IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Towards an Efficient and Accurate Speech Enhancement by a Comprehensive Ablation Study

Lidia Abad[1], *Fernando López*[2,3], *Jordi Luque*[3]

[1]Universidad Politécnica de Madrid
[2]Universidad Autónoma de Madrid
[3]Telefónica Innovación Digital

labadazcutia@gmail.com, {fernando.lopez, jordi.luque}@telefonica.com

Abstract

In the recent years, significant advancements in speech enhancement have been made through phase reconstruction, dual-branch methodologies or attention mechanisms. These methods produce exceptional results but at expenses of a high computational budget. This work aims to enhance the efficiency of the MP-SENet architecture by introducing MiniGAN, a generative adversarial network in the time-frequency domain. It features an encoder-decoder structure with residual connections, conformers, and parallel processing of signal magnitude and phase. We employ data augmentation techniques in training, investigate the impact of various loss terms, and examine architectural alterations to achieve lower operational costs without compromising performance. Our results on the VoiceBank+DEMAND evaluation set report that MiniGAN achieves competitive figures in objective metrics, obtaining a PESQ of 2.95, while maintaining low latency and reducing the computational complexity. The suggested MiniGAN system is ideally suited for real-time applications on resource-constrained devices, as it achieves a real-time factor of 0.24 and has a mere 373k parameters.

Index Terms: speech enhancement, time-frequency, magnitude, phase, efficiency, real time

1. Introduction

Speech enhancement (SE) models are crucial in current communications as they improve the quality and intelligibility of speech signals, by reducing background noise, reverberation, and other distortions [1]. SE can be understood as obtaining clean signals from original noisy signals. SE is crucial in multiple applications, such as telecommunications or hearing aids. In voice-controlled systems, it is used as a front-end to clean the speech for other technologies, such as wake-up word detection [2]. Traditional techniques for SE rely on statistical methods that assume the clean signal and noise follow specific statistical distributions. These methods have limitations. Deep learning has demonstrated significant improvements in SE. It can model complex patterns, handle large volumes of data, and adapt to diverse noise conditions. These advantages have greatly enhanced the ability of SE algorithms [3].

SE models can focus on different properties of speech signals, giving rise to time-domain models and time-frequency (TF) domain models. End-to-end models that enhance the speech signals based only on the waveform are known to be time domain models [4–6]. They face several difficulties. First, a lack of frequency features might result in distortions and artefacts. Second, the high sampling of the input space increases the computational cost of these models [7]. In contrast, TF models do not suffer from these constraints. The most successful models have recently been developed in the TF domain [8–13]. These models often transform the input signal through the Short-Time Fourier Transform (STFT). They obtain components of the signal, such as the magnitude and phase, but generally prioritize the former over the latter, as the enhancement of the phase is not without difficulties [13]. Nevertheless, recent studies have reported the importance of including phase components in SE processing tasks [14], demonstrating substantial performance improvements compared to studies that exclude phase information [13]. Therefore, several approximations have been proposed to solve the issues associated with estimating the phase [15]. Some models focus on the real and imaginary components of Fourier's complex spectrum, to enhance the signal's magnitude and phase [9–12]. A cross-cutting problem in SE remains: proposed architectures do not have a small enough footprint to be deployed in low-memory devices, creating a need to develop more lightweight approaches [16]. On top of that, models must be efficient enough to perform real-time SE [5]. Following previous research, several works in SE research have steered towards increasing the efficiency of SE models without significantly degrading their performance [17].

Given this context and motivated by the goal of defining a highly efficient model for SE, we propose MiniGAN. MiniGAN is a TF domain model inspired mainly by MP-SENet [13], but also by CMGAN [11] and TPTGAN [12]. The model consists of a generator and a discriminator. The noisy signal magnitude and phase are obtained in TF domain after performing the STFT. The generator employs as input the concatenated magnitude and phase, which are estimated from the noisy signal. In the latent space of the generator, two conformers are placed: one focuses on the time and the other on the frequency features. They are followed by two decoders used for parallel processing of the magnitude and phase of the noisy signal. The magnitude mask decoder outputs a mask that is applied to the noisy signal magnitude, obtaining the enhanced signal magnitude. Finally, by means of the phase and magnitude of the enhanced signal, the enhanced audio is reconstructed in time domain through the Inverse Short Time Fourier Transform (ISTFT). MiniGAN achieves nearly state-of-the-art metrics while being more efficient in terms of number of parameters, operations, model size, and real-time factor than MP-SENet. The high efficiency of our model is largely due to the simplification of the underlying architectures on which it is based. In addition to the ablation study on the state-of-the-art models, data augmentation techniques have been employed to mitigate overfitting in our model.

