ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746735

ATTENUATION OF ACOUSTIC EARLY REFLECTIONS IN TELEVISION STUDIOS USING PRETRAINED SPEECH SYNTHESIS NEURAL NETWORK

Tomer Rosenbaum[1] *Israel Cohen*[1] *Emil Winebrand*[2]

[1]Faculty of Electrical & Computer Engineering, Technion–Israel Institute of Technology, Haifa, Israel
[2]Insoundz Ltd., Tel Aviv, Israel

ABSTRACT

Machine learning and digital signal processing have been extensively used to enhance speech. However, methods to reduce early reflections in studio settings are usually related to the physical characteristics of the room. In this paper, we address the problem of early acoustic reflections in television studios and control rooms, and propose a two-stage method that exploits the knowledge of a pretrained speech synthesis generator. First, given a degraded speech signal that includes the direct sound and early reflections, a U-Net convolutional neural network is used to attenuate the early reflections in the spectral domain. Then, a pretrained speech synthesis generator reconstructs the phase to predict an enhanced speech signal in the time domain. Qualitative and quantitative experimental results demonstrate excellent studio quality of speech enhancement.

Index Terms— Acoustic early reflections, speech dereverberation, speech synthesis, generative adversarial networks.

1. INTRODUCTION

In television (TV) studios, the speech signal captured by the microphone is degraded by adverse effects such as additive noise and reverberation. Focusing on reverberation, the sound reaching the microphone consists of the desired direct sound, early acoustic reflections (which arrive roughly during the first 50 ms after the direct sound), and late reflections. It is known that late reflections cause significant degradation to the speech quality and intelligibility [1]. Early reflections, on the other hand, are traditionally considered desirable to boost speech coloration and intelligibility [2]. A study to determine the impact of early reflections on speech intelligibility by cochlear implant listeners showed that early reflections neither enhance or reduce listeners' speech perception [3]. However, for monitoring and evaluating audio devices, e.g., in TV studios, early reflections are considered undesirable and cause adverse effects to the sound quality [4].

Designing a studio to reduce early reflections is usually adequate, but expensive. Walker [5] presented a design methodology in sound monitoring rooms to control early reflections by redirection of early sound energy. This method was implemented in the design of new studios in BBC's broadcasting house. In [6], more design methodologies are described for the cases of monophonic, stereophonic, and also multichannel sound. Dunn and Protheroe [7] analyzed the early reflections in six control rooms and discussed how different room properties, such as room geometry, desk size, and materials, affect the early reflections. Shlomo and Rafaely [8, 9] presented a preliminary attempt to blindly estimate reflection amplitudes using an iterative estimator, based on maximum likelihood and alternating least squares.

Attenuation of late reflections and speech denoising have been extensively studied [10, 11, 12, 13], but attenuation of early reflections using digital signal processing is still a significant challenge. In this paper, our objective is to attenuate the early reflections in TV studios with arbitrary designs, using digital signal processing and machine learning. In our setup, the sound in the studio is captured using a fixed modern microphone array instead of the traditional methods (e.g., neck mic or boom mic) [14]. Once captured, the data is processed to reduce reverberation and background noise, and an enhanced speech signal is returned. This enhanced signal consists of the direct sound and the early reflections. We focus on generative adversarial network (GAN)-based speech synthesis generators that generate waveform speech signals given their Mel-spectrograms [15, 16, 17, 18]. Inspired by recent works in image processing [19], we propose that a generator, well-trained on clean speech signal synthesis, can be used as a prior for speech enhancement tasks. We show that artifacts caused by early reflections are very noticeable in the signal spectrogram magnitude. Therefore, enhancement in the time-frequency domain with a convolutional neural network is a natural choice. After the magnitude is enhanced, the pretrained generator reconstructs the spectrogram phase to get a studio-quality waveform speech signal.

