
DIGITAL AUDIO FILTER BY DECONVOLUTION WITH FREQUENCY BIN SPECIFIC DECAYED 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.

Keywords Spectral impulse response modification · digital dereverberation 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 [1, 2].

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 [2, 3].

Traditional methods for reverberation and dereverberation often rely on parametric models or sparse representations of the *impulse response* [3, 4]. However, convolution with a sparse, spectrally modified *impulse response*, that explicitly incorporates the T_{60} 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 approach differs from closely related state of the art work such as "Speech Dereverberation in the Time-Frequency Domain" [5], "Approximation of Real Impulse Response Using IIR Structures" [6] and "Representation and Identification of Systems in the Discrete-Time Wavelet Transform Domain" [7].

**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 [8]. The frequency of the positive chirp increases linearly from f_0 to f_1 within a duration T :

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

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

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

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

$$x_n = \bigoplus_{p=0}^{P-1} \dot{x}_{n-pN}. \quad (3)$$

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 $y_n = x_n * h_n$ [9]. 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:

$$h_n = y_n * x_n^{-1}. \quad (4)$$

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 $\eta = 0, 1, 2, \dots, N_{\text{frames}} - 1$ the frame boundaries are shifted over the signal with hop length $N_{\text{hop}} = \frac{N_{\text{DFT}}}{2}$, which makes up a overlap of $o = 50\%$. The frequency bins are counted by index $\mu = 0, 1, 2, \dots, N_{\text{DFT}} - 1$. 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:

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

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:

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

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

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

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

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

The *impulse responses* accumulate by averaging:

$$\bar{h}_{\nu} = \frac{1}{N_{\text{frames}}} \sum_{\eta=0}^{N_{\text{frames}}-1} h_{\nu,\eta}. \quad (9)$$

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

$$h_{\nu} = \frac{\bar{h}_{\nu}}{\max_{\nu} |\bar{h}_{\nu}|}. \quad (10)$$

These steps yield 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 decay certain frequency bins from blind $T60$ estimate.

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

The ISO 3382 standard [10] 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 [11]. Ratnam et al. propose an *Maximum Likelihood* (ML) algorithm for a blind single-channel $T60$ estimate[12]. 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:

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

The energy for each frequency bin is:

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

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):

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

The overlap factor is:

$$o = 1 - \frac{N_{hop}}{N_{\text{DFT}}}, \quad (14)$$

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

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

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: $\rho_\mu^{(T60's)} = \frac{T60_\mu(y_n)}{T60_\mu(x_n)}$.

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

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

The frame wise decay modified *impulse response* is:

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

The synthesis window (Hann) is applied by $\omega_\nu^{(hann)}$ [2] and furthermore the overlap has to be handled, done by *Constant Overlap Add Method* (COLA):

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

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:

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

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 living room, 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 Further 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.

4 Objectives

Several objective measures prove the applicability of the Method. Some Measures are used to compare the system impulse response with the filtered impulse response, and some are used to compare a filtered with a unfiltered signal.

The measures to compare the unfiltered and filtered signals can be found in the Results section table 1.

A *Logarithmic Signal-Power Distance* (LSD) was calculated to indicate the efficiency of the filter by taking the original signal z_n and the filtered audio signal z''_n into account:

$$LSD = 10 \log_{10} \left(\frac{(\overline{z''_n - z''_n})^2}{(\overline{z_n - z_n})^2} \right). \quad (20)$$

The *Short-Time Objective Intelligibility Measure* (STOI) measures the speech intelligibility [13]. 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 *Mean Opinion Score*, *Listening Quality Objective* (MOS-LQO), which is basically the clarity of an audio channel, which is also necessary for non-speech audio channels. Higher MOS-LQO indicate better quality of the audio channel. It is widely used in telecom, VoIP, and audio processing to measure speech enhancement or degradation [14].

Zwicker's time-varying *Loudness* model estimates how intense or loud a sound is perceived by human listeners [15].

Sharpness from DIN 45692 standard (Zwicker's model) quantifies how piercing or shrill sound feels, based on how present the frequencies above 3kHz are [16].

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) [17].

The *Definition Ratio*, D50, is the ratio of early to total energy within 50 ms [18]. It was used to indicate if early reverberations are present in the impulse response.

The Algorithms to measure perceptive sound quality can be used from the *Modular Sound Quality Index Toolbox* (MoSQITo) [19] [20], PESQ [21] and the STOI [22], respectively. Further measures have been evaluated by proprietary software, namely the Cubase 5.0.1 build 147 (Steinberg Media Technologies GmbH) [23].

