DIGITAL CONVOLUTION AUDIO FILTER BY FREQUENCY BIN-SPECIFIC DECAYED DISCRETE IMPULSE RESPONSE *

Stefan Ciba
https://www.researchgate.net/profile/Stefan-Ciba

ABSTRACT

The developed method focuses on the calculation and modification of a discrete *impulse response* in order to filter the characteristics from a known digital single channel recording setup and room characteristics such as early reflections and reverberations. The aim is a dryer and clearer signal reconstruction, which ideally would be the direct-path signal. The time domain *impulse response* is calculated from the cepstral domain and modified by means of frequency bin specific exponential decay in the spectrum. The decay rates are obtained by using the blind estimates of reverberation time ratio between recorded output and test signals for each frequency bin. The modified *impulse response* does filter a recorded audio-signal by deconvolution. The blind estimation stands out for its robustness to noise and non-idealities. This simple, yet powerful method, could potentially offer advantages in scenarios, where adjustment for early reflections and frequency bin specific reverberation characteristics is crucial, such as in audio production, acoustic simulation, or virtual reality applications, where acoustic characteristics don't change significantly during the record. This filter accounts for not optimal recording conditions and can be applied as one of the first steps in post-processing of a single audio channel. Some real-time applications can benefit from this method, but some might need to have real-time or spatial adaption, when the *impulse response* changes during the record and manual recalibration process would not be an option. The proposed method shows less early reflections and less reverberation compared to deconvolution with non modified *impulse response*. Estimation of a direct path signal is key to many applications.

***K*eywords** Spectral impulse response modification · digital early reflections and dereverberation convolution filter

1 Introduction

The field of digital audio signal processing has undergone significant development since the advent of digital techniques in the 1960s and 1970s, which enabled to manipulate sound with unprecedented precision and flexibility [7, 8].

Modeling and modification of the acoustic characteristics of room and recording setups, based on convolution with *impulse response*, is a key technique in digital audio processing. The Linear Time Invariant (LTI) system approach consisting of convolution of a audio signal with a measured or synthesized *impulse response*, makes it possible to e.g. simulate reverberation, filter out early reflections, and compensate for coloration introduced by the recording chain [8, 2].

Traditional methods for reverberation and dereverberation often rely on parametric models or sparse representations of the *impulse response* [2, 3]. However, convolution with a sparse, spectrally modified *impulse response*, that explicitly incorporates the $T60$ decay through exponential functions offers a direct and flexible approach for single-channel audio filtering. This method enables targeted suppression of early reflections, reverberation, and undesired recording characteristics, based on established signal processing principles. This novel approach differs from closely related state of the art work such as "Speech Dereverberation in the Time-Frequency Domain" [4], "Approximation of Real Impulse Response Using IIR Structures" [5] and "Representation and Identification of Systems in the Discrete-Time Wavelet Transform Domain" [6] and was not yet found in literature.

* *Citation*: **Authors. Title. Pages.... DOI:000000/11111.**

2 Signal processing

2.1 Periodic sine sweep for system identification

To characterize a system, a periodic linear sine sweep (chirp) can be used as the test signal [1]. The frequency of the positive chirp increases linearly from f_0 to f_1 within a duration T:

$$\acute{x}(t) = \sin\left(2\pi\left[f_0 t + \frac{f_1 - f_0}{2T}t^2\right]\right), \quad t \in [0, T]. \tag{1}$$

The signals are being sampled with the sample rate $f_s = 48$ kHz and thus contain $N = Tf_s$ samples, counting $n = 0, 1, 2, ...N-1$. By indexing with $t_n = \frac{n \bmod N}{f_s}$ the expression for the positive chirp becomes discrete and periodic:

$$\acute{x}_n = \sin\left(2\pi\left(f_0 t_n + \frac{f_1 - f_0}{2T}t_n^2\right)\right). \tag{2}$$

