HIFI-GAN-2: STUDIO-QUALITY SPEECH ENHANCEMENT VIA GENERATIVE ADVERSARIAL NETWORKS CONDITIONED ON ACOUSTIC FEATURES

Jiaqi Su[1], *Zeyu Jin*[2], *Adam Finkelstein*[1]

[1]Princeton University, USA [2]Adobe Research, USA

ABSTRACT

Modern speech content creation tasks such as podcasts, video voice-overs, and audio books require studio-quality audio with full bandwidth and balanced equalization (EQ). These goals pose a challenge for conventional speech enhancement methods, which typically focus on removing significant acoustic degradation such as noise and reverb so as to improve speech clarity and intelligibility. We present HiFi-GAN-2, a waveform-to-waveform enhancement method that improves the quality of real-world consumer-grade recordings, with moderate noise, reverb and EQ distortion, to sound like studio recordings. HiFi-GAN-2 has three components. First, given a noisy reverberant recording as input, a recurrent network predicts the acoustic features (MFCCs) of a clean signal. Second, given the same noisy input, and conditioned on the MFCCs output by the first network, a feed-forward WaveNet (modeled via multi-domain multi-scale adversarial training) generates a clean 16kHz signal. Third, a pre-trained bandwidth extension network generates the final 48kHz studio-quality signal from the 16kHz output of the second network. The complete pipeline is trained via simulation of noise, reverb and EQ added to studio-quality speech. Objective and subjective evaluations show that the proposed method outperforms state-of-the-art baselines on both conventional denoising as well as joint dereverberation and denoising tasks. Listening tests also show that our method achieves close to studio quality on real-world speech content (TED Talks and the VoxCeleb dataset).

Index Terms— speech enhancement, denoising, dereverberation, generative adversarial networks, acoustic features

1. INTRODUCTION

Speech enhancement methods typically focus on alleviating severe noise and reverberation from recordings and improving intelligibility for downstream tasks such as speech recognition. Modern content creation scenarios (e.g., podcasts, video voice-overs, and audio books) would benefit from improving consumer-grade recordings (which suffer from moderate noise, reverb, and EQ distortion) to professional studio quality. Therefore, this paper addresses the speech enhancement problem in a different context from that of previous work: to improve single-channel consumer-grade recordings to sound like professional studio recordings. To address this goal requires solving the combined problem of denoising, dereverberation and equalization matching, while targeting a studio-quality dataset.

Recent advances in machine learning have enabled significant progress on the long studied topics of speech enhancement, denoising and dereverberation problems. Typical methods tackle the problem by learning a spectral mapping [1, 2] or masking [3, 4] on the magnitude spectrogram, while inverse STFT process to recover waveform introduces audible artifacts due to missing or mismatching phase. Other methods predict phase alongside the spectrogram [5, 6], or learn complex ratio mask [7, 8]. Another approach focuses on enhancement directly in the waveform, for example, using WaveNet [9, 10] and Wave-U-Net [11], to avoid information loss or phase inversion. State-of-the-art methods like DEMUCS [12] and PoCoNet [13] have shown significant audio quality improvement, especially for hard denoising cases with low SNRs. Yet those methods learn from datasets like VoxCeleb [14], the Valentini dataset [15] and the DNS Challenge Dataset [16] that do not contain studio-quality target audio, thus limiting the capabilities of the learnt models. Moreover, these datasets do not simulate conditions matching typical consumer-grade recording environments, which limits their use in the context of the problem we address. As a result, such audio can be improved by these methods, but the results remain far from studio-quality.

Generative adversarial networks (GANs) have been widely shown effective in achieving high fidelity audio in speech processing and generation. Researchers in speech enhancement have explored GANs on spectral features [17, 18] as well as on waveform [19, 20]. HiFi-GAN [21] shows high fidelity results by applying discrimination in both the time domain and the time-frequency domain. Meanwhile, an emerging branch of research performs speech enhancement by re-synthesis [22, 23], given recent success in high-fidelity speech synthesis [24]. The idea is to extract speech features from the input audio and re-synthesize the clean waveform using neural vocoders. This approach aligns with our objective, as the synthesized audio is naturally free of noise and reverberation. The performance is however limited by the quality of existing vocoders, as most do not generalize well across speakers and tend to generate "robotic" voices. They are also susceptible to inaccurately estimated speech features, leading to speech content distortion and unnatural prosody.

