PERCEPTUALLY-MOTIVATED ENVIRONMENT-SPECIFIC SPEECH ENHANCEMENT

Jiaqi Su, Adam Finkelstein

Princeton University

Zeyu Jin

Adobe Research

ABSTRACT

This paper introduces a deep learning approach to enhance speech recordings made in a specific environment. A single neural network learns to ameliorate several types of recording artifacts, including noise, reverberation, and non-linear equalization. The method relies on a new perceptual loss function that combines adversarial loss with spectrogram features. Both subjective and objective evaluations show that the proposed approach improves on state-of-the-art baseline methods.

Index Terms— Denoising, de-reverberation, equalization matching, speech enhancement, perceptual loss.

1. INTRODUCTION

This paper introduces a unified method for enhancing the quality of recorded speech. Many factors in a typical environment can diminish the quality of a recording – including noise, reverberance, and undesirable equalization. We describe a data-driven method that ameliorates such effects by learning from example recordings made in a specific source environment. Following the acquisition approach proposed by Mysore [1], studio-quality recordings in a variety of voices are re-played and re-recorded in the source environment. This approach yields two parallel recordings – one studio quality and a second degraded in the source environment – from which our proposed method can learn a transfer function that can enhance new recordings made in the source environment towards studio quality.

Speech enhancement is a well-studied problem [2], and researchers have developed a variety of methods to remove noise [3, 4], remove reverberance [5, 6], and match equalization [7]. Classical approaches to these problems operate in the time-frequency domain and usually assume prior knowledge of the spectral structure of speech. Some methods estimate an inverse filter of a room impulse response so as to minimize weighted linear prediction error [8]. Researchers also apply non-negative matrix factorization on the spectrogram to remove both noise and reverberance [9, 10].

Our work builds on recent advances in speech enhancement using deep learning, which has demonstrated significant performance improvements over traditional approaches. Spectral methods learn to enhance speech by transforming the spectrogram of a distorted input signal to match that of a target signal. One approach is to estimate a non-linear mapping from a distorted input spectrogram directly to clean signal [11, 12]. A learned function can also predict a mask (binary or ratio) to be applied to the input spectrogram [13, 14, 15]. One common network architecture is the feed-forward neural network with several convolutional layers using a context window of several frames as input [11]. Kodrasi and Bourland [13] use a denoising auto-encoder to estimate late reverberation PSD from microphone signal PSD. Recurrent neural networks such as LSTM [14, 15] and GRU [16] are increasingly used in speech enhancement, because they better capture temporal dependency between frames.

A drawback common to spectral methods is that they rely on phase to recover the output waveform. The phase of a reverberant signal can deviate from that of the corresponding clean signal, and thus naively copying the input phase results in residual artifacts. Estimating phase via the widely-used Griffin-Lim algorithm [17] produces noticeable glitches. The method of Williamson and Wang [18] learns a complex ratio mask, but achieves only moderate performance gains because the phase component is less predictable than the magnitude component of the mask.

Deep learning methods that transform an input waveform directly to the output waveform obviate the phase problem. However, they require a relatively larger receptive field because the waveform itself has much higher temporal resolution than that of a spectrogram. WaveNet [19] and its variant for speech denoising [20] use dilated convolutions to enable large receptive fields. Recurrent neural networks have been used to establish long-term dependency [21]. Researchers have also tackled de-reverberation from the perspective of source separation in the time domain by conducting an LSTM on a latent space learned by an encoder-decoder structure [22].

A major challenge for waveform-based approaches is the design of a suitable loss function. L1 and L2 loss are brittle with respect to phase shifts and minor alignment errors, for example due to clock drift. Moreover, L1 and L2 loss fail to capture contextual information and perceptual qualities, as they penalize per-sample error in the waveform. Thus, while these loss functions work well for synthetic data which can be perfectly aligned, they tend to perform poorly for real world data. Generative adversarial networks (GANs) incorporate an implicit *adversarial loss* – the discriminator, which models whether a given audio example is real or fake, learns the distribution of the target signal and drives the generator to approximate the target. GANs can thus boost speech enhancement performance [15, 23, 24]. However, GANs can also suffer from mode collapse [25], and therefore researchers typically combine adversarial loss with some regularizing term like L1 loss [23].