The remainder of this paper is organized as follows: Section 2 presents and formulates the problem. Section 3 describes the proposed method. Section 4 details the experimental setup and shows the results. Section 5 concludes this work.

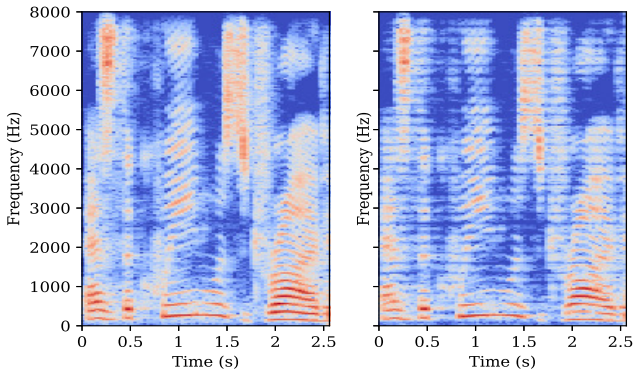

Fig. 1. Speech spectrograms: (a) Clean speech, (b) Speech with a single early reflection, $\alpha = 1$, $T = 2$ ms [Eq. (2)].

2. PROBLEM FORMULATION

2.1. Model

Let $x(t)$ be a speech source in the discrete-time domain in an arbitrary room. The signal is captured by a microphone array, and the observed data is processed using existing methods to reduce late reflections (e.g., using some variant of weighted prediction error (WPE) [13]) and to reduce background noises (e.g., using the minimum variance distortionless response (MVDR) beamformer [20]). The output of this processing is the observed single-channel signal, $y(t)$, which is modeled as:

$$y(t) = (h * x)(t) \tag{1}$$

where $h(t)$ is the "enhanced" room impulse response (RIR), satisfying $T_{60} \leq 50$ ms (meaning $20 \log_{10} |h(t)| \leq -60$ dB for $t \geq 50$ ms) and $*$ stands for linear convolution. Let $X = \text{STFT}(x) = X(t, f)$ denote the short-time Fourier transform (STFT) of $x(t)$, and let $|X_M| = \text{Mel}(|X|) = |X_M(t, c)|$ denote the Mel-spectrogram of $x(t)$. Given the observed degraded signal $y(t)$, the goal is to design a system f that returns an estimate of the source signal $f(y(t)) = \hat{x}(t) \approx x(t)$.

2.2. Time-Frequency Domain

As shown in Figure 1, early reflections in speech signals cause notches in certain frequencies, which are very noticeable in the STFT magnitude (the blue horizontal stripes). To get some intuition regarding this phenomenon, assume a simple model of a single early reflection:

$$h_{\text{single}}(t) = \delta(t) + \alpha\delta(t - T) \tag{2}$$

where α and T are the amplitude and the delay of the reflection, respectively. Now, assume $\alpha = 1$ and $x(t) = \cos(2\pi f t)$ for some frequency $f \in \mathbb{R}^+$. Then, the degraded signal is given by:

$$\begin{aligned} y(t) &= \cos(2\pi f t) + \cos(2\pi f(t - T)) \\ &= 2\cos(2\pi f T)\cos(2\pi f(2t - T)), . \end{aligned} \tag{3}$$

Hence, $y(t) \equiv 0$ if $\cos(2\pi f T) = 0$, which is true if $f = \frac{1}{4T} + \frac{1}{2T}k$ for some $k \in \mathbb{Z}^+$. For, e.g., $T = 2$ ms, we will observe notches in the frequencies $125, 375, 625, \ldots$ Hz.

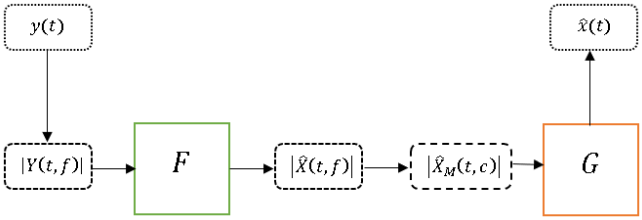

Fig. 2. Proposed system. The module F has trainable weights while the weights of G are fixed.