This paper proposes HiFi-GAN-2, which builds on our previous HiFi-GAN method [21] and targets studio-quality output. The previous HiFi-GAN uses a feed-forward WaveNet together with deep feature matching in multi-domain and multi-scale discriminators. HiFi-GAN-2 incorporates a separate recurrent neural network to predict the acoustic features of a clean target from those of noisy input. The WaveNet then conditions on the predicted acoustic features to generate the clean audio. This modification significantly improves output audio quality. We believe that the acoustic features, estimated from the entire input audio sequence, help the WaveNet (which has limited receptive field) to generate audio that more faithfully matches the original speaker and content. We evaluate the proposed method using objective and subjective tests in three application scenarios: (1) joint denoising and dereverberation for real-world recordings, (2) enhancement for real-world speech content at full bandwidth, and (3) conventional denoising. We also show in subjective evaluation that conventional denoising datasets that are of low quality can hinder model performance, and thus encourage use of studio-quality datasets in future research.

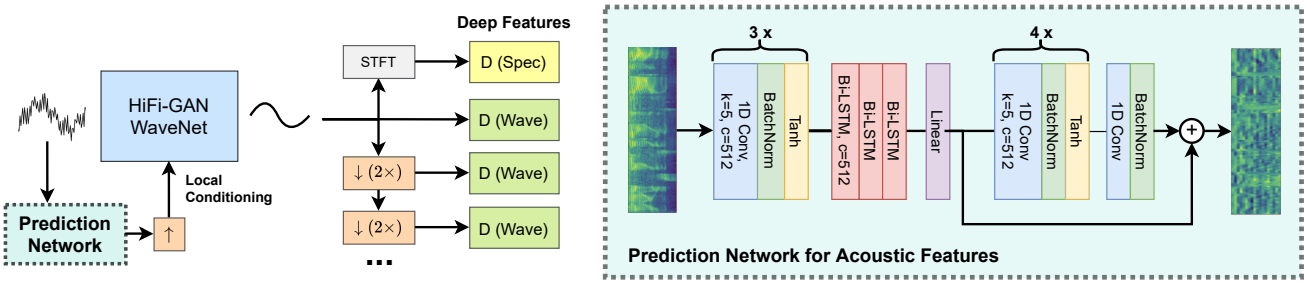

Figure 1: Architecture. A pre-trained network (*right*) predicts acoustic features (MFCCs) of clean speech based on a noisy input spectrogram. A WaveNet (*left*) generates clean speech from the same noisy input, locally conditioned on the predicted MFCCs. Adversarial training with deep feature matching involves a spectrogram discriminator and multiple waveform discriminators for the signal at different resolutions.

2. METHOD

HiFi-GAN-2 builds on top of our previous work HiFi-GAN [21] for speech denoising and dereverberation, to further push towards studio quality. HiFi-GAN uses an end-to-end feed-forward WaveNet together with deep feature matching in multi-scale multi-domain discriminators. Although HiFi-GAN is shown successful for obtaining clean high-fidelity audio recordings from noisy reverberant conditions, we observe inconsistency in speaker identity when noise and reverb are strong. This is likely caused by the ambiguity in disentangling speech content and speaker identity from environment effects (EQ and reverb). Moreover, the feed-forward WaveNet is not able to enforce consistent speaker identity due to limited receptive field and lack of global context. Thus, the network would benefit from extra information that helps it to infer speaker identity and content, i.e. the clean speech. One possible solution is to use speaker embedding as global conditioning, similar to that of multi-speaker speech synthesizer [25], but we did not observe quality improvement, possibly due to the utterance-level fuzziness of the embedding space. Instead, we propose conditioning the WaveNet on acoustic features that contain clean speaker identity and speech content information. Hence, we incorporate a separate recurrent neural network to predict clean acoustic features from the input noisy reverberant audio, which is then used as time-aligned local conditioning for HiFi-GAN. Such design combines benefits from waveform-to-waveform conversion, which avoids information loss and artifacts in STFT/ISTFT processes, and the effectiveness of acoustic features in modeling human perception of speech over a long period of context. The overall architecture is shown in Figure 1.

