mask and the objective measures, dictates the improvement in the speech quality, while simultaneously reducing the speech distortion observed in the noisy mixture.**

Index Terms: speech enhancement, generative adversarial network, convolutional neural network, inherent mask estimation.

I. Introduction

Development of Speech Enhancement (SE) algorithms for suppression of the additive noise present in the noisy mixture, has always been a long-standing goal among the signal processing research community. A significant reduction in the speech intelligibility is observed under the presence of noisy background interferences [1]. SE aims to improve the speech intelligibility and quality of the noisy mixture. The SE technique finds its application in generating the noise-robust speech-specific features, that capture the significant information present in the speech and in suppressing the additive noise [1]. The robust features generated after applying the SE technique as a pre-processing tool can be used in speech and speaker recognition task [2], [3]. The enhanced speech waveform can be used in designing the cochlear implant (CI) and hearing aid devices, that essentially improve the human speech perception in the noisy environment [4].

Traditional SE methods, such as spectral subtraction [5] and Wiener filtering [6], are less effective in low Signal-to-Noise Ratio (SNR) and non-stationary noise conditions. The more recent technique focus on the mask learning and feature mapping approaches [7]–[11]. Estimation of the Time-Frequency (T-F) mask using data-driven supervised learning is the state-of-the-art technique in SE [7]–[11], due to the sufficient availability of the prior knowledge, such as speaker's identity or the noise type [8]. These techniques make no statistical assumption and suppress the noise by observing a large number of representative pairs of noisy and noise-free speech samples. A Deep Neural Network (DNN) is the most preferred supervised learning algorithm, trained to learn a mapping function between the noisy spectral features and the mask [8]–[10].

Speech is a sequential data that can be used for modeling the temporal characteristics and capturing the long-term dependencies. Hence, learning the DNN parameters from the spectral context helps in better understanding the information captured by the neighboring frames [12]. However, a fully connected DNN do not take in account the temporal information of the input data [13], [14]. A Convolutional Neural Network (CNN) provides a viable solution by taking the advantage of data locality and is comparatively less computationally expensive, due to its weight sharing property [13]–[16]. CNNs have shown comparable results in noise elimination [13] and generating robust speech-specific features in speech recognition task [14], [15]. Very recently, the authors in [16], have proposed to use a fully CNN for SE task, by finding a mapping function between the noisy and enhanced speech spectra. Such an approach may not be able to enhance the speech components and improve the speech intelligibility, where T-F masking-based enhancement methods are proven to be effective for removing the background interferences and reducing the speech distortion. The traditional supervised learning algorithm, such as DNN, CNN, uses the Maximum Likelihood (ML)-based optimization function, to predict the T-F mask. The Mean Square Error (MSE, an ML optimization) only optimizes the numerical estimates between the groundtruth and the estimated. This numerically estimated error may not always lead to perceptually intelligible speech [17]. Moreover, this approach is data-dependent and prevents the network from learning the perceptually optimal parameters, because the ML criteria put prior assumption on the data distribution [18].

Generative Adversarial Network (GAN), provides an alternative to the ML-based optimization criteria [18]. GAN learns a mapping function through a discriminative process, by minimizing the distributional divergence between the model and data distribution [17]. Since, convolutional network-based approaches have shown to perform better in SE task [16], [19], [20], in this paper, we attempt to use CNN-based GAN for T-F masking-based SE. Similar attempts for SE were proposed in [20], however the significance of T-F mask estimation for noise suppression was not explored. First, we show the need for estimating the T-F mask in SE by comparing the enhanced T-F representation obtained with and without mask estimation. Thereafter, a fully CNN is employed for T-F mask estimation, where we have shown that CNN alone is not sufficient to predict the T-F mask accurately. To address this limitation, we propose to use CNN-GAN framework for T-F mask estimation. The above network can also utilize the $\mathcal{L}_2$ distance computation, as a regularizer, between the groundtruth and the estimated [21]. This framework is applicable to any T-F representation and effectively reduces the speech distortion, removes the background interferences and improves the speech intelligibility.