3. PROPOSED METHOD

The proposed system is illustrated in Figure 2, which comprises two main modules: An attenuation module, F, of early reflections, and a speech synthesis generator, G. The module F is a U-Net neural network, and the module G is a pretrained speech synthesis generator HiFi-GAN [15] that synthesizes the speech waveform from its Mel-spectrogram. Given a speech waveform with early reflections, $y(t)$, the network F is fed with the STFT magnitude $|Y(t, f)|$. The enhanced spectrogram $|\hat{X}(t, f)| = F(|Y(t, f)|)$ is transformed to Mel domain $|\hat{X}_M|$ and then fed into the generator G to generate the enhanced waveform $\hat{x} = G(|\hat{X}_M|)$. In the training phase, given a reference clean speech signal, x, we acquire the corresponding signal y according to (1). The RIR is generated as follows:

$$h_{\text{multi}}(t) = \delta(t) + n(t)b(t)\exp(-\alpha_{T_{60}} t) \tag{4}$$

where $\delta(t)$ is Kronecker delta and:

$$n(t) \sim \mathcal{N}(0, 1) \tag{5}$$

$$b(t) = \begin{cases} 1 & \text{w.p. } 0.05 \\ 0 & \text{w.p. } 0.95 \end{cases} \tag{6}$$

$$T_{60} \sim \mathcal{U}[15, 50] \text{ ms} \tag{7}$$

$$\alpha_{T_{60}} = \frac{3 \log 10}{T_{60}}. \tag{8}$$

Note that the left term of (4) (i.e., the Kronecker delta) corresponds to the direct speech signal, and the right term corresponds to the early reflections.

The objective function for optimization is given by:

$$\mathcal{L}(x, y) = \mathcal{L}_{\text{s}}(|X|, F(|Y|)) + \mathcal{L}_{\text{s}}(|X|, |\text{STFT}(\hat{x})|) \tag{9}$$

Table 1. DNSMOS scores (↑) on LJSpeech validation set. Degraded signals are generated using RIR from Eq. (4).

T_{60} [ms]	**Ref.**	**Deg.**	F	G	$F+G$
20	4.11	3.85	4.01	3.95	**4.02**
30		3.84	4.00	3.95	**4.02**
40		3.84	3.98	3.94	**4.01**
50		3.84	3.98	3.96	**4.01**

where the spectral loss, $\mathcal{L}_{\text{s}}$ is defined as:

$$\mathcal{L}_{\text{s}}(|X|,|Y|) = \sum_{t,f} ||X(t,f)| - |Y(t,f)|| + \lambda_{\text{s}} \sum_{t,f} \left| \log \left| \frac{X(t,f)}{Y(t,f)} \right| \right| . \tag{10}$$

Note that the first term of (9) is computed with respect to the output of F, and the second term is computed with respect to the output of G. During training, the weights of the pretrained generator are fixed, and the optimization of $\mathcal{L}$ is with respect to the weights of F.

4. EXPERIMENTAL RESULTS

4.1. Data and Implementation Details

To train the module F, we use the LJSpeech dataset [21], sampled at $f_{\text{s}} = 22.05$ kHz. The network is fed with STFT magnitudes with 512 frequency bands. Its architecture is U-Net, including 6 convolution layers with kernel size 5×5 and stride 2, followed by 6 transposed convolution layers. For G, we use the official implementation of HiFi-GAN [15] (configuration V1), which generates a speech waveform its Mel spectrogram with 80 bands, pretrained on LJSpeech dataset [21]. In every epoch, given a clean speech sample $x(t)$ from the dataset, we randomly draw $h_{\text{multi}}(t)$ according to (4) and generate the corresponding $y(t)$. The system is trained for 100 epochs with batch size of 2 using AdamW optimizer [22]. We set $\lambda_{\text{s}} = 1$ in (10) when optimizing the objective function (9).