2.1. Acoustic feature prediction network

We propose a network inspired by Tacotron 2 [25] for acoustic feature prediction. It consists of three pre-processing convolution blocks (1D convolution, batch normalization and ReLU), three layers of bi-directional LSTMs, a linear projection layer, and a post-net of five convolution blocks (1D convolution, batch normalization and Tanh activation except for the last block). We use channel size of 512 across all the layers, kernel size of 5 for the convolutions, momentum of 0.9 for the batch normalization layers, and dropout of 0.2 for the recurrent layers. This network is trained using the acoustic feature of simulated noisy reverberant audio as input and that of clean audio as target. It minimizes the MSE losses of acoustic feature as well as the delta (first order difference) of the feature, for the outputs both before and after post-net.

To select a proper acoustic feature, we examined log mel spectrogram (Mel) and Mel-frequency Cepstral Coefficients (MFCCs). While Mel has higher frequency resolution, the MFCCs is more robust to noise. Our experiments found that predicting 18-coefficient MFCCs of the target clean audio from the 80-coefficient Mel of the input audio yields the best result. Since each cepstral coefficient has a different range of values, the target MFCCs is also globally normalized by subtracting each coefficient with the mean and dividing by four times the standard deviation using the clean audio dataset's statistics, following the practice of Qian et al. [26]. We did not observe statistically significant improvement in changing the number of cepstral coefficients to 24; yet the performance drops with 30 or 12 coefficients. Thus, we stick to 18 coefficients for further experiments. Our ablation study is discussed in details in Section 3.1.

2.2. Conditional WaveNet

The waveform denoising network is a feed-forward WaveNet [9] with local conditioning [27]. It uses non-causal dilated convolutions with dilation rate as a power of two to enable large receptive field. We use three WaveNet stacks (totaling 30 layers) and a channel size of 128 across the network. Our early experiments show vanishing benefit to further increasing the number of WaveNet stacks, as well as degraded performance with other channel sizes (64 or 256). We use weight normalization on all layers to accelerate convergence.

The prediction from the pre-trained acoustic feature prediction network is up-sampled using linear interpolation along time axis to match the length of the input waveform and is applied via additive local conditioning as is described in the original WaveNet design [27]: in each WaveNet layer, it is convolved with a 1×1 convolution before being added to the filter activation; same process is done for the gate activation. We hypothesize that the WaveNet can utilize the local conditioning in two ways: (1) if the acoustic features contain sufficient information, the WaveNet may serve like a vocoder where it re-synthesizes speech using the phase of the input waveform; (2) or, the WaveNet utilizes this auxiliary information to gain access to a cleaner representation of speech content as well as larger temporal context. Our experiment shows that (2) is more likely the case as the WaveNet with randomized acoustic features can still generate intelligible speech but it sounds muffled and less recognizable as the original speaker.

2.3. Adversarial training and loss functions

The adversarial training helps to improve perceptual quality and removes artifacts and noises. We follow the same design as HiFi-GAN, using a spectral discriminator and a set of waveform discriminators. The spectral discriminator takes in the 128-coefficient log mel-spectrogram. It consists of four stacks of 2D convolution layer, batch normalization and Gated Linear Unit (GLU), and lastly a convolution layer followed by global average pooling, similar to the one

used in StarGAN-VC [28]. It uses kernel sizes of (7, 9), (5, 8), (4, 8), (4, 6) and stride sizes of (1, 1), (1, 2), (2, 2), (2, 2) for the stacks, and the last convolution layer uses a kernel size of (32, 5). The channel sizes is 32 across all the layers. Meanwhile, a set of three waveform discriminators respectively operate at the output signal down-sampled by different ratios as a power of two, following the design in MelGAN [24]. Each waveform discriminator is composed of a set of grouped convolutions and global average pooling at the end, with Leaky ReLU between the layers. Specifically, the kernel sizes are 15, 41, 41, 41, 41, 5, 3; stride sizes 1, 4, 4, 4, 4, 1, 1; channel sizes 16, 64, 256, 1024, 1024, 1024, 1; and group sizes 1, 4, 16, 64, 256, 1, 1. The adversarial losses take hinge loss formulation.