A. *Recent Work*

GAN is a generative modeling technique applied initially in the field of computer vision [18], [19], [22]. Recent studies have shown the potential of exploiting GAN in the speech technology-related applications, that aim to learn a suitable mapping function and accurately reconstructs the enhanced speech while maintaining the speech quality and intelligibility [17]. A GAN-based postfilter proposed in [23] reconstructs the spectrogram that resembles the true data in the high-dimensional Short-Time Fourier Transform (STFT)-domain. Notable improvement has been observed in the SE [20], [21] and Voice Conversion (VC) [24], [25] task. The conditional GAN (cGAN) architecture investigated in [21], adopted a pixel-to-pixel (Pix2Pix) framework for SE, by learning the mapping function between the spectra of noisy speech and its enhanced version. The end-to-end SE method proposed in [20] has implemented an encoder-decoder CNN-based architecture within the GAN framework. These approaches directly predict, either the enhanced spectra [21] or the enhanced speech waveform [20], and do not exploit the significance of T-F mask, that has shown better objective measures and improved perception of the enhanced speech. In this paper, we attempt to analyze the significance of (a) CNN, (b) CNN-GAN, and (c) Pix2Pix, architecture for the SE task.

B. *Generative Adversarial Network (GAN)*

The aim of the generative network is to model the data distribution $\mathcal{X}$ explicitly and generate the samples from the estimated model distribution $\hat{\mathcal{X}}$. GANs are a structured probabilistic model that set up a min-max game between the generator (G) and the discriminator (D) [18]. The network G learns the mapping between the samples y from some prior distribution $\mathcal{Y}$ to samples x following $\mathcal{X}$. The network D is a binary classifier, with input as real samples belonging to $\mathcal{X}$ and fake samples generated by the network G. As training proceeds, the network D maximizes the likelihood of samples belonging to $\mathcal{X}$ as real and minimizes the likelihood of generated samples belonging to $\hat{\mathcal{X}}$ (generator's output) as fake. The adversarial characteristics force the network G to generate the realistic samples that closely follows $\mathcal{X}$, essentially developing the Nash equilibrium and leaves the network D unable to differentiate between the $\mathcal{X}$ and $\hat{\mathcal{X}}$ [18]. The objective function can be mathematically formulated as shown in [18]:

$$\min_D V(D) = -\mathbb{E}_{x\sim\mathcal{X}}[\log D(x)] - \mathbb{E}_{y\sim\mathcal{Y}}[1 - \log(D(G(y)))], \tag{1}$$

$$\min_G V(G) = -\mathbb{E}_{y\sim\mathcal{Y}}[\log D(G(y))], \tag{2}$$

where $\mathbb{E}_{y\sim\mathcal{Y}}$ denotes the expectation over all the samples y belonging to the distribution $\mathcal{Y}$.

II. T-F MASKING-BASED SE

A. *Analysis of masking and non-masking approach*

Supervised SE aiming to estimate a suitable target (mask) has shown a lot of promise. Mask estimation for SE task has shown large speech intelligibility improvement in noise for both hearing impaired and normal listeners [8]. This method is amenable to real-time implementation, generalizes well to the different T-F representations, given the sufficient amount of training data and produces faster inference [8]. On the other hand, learning the mapping function without mask estimation may not be able to enhance the noisy speech, with better human perception and intelligibility improvements, than the speech enhanced using masking-based approaches [20], [21], [26]. Fig. 1(c) shows one instance of such a failure. The estimated mask used for obtaining the enhanced T-F representation in Fig. 1(b), is learnt using the proposed CNN-GAN architecture. The solid circle shows the presence of unwanted background interferences and the dash-dot circle shows the inability of non-masking approach in preserving the higher frequency harmonics. This fails to uphold the noise suppression property of SE task. However, the masking approach proves to be a viable SE technique.