5 Results

The filter has several well defined effects, which appear in the objective results table 1. Because reverberations are filtered from the signal, the signal amplitude and power decreases. The LSD measure shows that effect in one figure and gives it a name. The most significant estimated pitch of the audio increases, which is due to filtered room characteristic that allowed deeper frequencies to decay slowly. The filter decreases the T60 as expected, at least for those frequencies that showed reverberation. Loudness, which can be unpleasant, decreases, but sharpness increases, as result of the filter. Roughness is almost untouched by the filter, a slight decrease can be observed. The D50 shows, that reflections after 50ms disappeared.

The original and the modified impulse response of the speech audio recording setup are shown in numerical plot 2. A slow decay, early reflections and some unwanted behavior, that account for late reflections, can be seen in the original unmodified impulse response. The filter decreases the unwanted behaviors selectively for the frequency bins by attenuation factor based on the T60 decay matrix in numerical plot 1. Because the impulse response from the speech record was different from the drum record, filter results differ. The STOI measure stays untouched by the filter: No speech enhancement and no speech degradation due to filtering: No noise was added or subtracted to the speech sample due to filtering. Unexpectedly, but for obvious reason, the T60 decreased to zero after filtering in the speech audio.

6 Conclusion

The filter is designed aggressively on purpose and filters out reverberation and produces a dry sound with a minimum of room and recording setup characteristic, allowing the direct path signal to be dominant. It might be necessary to weight the output and mix it into the original signal, if it appears too sharp for the listener. However, as the filter does exactly what it should, the output might need further processing for more pleasant sound characteristic, but is obviously a excellent start for further audio processing.

7 Supplements

The Python-code of the studied method and audio samples are to be found at GitHub: <https://github.com/fanci90/Digital-Deconvolution-Audio-Filter-repo.git>.

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

	Drum Record degraded (z_n)	Drum Record filtered (z''_n)	Speech degraded (z_n)	Speech filtered (z''_n)
<i>min sample value*</i>	-1	-0.919	-0.045	-0.045
<i>max sample value*</i>	+1	+0.948	0.046	0.043
<i>peak amplitude*</i>	0dB	-0.46dB	-26.75dB	-27.43dB
<i>DC offset*</i>	$-\infty$ dB	$-\infty$ dB	-85.30dB	-86.44dB
<i>Estimated Pitch*</i>	2248.0Hz / C#6	9649.6Hz / D8	1445.2Hz / F#5	1530.5Hz / G5
<i>Min RMS Power*</i>	$-\infty$ dB	$-\infty$ dB	-61.03dB	-63.28dB
<i>Max RMS Power*</i>	-1.84dB	-12.54dB	-32.82dB	-32.61dB
<i>Average*</i>	-14.38dB	-28.84dB	-41.34dB	-41.27dB
LSD	-16.29dB		$7.63 \cdot 10^{-5}$ dB	
T60	5.475ms	0.186ms	0.1816ms	0ms
STOI	N/A	N/A	0.9978	0.9978
Δ STOI	N/A		0	
MOS-LQO	4.4202	4.5314	4.6308	4.6337
Δ MOS-LQO	0.1112		$2.9 \cdot 10^{-3}$	
<i>Loudness</i>	0.90320	0.30145	0.14764	0.14393
<i>Roughness</i>	0.23022	0.22169	0.21995	0.21946
<i>Sharpness</i>	3.26095	3.95266	3.09846	3.16039
D50	42.9%	100.0%	71.7%	100.0%