The supervised loss function of the generator is composed of L1 waveform loss, and L1 losses of multiple log spectrograms with different FFT window sizes (i.e 512, 1024, and 2048 for 16kHz audio, each with one-fourth as its hop size). In addition, we apply the adversarial losses, as well as the feature matching losses [24] of the discriminators which are computed as L1 difference of the deep features between the generated audio and the ground-truth clean audio. The feature matching loss helps to stabilize GAN training and prevents the generator from mode collapse.

3. EXPERIMENTS

We evaluate our method, ablations and various baselines over studio-quality speech enhancement task as well as conventional denoising task. The term "studio-quality" implies that the clean audio used in training are recorded and professionally edited in an anechoic studio, at a sample rate >=44.1kHz. The "clean" category of the Device and Produced Speech (DAPS) Dataset [29] fits into this requirement. Due to limited bandwidth of baseline methods, we first conduct a comparative study at 16kHz, on joint denoising and dereverberation task on the DAPS dataset. Then we expand the experiment to real-world recordings used in content creation, evaluated at full 48kHz. Finally, we apply our method to conventional denoising task to show its broad applicability.

We used the architecture described in Section 2 for experiments. We compute Mel and MFCC using FFT length of 512 and hop size 160 at 16kHz. We first train the acoustic feature prediction network (24M params) for 100k steps using Adam optimizer with a batch size of 64 and input length of 256 frames. The learning rate starts with 0.001 and gets halved every 20k steps. Then we train WaveNet (10M params) with the weights of acoustic feature prediction network fixed. The WaveNet first trains for 1000k steps with learning rate 0.001, using the waveform and the spectrogram losses. Next we add randomly initialized discriminators to the output of the WaveNet (generator). We use learning rate 1e-05 for the generator (adversarial loss, feature matching loss and previously used loss), and 0.001 for the discriminators, for 100k steps. A batch size of 6 and a sample length of 22K are used throughout training. On a Tesla V100, each of the three training stages takes seven days, and inference takes 0.5 seconds per second of input audio. Audio samples for our experiments are available at:
`https://pixl.cs.princeton.edu/pubs/Su_2021_HSS/`

3.1. Joint denoising and dereverberation

The DAPS Dataset provides pairs of recordings of the same set of studio-quality speech re-recorded under twelve different room environments, and thus aligns with our goal of converting real-world recordings to studio-quality recordings. One male voice (m10) and one female voice (f10) are held out for evaluation purpose. We also hold out 2 minutes of audio per training voice for validation purpose. Our training set is constructed around the rest of the DAPS Dataset's clean set following the same data simulation and augmentation procedure as described in HiFi-GAN [21]. We convolve these studio-quality speech recordings with the 270 impulse responses from the MIT Impulse Response Survey Dataset [30], and then add noise from the REVERB Challenge database [31] and the ACE Challenge database [32]. Data augmentation of HiFi-GAN is used on all of speech, impulse responses and noise samples.

Table 1: Objective measures on the DAPS dataset.

Method	PESQ	STOI	SRMR	FW-SSNR
Noisy	1.41	0.87	4.79	3.04
HiFi-GAN [21]	2.00	0.89	7.67	7.62
HiFi-GAN (3×10)	1.92	0.89	7.61	8.52
FullSubNet [8]	2.14	0.89	7.23	4.50
DEMUCS [12]	2.16	**0.92**	7.51	**10.15**
HiFi-GAN-2 (ours)	2.23	0.92	7.83	9.98
Ablation Models				
A (no GAN)	2.32	0.92	7.77	10.08
A-GT (no GAN, GT)	**2.33**	0.92	8.23	10.03
B (mfcc2mfcc, no GAN))	2.23	0.92	7.57	9.94
C (local norm, no GAN)	2.25	0.92	7.54	9.96
D (mel2mel, no GAN)	2.28	0.92	7.82	9.89
D-GC (mel2mel, no GAN, GT)	2.21	0.91	**8.32**	9.50
A-GT-GAN (GT)	2.26	0.92	8.45	9.77
D-GAN (mel2mel)	2.25	0.91	7.87	9.60