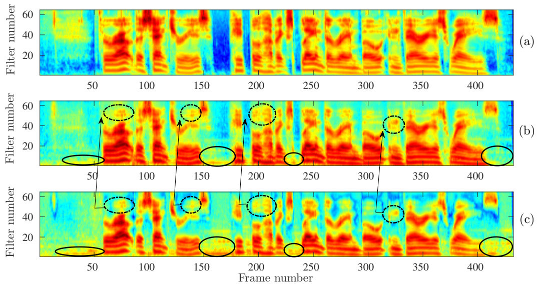

Fig. 1: Non-masking approach fails to properly reconstruct the enhanced T-F representation (a) clean T-F representation, (b) masking approach, and (c) non-masking approach of obtaining enhanced T-F representation: the solid-circle shows the presence of noise and the dash-dot-circle shows the inability in preserving the higher frequency harmonics.

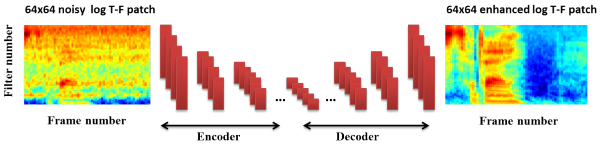

Fig. 2: Basic framework of the fully convolutional encoder-decoder architecture for the SE. The encoder and decoder has strided and fractional strided convolutions, respectively, followed by batch normalization and ReLU as activation function.

B. T-F masking using CNN

A CNN can be trained to learn a mapping function between the noisy and enhanced T-F representation while learning the T-F mask-like representation implicitly. This method termed as a task-dependent masking, significantly improves the objective quality and the speech intelligibility [26]. Motivated by the study reported in [26], we aim to optimize the error between the log T-F representation of clean and enhanced speech. In a way, the output of the last layer of the network can be treated as a T-F mask, that is further multiplied with the noisy T-F representation to yield an enhanced T-F representation. If values of the learned T-F mask is constrained between 0 to 1, then the T-F mask should represent to an Ideal Ratio Mask (IRM) [26].

Fig. 2 shows the basic framework of the fully CNN network, with only convolutional layers, designed in a way similar to an autoencoder, as explored by other studies [19], [20]. Such an architecture learns the spatial downsampling and upsampling operations, that benefits the mapping from the network G to network D [19]. In the encoding stage, the noisy T-F patch is projected and compressed through a series of strided convolution [19], followed by batch normalization [27] and rectified linear units (ReLUs) [28]. These layers are followed till the network captures the low-level details [19]. Once the compression bottleneck is achieved, in the decoding stage, the encoding process is reversed through a sequence of fractional-strided transposed convolutions (deconvolutions) [19], followed again by batch normalization and ReLUs. Such an architecture produces high quality enhanced T-F representation and avoids overfitting training patches, as suggested in [19].

C. T-F masking using CNN-GAN

The initial experiment of estimating T-F mask using CNN, suggests that this framework fails to learn an accurate T-F mask. A discriminative model, such as a CNN is prune to the unseen adversarial examples, leaving the network unable to discriminate between the original and deformed (original with additive noise) image at the input [18]. Moreover, the discriminative model-based architecture may sometimes fail in learning the optimal parameters [18]. However, the generative model can take the advantage of producing samples that are intended to come from the training data distribution [18]. Hence, CNN can be modeled more reliably on employing GAN, which is both a generative and a discriminative modeling technique [19]. The objective of the network G is to generate an enhanced T-F mask or representation, given the noisy T-F representation. The network G inherently estimates the T-F mask, while the network D accurately learns to discriminate between the clean and enhanced (output of G) T-F representation. In addition to the adversarial loss, the network G can also exploit the significance of minimizing the $\mathcal{L}_2$ distance between the enhanced (output of G) and clean T-F representation [29]. The objective function of network G can be mathematically formulated as in [29]:

$$\begin{aligned}\min_G V(G) = -\mathbb{E}_{y\sim\mathcal{Y}}[\log(D(G(y)))] + \\ \frac{1}{2}\mathbb{E}_{x\sim\mathcal{X},y\sim\mathcal{Y}}[\log(x) - \log(G(y))]^2.\end{aligned} \tag{3}$$

In this paper, we analyze the significance of modeling GAN using CNN-based G and D network, for T-F mask estimation. The network G is modeled similarly as the fully CNN (Fig. 2 and Sec. 2.2). The network D, which is a binary classifier can be modeled similarly to an encoder in the fully CNN.