The chirp vector can be further extended for P periods by array tiling:

$$x_n = \bigoplus_{p=0}^{P-1} \acute{x}_{n-pN}. \tag{3}$$

2.2 impulse response estimation

Given discrete time series x_n and its recorded signal y_n with N samples, their relation can be described by a discrete LTI System with $y_n = x_n * h_n$. Therein the *impulse response* h_n describes the signal chain, that can include unpleasant room characteristics that cause e.g. early reflections, reverberation, resonance and add non-linearity of the sensor and the loudspeaker.
The *impulse response* can be decorrelated by deconvolution from the related signals for system identification:

$$h_n = y_n * x_n^{-1}. \tag{4}$$

The deconvolution method of choice is the frame-wise subtraction of the test signal and the recorded signal in cepstral domain, followed by back transformation into time domain. The *impulse responses* are subsequently combined and normalized to get one time domain representation.
The discrete time signals are split into N_{frames} frames with a frame length of $N_{\text{DFT}} = 5 \cdot f_s$, so the DFT size was chosen in relation to the sample rate f_s, for the *impulse response* to have a time duration that respects usual reverberation and delay time. For each frame index $\eta = 0, 1, 2, ..., N_{\text{frames}} - 1$ the frame boundaries are shifted over the signal with hop length $N_{hop} = \frac{N_{\text{DFT}}}{2}$, which makes up a overlap of $o = 50\%$. The frequency bins are counted by index $\mu = 0, 1, 2, ..., N_{\text{DFT}-1}$. In the same manner, the discrete time step is respected by the index ν, to count the samples in a frame. Subbands can be expressed as angular frequency $\Omega = \frac{2\pi}{N_{\text{DFT}}}$ times index μ. DFT spectra are processed frame by frame with rectangular window $\omega_\nu^{(\text{rect})}$.
The complex windowed short-time spectrum is:

$$\underline{Y}_{\mu,\eta} = \sum_{\nu=0}^{N_{\text{DFT}}-1} y_{\eta N_{hop}-\nu} \cdot \omega_\nu^{(\text{rect})} e^{-j\Omega\nu}. \tag{5}$$

The original signal short time spectrum is $\underline{X}_{\mu,\eta}$. The recorded short time spectrum is $\underline{Y}_{\mu,\eta}$. They are regularized by $|\underline{X}_{\mu,\eta}|_\varepsilon = \max(|\underline{X}_{\mu,\eta}|, \varepsilon)$ for numerical stability.

The real Cepstrum is obtained by transformation as follows:

$$C_{\mu,\eta}^{(x)} = \Re\left\{\frac{1}{N_{\text{DFT}}} \sum_{\mu=0}^{N_{\text{DFT}}-1} \ln|\underline{X}_{\mu,\eta}|_\varepsilon e^{j\Omega\mu}\right\}. \tag{6}$$

In cepstral domain the deconvolutions are accomplished by subtractions, in order to obtain the *impulse responses*:

$$C_{\mu,\eta}^{(h)} = C_{\mu,\eta}^{(y)} - C_{\mu,\eta}^{(x)}. \tag{7}$$

The *impulse responses* transform back into the time domain vise versa. Although the imaginary part should be zero, only the real part is used, to prevent numerical errors:

$$h_{\nu,\eta} = \Re\left\{\frac{1}{N_{\text{DFT}}} \sum_{\mu=0}^{N_{\text{DFT}}-1} \overbrace{e^{\left(\sum_{\mu=0}^{N_{\text{DFT}}-1} C_{\mu,\eta}^{(h)} e^{-j\Omega\mu}\right)}}^{H_{\mu,\eta}} e^{j\Omega\mu}\right\}. \tag{8}$$

The *impulse responses* accumulate by averaging:

$$\overline{h}_\nu = \frac{1}{N_{\text{frames}}} \sum_{\eta=0}^{N_{\text{frames}}-1} h_{\nu,\eta}. \tag{9}$$

To ensure the *impulse response* has a consistent scale, with maximum absolute value 1, it is normalized:

$$h_\nu = \frac{\overline{h}_\nu}{\max_\nu |\overline{h}_\nu|}. \tag{10}$$

This yields a single, normalized *impulse response*, that is suitable as a good overall representation.

2.3 Spectral *impulse response* modification

The time domain *impulse response*, e.g. from equation (10), is transformed by STFT, to add decay to certain frequency bins.

2.3.1 Blind estimation of $T60$ reverberation times for each bin

The ISO 3382 standard [13] formally defines $T60$ as the time required for spatial sound energy to decay by 60 dB. This work uniquely uses blind estimate of $T60$, despite the *impulse response* is known, to shape the decay of each frequency bin in the *impulse response*, targeting specific acoustic and system artifacts for convolution filter, which is an approach not yet found in literature. Furthermore the $T60$ reverberation time is a quantity that is traditionally used for measurement, characterizing and shaping room acoustics usually as a global or band-averaged parameter [**?**]. Ratnam et al. provide the maximum-likelihood framework for single-channel blind $T60$ estimation [9]. To do the frequency bin specific modifications on the *impulse response*, the necessary array of decay parameters is obtained by the following procedure.
The Power Spectral Density (PSD) for a discrete signal x_n is its STFT magnitude squared:

$$S_{\mu,\eta} = |\underline{X}_{\mu,\eta}|^2. \tag{11}$$

The energy for each frequency bin is:

$$E_{\mu,\eta} = \frac{\sum_{\eta'=\eta}^{N_{\text{frames}}-1} S_{\mu,\eta'}}{\sum_{\mu'=0}^{N_{\text{DFT}}-1} S_{\mu,\eta'}}. \tag{12}$$