4.2. Performance Evaluation

To show the contribution of the system, we compare the speech quality of 4 waveform signals:

1. Deg. (degraded) – the input signal $y(t)$.
2. F – signal is formed by inverse STFT using the magnitude of $\hat{X}(t,f)$ and the phase of $Y(t,f)$ (denoted as $z(t)$).
3. G – signal is formed using only the generator without F (i.e., the signal $G(|Y_M|)(t)$).

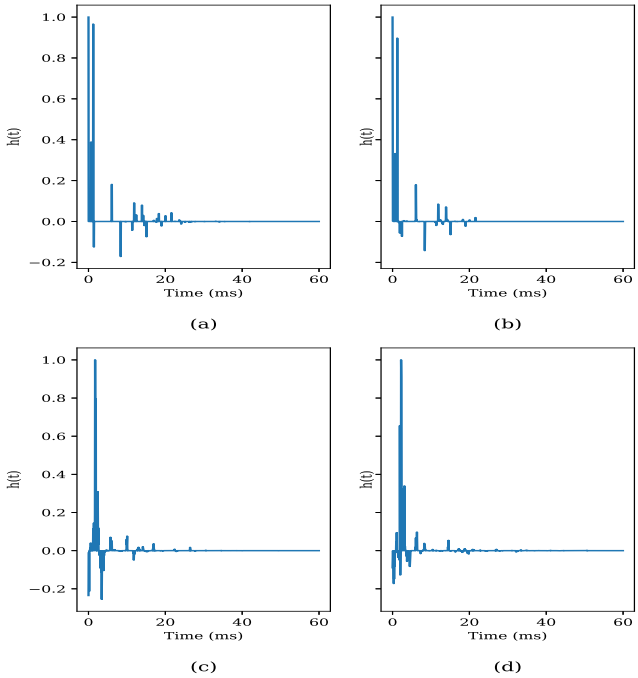

Fig. 3. RIR and estimations: (a) RIR generated according to (4) with $T_{60} = 40$ ms, (b) RIR estimate from the output of F, (c) RIR estimate from the output of the generator (G), (d) RIR estimate from the output of the system ($F+G$).

4. $F+G$ – the output of the system $\hat{x}(t)$.

Table 1 shows the mean DNSMOS score [23], which is highly correlated with human perception, over the LJSpeech validation set (a larger score indicates better perceptual speech quality). As can be seen, $F+G$ achieves the best score.

To evaluate the system's performance, we propose to use the C_2 clarity index. Given an RIR $h(t)$, the C_2 clarity index is defined as:

$$C_2(h) = 10 \log_{10} \frac{\sum_{t=0}^{\tau_2} h^2(t)}{\sum_{t=\tau_2+1}^{\infty} h^2(t)} \tag{11}$$

where τ_2 is the timestamp corresponding to 2 ms after the peak of the direct speech (the Kronecker delta in (4)). Larger values of C_2 imply that reflections later than 2 ms after the direct speech are insignificant. It is important to mention that usually, the C_{50} clarity index (i.e., taking τ_{50} instead of τ_2) is used for evaluation of dereverberation methods. Still, since we try to measure the attenuation of reflections that arrive much earlier than 50 ms, it makes more sense to use the C_2 clarity index [7]. Note that the RIR is required for the calculation of the clarity index. However, because of the non-linearity of F and G, the relation between the clean signal $x(t)$ and the reconstructed signal $\hat{x}(t)$ is not necessarily linear. For proof-of-concept, we approximate the RIR using

