ATTENUATION OF ACOUSTIC EARLY REFLECTIONS IN SIMULATED ROOM USING PRETRAINED SPEECH SYNTHESIS NEURAL NETWORK

Dan Ilan Ben David

June 2022

Abstract

The layout of a standard room creating early reflection in the room, which convert to a distortions of the audio signal. ideally we desire to achieve a audio speech level of an acoustic room, even if the record have been recorded in a less then ideal setting. methods to enhance the audio are varied, for example machine learning and digital signal processing have been extensively used. and yet, most methods have not taken to account the physical layout and characteristics of the room, witch have a high correlation to the early reflection and can help reduce them. in this paper we propose a two-stage machine learning methods, that have been trained on a simulated room in order to extract those characteristics. the model have been trained using a simulated room impulse respond convoluted with a acoustic room level audio signal. the model compose of a trainable U-net convolution neural network on the spectral domain of the audio in order to attenuate the early reflection. fallowed by a pretrain speech synthesis generator that work on the mel-spectral of the audio, in order to predict the enhanced audio speech signal in the time domain. In this paper we have shown the value of using a true room impulse response, in order to improve on already existing methods.

1 Introduction

In a recording room, the sound signal captured by a microphone is degraded do to the effects of noise and the reverberation of the signal in the room. unlike noise, there is a correlation in the reverberation we can utilized in order to extract of original clean sound signal. the reverberation can be split into the desired direct sound, early acoustic reflections (which arrive in the first 50 ms after the direct sound) and late reflections (anything later then 50 ms). although early reflections mostly considered a desirable effect, and can boost speech coloration and intelligibility [2], a clean and direct sound signal can be use for the tasks on monitoring and evaluation of a audio signal and equipment [4]. this is why TV and radio station use an acoustic recording room in order to cancel these reflection.

Designing a recording room in a way that reduces early reflections is usually expensive, this paper propose an inexpensive method to salve this solution. the method is a generative adversarial network (GAN)-based speech synthesis generators that generate waveform speech signals given their Mel-spectrograms [5, 6, 10, 3]. with convolutions neural network enhancement in time-frequency domain beforehand, to better use the early reflections correlation in the room [7]. in this paper, we will focus on the simulation path of the room, to better represent a recording room, and even used a sensors array to test the method.

this paper is organized as follows: Section 2 presents and formulates the problem and the room simulation. Section 3 describes the method. Section 4 details the experimental setup and presents the results. Section 5 concludes the paper.

2 Problem Formulation

2.1 model

let $x(t)$ be a clean speech source with no reflection and no background noise. the signal-channel in the room is modeled as:

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

when $h(t)$ represent the room impulse response (RIR), and $(*)$ stand for the standart linear convolution. in order to take into account only the early reflection, we going to assume h satisfy $T_{60} \leq 50ms$ (meaning $20 \log_{10} |h(t)| \leq -60dB$ for the first $50ms$), every thing after that is consider a noise, which this paper don't delve into. the goal of our method is to design a system f that can estimate the source signal $x(t)$ from the noisy signal $y(t)$.

$$f(y(t)) = \hat{x}(t) \cong x(t) \tag{2}$$

2.2 Room Simulation

in order to simulate a room, we first need to understand how the power energy flux (PEF) of the sound wave (SW) are traveling in our module [8]. assume we can divide the walls in the enclosure into small discrete panels. for an incoming SW toward a panel, we can calculate the outgoing PEF toward any direction [Ω_e in figure 1]. for an incoming SW toward a panel [Ω_i in figure 1], the SW get reflected in all direction, this can be split into two types of reflection [1]. part of it will get reflected along the main reflection line [Ω_l in figure 1], called specular reflection, this is where most of the energy will transfer to. and the other type will diffuse in all direction, called diffuse reflection. and the rest is getting absorbed by the panel itself, and will be counted as a lost of energy.

let base all our calculation from the point view of the panel (orientation β_p). first we need to calculate the main reflection line base on an incoming PEF toward the panel in direction Ω_i (altitude angle of θ_i, and azimuth ϕ_i). when the PEF hit the panel, the main reflection line get bounce off it with the same altitude, but with azimuth in the opposite direction. if Ω_l is the direction of the line, then $\theta_l = \theta_i, \phi_l = \phi_i + \pi$. now in order to find the out going reflected PEF