2. MiniGAN

MiniGAN is a TF domain model, in which the magnitude and phase of the noisy signal are obtained through the STFT. Feature extraction is done with 512 points for Fourier Transform,

10.21437/IberSPEECH.2024-21

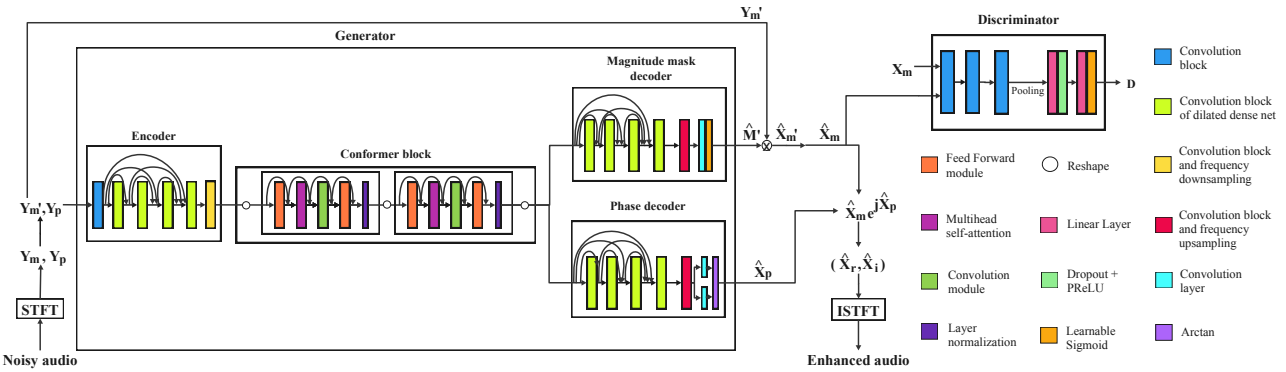

Figure 1: *MiniGAN structure. Y is the noisy signal, X the clean signal and $\hat{X}$ the enhanced signal. $\hat{M}$ is the mask of the magnitude. The subscripts m, p, r and i refer to the magnitude, phase, real and imaginary components respectively. The apostrophe indicates compressed feature.*

a windows size of 32ms and a hop size of 6.25ms. A power compression is applied to the magnitude of the noisy signal, in order to adjust the importance of quiet and loud sounds, thus modelling the human perception that is more sensitive to loud sounds than to quiet ones [18]. The compression factor is set to 0.3. Therefore, $Y'_m = Y_m^{0.3}$. The model inputs the compressed magnitude Y'_m and phase Y_p of the noisy signal. The estimated magnitude mask $\hat{M}'$ is applied to the compressed magnitude of the noisy signal Y'_m, obtaining the compressed magnitude of the enhanced signal $\hat{X}'_m$. This is followed by a power decompression, where $\hat{X}_m = \hat{X}_m'^{1/0.3}$. The magnitude $\hat{X}_m$ and phase $\hat{X}_p$ of the enhanced signal are used to obtain the real and imaginary components of the spectrum, $\hat{X}_r$ and $\hat{X}_i$, on which the ISTFT is applied to estimate the enhanced speech in time domain. Note that just the magnitude of the enhanced and clean signals, $\hat{X}_m$ and X_m, are employed as input for the discriminator.

2.1. Deep Learning Architecture

MiniGAN, which structure is depicted in figure 1, is based on a generative adversarial network scheme. The generator is composed of an encoder, a conformer block in the latent space, and two decoders, one for the magnitude and other for the phase of the signal. A noteworthy simplification of MP-SENet takes place in the latent space, as the original model comprises four conformer blocks.