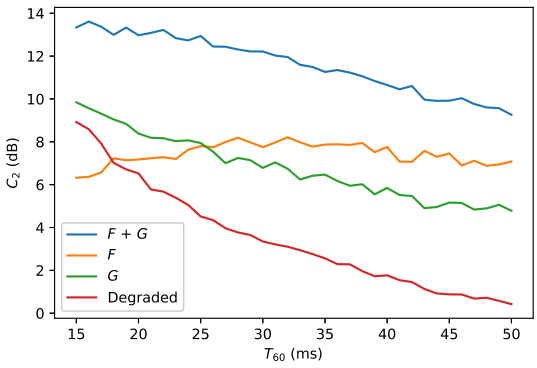

Fig. 4. Clarity index C_2 (↑) for different values of T_{60}.

least square regression with L_1 regularization (LASSO [24]). More specifically, given input-output pair $\{x, y\}$, we approximate h by solving the optimization problem:

$$\hat{h} = \arg\min_{h} \|Xh - y\|_2^2 + \lambda\|h\|_1 \qquad (12)$$

where X is a Toeplitz matrix formed from the signal $x(t)$, and λ is a hyper-parameter that controls the weight of the regularization. We choose L_1 regularization instead of L_2 because we expect h to be sparse. We assume that the length of h is 60 ms. Figure 3 shows an example of RIR generated according to the model in (4), its estimate using (12), and RIR estimates using the output of the generator G, and the output of the system $F + G$. Note that for obtaining RIR estimates from the outputs of the generator or the system, the Toeplitz matrix X in (12) is formed from the output signal instead of $x(t)$ (i.e., we reconstruct the clean speech using G or $F+G$). In all cases, we choose $\lambda = 20$.

We evaluate the C_2 clarity index on LJSpeech validation set based on the approximation of h in the following way:

- Given a clean speech signal $x(t)$ and a fixed T_{60} value, generate 10 RIRs according to the model in (4), the corresponding $y_i(t)$ $(i = 1, \ldots, 10)$ and the corresponding outputs $\hat{x}_i$ for $F + G$, $G(|Y_M|_i)$ for G, and z_i for F.
- Estimate RIRs with respect to the pairs $\{G(|X_M|_i), \hat{x}_i\}$ for $F + G$, $\{G(|X_M|_i), G(|Y_M|_i)\}$ for G, and $\{x_i, z_i\}$ for F.
- Calculate C_2 and take the mean over i.
- Take the mean over all files in the validation set (150 records).

Results for different T_{60} values are shown in Figure 4. As can be seen, $F + G$ achieves the largest clarity index, which matches the DNSMOS scores in Table 1. Interestingly, the clarity index of F is approximately constant for different values of T_{60}. The score of G is better than the score of the degraded signals, which indicates that even just reconstructing a speech waveform using a pretrained generator might improve the speech quality (Table 1) and reduce early reflections (Figure 4).