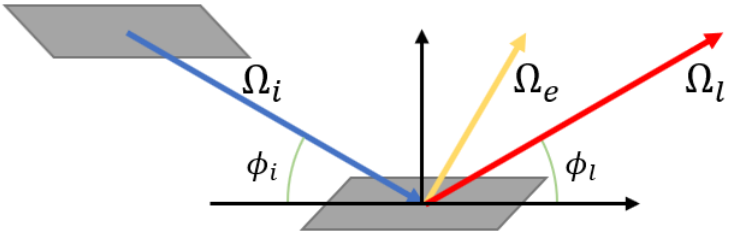

Figure 1: SW from source panel Ω_i, get reflected in direction Ω_e with power base of the distance from the main panel Ω_l

in direction Ω_e, we want to calculate how "close" Ω_e the wanted direction from the main reflection line Ω_l, the "closeness" d can be calculated as.

$$\Theta = \theta_e - \theta_l, \Phi = \phi_e - \phi_l \tag{3}$$

$$x = cos(\Phi) \cdot sing(\Theta), y = sin(\Phi) \cdot sing(\Theta) \tag{4}$$

$$d = \sqrt{x^2 + y^2} \tag{5}$$

the closer Ω_e is to Ω_l, the smaller d will get $(0 \leq d \leq 1)$, and the more energy we expected to get reflected. in order to account for the two reflection type and the energy lost, we propose a symmetric function base of d. where the maximum of it energy will be around 0, and fade when getting close to 1. we also want the maximum energy to be less then or equal to 1, in order to account for the losses. for the sake of simplicity, we picked a base *gaussian* function like so:

$$\rho(d) = gaussian_{\mu=0,\eta=0.2}(d) * \frac{1}{3} + 0.2 \approx \frac{2}{3} \cdot e^{-12.5 \cdot d^2} + 0.2 \tag{6}$$

$$\rho(\Omega_i, \Omega_e; \beta_p) = \rho(d_{i,e;p}) \tag{7}$$

where Ω_i is the incoming PEF and Ω_e is the outgoing PEF directions, bases on the orientation of the panel β_p. as can be seen in the figure 2, the specular reflection reflect over 80% of it max reflection, and the diffuse reflection reflect only 20%.

in this paper, must of the PEF calculation will come from an incoming SW at location x_s, that travel to the panel at location x_p and get reflected in direction Ω_e. from the two point x_s, x_p, we can extract the incoming direction Ω_i of the PEF, and the distance the SW travel r_i. we can calculate the outgoing PEF like so.

$$PEF(x_s, \Omega_e; x_p, \beta_p) = \rho(\Omega_i, \Omega_e; \beta_p) \cdot g(r_i) \cdot \nu(x_s, x_p) \tag{8}$$

where $\rho(\Omega_i, \Omega_e; \beta_p)$ is the reflection equation (7), $g(r_i) = 1/4\pi r_i^2$ is the SW propagation, r_i is the distance between x_s, x_p, and $\nu(x, x_p)$ is a function that checks for obstacles between the two points ($\nu(x_s, x_p) = 1$ if there is none, else $\nu(x_s, x_p) = 0$). finally, by adding a time dimension to the PEF to (8), we get the reflection kernel [RK] from point x_s to panel p, in direction Ω_e.

$$R(x_s, \Omega_e, t; x_p, \beta_p) = PEF(x_s, \Omega_e; x_p, \beta_p) \cdot \delta(t - V_{sound} \cdot r_i) \tag{9}$$

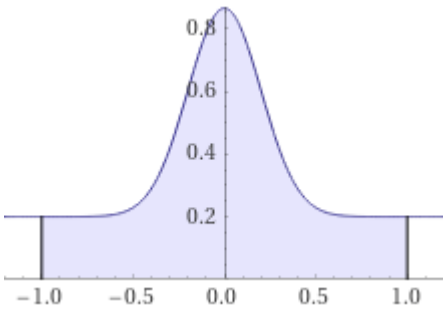

Figure 2: Reflection power function from the main line

and in discreet form.

$$R(x_s, \Omega_e, t; x_p, \beta_p) = \begin{cases} PEF(x_s, \Omega_e; x_p, \beta_p), & \text{if: } t = V_{sound} \cdot r_i \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $\delta(t)$ is the direct delta function, and V_{sound} speed of sound in the enclosure.