The encoder has six convolution blocks, each of them comprised of a convolution layer, instance normalization [19] and PReLU activation [20]. The first block increases the number of channels from two, corresponding to the compressed magnitude and phase, to a total of thirty two. Note that in the original MP-SENet architecture, sixty four channels are employed instead. The four following blocks use a dilated dense net [21] with residual connections. The dilated dense net has dilation rates of 1, 2, 4 and 8. The residual connections increase the number of input channels by thirty two in each block, while the number of output channels is always thirty two. The last convolution block is used to downsample the frequency channel to its half. By doing so, the complexity in the latent space is reduced. The conformer block of the latent space has two conformers with an encoder-only structure. The first conformer focus on time features and the second on frequency ones. Before each conformer, reshaping the data is necessary in order to focus on the desired dependencies. Two decoders are subsequently used in parallel: one to process the magnitude and the other to handle the phase of the noisy signal. Both of them use a dilated dense net with the same dilation rates as in the encoder. Then, a convolution block is used to upsample the frequency dimensionality to its original one. In the magnitude mask decoder a last convolution layer is followed by a learnable sigmoid. The phase decoder has two convolution layers at the end, that separately focus on the real and the imaginary components of the complex spectrum. Finally, their outputs are combined using the arctangent function to obtain the phase of the enhanced signal.

The discriminator is conformed by three convolution blocks that include spectral normalization, which helps stabilizing the discriminator training [22]. After that, a pooling layer is used to reduce the dimensionality and two feed forward layers are used to perform the classification between clean and enhanced signals. The first layer is followed by PReLU and drop out, while the second one by a learnable sigmoid. This activation function helps discriminating between clean and enhanced signals.

2.2. Training criteria

As in MP-SENet, the generator loss is multiterm, in order to focus on different elements of the clean and enhanced signals. The loss of the generator is defined as a three term loss:

$$L_G = L_{TF} + \alpha_5 L_{Time} + \alpha_6 L_{GAN}, \tag{1}$$

where the weights α_5 and α_6 are hyperparameters. The real-imaginary term of the MP-SENet loss function is not included in MiniGAN to reduce computations. In fact, the ablation study reported by the MP-SENet work [13] shows that excluding this term has little impact on performance. The loss in the time-frequency domain, L_{TF}, is defined as:

$$L_{TF} = \alpha_1 L_{Mag} + L_{Phase}, \tag{2}$$

where L_{Mag} is the loss in the magnitude space with weight α_1 and L_{Phase} the loss in the phase space. They are expressed as:

$$L_{Mag} = \mathbb{E}_{X_m,\hat{X}_m}[||X_m - \hat{X}_m||_2^2], \tag{3}$$

$$L_{Phase} = \alpha_2 L_{IP} + \alpha_3 L_{GD} + \alpha_4 L_{IAF}, \tag{4}$$

where X_m and $\hat{X}_m$ are the magnitude of the clean and enhanced signals respectively. $\mathbb{E}$ refers to expectancy terms and $||\cdot||_2$ refers to the L_2 norm. The three terms of the phase loss are multiplied by their corresponding weights α_2, α_3 and α_4. L_{IP}, L_{GD} and L_{IAF} stand for the instantaneous phase loss, the group delay loss, and the instantaneous angular frequency loss respectively, which are expressed as:

$$L_{IP} = \mathbb{E}_{X_p,\hat{X}_p}[||f_{AW}(X_p - \hat{X}_p)||_1], \tag{5}$$

$$L_{GD} = \mathbb{E}_{\Delta_{DF}(X_p,\hat{X}_p))}[||f_{AW}(\Delta_{DF}(X_p - \hat{X}_p))||_1], \quad (6)$$

$$L_{IAF} = \mathbb{E}_{\Delta_{DT}(X_p,\hat{X}_p))}[||f_{AW}(\Delta_{DT}(X_p - \hat{X}_p))||_1], \quad (7)$$