Our approach combines some aspects of the spectral methods and waveform methods. In particular, we use a waveform-to-waveform generator in a GAN framework, regularized by a spectrogram-based loss function. The method is robust to phase shifts in the data because it works directly on the waveform, rather than recovering it from a spectrogram. The loss function captures some perceptual qualities directly from the spectrogram, while other perceptual qualities that are more difficult to characterize are modeled by the adversary. Thus, our main contributions are: (1) a unified speech enhancement method that works with both synthetic and real data, involving noise, reverberance, non-linear microphone response, and dynamic gain adjustment – and wherein the data is only weakly aligned due to phase shift, clock drift, etc; (2) a perceptually-motivated spectrogram loss function, coupled with a GAN, with application to speech enhancement; and (3) experiments (MOS tests and objective evaluations) demonstrating improved performance over baseline methods for both synthetic and real data.

2. APPROACH

2.1. Feed-forward WaveNet

WaveNet [19] is a waveform-to-waveform autoregressive model that models the probability distribution of a sample conditioned on all previous samples. Its key design features have proven successful in dealing with waveform synthesis and conversion: dilated convolutions yield a receptive field that is exponential in the number of layers; causal convolutions model temporal dependencies; residual and skip connections accelerate learning; Gated Activation Units (GAU) help information flow as in LSTMs; and Softmax prediction supports multi-modal distributions. As a generative approach, although it trains in parallel, it requires generating one sample at a time, which is inefficient at inference time. Later, Parallel-WaveNet [26] was proposed to reduce inference time on GPUs, at the cost of higher complexity in the learning procedure.

Inspired by WaveNet, Speech Denoising WaveNet [20] uses the WaveNet architecture but modifies several design choices. It uses non-causal dilated convolutions and real-valued regression outputs to predict samples in parallel. It uses 3×1 convolution as a post-processing step to avoid sporadic point discontinuities. For the purpose of speech enhancement, it has L1 loss on both the clean speech prediction branch and the noise prediction branch. As shown in the generator part of Figure 1, we use the same architecture as Speech Denoising WaveNet for our generator, but with a different loss function. Also, we do not use speaker identity conditioning, which is present in WaveNet and Speech Denoising WaveNet, because we target at a speaker-independent model.

2.2. Perceptually-motivated Loss

One limitation of Speech Denoising WaveNet is that training requires sample-to-sample L1 loss, and thus the approach only trains on simulated data obtained by adding background noise to clean speech. The effectiveness of L1 loss falls off when one waveform is time-shifted from the other by even one sample. Moreover, a significant performance gap is observed at inference time on real data, since real world noise can interact with speech in complicated ways. The approach is also not applicable for types of noise that cannot be produced in simulation. Re-examining L1 loss, another issue is that it only considers per-sample accuracy, but not perceptual quality. Humans hear sound not as individual samples, but rather the frequencies inherent in sequences of samples. Therefore, we seek an objective function that is more closely related to human perception and can withstand a certain degree of misalignment between data.

2.2.1. Spectrogram Loss

Perceptual loss was proposed in computer vision research for neural style transfer [27]. Feature maps are extracted from a trained recognition network such as VggNet, as a semantic representation of the input image. Content loss is computed on the mean squared error (MSE) of the feature maps, whereas style loss is computed on the MSE of the Gram matrix of the feature maps. Researchers have found that the learned feature maps capture information more closely related to human perception than input pixels [28]. In the audio domain, Parallel-WaveNet [26] takes a similar approach using feature maps of a trained phoneme recognition network with the WaveNet structure. They found that applying perceptual loss has significantly improved their performance. However, one difficulty for using perceptual loss in the audio domain is that there is no well established pre-trained recognition model from which researchers can extract deep features. Meanwhile, there are many well studied traditional acoustic features, such as log spectrogram, log mel-spectrogram, mel-frequency cepstral coefficients (MFCC), and mel-generalized cepstral coefficients (MGC). They work well as an informative representation of audio, especially for human speech.

In this paper, we propose to use a spectrogram-based loss function for training, as a kind of perceptual loss. In addition, we also use L1 or L2 loss on the waveform, when sample level alignment is possible, for example with synthetic data. This combines benefits from both waveform-to-waveform conversion, which has no need for inverse STFT, as well as the spectrogram's effective modeling of human perception. For the particular choice of spectrogram, we experimented with traditional spectrogram, log spectrogram, mel-spectrogram and log mel-spectrogram, and found L1 loss on log spectrogram works best in our application. Finally we note that the steps of the STFT process are differentiable, and thus we can perform end-to-end deep learning with spectrogram-based loss.