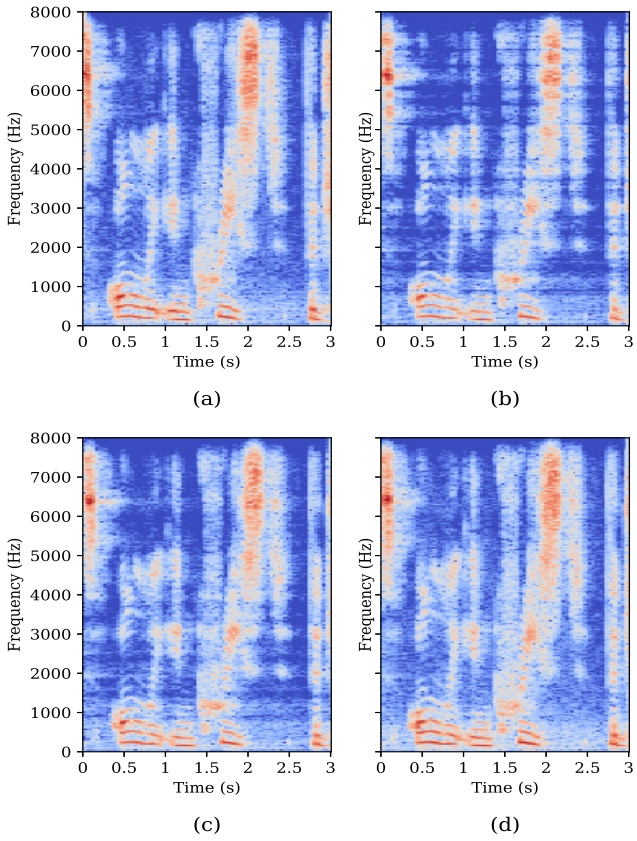

Fig. 5. Speech spectrograms: (a) Clean speech, (b) degraded speech ($T_{60} = 40$ ms), (c) output of G, (d) output of $F + G$.

For demonstration, we present typical speech spectrograms of the outputs in Figure 5. We see that G enables to eliminate the high-frequency notches, while $F+G$ enables to eliminate both the low and high-frequency notches.

5. CONCLUSIONS

We have presented a method for attenuating early acoustic reflections in TV studios using digital signal processing and machine learning. Experimental results show that the proposed method reduces early reflections and that algorithmic solutions might be an alternative for traditional methods to control early reflections in studios. The method may be extended to the multichannel case in future work, where spatial information can be exploited to improve performance. Furthermore, an additional bandwidth extension module may be explored to get studio-quality speech enhancement sampled at 48 kHz.

6. REFERENCES

[1] Arthur Boothroyd, "Room acoustics and speech perception," *Seminars in Hearing*, vol. 25, pp. 155–166, 2004.

[2] J. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, 2003.

[3] Y. Hu and K. Kokkinakis, "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. EL22–EL28, 2014.

[4] S. Kishinaga, Y. Shimizu, S. Ando, and K. Yamaguchi, "On the room acoustic design of listening rooms," *Journal of The Audio Engineering Society*, 1979.

[5] R. Walker, "Early reflections in studio control rooms: The results from the first controlled image design installations," in *Audio Engineering Society Convention 96*. Audio Engineering Society, 1994.

[6] B. Walker, "Room acoustics for multichannel listening: Early reflection control," in *Audio Engineering Society Conference: UK 22nd Conference: Illusions in Sound*. Audio Engineering Society, 2007.

[7] M. Dunn and D. Protheroe, "Visualization of early reflections in control rooms," in *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[8] T. Shlomo and B. Rafaely, "Blind localization of early room reflections from reverberant speech using phase aligned spatial correlation," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2547–2547, 2020.

[9] T. Shlomo and B. Rafaely, "Blind amplitude estimation of early room reflections using alternating least squares," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 476–480.

[10] O. Ernst, S. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 390–394.

[11] L. Pfeifenberger and F. Pernkopf, "Blind speech separation and dereverberation using neural beamforming," *arXiv preprint arXiv:2103.13443*, 2021.

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

[13] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation.," in *Interspeech*, 2017, pp. 384–388.

[14] A. Farina, A. Capra, L. Chiesi, and L. Scopece, "A spherical microphone array for synthesizing virtual directive microphones in live broadcasting and in post production," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. Audio Engineering Society, 2010.

[15] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 17022–17033, Curran Associates, Inc.

[16] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[17] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, "VocGAN: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," in *Proc. Interspeech 2020*, 2020, pp. 200–204.

[18] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram," *arXiv preprint arXiv:1904.03976*, 2019.

[19] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 2437–2445.

[20] EAP Habets, J. Benesty, S. Gannot, and I. Cohen, "The MVDR beamformer for speech enhancement," in *Speech Processing in Modern Communication*, pp. 225–254. Springer, 2010.

[21] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset, 2017.

[22] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in Adam," *CoRR*, vol. abs/1711.05101, 2017.

[23] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6493–6497.

[24] M. Schmidt, "Least squares optimization with L1-norm regularization," *CS542B Project Report*, vol. 504, pp. 195–221, 2005.