where X_p and $\hat{X}_p$ are the phase of the clean and enhanced signals, f_{AW} is the antiwrapping function, and Δ_{DF} and Δ_{DT} the differential operators in the frequency and time axis respectively. The antiwrapping function is used to avoid the error expansion caused by phase wrapping, so ensuring that the angle x is within the range $[0, 2\pi]$ [15]. The $||\cdot||_1$ refers to the L_1 norm. The time domain loss, L_{Time}, is expressed as:

$$L_{Time} = \mathbb{E}_{x,\hat{x}}[||x - \hat{x}||_1], \quad (8)$$

where x refers to the waveform of the signals. Finally, the loss of the generator with respect to the discriminator is defined as:

$$L_{GAN} = \mathbb{E}_{X_m,\hat{X}_m}[||D(X_m, \hat{X}_m) - 1||_2^2]. \quad (9)$$

Overall, the generator loss is:

$$L_G = \alpha_1 L_{Mag} + \alpha_2 L_{IP} + \alpha_3 L_{GD} + \\ + \alpha_4 L_{IAF} + \alpha_5 L_{Time} + \alpha_6 L_{GAN}, \quad (10)$$

The coefficients are set as $\alpha_1 = \alpha_2 = 0.2$, $\alpha_3 = \alpha_4 = \alpha_5 = 0.1$, and $\alpha_6 = 0.05$, based on faster convergence in the validation split. No exhaustive optimization is performed. The discriminator loss is defined as:

$$L_D = \mathbb{E}_{X_m}[||D(X_m, X_m) - 1||_2^2] + \\ + \mathbb{E}_{X_m,\hat{X}_m}[||D(X_m, \hat{X}_m) - 0||_2^2]. \quad (11)$$

3. Methodology

3.1. Dataset

The data used for MiniGAN training and testing comes from the public database VoiceBank+DEMAND [23]. It contains 23899 audios of English speakers with different accents. Most of the audios last less than five seconds, although some of them have a duration up to seventeen seconds. The database is conformed by clean and noisy audios, obtained by merging clean audios of the Voice Bank Corpus [24] and noises from the DEMAND [25] database. Two additional synthetic noises are used to simulate human babble and male speech frequency in the VoiceBank+DEMAND or Valentini database. Different Signal-to-Noise Ratio (SNR) in the range of [0, 20] dB define the noisy audios. The Valentini database is already divided into train and test set and two train sets are made available. From these, the train set composed of fifty six different speakers is selected and a 10% is held out for validation purposes. The train and validation sets keep the same types of noise and SNRs ranges, but different speakers. The test set is not modified in order to perform a consistent comparison of the results with other SE models in the literature. The original test set is comprised of different speaker identities, noise types and SNRs than those of the train and validation sets.

3.2. Data augmentation

To avoid overfitting of the model, data augmentation techniques are implemented, increasing the volume and data diversity. For doing so, the model is exposed during the training process to a higher number of speech perturbations and different backgrounds, hence making the learning process to better generalize. The test set remains unmodified to facilitate easy comparison of our results. The noisy audios that are fed to the model, are derived from the Valentini train set. First, the clean audios and noises are separated. They are adjusted to a window of 2s. If the segment is smaller than 2s, zero padding strategies are used. After adjusting the audios to a window size, stationary white noise can be added to the clean audios probabilistically. White noise is only used in one of the MiniGAN variations. The SNR between the clean audios and the white noise is uniformly distributed within the range [0, 20] dB. After that, random background noises from the Valentini dataset, previously adjusted to the 2s window, are added to the clean audios. In this case the SNR uniform distribution is defined within the range [5, 30] dB. On top of merging clean audios and random noises from the Valentini dataset, reverberation is introduced in the audios. This is only done probabilistically, with a probability of 0.5. Finally, for all the audios, BandMasking from [5] is employed, masking some frequency bands in order to focus on low frequency regions, where most of the human speech concentrates.

3.3. Training and evaluation

We train the different approaches on a server that has a GeForce GTX 1080 Ti, an Intel Core i7-7700K CPU running at 4.20GHz, and 32GB of RAM. The initial learning rates of the generator and discriminator are set to 1×10^{-3} and the number of epochs is set to 120, with patience of 75 epochs. AdamW is selected as the optimizer [26], halving the learning rate every 30 epochs. Because of hardware limitations, the batch size used is 1 but a gradient accumulation of 16 is employed to update the gradients. In extremely unstable experiments, the initial learning rate of the generator is set to 5×10^{-4}, Kaiming initializers [20] are employed and clipping gradient is implemented with a clipping factor of 0.5 for both the generator and discriminator.