2.2.2. Generative Adversarial Training

In our experiments, we found that a feed-forward network like WaveNet coupled with spectrogram loss could mostly remove reverberation, but the resulting audio sounds noisy. The noise is most severe for models trained on real data, and particularly for either unvoiced sounds or audio containing correlated noise. It has been widely observed in computer vision that L1 and L2 loss functions have the problem of blurring image results [29]. The analogous over-smoothing effect on spectrograms can give rise to artifacts like poorly modeled unvoiced sounds. Therefore, we supplement our spectrogram loss with a Generative Adversarial Network (GAN) training process to encourage finer detail in the spectrogram.

GANs are an important tool for generative and conversion deep learning problems. They have been used in speech enhancement (e.g., SEGAN [23]) and coupled with WaveNet structure for synthesis [30]. The main idea is that two adversaries – the generator and the discriminator – are trained together. The discriminator tries to distinguish between real data and the output of the generator, and will learn to identify any telltale artifacts from generation. Meanwhile the generator improves its fidelity as it learns to fool the discriminator (in some sense mimicking human perception). Thus, GANs help produce more naturalistic speech, and reduce telltale artifacts like noise [15, 23, 24] However, GANs can be hard to train because of mode collapse [25], and thus studies have pointed out that GANs should be trained together with some auxiliary loss [29, 31]. Therefore, we use the proposed spectrogram loss as auxiliary supervision when training with a GAN.

We adopt the discriminator structure from StartGAN-VC [32], which takes in a segment of log mel-spectrogram and predicts joint probability of the speech being real. It is a gated CNN, with several stacks of convolutional layer, batch normalization layer and Gated Linear Unit (GLU). The discriminator is fully convolutional, thus allowing inputs of arbitrary size. Figure 1 shows the detailed structure of the discriminator. During training, the generator G optimizes loss L_G which is the sum of adversarial loss and spectrogram loss:

$$\begin{aligned} L_G(x, x') &= \alpha|\text{LogSpec}(G(x)) - \text{LogSpec}(x')| \\ &\quad + (1-\alpha)(1 - D(\text{LogMel}(G(x))) \\ L_D(x, x') &= D(\text{LogMel}(G(x))) + 1 - D(\text{LogMel}(x')) \end{aligned}$$

The discriminator D optimizes loss L_D. Here LogSpec represents log spectrogram, and LogMel represents log mel-spectrogram. The tuple (x, x') denotes the pair of input audio x and target audio x' from the training dataset.

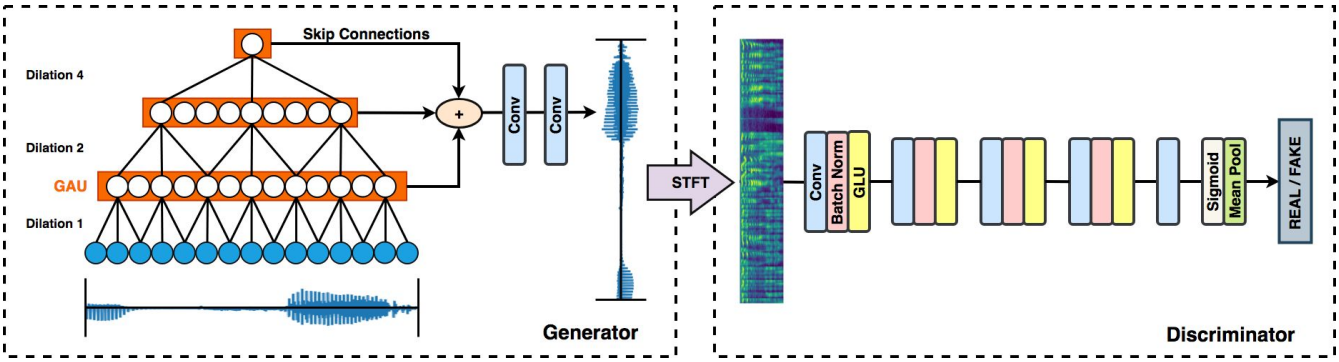

Fig. 1: Network Architecture. The generator is a feed-forward WaveNet architecture that outputs a waveform based on an input waveform. The discriminator takes in the log mel-spectrogram of either a real waveform or a generated waveform, and outputs a prediction (real or fake).