The $T60$ time for each frequency bin is defined as the time it takes for the cumulative energy to fall below a threshold (e.g., 0.001, corresponding to -60 dB of magnitude):

$$\eta_\mu^{(\text{T60})} = \min\left\{\eta : E_{\mu,\eta} < 0.001\right\}. \tag{13}$$

The overlap factor is:

$$o = 1 - \frac{N_{hop}}{N_{\text{DFT}}}, \tag{14}$$

the $T60$ estimate in seconds for every frequency bin is:

$$T60_\mu = \frac{o \cdot \eta_\mu^{(\text{T60})}}{f_s}. \tag{15}$$

2.3.2 Application of the decay function

The ratio ρ between the $T60$'s of the original test signal x_n and its recorded signal y_n, weights the exponential decay for every frequency bin: $\rho_\mu^{(\text{T60's})} = \frac{T60_\mu(y_n)}{T60_\mu(x_n)}$.

With the frame wise time duration $\tau = \eta \cdot \frac{N_{hop}}{f_s}$, the exponential decay function is:

$$D_{\mu,\eta} = e^{\left(-\frac{\tau}{\rho_\mu^{(\text{T60's})}}\right)}. \tag{16}$$

The frame wise decay modified *impulse response* is:

$$h'_{\nu,\eta} = \frac{1}{N_{\text{DFT}}} \sum_{\mu=0}^{N_{\text{DFT}}-1} \underline{H}_{\mu,\eta} \cdot D_{\mu,\eta}\, e^{-j\Omega\mu}. \tag{17}$$

The synthesis window (Hann) is applied by $\omega_{\nu}^{(\text{hann})}$ [8] and furthermore the overlap has to be handled, done by *Constant Overlap Add Method* (COLA):

$$h'_n = \sum_{\substack{\eta \\ 0 \le n-\eta N_{hop} < N_{\text{DFT}}}} \omega_{n-\eta N_{hop}}^{(\text{hann})} \cdot h'_{n-\eta N_{hop},\eta}. \tag{18}$$

2.4 Filterbank application

Finally, additional overall exponential decay dk can be applied to the modified *impulse response* $h''_n = h'_n \cdot e^{-dk \cdot n}$ to avoid echoes. The *impulse response* is applied as spectral filter bank $\underline{H}''_{\mu,\eta}$ for any recorded signal spectrum $\underline{Z}_{\mu,\eta}$ by deconvolution:

$$\underline{Z}''_{\mu,\eta} = \frac{\underline{Z}_{\mu,\eta}}{\underline{H}''_{\mu,\eta}}. \tag{19}$$

Hann-windowing applies for STFT and iSTFT. The COLA method applies for iSTFT, respectively. The frame size was set to number of samples of the *impulse response* and gain was normalized by strictly preventing clipping if necessary to get the desired filtered signal $\hat{z}''_n = z''_n / max(|z'_n|)$.

3 Test setup

A sine-sweep and a speech audio was recorded by integrated microphone from a *Convertible Workstation* (HP ZBook Studio x360 G5 (6TW61EAABD)), on a desk, located in a pitched roof area corner of a livingroom, that has about 1 Ar. The sine-sweep sound was played back from the internal loudspeakers of the *Convertible Workstation*. The integrated noise reduction was turned off. The goal was to improve the speech sound quality with the aim of having dry and clear sound when speaker sits in front of the *Convertible Workstation*. A Futher assumption was, that degraded speech intelligibility could get better.

In the same manner the sine-sweep and a drum-set groove have been recorded from the same location. The sine-sweep was played back on a Alesis Active M1 MKII (8N) near-field monitor beside the drum-set. The goal was to have dry and clear studio-like raw recording quality despite poor recording-conditions.

3.1 Objective

Several objective measures prove the applicability of the Method. They can be found in the Results section table 1. The $T60$ of the *impulse response*-filtered signal and the $T60$ of the filtered signal by modified *impulse response* can be compared.