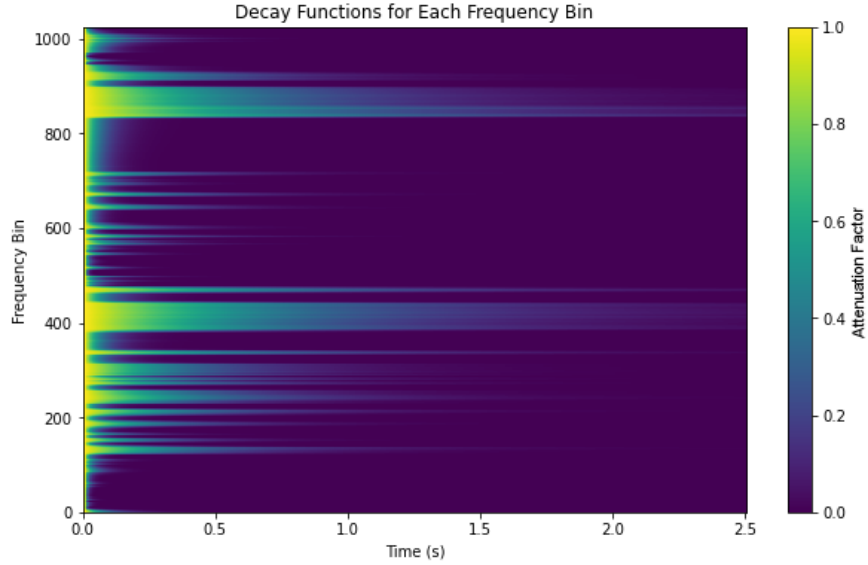
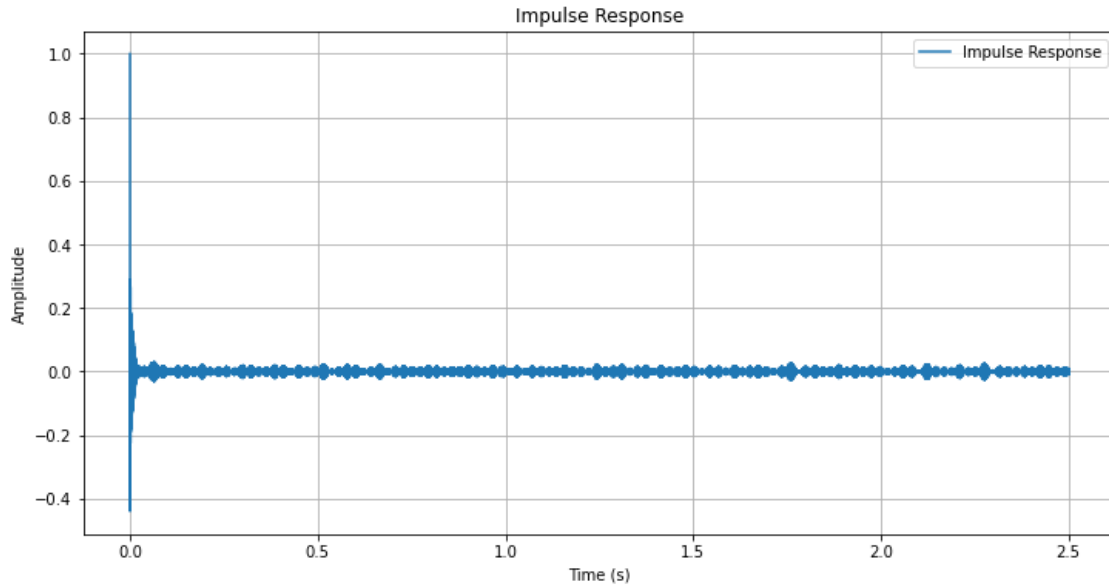


Figure 1: The $T60$ decay matrix.