3. EVALUATION

We evaluate our method using a 20-layer feed-forward WaveNet with two stacks of dilated convolutions. The channel size is 256 across the entire network. We use two convolutional layers of 3×1 filters at the post-processing step. The STFT for the spectrogram loss relies on a window size of 2048 and hop size of 512, with a sampling rate of 16 kHz. We give equal weights to sample-level L1 loss and spectrogram loss when necessary. For the discriminator, we use kernel sizes of (3, 9), (3, 8), (3, 8), (3, 6), stride sizes of (1, 2), (1, 2), (1, 2), (1, 2), and channel sizes of (1, 32), (32, 32), (32, 32), (32, 32) for the sequence of the network layers, following the discriminator structure of StarGAN-VC [32]. The input is computed as the log of 80 coefficient mel-spectrogram ranging from 20Hz to 8000Hz, using the same STFT parameters as before. For training with both L1 loss and perceptual loss, we train the network for 100K iterations with a batch size of 10 using the ADAM optimizer with learning rate 0.0001, reduced by a factor of ten after 50K iterations. Training with only perceptual loss is more sensitive to the learning rate, and thus we reduce the learning rate by a factor of ten after 10K iterations, and again after 50K iterations. For GANs, we take a generator model pre-trained with spectrogram loss, and train the discriminator from scratch for 5K iterations while keeping the generator fixed. Finally we perform joint training on both generator and discriminator.

Experiments were conducted on both simulated room settings and real room settings. For real data, we used the Device and Produced Speech (DAPS) Dataset [1], which provides pairs of studio quality recordings and distorted ones of the same speech. DAPS is produced by replaying studio quality audio in certain indoor or outdoor environments and re-recording them with consumer devices. Thus, the audio pairs are weakly aligned, and the low quality recordings incorporate interactions of many acoustic factors in the real world scenario, making it suitable for our purposes. For simulated data, we obtained noisy reverberant recordings by applying room impulse response filters from the SimData category of the 2014 REVERB Challenge dataset [6] to the studio quality recordings in DAPS. Background noise was added to the reverberant speech with 20dB SNR. This gave us perfectly aligned inputs and targets. Twenty voices, ten male and ten female, from DAPS dataset, were used in our experiments. Each of the voices narrates the same set of five scripts, each around two minutes. Across all the experiments, the first nine voices for each gender and the first four scripts were used for training, and the remaining unseen speakers and unseen script were held out for evaluation. Each experiment handles one specific room setting with distinct acoustic conditions.

We conducted subjective and objective evaluations of our methods in comparison with three baseline methods from the literature.

1. WPE [8]: A traditional inverse filtering method that does not require training. It addresses de-reverberation specifically, but not denoising or equalization matching.
2. BLSTM [14]: A learning-based spectral masking method, using two layers of Bidirectional LSTM to predict ideal ratio mask over magnitude spectrogram. It copies the input phase for the prediction to get waveform back. We compare with this method on both simulated and real data.
3. Speech Denoising WaveNet (WN) [20]: The architecture used in our method, but with a different loss. It uses L1 losses on both the speech prediction branch and the noise prediction branch, which restricts it to training on perfectly aligned data, so we will show comparison with it only on simulated data.

We compare those baselines with two variants of our method:

1. SPEC: Uses spectrogram loss solely.
2. SPEC-GAN: Combines spectrogram loss and GAN loss.

In the scenarios of simulated data, we always used L1 loss alongside our loss, because L1 is an easier training signal than perceptual loss and helps training loss to drop fast. In the scenarios of real room, we solely used perceptual loss.

3.1. Subjective Evaluation

We conducted a Mean Opinion Score (MOS) test which asks subjects to rate the quality of the provided audio pieces. Subjects were recruited on Amazon Mechanical Turk (AMT), a crowd-sourcing platform commonly used for such experiments [33]. We designed four room settings of tests, three simulated rooms with varying reverberation time and microphone-to-speaker distance (Table 1), and one real room from the DAPS dataset – a typical office room with moderate noise and reverberance. The models for comparison were trained and evaluated on each specific room setting respectively. There are 26 utterances, each spoken by one male and one female. A subject is presented with 18 different rating questions, two for each method condition (five methods, clean speech and reverberant speech), and additionally four validation questions in the same test voice. We accept the answers only if the subject rates all validation questions correctly, resulting in more than 100 ratings per tuple (condition, room, voice). The audio clips and their MOS scores can be found at our project website.[1]

[1] https://pixl.cs.princeton.edu/pubs/Su_2019_PM/

Table 1: Simulated Room Conditions. RT60 is the time that takes for reverberation to decay by 60dB. Distance is the distance between speaker and microphone when room impulse response is captured.