The logarithmic *Signal to Noise Ratio* (SNR) [15] was used to indicate the efficiency of the filter by taking the original signal z_n and the filtered audio signal z''_n into account. The averages of signal and noise power are calculated by $S = \overline{z''^2_n}$ and $N = \overline{(z_n - z''_n)^2}$. The logarithmic fraction of them is:

$$SNR = 10 log_{10} \left(\frac{S}{N} \right). \tag{20}$$

The STOI measures the speech intelligibility. It is only applicable to speech audio.

The *Perceptual Evaluation of Speech Quality* (PESQ) is an algorithm that models human auditory perception, capturing effects like distortion, background noise, clipping, and delay. It measures clarity of an audio channel, which is also necessary for non-speech audio channels. Higher PESQ scores indicate better quality of the audio channel. It is widely used in telecom, VoIP, and audio processing to measure speech enhancement or degradation cite.... The PESQs of the original signal and the filtered signal are compared due to their difference. Therein the PESQ from a

degraded signal was estimated from its filtered signal as reference and vise versa.

Zwicker's time-varying *loudness* model estimates how intense or loud a sound is perceived by human listeners cite....

Sharpness from DIN 45692 standard (Zwicker's model) quantifies how piercing or shrill sound feels, based on how present the frequencies above 3kHz are cite....

Roughness measures how harsh or "fluttery" a sound feels, linked to rapid, audible amplitude fluctuations in the range most noticeable to human ears (20–300Hz modulation).

The Algorithms to measure perceptive sound quality can be used from the *Modular Sound Quality Index Toolbox* (MoSQITo) [16], PESQ [17] and the STOI [18], respectively. Further measures have been evaluated by proprietary software, namely the Cubase 5.0.1 build 147 (Steinberg Media Technologies GmbH) [19].

3.2 Subjective

Furthermore subjective hearing tests where performed. Some listener notice the difference of acoustic signal with less reverb, at least the experts do.

4 Results

Table 1: Objective measures. Statistics with asterisk (*name**) derived from Cubase 5.0.1 build 147.

	Drum-Set degraded (z_n)	Drum-Set filtered (z_n'')	Speech degraded (z_n)	Speech filtered (z_n'')
min sample value*	-1	-0.961	-0.045	-0.052
max sample value*	+1	+1	0.046	0.045
peak amplitude*	0dB	0dB	-26.75dB	-25.73dB
DC offset*	-87.70dB	-91.67dB	-85.31dB	-88.32dB
Estimated Pitch*	2207.5Hz / C#6	5057.2Hz / D#7	1452.2Hz / F#5	1834.9Hz / A#5
Min RMS Power*	-67.68dB	-96.45dB	-61.02dB	-63.10dB
Max RMS Power*	-4.58dB	-11.67dB	-32.82dB	-32.08dB
Average*	-19.92dB	-24.63dB	-41.34dB	-41.29dB
SNR	-9.648dB		-9.507dB	
T60	31.537ms	26.679ms	23.793ms	20.655dB
STOI	N/A	N/A	0.9807	0.9807
Δ STOI	N/A		0	
PESQ	3.6052	4.1774	4.4319	4.4410
Δ PESQ	0.5722		9.1E-3	
Loudness	0.90320	0.60329	0.14764	0.14393
Roughness	0.23022	0.22802	0.21995	0.21946
Sharpness	3.26095	3.50446	3.09846	3.16039

The results are shown in figure 2. ....

5 Conclusion

Future work will focus on a cepstral lifter approach during the *impulse response* estimation. Instead of the T60 coefficient estimation a artificial neural network might learn to optimize features from the impulse response.

6 Supplements

The Python-code of the studied method and audio samples are to be found at: https://git....