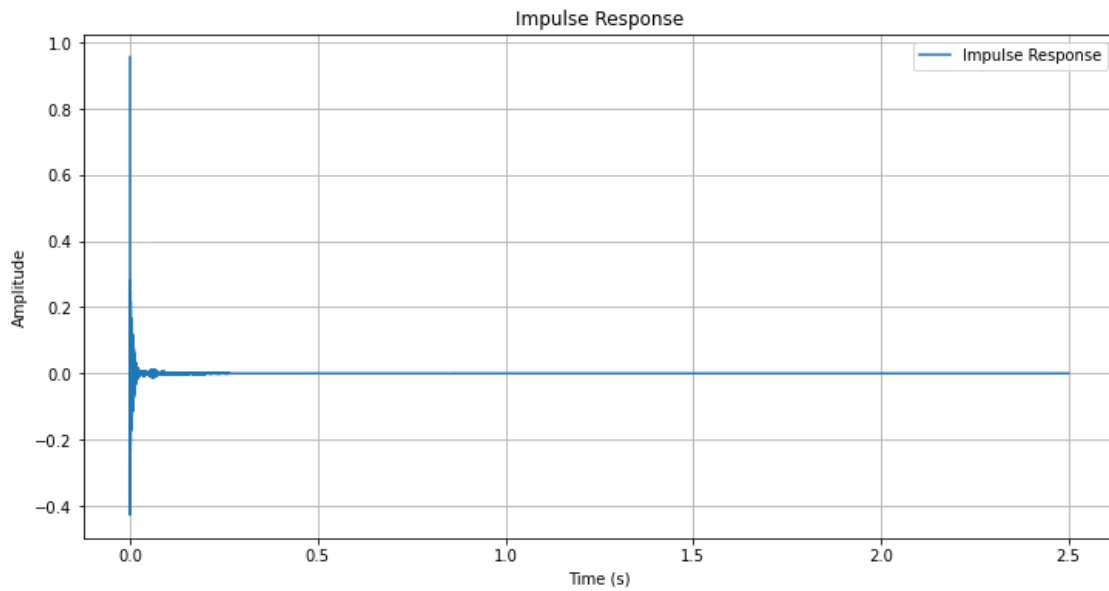
Acknowledgments

The author gratefully acknowledges Simon Ciba, M.A., known from e.g. "WhisPER – A New Tool for Performing Listening Tests", for his careful review of this work and his valuable comments.

Furthermore the author acknowledges the use of the AI language model Perplexity, developed by Perplexity AI, for assistance in literature and code review in this work.



(a) Original.



(b) Modified.

Figure 2: *Impulse response.*

References

- [1] A. V. Oppenheim, R. W. Schaffer, *Discrete-Time Signal Processing*, Prentice Hall, 1989. ISBN: 0-13-216292-X
- [2] Julius O. Smith III, *Spectral Audio Signal Processing*, W3K Publishing, 2011. Available online: <https://ccrma.stanford.edu/~jos/sasp/>.
- [3] J. A. Moorer, "About this reverberation business," *Computer Music Journal*, vol. 3, no. 2, pp. 13–28, 1979. doi:10.2307/3680287.
- [4] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, The Netherlands, 2007.

- <https://research.tue.nl/files/1972985/200710970.pdf>.
- [5] 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.
 - [6] 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>.
 - [7] 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>.
 - [8] 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>.
 - [9] N. Fliege, *Systemtheorie*. B. G. Teubner Stuttgart 1991 (Informationstechnik). ISBN: 3-519-06140-6
 - [10] 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).
 - [11] H. Kuttruff, *Room Acoustics* (6th ed.), CRC Press, 2016. (Open access chapters available) ISBN: 978-1498740436
 - [12] 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>
 - [13] Taal, Cees H. and Hendriks, Richard C. and Heusdens, Richard and Jensen, Jesper, *A short-time objective intelligibility measure for time-frequency weighted noisy speech*, 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. Pages 4214-4217, DOI: 10.1109/ICASSP.2010.5495701.
 - [14] (ITU), I. T. U.: Rec. p.862: Perceptual evaluation of speech quality(pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speechcodecs. 2001.
 - [15] ISO 532-2:2017(E), *Acoustics — Methods for calculating loudness — Part 2: Moore-Glasberg method*, International Organization for Standardization, Geneva, 2017. <https://cdn.standards.iteh.ai/samples/63078/a42f5c5bd7014d5f99cc3a7d4fa0c0c9/ISO-532-2-2017.pdf>.
 - [16] HEAD Acoustics, “Psychoacoustic Analyses I: Loudness and Sharpness Calculation,” Application Note, 201x. https://cdn.head-acoustics.com/fileadmin/data/global/Application-Notes/SVP/Psychoacoustic-Analyses-I_e.pdf.
 - [17] H. Fastl, “Psychoacoustic Basis of Sound Quality Evaluation,” Technical University of Munich, 2006. <https://mediatum.ub.tum.de/doc/1138486/729841.pdf>. (Free TUM repository source covering roughness and fluctuation strength models.)
 - [18] T. Houtgast and H. J. M. Steeneken, “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility,” *Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.
 - [19] Green Forge Coop. MOSQUITO (Version 1.1.1). <https://doi.org/10.5281/zenodo.10629475>
 - [20] Eomys Engineering, Modular Sound Quality Index Toolbox (MoSQITo), 2025. <https://github.com/Eomys/MoSQITo>
 - [21] C. Ludlow, python-pesq: Perceptual Evaluation of Speech Quality (PESQ), 2025. <https://github.com/ludlows/python-pesq.git>
 - [22] M. Pariente, pystoi: Short-Time Objective Intelligibility (STOI), 2025. <https://github.com/mpariente/pystoi>
 - [23] Steinberg Media Technologies GmbH, Cubase 5.0.1 build 147 [Computer Software], 2009. <https://o.steinberg.net/index.php?id=1782&L=1>