2.3 Room Acoustic Rendering

in order to simulate the RIR, we use the acoustic radiance transfer **technique** (ART) as seen in equation (10). let $L(x, \Omega, t)$ be the time-dependent outgoing PEF at point x along direction Ω. the general equation energy flux can be write as.

$$L(x, \Omega, t) = L_0(x, \Omega, t) + \int_G R(x', \Omega, t; x, \beta) L(x, \frac{x - x'}{|x - x'|}, t) dx' \tag{11}$$

Where L_0 is the first reflection between the (source, panels and sensor), G is the set of all surface points in the room, and $R(x', \Omega, t; x, \beta)$ is the RK of the surface. in order to solve the ART problem, the room surface G is divided into N small discrete panels. then the problem can be solve in a numerical way. Most of our calculation occurs between the panels them self. starting from a panels j the PEF travel toward panel i, and reflected in direction to panel k. we can write a RK matrix $F^i(t)$ for every panel i, that calculate the reflection kernel between every 2 at size (N matrix of NxN). to maintain consistency in the following equations, i will represent the panel we working on, j will represent the RK coming from panel j, and k will represent the PEF reflected toward panel k.

$$F^i[j, k](t) = R(x'_j, \Omega_k, t; x_i, \beta_i) \cdot A_i \tag{12}$$

$R(x'_j, \Omega_k, t; x_i, \beta_i)$ is the RK from (10), x_i, β_i is the location and orientation of panel i, the location the RK coming from is x'_j (from panel j to i), Ω_k is the direction for the out going PEF (from panel i to k), and A_i is the small area of the discrete panel i. To limit the size of the time domain, we want too use the discrete RK 10 with a fixed length, base on the sample rate of the system. we want the length to be $60ms$ to account for the first $50ms$ of the early reflection in the room, and the distance between the source and the sensor.

now we going to calculate the outgoing RK between all the panels in ascending order of reflections, if we know the $(n-1)^{th}$ RK in the room, we can calculate the n^{th} RK. let define $L_n(t)$ as the incoming RK matrix of the room, from every panel to every panel, in the n^{th} reflection. if we pick an outgoing RK matrix of panel i, we can use $L_{n-1}(t)$ and F^i to calculate all the incoming RK of $L_n(t)$ from panel i, like so.

$$L_n[i,k](t) = \sum_{a=1}^{N} conv(F^{[i]}[a,k](t), L_{n-1}[a,i](t)) = F^{[i]}[:,k](t) *' L_{n-1}[:,i](t) \tag{13}$$

when L_1 represent the first incoming RK matrix from the source, *conv* represent a linear convolution with $pad[2N-1,0]$ in order to keep the same size of the time domain and discard any future effect after $60[ms]$. this problem resemble a matrix multiplication on the first dimensions of the vector/matrix, but with *conv* instate of *mult* in the time domain. in this page we will define $(*')$ as a special multiplication base on vector and matrix multiplication, we apply the vector/matrix multiplication on the first dimensions, but with convolution on the time domain. with $(*')$ we can efficiently calculate the a full column in L_n. in order to calculate column i, we need only F^i and $L_{n-1}[:,i]$, and it can be calculated using multiplication matrices we had defined.

$$L_n[i,:](t) = F^i *' L_{n-1}[:,i]^T(t) \tag{14}$$

when $(^T)$ is transposed function on the first dimensions, not including the time depth. because each column i can be calculated individually (with it own F^i and it own row from $L_{n-1}[:,i]$), we can combine the list of F^i into a single matrix F, flatten L_n and salve the problem in a matrix way.

$$L_n = F *' L_{n-1} = F^2 *' L_{n-2} = ... = F^n *' L_1 \tag{15}$$

Although we don't use this form to calculate L_n, we will use (15) to symbolize the full calculation of the method. the true calculation we be individually like (14). Finally in-order to calculate the main RK matrix of the room, we just need to sum all thenth different reflection energy flux matrices, for every n^{th}.

$$L = \sum_{n=1}^{\infty} L_n = \left(I + \sum_{n=1}^{\infty} F^n \right) *' L_1 = Q *' L_1 \tag{16}$$

L represent the full RK matrix from every panel to every panel in the room, $Q = I + \sum_{n=1}^{\infty} F^n$, L_1 is the first incoming RK matrix from the source. as we can see, we can find Q independently from L_1.