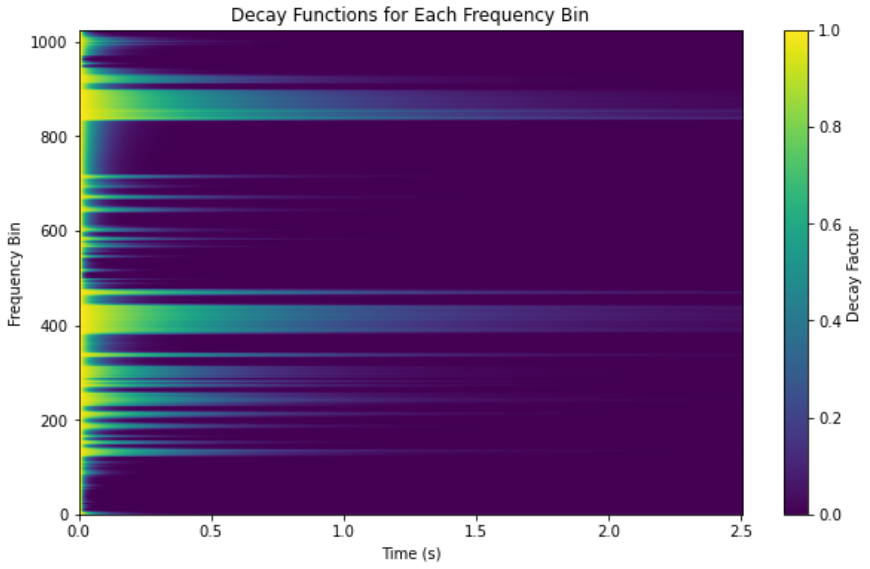


Figure 1: The $T60$ decay matrix.

Acknowledgments

The author(s) acknowledge the use of the AI language model Perplexity, developed by Perplexity AI, for assistance in literature and code review in this work.

References

[1] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," *AES Convention 108*, 2000.
https://www.aes.org/e-lib/browse.cfm?elib=10211

[2] J. A. Moorer, "About this reverberation business," *Computer Music Journal*, vol. 3, no. 2, pp. 13–28, 1979. doi:10.2307/3680287.

[3] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, The Netherlands, 2007. [Online]. Available: https://research.tue.nl/files/1972985/200710970.pdf

[4] A. Oyzerman, I. Cohen, "Speech Dereverberation in the Time-Frequency Domain," M.Sc. Thesis, Technion, 2012.
https://israelcohen.com/wp-content/uploads/2018/05/AnnaOyzerman_MSc_2012.pdf

[5] A. Primavera et al., "Approximation of Real impulse response Using IIR Structures," Proc. EUSIPCO, 2011.
https://www.eurasip.org/Proceedings/Eusipco/Eusipco2011/papers/1569422051.pdf

[6] I. Cohen, "Representation and Identification of Systems in the Discrete-Time Wavelet Transform Domain," IEEE Trans. Signal Processing, 2007. https://israelcohen.com/wp-content/uploads/2018/05/ASM2007.pdf

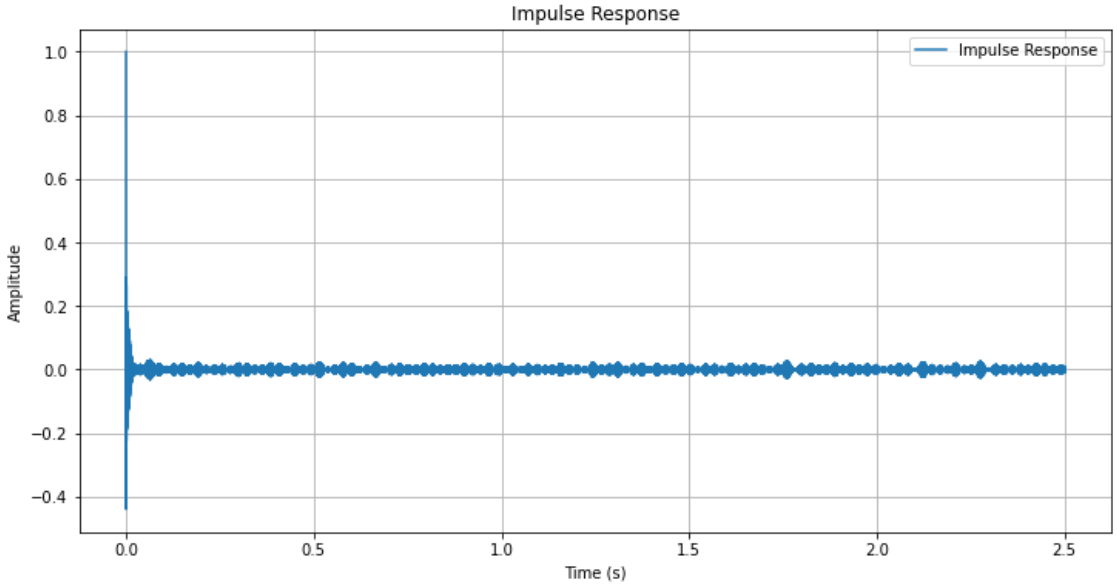
[7] A. V. Oppenheim, R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 1989.
ISBN: 0-13-216292-X