D. SE using Pix2Pix framework (Non-masking approach)

The cGAN architecture in [21] adapted the Pix2Pix framework for SE task. The Pix2Pix model optimizes the error between the noisy and clean T-F representation, without learning the T-F mask. To compare the Pix2Pix (non-masking) approach with the T-F masking technique using CNN-GAN, we adapt the similar architecture as discussed in Sec. 2.3, in a Pix2Pix framework. We call the above model as Pix2Pix-$\mathcal{L}_2$, having Eq. (3) as the objective function. Moreover, we also analyze the Pix2Pix architecture with the least squares (LS) adversarial loss in addition to the $\mathcal{L}_1$ distance computed between the enhanced and clean T-F representation, as in [21]. We call this model as Pix2Pix-$\mathcal{L}_1$.

III. Experimental Setup

A. Dataset

We use the dataset released by Valentini *et. al.* [30], which contains 30 speakers from the Voice Bank corpus [31] under mismatched conditions. The dataset with similar condition as in [20] is selected, as one of the purpose of the study is to signify the importance of T-F masking-based approach using an encoder-decoder convolutional GAN architecture. The training set contains 28 English speakers and 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 of 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 [32]. The train and test set contains 11572 and 824 utterances, respectively.

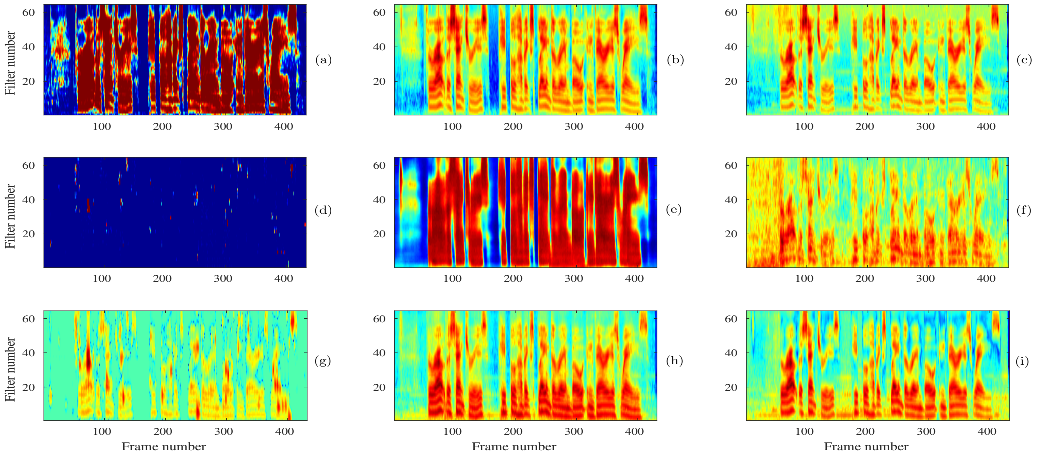

Fig. 3: (a) Oracle mask, Gammatone spectrum of (b) clean speech, (c) noisy speech. Predicted mask using (d) CNN, (e) proposed CNN-GAN. Gammatone spectrum of enhanced speech using (f) Pix2Pix-$\mathcal{L}_1$ (g) CNN, (h) proposed CNN-GAN, (i) Pix2Pix-$\mathcal{L}_2$.