Our best full approach **HiFi-GAN-2** consists of an acoustic feature prediction network that predicts globally normalized 18-coefficient MFCCs of clean target from 80-coefficient log mel spectrogram of noisy input, the WaveNet conditioning on the predicted 18-coefficient MFCCs, and GAN training. We conducted ablation experiments to address the following four design questions, and accordingly eight variants of our approach: **Q1:** *Should we train the WaveNet with ground truth acoustic features or generated ones?* **Q2:** *Should we use MFCCs or other acoustic features (e.g. log mel spectrogram) for conditioning?* **Q3:** *Should we predict clean MFCCs from input audio's MFCCs directly or from its log mel spectrogram?* **Q4:** *Should we apply global normalization, local normalization or no normalization for the conditioning?*

Model A: Same as **HiFi-GAN-2**, but no GAN training ("no GAN")
Model A-GT: Same as **Model A**, but conditioning on ground-truth clean acoustic features for training ("GT").
Model B: Same as **Model A**, but the prediction network takes globally normalized 18-coefficient MFCCs as input ("mfcc2mfcc").
Model C: Same as **Model A**, but the MFCCs are locally normalized using instance statistics ("local norm").
Model D: Same as **Model A**, but the prediction network outputs 80-coefficient log mel spectrogram ("mel2mel").
Model D-GT: Same as **Model D**, but conditioning on ground-truth clean acoustic features for training ("GT").
Model A-GT-GAN: Model A-GT with GAN training.
Model D-GAN: Model D with GAN training.

We also compare to four state-of-the-art baselines: our previous **HiFi-GAN** [21] with two WaveNet stacks and **HiFi-GAN (3×10)** with three stacks (as in this work), a spectral-domain method using complex ratio masking [8] (**FullSubNet**), a time-domain method using encoder-decoder structure [12] (**DEMUCS**), and a speech enhancement by resynthesis method [23] (**Regen**). **DEMUCS** and **FullSubNet** originally targeted at speech denoising, so we re-train their released models on our training set. Meanwhile, since speech re-synthesis may completely change the appearance of the signal, **Regen** is compared in the subjective evaluation only, using its

released audio samples re-sampled to 16kHz.

Table 1 shows the objective metrics [31] for speech denoising, dereverberation and enhancement. All variants of our proposed methods outperform all the baselines in PESQ and SRMR. **Model A-GT**, **Model A** and **Model D** scores the top three. Though **HiFi-GAN-2**'s objective score is lower, it has the highest perceptual quality shown in subjective evaluations. Adding one WaveNet stack brings moderate improvement to the perceptual quality but not the objective measures. Adding conditioning to the WaveNet improves the objective scores universally by a large margin. Globally normalized 18-coefficient MFCCs scores the best as conditioning (Q2, Q4), and it can be more accurately predicted from Mel than from MFCCs (Q3). Training with ground-truth conditioning can degrade test performance due to mismatch of training and inference conditions, as is the case in **Model D** and **D-GT**. However, training with ground truth MFCCs (**Model A-GT**) outperforms generated ones (**Model A**) (Q1). This may be due to that the prediction of MFCCs as a compact representation is sufficiently close to the ground truth. Although GAN training lowers objective scores, we observe significant perceptual quality improvement in the listening tests. GAN helps the output to match the clean audio's data distribution (hence sounds realistic) rather than direct approximation to ground truth.

Since the objective scores may not correlate with perceptual quality well [33], we also conduct Mean Opinion Score (MOS) tests using Amazon Mechanical Turk (AMT) on the baselines and our top performing methods. Using a studio-quality recording as high anchor and audio with noise (0dB SNR) as low anchor, a subject is asked to rate the sound quality of an audio recording on a scale of 1 to 5, with 1=*Bad*, 5=*Excellent*. We collected 449 valid HITs with 208 unique workers, totalling 11674 ratings. The MOS scores are shown in Figure 2(a). Our methods outperform all the baselines, and **HiFi-GAN-2** achieves the best average rating of 3.90 (± 0.03, $p < 0.05$ over second best in unpaired t-test). Therefore, adding conditioning and adding GAN training respectively bring steady perceptual quality improvement. While **Model A** and **Model A-GT** are rated the same, training on generated MFCCs receive more improvement from adversarial training than on ground-truth ones, as the former exposes artifacts to the discriminators caused by inaccuracy in MFCC prediction. **Model A-GT-GAN** can be an efficient alternative to **HiFi-GAN-2** as training on GT is easier.