[8] Julius O. Smith III, *Spectral Audio Signal Processing*, W3K Publishing, 2011.
Available online: https://ccrma.stanford.edu/~jos/sasp/

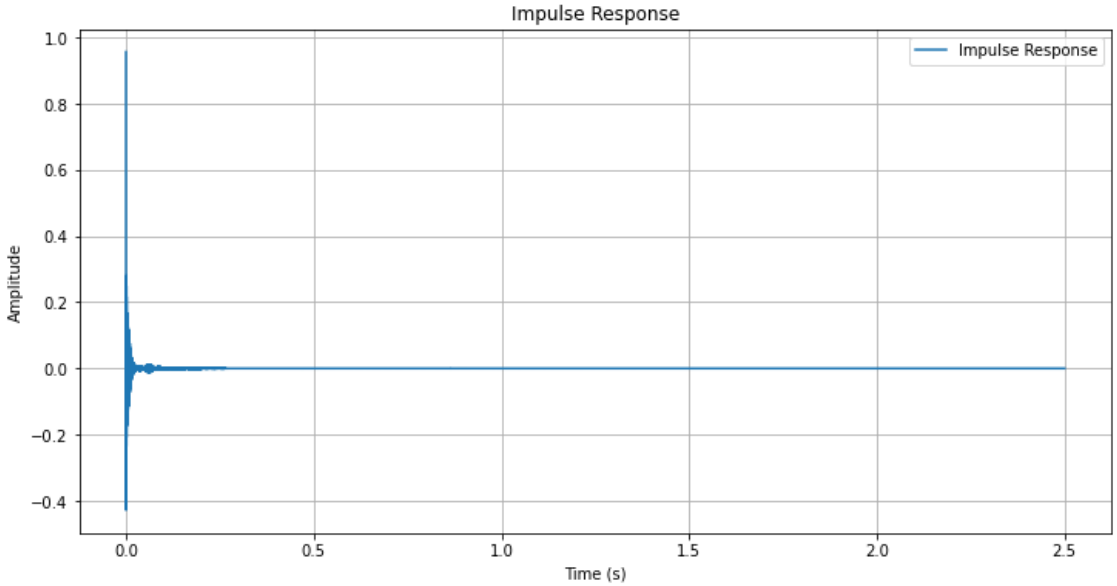
[9] R. Ratnam et al., *Blind Estimation of Reverberation Time*, J. Acoust. Soc. Am., 114(5), pp. 2877-2892, 2003.
https://www.ee.columbia.edu/~dpwe/papers/Ratnam03-reverb.pdf

[10] C. Doire et al., *Single-Channel Blind Estimation of Reverberation Parameters*, Proc. EUSIPCO, 2015.
https://www.commsp.ee.ic.ac.uk/~sap/wp-content/uploads/2015/07/Doire2015.pdf

[11] Y. Liu et al., *A Composite T60 Regression and Classification Approach for Speech Dereverberation*, arXiv:2302.04932, 2023.
https://arxiv.org/abs/2302.04932



(a) Original.



(b) Modified.

Figure 2: *Impulse response*.

[12] H. Kuttruff, *Room Acoustics* (6th ed.), CRC Press, 2016. (Open access chapters available) ISBN: 978-1498740436

[13] ISO 3382-1:2009, *Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces*. `https://www.iso.org/standard/40979.html` (Standard available for purchase, but methods widely referenced in open literature)

[14] Green Forge Coop. MOSQITO (Version 1.1.1). `https://doi.org/10.5281/zenodo.10629475`

[15] N. Fliege, *Systemtheorie*. B. G. Teubner Stuttgart 1991 (Informationstechnik). ISBN: 3-519-06140-6

[16] Eomys Engineering, Modular Sound Quality Index Toolbox (MoSQITo), 2025. Available at: `https://github.com/Eomys/MoSQITo`

[17] C. Ludlow, python-pesq: Perceptual Evaluation of Speech Quality (PESQ), 2025.
Available at: https://github.com/ludlows/python-pesq.git

[18] M. Pariente, pystoi: Short-Time Objective Intelligibility (STOI), 2025.
Available at: https://github.com/mpariente/pystoi

[19] Steinberg Media Technologies GmbH, Cubase 5.0.1 build 147 [Computer Software], 2009.
Available at: https://o.steinberg.net/index.php?id=1782&L=1