Room	RT60	Distance
Sim Room 1	0.25s	200cm
Sim Room 2	0.25s	50cm
Sim Room 3	0.50s	50cm

Figure 2 shows the MOS results. Our methods, both SPEC and SPEC-GAN, outperform all the baselines in all four room settings. Our methods achieve ratings close to clean speech in Sim Room 1 and Sim Room 2, which are relatively easy cases. SPEC-GAN has varying performance, and its average rating does not always outperform SPEC's. Further ANOVA test revealed that our method beats all baseline methods with p-value lower than 10^{-5} except for Sim Room 1 when our method is compared to WaveNet which has p-value below 0.002 (still significant). This means our method significantly improves over all baselines. However, it is not statistically significant when SPEC-GAN is compared with SPEC. Our belief is that this may be due to just-noticeable differences, and requires further study.

WPE and BLSTM reduce reverberation but do not totally remove it. Thus, they do not work with the target of generating studio quality speech. However, they are relatively robust to a wide range of reverberation times, and thus their performances do not drop much with the increasing difficulty of the room settings. Speech denoising WaveNet has the closest results to ours, but it exhibits substantially more noise, especially correlated noise, relative to our methods.

We also observe gender-based differences in the results. In Sim Room 3, where SPEC-GAN improves the female voice, the male voice score worsens. This may imply that speech enhancement requires optimizing different aspects for different genders.

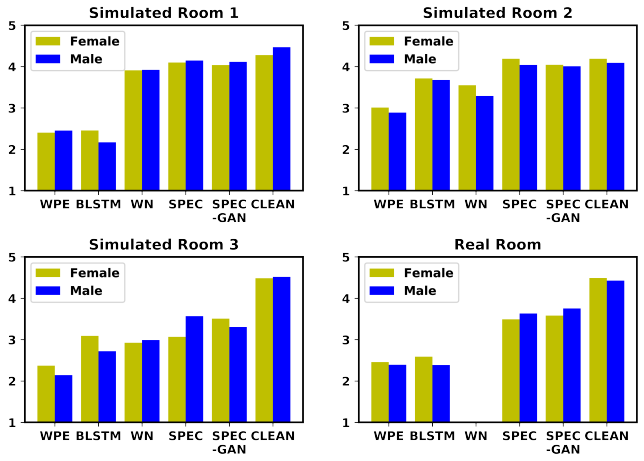

Fig. 2: MOS Test conducted on four different room settings. Models are trained to be gender independent cross-speaker, but we split data during evaluation to reveal the performance difference on voices of different genders. The methods compared are Weighted Prediction Error (WPE), two-layer bidirectional LSTM (BLSTM), Speech Denoising WaveNet (WN), our spectrogram loss (SPEC), our spectrogram loss plus generative adversarial training (SPEC-GAN) and clean studio recordings (CLEAN). WN is not used in "Real Room" experiment and thus not plotted.

Table 2: Objective Evaluation Scores. PESQ, FESEGSNR and SRMR are the higher the better. CD is the lower the better.

Method	PESQ	FWSEGSNR	SRMR	CD
CLEAN	4.64	35.0	8.41	0.0
REVERB	1.24	-0.63	5.82	7.02
DN-WN	2.17	-1.55	**8.18**	6.94
BLSTM	2.10	5.87	6.90	3.87
WPE	1.39	0.01	7.03	6.98
SPEC	2.45	6.34	7.45	4.70
SPEC-GAN	**2.61**	**12.53**	8.17	**3.12**

3.2. Objective Evaluation

We report in Table 2 objective evaluation scores for metrics Perceptual Evaluation of Speech Quality (PESQ), Frequency-weighted Segmental SNR (FWSEGSNR), Speech-to-reverberation Modulation Energy Ratio (SRMR), and Cepstrum Distance (CD) based on evaluation tools provided by the 2014 REVERB Challenge [6]. These are commonly used metrics for evaluating enhancement over noisy reverberant speech.

SPEC-GAN significantly outperforms all other baselines in PESQ, FWSEG-SNR and CD, and achieves performance close to clean speech on SRMR. This suggests that the improvement of the GAN is just-noticeable, and thus not captured by human Turkers in the MOS test. Employing a GAN does reduce the interruptions of background noise and ghosting of residual reverberation.