For testing the models, the original Valentini test set is employed. Entire audio files are enhanced without slicing. The objective metrics employed to evaluate the model include the PESQ [27], STOI [28], segmental SNR (SSNR) [29], and computed SIG, BAK and OVL (CSIG, CBAK and COVL) [30]. To analyse the efficiency of the approaches, the memory fingerprints as the model size and the number of parameters are measured. On top of that, the Floating Point Operations per Second (FLOPS) and Real Time Factor (RTF) for CPU inference are also reported. The RTF is the ratio between the time taken to process an audio and its duration. For a model to be used in real time, its RTF should not be higher than 1 [31].

4. Results

The motivation behind this work is to develop a model that excels in efficiency. Therefore, we do not only measure the results in terms of SE objective metrics, but also study the efficiency of our model, obtaining outstanding results. Apart from MiniGAN, some modifications to it are studied. We provide a dedicated page for the demonstration of the models[1]. MiniGAN-FT refers to a variation of MiniGAN in which the time and the frequency conformers are switched. For the MiniGAN-48 we increase up to forty eight the number of channels, compared to thirty two used in MiniGAN. In MiniGAN-NDA no data augmentation is applied during training. MiniGAN-WN stands for the addition of white noises during MiniGAN training. Finally, MiniGAN-ED refers to MiniGAN with an Encoder-Decoder

[1]MiniGAN demo - https://lidiaabad.github.io/MiniGAN/index.html

Table 1: *Number of parameters and objective metrics obtained on the Valentini test set for the noisy audios, state-of-the-art models and MiniGAN and its variations. Avg. refers to the average value obtained for the Valentini test set, and 95% CI to the range values of the 95% confidence intervals.*

Model	Param.	PESQ↑		CSIG↑		CBAK↑		COVL↑		SSNR↑		STOI↑	
	(M)	Avg.	95%CI	Avg.	95%CI	Avg.	95%CI	Avg.	95%CI	Avg.	95%CI	Avg.	95%CI
Noisy	-	1.97	-	3.48	-	2.55	-	2.74	-	1.68	-	0.92	-
SEGAN [4]	97.47	2.16	-	3.48	-	2.94	-	2.80	-	7.73	-	0.92	-
MetricGAN [8]	-	2.86	-	3.99	-	3.18	-	3.42	-	-	-	-	-
DEMUCS [5]	18.87	3.07	-	4.31	-	3.40	-	3.63	-	-	-	0.95	-
MANNER [6]	-	3.21	-	4.53	-	3.65	-	3.91	-	-	-	0.95	-
DB-AIAT [9]	2.81	3.31	-	4.61	-	3.75	-	3.96	-	10.79	-	**0.96**	-
DPT-FSNet [10]	**0.91**	3.33	-	4.58	-	3.72	-	4.00	-	-	-	**0.96**	-
CMGAN [11]	1.83	3.41	-	4.63	-	3.94	-	4.12	-	11.10	-	**0.96**	-
TPTGAN [12]	1.03	3.35	-	4.59	-	3.83	-	4.02	-	**11.63**	-	-	-
MP-SENet [13]	2.05	**3.50**	-	**4.73**	-	**3.95**	-	**4.22**	-	10.64	-	**0.96**	-
MiniGAN	0.373	2.95	(.05, .05)	**4.36**	(.04, .04)	3.46	(.03, .05)	**3.72**	(.05, .05)	7.49	(.20, .37)	0.94	(.00, .01)
MiniGAN-FT	0.373	2.93	(.05, .05)	**4.36**	(.05, .05)	**3.52**	(.03, .05)	3.71	(.05, .05)	8.43	(.16, .36)	0.94	(.00, .01)
MiniGAN-48	0.801	2.95	(.05, .04)	4.34	(.04, .05)	3.46	(.03, .05)	3.71	(.05, .05)	7.62	(.17, .41)	**0.95**	(.01, .00)
MiniGAN-NDA	0.373	**2.96**	(.04, .06)	4.31	(.05, .04)	3.43	(.03, .05)	3.70	(.05, .05)	7.06	(.21, .38)	0.94	(.00, .01)
MiniGAN-WN	0.373	2.82	(.06, .04)	4.27	(.05, .03)	3.47	(.03, .05)	3.59	(.04, .04)	**8.48**	(.14, .39)	0.94	(.00, .00)
MiniGAN-ED	**0.325**	2.05	(.02, .05)	3.64	(.04, .04)	2.64	(.01, .03)	2.88	(.03, .04)	1.50	(.05, .07)	0.93	(.00, .01)