B. *Network Setup*

We train four networks to compare the results. The first network is a fully encoder-decoder CNN architecture, whose parameters are optimized using the MSE criteria between the enhanced and clean T-F representation. The encoder has four 4x4 strided convolutions, with 2x2 stride length stacked after each other, with the filter size of [64, 128, 256, 512] dimension. The convolution bottleneck has the dimension of 100. The decoder has four 4x4 fractional strided convolutions, with 2x2 stride length and the filter size of [512, 256, 128, 64] dimension. The second network is a CNN-GAN with $\mathcal{L}_2$ regularization. The network G of CNN-GAN is identical to the fully CNN (network 1). The network D has six 4x4 strided convolutions, with 2x2 stride length and the filter size of [64, 128, 256, 512, 64, 1] dimension. Both the CNN and CNN-GAN architecture inherently estimates the T-F mask. The last hidden layer in CNN and network G has sigmoid activation to limit the mask values between 0 and 1. The third network is a Pix2Pix-$\mathcal{L}_2$ architecture with $\mathcal{L}_2$ regularization and the forth network being a Pix2Pix-$\mathcal{L}_1$ architecture with $\mathcal{L}_1$ regularization and LS objective function. Both of this architecture has got the same setup as CNN-GAN, except they do not estimate the T-F mask, making them a non-masking approach for SE task. Each convolutional layer is followed by batch normalization and ReLU activation as in [16], except the last hidden layer in network D that uses the sigmoid activation.

The original utterances are downsampled from 48 kHz to 16 kHz and then pre-emphasized with a factor of 0.95 [20]. The 64-channel Gammatone features are extracted with 20 ms Hamming window and 10 ms overlap between consecutive frames. During training, we create the patch with a sliding window of 50 % overlap, whereas during testing, we slide the window with no overlap [20]. The input to the network is a 64x64 patch of the noisy log-Gammatone spectrum. The 64x64 patch of the clean log-Gammatone spectrum is extracted for optimizing the network parameters. All the four networks are trained for 20 epochs with Adam optimizer [33] and 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. Once a particular network is trained, the epoch showing the least MSE on the validation set is chosen for the testing purpose.

C. *Experimental Results*

The predicted mask and the Gammatone spectrum of the enhanced speech for four different architectures are shown in Fig. 3. The visual inspection indicates the T-F mask predicted by the proposed CNN-GAN is significantly better than the mask predicted by the CNN alone. The T-F mask predicted by the proposed CNN-GAN especially preserves the finer structures and crucial harmonics. The Gammatone spectrum reconstructed using non-masking approaches, such as Pix2Pix-$\mathcal{L}_2$ and Pix2Pix-$\mathcal{L}_1$, illustrates the presence of background interferences and their inability in preserving the harmonics in higher frequency regions. 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 [34]. Perceptual Evaluation of Speech Quality (PESQ) (from -0.5 to 4.5) is a wideband version recommended in ITU-T P.862.2 [35]. These metrics are computed using the implementation given in [1].

TABLE I: Performance comparisons between the noisy signal, Pix2Pix-$\mathcal{L}_2$, Pix2Pix-$\mathcal{L}_1$, CNN, CNN-GAN, SEGAN and the Wiener filter-based enhancement

Metric	**Noisy**	**Pix2Pix-$\mathcal{L}_2$**	**Pix2Pix-$\mathcal{L}_1$**	**SEGAN [20]**	**Wiener [20]**	**CNN**	**CNN-GAN (proposed)**
CSIG	3.35	2.81	1.98	3.48	3.23	1.64	**3.55**
CBAK	2.44	2.57	1.63	2.94	2.68	1.72	**2.95**
COVL	2.63	2.42	1.54	2.8	2.67	1.31	**2.92**
PESQ	1.97	2.15	1.29	2.16	2.22	1.12	**2.34**
STOI	0.91	0.88	0.74	**0.93**	-	0.62	**0.93**

Pix2Pix-$\mathcal{L}_2$, Pix2Pix-$\mathcal{L}_1$, SEGAN, Wiener are the non-masking approaches, whereas CNN and CNN-GAN are the masking approaches. '-' indicates data is not available.