4. CONCLUSION

In this paper, we present a deep learning method to enhance speech recordings made in a specific environment. The method handles denoising, de-reverberation, and equalization matching in one network. We introduce a new perceptually motivated loss function that combines adversarial loss with spectrogram features. We show that the method offers an improvement over state-of-the-art baseline methods in a mean opinion score test and objective evaluation tests.

In general, DNN-based speech enhancement methods achieve optimal performance when training and inference have matched acoustic conditions. Our method also learns for a specific environment. The receptive field of a neural network usually needs to scale up with the reverberation time to cover the corresponding span of the original clean signal. To extend our method to cross-environments, future work could explore various strategies, such as the method of Wu et al. [34] which selects the optimal temporal and spatial contexts based on estimated reverberation time before inference, and Lee et al. [35] which integrates the trained DNN models of various acoustic conditions online. In the meantime, the idea of learning a feature embedding can be integrated to enable environment-independent learning. Also, future research could employ some kind of bottleneck structure to learn features that describe what is being spoken, and force all different types of noise or environments to the same spot in the embedding.

Although we demonstrate on the speech enhancement problem, the method presented in this paper is general. We believe it could be applied to the broader problem of *acoustics matching* where recordings from one environment are transformed to sound like they were recorded in another, in a fashion similar to the work of Germain et al. [7] (except that they just focus on equalization). This could be useful for sound effect production.

5. REFERENCES

[1] Gautham J Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? A dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2015.

[2] Philipos C Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2 edition, 2013.

[3] Yariv Ephraim and David Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] Pascal Scalart et al., "Speech enhancement based on a priori signal to noise estimation," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. IEEE, 1996, vol. 2, pp. 629–632.

[5] Patrick A Naylor and Nikolay D Gaubitch, *Speech dereverberation*, Springer Science & Business Media, 2010.

[6] Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Armin Sehr, Walter Kellermann, and Roland Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[7] François G Germain, Gautham J Mysore, and Takako Fujioka, "Equalization matching of speech recordings in real-world environments," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 609–613.

[8] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[9] Zhiyao Duan, Gautham J Mysore, and Paris Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in nonstationary noise environments," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[10] Hideaki Kagami, Hirokazu Kameoka, and Masahiro Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 31–35.

[11] Kun Han, Yuxuan Wang, DeLiang Wang, William S Woods, Ivo Merks, and Tao Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 6, pp. 982–992, 2015.

[12] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[13] Ina Kodrasi and Hervé Bourlard, "Single-channel late reverberation power spectral density estimation using denoising autoencoders," *Proc. Interspeech 2018*, pp. 1319–1323, 2018.

[14] Wolfgang Mack, Soumitro Chakrabarty, Fabian-Robert Stöter, Sebastian Braun, Bernd Edler, and Emanuël Habets, "Single-channel dereverberation using direct mmse optimization and bidirectional lstm networks," *Proc. Interspeech 2018*, pp. 1314–1318, 2018.

[15] Chenxing Li, Tieqiang Wang, Shuang Xu, and Bo Xu, "Single-channel speech dereverberation via generative adversarial training," *Proc. Interspeech 2018*, pp. 1309–1313, 2018.

[16] Joao Felipe Santos and Tiago H Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, 2018.

[17] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[18] Donald S Williamson and DeLiang Wang, "Speech dereverberation and denoising using complex ratio masks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5590–5594.

[19] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio.," in *SSW*, 2016, p. 125.

[20] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[21] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn 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*. Springer, 2015, pp. 91–99.

[22] Yi Luo and Nima Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," *Proc. Interspeech 2018*, pp. 342–346, 2018.

[23] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[24] Meet H Soni, Neil Shah, and Hemant A Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5039–5043.

[25] Ian Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[26] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al., "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[27] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[30] Qiao Tian, Bing Yang, Jing Chen, Benlai Tang, and Shan Liu, "Generative adversarial network based speaker adaptation for high fidelity wavenet vocoder," *arXiv preprint arXiv:1812.02339*, 2018.

[31] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[32] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks," *arXiv preprint arXiv:1806.02169*, 2018.

[33] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?," *Perspectives on psychological science*, vol. 6, no. 1, pp. 3–5, 2011.

[34] Bo Wu, Kehuang Li, Minglei Yang, and Chin-Hui Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, 2017.

[35] Wei-Jen Lee, Syu-Siang Wang, Fei Chen, Xugang Lu, Shao-Yi Chien, and Yu Tsao, "Speech dereverberation based on integrated deep and ensemble learning algorithm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*. IEEE, 2018, pp. 5454–5458.