modification, where the encoder and decoders include five convolution blocks instead of six. No dilated dense net with residual connections is used. The number of channels is doubled in each convolution block of the encoder and halved in the decoders. The last convolution block of each decoder reduces the number of channels to a fourth. The number of channels used in the latent space is sixty four. The frequency is not downsampled before the latent space, nor upsampled in the decoders. The decoders, in order to be symmetrical to the encoder, use dilation rates in inverted order to those of the encoder.

Table 1 shows the objective metrics obtained for the noisy audios of the test set, the results reported by SE models of the literature in their respective works and the results of the models developed in this paper. The number of parameters in millions is also shown. Bold numbers highlight the best values both reported for the state-of-the-art approaches and for our models. For our experiments, the average value (Avg.) is computed along with the 95% confidence intervals (95% CI), which were estimated using 10,000 bootstrap samples and $\alpha = 0.05$. Note that the MiniGAN models are more lightweight in terms of number of parameters than most popular models. With respect to the performance, MiniGAN does not outperform the current state-of-the-art but achieves competitive results while considerably reducing the number of parameters. The smallest model corresponds to MiniGAN-ED. Nevertheless, this architecture obtains poor result on the objective metrics, similar to those of noisy audios, likely due to the absence of the dilated dense net with residual connections and the frequency downsampling. The similarity in objective metrics obtained for the remaining MiniGAN variations, indicates that the model performance is not significantly determined by the elements studied.

Table 2 shows the efficiency figures for MP-SENet, MiniGAN and MiniGAN-ED. The remaining models developed in this work are not studied in terms of efficiency because of their high similarity with the original MiniGAN, both in structure and in objective metrics results. Analysing MiniGAN-ED is of interest because, despite achieving the lowest scores in objective metrics, it is the smallest model. The average value (Avg.) and the 95% confidence intervals (95% CI) are reported for the FLOPS and RTF. MiniGAN is more efficient than MP-SENet in number of parameters, operations, model size and RTF. MiniGAN and MiniGAN-ED are real time models as their RTF $\leq$1. The smallest model and the model with lower number of parameters is MiniGAN-ED. However, the number of FLOPS and RTF are smaller for MiniGAN. The original MiniGAN shows lower computational overload and is faster than MiniGAN-ED. This and the fact that MiniGAN obtains significantly higher objective metrics, make our original model outstanding in terms of performance and efficiency. Hence, it can be derived that the frequency downsampling, the lower number of channels and the encoder-decoder structure of MiniGAN generator are determinant elements for its efficiency and performance. Regarding the discriminator, we believe that incorporating phase information along with the magnitude is a promissing avenue for future work.

Table 2: *Efficiency comparison of MP-SENet [13], MiniGAN, and MiniGAN-ED. Avg. refers to the average value obtained for the Valentini test set, and 95% CI to the range values of the 95% confidence intervals.*

Model	Param.	FLOPS (M) ↓		Size	RTF ↓	
	(M)	Avg.	95%CI	(MB)	Avg.	95%CI
MP-SENet [13]	2.05	47489	(6, 5)	8.23	1.15	(.01, .01)
MiniGAN	0.373	**9023**	(1, 1)	1.43	**0.24**	(.00, .00)
MiniGAN-ED	**0.325**	10958	(1, 1)	**1.25**	0.56	(.01, .00)