Moreover, the Short-Time Objective Intelligibility (STOI) that records the improvement in speech intelligibility [36] is also computed.

Table 1 shows the computed metric scores for different architectures. Optimizing the Pix2Pix architecture with $\mathcal{L}_2$ regularization is observed to perform better (in terms of both the resynthesized Gammatone spectrum and objective metrics) than the network regularized using $\mathcal{L}_1$ regularization. The scores using other non-masking technique, such as SEGAN and Wiener filtering are directly taken from [20], as their evaluation is on the same dataset. The quality scores and the enhanced speech's spectrum, suggest that the proposed masking approach (CNN-GAN) clearly outperform the non-masking-based approaches for SE. Moreover, 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. In addition, 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 [17]. The STOI score, found to be highly correlated with the intelligibility, reflects almost the similar perceptual intelligibility gain using CNN-GAN and SEGAN.

IV. Summary and Conclusions

In this study, we analyzed the various masking and non-masking approach for Speech Enhancement (SE). The proposed masking approach (CNN-GAN) learns a suitable mapping function between the noisy and clean T-F representation, by inherently learning the T-F mask. The CNN-GAN improves the speech quality and intelligibility of the enhanced speech, over the state-of-the-art Pix2Pix, SEGAN, and Wiener filtering-based non-masking approaches. The study also shows that the $\mathcal{L}_2$ regularization reconstructs the spectrum better than the $\mathcal{L}_1$ regularization for the Pix2Pix architecture, in the SE task. Our further work will involve a comprehensive evaluation of the proposed framework in more critical SNR situations and explore different deep learning architectures, such as Recurrent Neural Network (RNN) and deep Bidirectional Long Short Term Memory (BLSTM) in the GAN framework, that might suppress the noise more efficiently for the SE task.

References

[1] P. C. Loizou, "Speech enhancement: Theory and practice," 2^{nd} 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] Z. Chen, Y. Huang, J. Li, and Y. Gong, "Improving mask learning based speech enhancement system with restoration layers and residual connection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3632–3636.

[8] 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.

[9] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7092–7096.

[10] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America (JASA)*, vol. 139, no. 5, pp. 2604–2612, 2016.

[11] Y. Wang and D. Wang, "A structure-preserving training target for supervised speech separation," in *IEEE ICASSP*, Florence, Italy, 2014, pp. 6127–6131.

[12] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Springer Science & Business Media, 2012, vol. 247.

[13] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain, 2016, pp. 2802–2810.

[14] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM TASLP*, vol. 22, no. 10, pp. 1533–1545, 2014.

[15] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *International Conference on Machine Learning (ICML)*, New York, USA, 2016, pp. 173–182.

[16] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1993–1997.

[17] Y. Saito, S. Takamichi, H. Saruwatari, Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM TASLP*, vol. 26, no. 1, pp. 84–96, 2018.

[18] 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.

[19] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in

IEEE International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2016, pp. 1–16.

[20] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech Enhancement Generative Adversarial Network,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3642–3646.

[21] D. Michelsanti and Z.-H. Tan, “Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3642–3646.

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, USA, 2017, pp. 1125–1134.

[23] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, “Generative adversarial network-based postfilter for STFT spectrograms,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3389–3393.

[24] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, “Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1283–1287.

[25] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, “Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3364–3368.

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

[27] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning (ICML)*, Lille, France, 2015, pp. 448–456.

[28] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1026–1034.

[29] 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.

[30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech,” http://dx.doi.org/10.7488/ds/1356, Available Online; Last accessed 17-January-2018.

[31] 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 (O-COCOSDA/CASLRE)*, Gurgaon, India, 2013, pp. 1–4.

[32] 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.

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

[34] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE/ACM TASLP*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[35] “P.862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs,” *Geneva: International Telecommunication Union*, 2007.

[36] C. 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.