After we found a way to calculation a the reflection between the panels in the room (17), we can calculate the true RIR. First we need to pick a locations for the source S, and the sampling point M (P will stand for the panels). the RIR can be divided into 3 types of SW:

- D - direct movement from $(S \to M)$
- L_0- RK that hit P once, and continue to the M $(S \to P \to M)$
- E - RK that come from echo between the panels P $(S \to L_1 \to Q \to M)$

We can calculate all of them like so:

$$D\left[t=\frac{r_{_{S,M}}}{V_{sound}}\right]=g(r_{_{S,M}})\cdot\nu(x_S,x_D) \tag{17}$$

$$L_0=\sum_{i=1}^{N}g(r_{x_i,M})\cdot\nu(x_i,x_M)\cdot R(x_i,S,\Omega_M,t) \tag{18}$$

When $r_{_{S,M}}$ is the distance between S to M, V_{sound} is the speed of sound, and $g(r)\cdot\nu(x_S,x_D)$ are the same from (8). to calculate the echo reflections between the panels, we first need to create matrix L_1.

$$L_1[i,k]=R(x_i,S,\Omega_{x_k},t) \tag{19}$$

$$E=\sum_{i=1}^{N}\sum_{k=1}^{N}(Q*'L_1)[i,k]*R(x_k,x_i,\Omega_M,t) \tag{20}$$

And by summing the different RK, end up with:

$$RIR=D+L_0+E \tag{21}$$

3 Proposed Method

To create our dataset $y(t)$ of speech signals with simulated early reflections, we will use the generated RIR, $h(t)$ from (22), with a clean audio speech $x(t)$. like so:

$$y(t)=x(t)*h(t) \tag{22}$$

where (*) is the same *conv* , that discard any future changes after the first $60[ms]$. Must of our calculation occurs will simulating the RIR, and specific when we simulate the reflection matrix Q in (17). note that after calculating Q, we can simulate multiples RIR by picking a source location for an audio signal, and a location for the sensors. thus for every Q we can simulate multiple different RIR in this simulated room, within a linear time complexity.

The simulated room was chosen to be a small square shape room with a floating panel in the middle of it, in order to simulate an existing table in the room. which will most likely be located in a recording room (figure 3). and so we simulated the RIR of multiple sensors, in a sensor array $\bar{y}(t)$, from a single sorce location in the room. for the system, we use the proposed module (figure 4) from [7]. this system comprises of two main modules. a pretrained speech synthesis generator G [3], that was train to remove noises. with a trainable U-Net structure F [11] that come before G. In this implementation, model F input convolution is expanded to receive multiple input from a sensor array $\bar{y}(t)$. First we take the STFT power spectrum of a an input array $\bar{y}(t)$, $|\bar{Y}(t,f)|$. then fed it into F, to receive a single enhanced signal spectrum $|\hat{X}(t,f)|$. the enhanced spectrum is transform to the Mel-spectrum domain $|\hat{X}_M(t,c)|$, then fed to the generator G, to get the generated output signal $\hat{x}(t)$. The U-NET [11] architecture of module F is build with a input size of the $\bar{y}(t)$ STFT array and a output of a single STFT channel. the structure is compose of 6 convolution layers, with kernel-size of $5x5$, fallowed by batch-norm and leaky-ReLU. fallow by 6 transpose-convolution at the same size, with batch-Norm and ReLU and