5. Conclusions

In this paper, we presented MiniGAN, a novel TF model for speech enhancement that excels in efficiency. The overall architecture of our model is a generative adversarial network with an encoder-decoder structure in the generator. The latent space is based on two conformers and the decoders allow for parallel processing of the magnitude and phase of noisy signals. By focusing on a lightweight architecture, MiniGAN delivers high-quality speech enhancement suitable for real-time applications on devices with limited resources. Experimental results on the VoiceBank+DEMAND dataset confirm that MiniGAN strikes an optimal balance between performance and computational demands. Specifically, MiniGAN achieves a 5.5-fold reduction in the number of parameters and a 4.8-fold decrease in RTF compared to MP-SENet, while maintaining a PESQ score of 2.95. Furthermore, the study of several MiniGAN variations reveals that the encoder-decoder is a critical element influencing model performance and efficiency.

6. References

[1] T. Bäckström, O. Räsänen, A. Zewoudie, P. P. Zarazaga, L. Koivusalo, S. Das, E. G. Mellado, M. B. Mansali, D. Ramos, S. Kadiri, and P. Alku, *Introduction to Speech Processing*, 2nd ed., 2022. [Online]. Available: https://speechprocessingbook.aalto.fi

[2] G. Cámbara, F. López, D. Bonet, P. Gómez, C. Segura, M. Farrús, and J. Luque, "TASE: Task-aware speech enhancement for wake-up word detection in voice assistants," *Applied Sciences*, vol. 12, no. 4, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/4/1974

[3] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. Moore, "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," *Trends in Hearing*, vol. 27, p. 23312165231209913, 2023.

[4] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[5] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[6] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "MANNER: Multi-view attention network for noise erasure," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7842–7846.

[7] P. Ochieng, "Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis," *Artificial Intelligence Review*, vol. 56, no. Suppl 3, pp. 3651–3703, 2023.

[8] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[9] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7847–7851.

[10] F. Dang, H. Chen, and P. Zhang, "DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6857–6861.

[11] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric gan for speech enhancement," *arXiv preprint arXiv:2203.15149*, 2022.

[12] Z. Liu, Z. Jiang, W. Luo, Z. Fan, H. Di, Y. Long, and H. Wang, "TPTGAN: Two-path transformer-based generative adversarial network using joint magnitude masking and complex spectral mapping for speech enhancement," in *Neural Information Processing*, B. Luo, L. Cheng, Z.-G. Wu, H. Li, and C. Li, Eds. Singapore: Springer Nature Singapore, 2024, pp. 48–61.

[13] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *INTERSPEECH 2023*. ISCA, 8 2023.

[14] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639310002086

[15] Y. Ai and Z.-H. Ling, "Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses," 2023.

[16] L. Yang, W. Liu, R. Meng, G. Lee, S. Baek, and H.-G. Moon, "Fspen: an ultra-lightweight network for real time speech enhancement," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 671–10 675.

[17] I. Fedorov, M. Stamenovic, C. Jensen, L.-C. Yang, A. Mandell, Y. Gan, M. Mattina, and P. N. Whatmough, "Tinylstms: Efficient neural speech enhancement for hearing aids," in *Interspeech 2020*. ISCA, Oct. 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1864

[18] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon, "Exploring tradeoffs in models for low-latency speech enhancement," 2018.

[19] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2017. [Online]. Available: https://arxiv.org/abs/1607.08022

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015. [Online]. Available: https://arxiv.org/abs/1502.01852

[21] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6629–6633.

[22] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018.

[23] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and tts models," *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*, 2017.

[24] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[25] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust*, 2013, pp. 1–6.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[29] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," *5th International Conference on Spoken Language Processing (ICSLP 1998)*, 1998. [Online]. Available: https://api.semanticscholar.org/CorpusID:7794824

[30] H. Ding, T. Lee, Y. Soon, C. K. Yeo, P. Dai, and G. Dan, "Objective measures for quality assessment of noise-suppressed speech," *Speech Communication*, vol. 71, pp. 62–73, 2015.

[31] V. Pratap, Q. Xu, J. Kahn, G. Avidov, T. Likhomanenko, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Scaling up online speech recognition using convnets," 2020.