3.2. Real-world speech at full bandwidth

We gather real-world customer-grade recordings from TED Talks (`www.ted.com`) and VoxCeleb1 [14] to further evaluate if our method can suffice speech content creation needs. We selected 10 male and 10 female speakers from TED 2004-06 and sampled two random sentences (5-6 seconds) per episode. For VoxCeleb1, we used Speech Transmission Index (STI) [34] to label each recording, and randomly sampled 50 audio clips to cover an STI range of 0.75-0.99 uniformly. Details are on our experiment result website.

We used the same trained models from Section 3.1, and extended output sample rate from 16kHz to 48kHz using the bandwidth extension model of Su et al. [35] trained also on the DAPS Dataset. We conducted the same MOS test as in Section 3.1, including 348 valid HITs with 128 unique workers and 7656 ratings. The result in Figure 2(b) shows that **HiFi-GAN-2** performs the best, and works well together with the bandwidth extension algorithm, achieving close to studio quality for the resulting 48kHz audio (4.27 ± 0.03, $p < 0.05$ over second best, $p < 0.0001$ over all baselines).

Furthermore, we conducted experiments to show that datasets used as clean audio in conventional speech enhancement can be low quality and thus hinders the performance of algorithms trained on them. For example, CREMA-D [36], a crowd-sourced emotional dataset, contains similar amount of reverb and noise to customer-grade recordings. We conducted the same MOS test as above on CREMA-D dataset using first 62 speakers speaking sentence labelled "IWL" in neutral emotion (84 unique workers, 4496 ratings); the result shows HiFi-GAN-2 at full bandwidth scores 4.254 while the original dataset only scores 2.349; it also worth mentioning that DEMUCS trained on the DNS Challenge dataset (which uses CREMA-D as clean data) scored 2.813 that is far lower than HiFi-GAN-2 trained on the DAPS dataset.

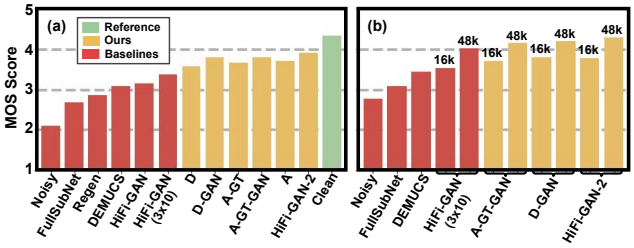

Figure 2: MOS scores: (a) joint denoising and dereverberation on the DAPS dataset; (b) enhancement on real-world speech content.

Table 2: Objective measures on the Valentini Dataset.

Method	PESQ	CSIG	CBAK	COVL
Noisy	1.97	3.35	2.44	2.63
HiFi-GAN [21]	2.94	4.07	3.07	3.49
DEMUCS [12]	3.07	4.31	3.4	3.63
Model A-GT	**3.18**	**4.49**	**3.60**	**3.84**
Model A-GT-GAN	3.14	4.41	3.56	3.78
Model A	3.15	4.37	3.60	3.76
HiFi-GAN-2	3.11	4.37	3.54	3.74

3.3. Conventional denoising

To show that the proposed methods also works in conventional setting, we experimented with the common benchmark Valentini dataset [15] for speech denoising. We follow the standard split of 28 speakers for training and 2 speakers for test. Table 2 shows our methods outperform all the other state-of-the-art methods on the objective measures, and **Model A-GT** achieves the highest scores so far to our knowledge. It is consistent with our previous observations that training with ground-truth conditioning without GAN is most favored by the objective measures.