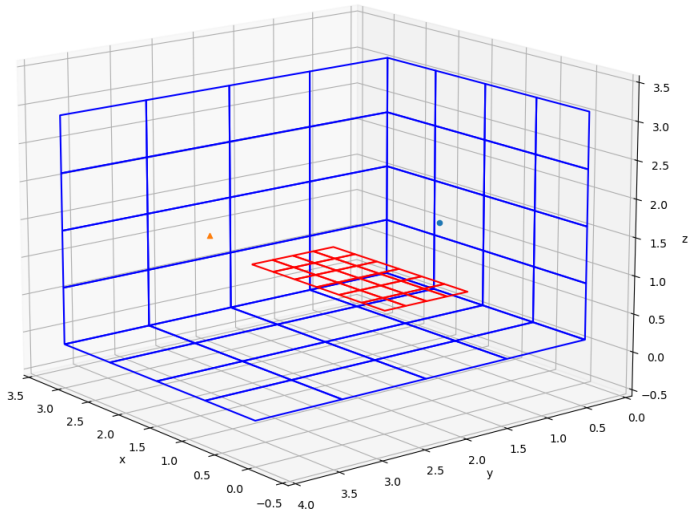

Figure 3: simulated room

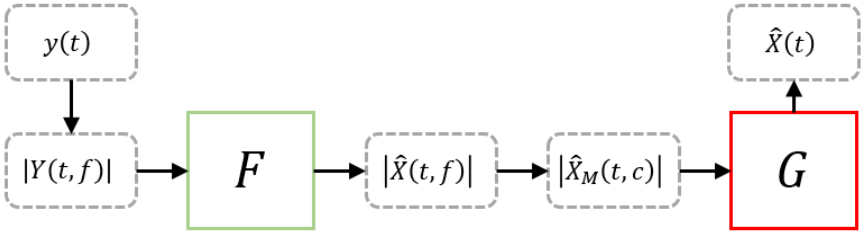

Figure 4: Proposed system. module F has trainable weights while the weights of G are fixed.

dropout of 0.5. unlike the original U-NET [11], the structure does not contain max-pooling between the layers.

for the training pass, we train only the weights of F, will G weights were fix. for the loss function we use:

$$L(x, \bar{y}) = L_s(|X|, F(|\bar{Y}|)) + L_s(|X|, |\texttt{STFT}(\hat{x})|) \tag{23}$$

note that $F(|\bar{Y}|)$ output a single STFT channel. L_s defined as:

$$L_s(|X(t,f)|, |Y(t,f)|) = \sum_{t,f} ||X(t,f)| - |Y(t,f)|| + \lambda_s \sum_{t,f} \left| log \left| \frac{X(t,f)}{Y(t,f)} \right| \right| \tag{24}$$

4 Experimental Results

4.1 Data and Implementation Details

first we simulated a dataset of RIR (21). we created a square shape room, with a uniform variance at size of $(2-4, 3-5, 2-4)$ meters, and we adding to it a floating square table at size of$(1.5-2.5, 0.5-1.5)$ meters. the middle location of the table locate between $(0.5-2.5, 1-3, 0.5-1.5)$ (figure 3). every wall in the room is spilt into $(4, 4)$ grid of panels, and the table into $(6, 4)$, for a total of 120 panels. for every simulated room, we pick 20 location for the source sound, and 50 location for a sensor array. for a total of 1000 simulated RIR for every room. The database we created was build with a sensor array of 2 sensors, with a fix distance of 0.3 meters between them, and with a changing height and distance from the source signal. we simulate $12k$ pars of RIR.

To train the module, we use the LJSpeech dataset that sampled at $f_s = 22.05[\texttt{kHz}]$. the STFT magnitude are at 512 frequency band. we chose $\lambda_s = 1$ for $[L_s]$ and AdamW as our optimizer. the system was train for 100 epochs with a batch size of 2. We train 2 models of F overall. one was train on the dataset with two inputs channels (F_2), and the other only on a single channel input from the dataset (F_1). both models used a RIR that was rolled into the beginning of the signal, and was normalise to maintain $RIR[0] = 1$.

Note that the model was also train on 2 sensors input with random orientation between the sensors, which caused a time shift between the sensors. it seem that the propose model have hard time to work with it, and the model preform worst then the signal sensor input model.

4.2 Performance Evaluation

In order to evaluate the system, we going to compere the propose model against the HIFI-gen model [3], the U-NET-HIFI-gen model from [7], with the original speech signal, plus it degraded one. We will use the propose model in [9] to score them, which ranks given the signal on a scale from 1 to 5. where 5 is the highest score and 1 is the lower. this model was train to mimic how human score a sound signal.