4. CONCLUSION

In this paper, we characterize the difference between conventional speech enhancement and studio-quality audio enhancement, and present HiFi-GAN-2, a waveform-to-waveform enhancement method that improves the quality of real-world amateur recordings to studio quality. HiFi-GAN-2 consists of a recurrent neural network that predicts acoustic features (i.e. MFCCs) of the clean target from the input audio, and a feed-forward WaveNet for waveform enhancement that conditions on the predicted acoustic features, together with multi-domain multi-scale adversarial training. A pre-trained bandwidth extension network can be optionally applied to generate the final 48kHz studio quality signal from the output of HiFi-GAN-2. Extensive evaluations show that the proposed method outperforms all the other state-of-the-art baselines in both objective metrics and subjective metrics on joint dereverberation and denoising tasks as well as conventional denoising task.

5. REFERENCES

[1] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM TASLP*, vol. 23, no. 6, pp. 982–992, 2015.

[2] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 7–19, 2015.

[3] D. S. Williamson and D. Wang, "Speech dereverberation and denoising using complex ratio masks," in *ICASSP 2017*, pp. 5590–5594.

[4] W. Mack, S. Chakrabarty, F.-R. Stöter, S. Braun, B. Edler, and E. Habets, "Single-channel dereverberation using direct mmse optimization and bidirectional lstm networks," *Proc. Interspeech 2018*, pp. 1314–1318.

[5] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM TASLP*, vol. 28, pp. 380–390, 2019.

[6] H.-S. Choi, H. Heo, J. H. Lee, and K. Lee, "Phase-aware single-stage speech denoising and dereverberation with u-net," *arXiv preprint arXiv:2006.00687*, 2020.

[7] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *Interspeech 2020*.

[8] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," *arXiv:2010.15508*, 2020.

[9] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *ICASSP 2018*.

[10] J. Su, A. Finkelstein, and Z. Jin, "Perceptually-motivated environment-specific speech enhancement," in *ICASSP 2019*, pp. 7015–7019.

[11] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[12] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *Proc. Interspeech 2020*, pp. 3291–3295.

[13] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, *et al.*, "PoCoNet: better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," *Proc. Interspeech 2020*, pp. 2487–2491.

[14] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620.

[15] C. Valentini-Botinhao *et al.*, "Noisy speech database for training speech enhancement algorithms and tts models," 2017.

[16] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, *et al.*, "The Interspeech 2020 deep noise suppression challenge," *Proc. Interspeech 2020*, pp. 2492–2496, 2020.

[17] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *ICASSP 2018*, pp. 5024–5028.

[18] 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 *ICML*, 2019, pp. 2031–2041.

[19] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *Proc. Interspeech 2017*, pp. 3642–3646.

[20] S. Pascual, J. Serrà, and A. Bonafonte, "Towards generalized speech enhancement with generative adversarial networks," *Proc. Interspeech 2019*, pp. 1791–1795.

[21] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: high-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Interspeech 2020*.

[22] S. Maiti and M. I. Mandel, "Speaker independence of neural vocoders and their effect on parametric resynthesis speech enhancement," in *ICASSP 2020*. IEEE, pp. 206–210.

[23] A. Polyak, L. Wolf, Y. Adi, O. Kabeli, and Y. Taigman, "High fidelity speech regeneration with application to speech enhancement," *arXiv preprint arXiv:2102.00429*, 2021.

[24] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *NeurIPS 2019*.

[25] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *ICASSP 2018*.

[26] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," *In ICASSP 2020*, pp. 6284–6288.

[27] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, *et al.*, "Wavenet: A generative model for raw audio." in *SSW*, 2016, p. 125.

[28] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks," *arXiv:1806.02169*, 2018.

[29] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments..." *IEEE Signal Proc. Letters*, vol. 22, no. 8, 2015.

[30] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[31] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *WASPAA 2013*, pp. 1–4.

[32] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ace challenge," *IEEE/ACM TASLP*, vol. 24, no. 10, pp. 1681–1693, 2016.

[33] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin, "A differentiable perceptual audio metric learned from just noticeable differences," *Proc. Interspeech 2020*, pp. 2852–2856.

[34] P. Seetharaman, G. Mysore, B. Pardo, P. Smaragdis, and C. Gomes, "Voiceassist: Guiding users to high-quality voice recordings," in *ACM CHI 2019*, pp. 1–6.

[35] J. Su, Y. Wang, A. Finkelstein, and Z. Jin, "Bandwidth extension is all you need," in *ICASSP 2021*.

[36] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.