1. *original* - the original clean signal $x(t)$.
2. *noisy* - the degraded signal $y(t)$.
3. G - the reconstructed signal using only the HIGI-gen signal, $G(|Y_M|)$.
4. $F_{old} + G$ - the proposed model in [7], that was train with a randomly drawn RIR.
5. $F_1 + G$ - our proposed model [chapter 3], on a single input channel.
6. $F_2 + G$ - our proposed model [chapter 3], with two input channels.

First of all, we can see that our model (the both green plot) did better then the old algorithm on the score graph, when the double channels model did better on average from the single channel. we can conclude that the model was able to extract more data from the two channel input, and better improve the signal. We can also see that both the original and the degraded signal have on average

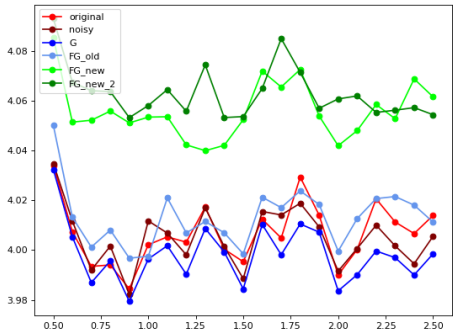

Figure 5: plot of the scores vs distance from the scores

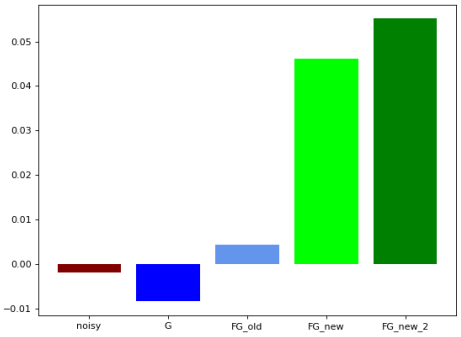

Figure 6: bar of the average scores from the average original signal score

a simulate score. that not surprising because as we say before, early reflection don't really effect the score of how humans hear the sound. but on the other hand, the HIFI-gen model [3] have hard time to improve the signal from it degraded form, and did worst then it. while the old U-NET-HIFI-gen model [7] improve only slightly over the original signal.

5 Experimental Results

We have presented a method to simulate a true RIR and use it to train a method attenuating acoustic early reflections in record studios. the experimental results shown that using a true RIR simulation give better result over using a simulate method that was train on random drawn RIR, and have the potential of using multiple sensors and take advantage of them to better enhance the signal. For future work, we looking to find a model that can better extract the data from a two channel input, even when the sensors are not keep in parallel to the source.

References

[1] Hequn Bai, Gael Richard, and Laurent Daudet. Modeling early reflections of room impulse responses using a radiance transfer method. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, 2013.

[2] John S Bradley, Hiroshi Sato, and Michel Picard. On the importance of early reflections for speech in rooms. *The Journal of the Acoustical Society of America*, 113(6):3233–3244, 2003.

[3] Lauri Juvela, Bajibabu Bollepalli, Junichi Yamagishi, and Paavo Alku. Gelp: Gan-excited linear prediction for speech synthesis from mel-spectrogram. *arXiv preprint arXiv:1904.03976*, 2019.

[4] Shinji Kishinaga, Yasushi Shimizu, Shigeo Ando, and Kiminori Yamaguchi. On the room acoustic design of listening rooms. In *Audio Engineering Society Convention 64*. Audio Engineering Society, 1979.

[5] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[6] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

[7] Tomer Rosenbaum, Israel Cohen, and Emil Winebrand. Attenuation of acoustic early reflections in television studios using pretrained speech synthesis neural network. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7422–7426. IEEE, 2022.

[8] Samuel Siltanen, Tapio Lokki, Sami Kiminki, and Lauri Savioja. The room acoustic rendering equation. *The Journal of the Acoustical Society of America*, 122(3):1624–1635, 2007.

[9] Emil Winebrandb Ofri Gabsob Tomer Rosenbauma, Israel Cohena. Differentiable mean opinion score estimation and application in speech enhancement. 2022.

[10] Jinhyeok Yang, Junmo Lee, Youngik Kim, Hoonyoung Cho, and Injung Kim. Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network. *arXiv preprint arXiv:2007.15256*, 2020.

[11] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.