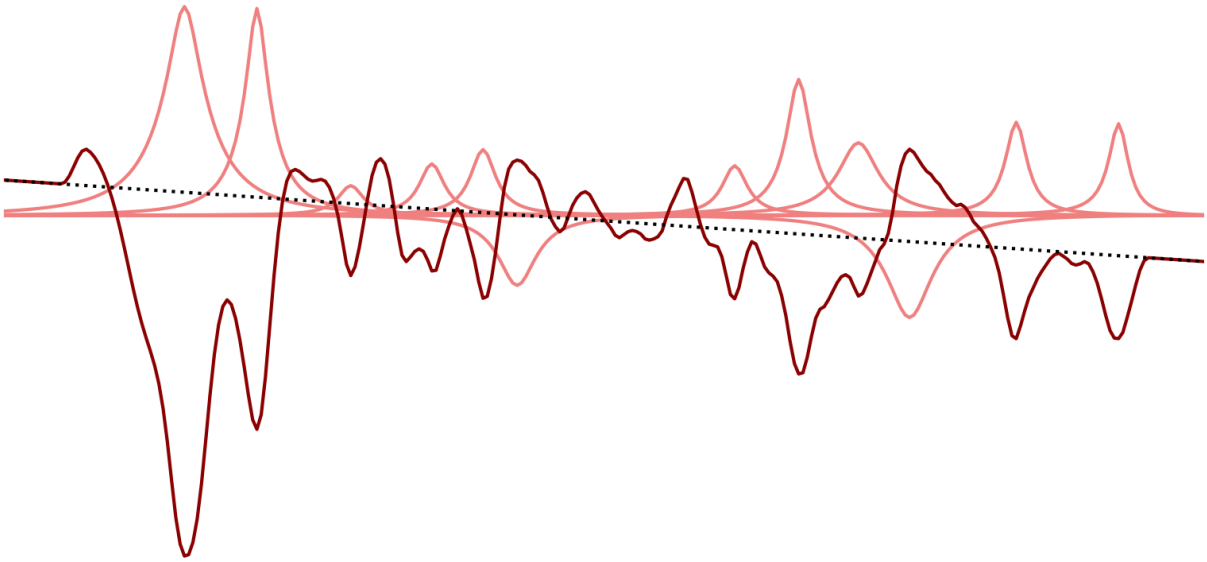


Loudspeaker–Room Response Equalization Using a Smartphone Microphone

Master's thesis in the Master's Programme Sound and Vibration,
at the Division of Applied Acoustics

JAN HANUŠ

DEPARTMENT OF ARCHITECTURE AND CIVIL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Loudspeaker–Room Response Equalization Using a Smartphone Microphone

JAN HANUŠ

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
Gothenburg, Sweden 2026

Loudspeaker–Room Response Equalization Using a Smartphone Microphone

Supervisor: Jens Ahrens, Chalmers University of Technology
Examiner: Jens Ahrens, Chalmers University of Technology

Master's Thesis 2026
Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Pre-processed measured magnitude room response curve (red) and individual peak filters (pink) that together form a compensation filter.

Typeset in LaTeX
Gothenburg, Sweden 2026

Abstract

The thesis investigates the topic of loudspeaker-room response equalization (LRRE) and tries to find a suitable method for integration into a pair of compact studio monitors. LRRE is a method for improving a loudspeaker system's performance in a room by measuring an impulse response in the listening position, based on which compensation filters are computed and applied to the loudspeaker system. The aim is also to assess whether such impulse response measurements can be performed using today's smartphones with their built-in microphone.
The whole procedure, from impulse response measurement to the computation of the compensation filter, is described, and techniques that correct only the magnitude response (minimum-phase compensation) and those correcting both the magnitude and phase responses (mixed-phase compensation) are compared. Three particular equalization functions are implemented in Python: a minimum-phase FIR filter, a mixed-phase FIR filter obtained by the x-filtered normalized least mean square algorithm, and a minimum-phase IIR filter comprising a cascade of biquad peak filters. It is found that the latter performs the best and is the most suitable option for integration into compact studio monitors.
A theoretical performance of a smartphone used as the measurement device is discussed by presenting the expected behaviour of the present MEMS microphone, and the application of digital signal processing by the manufacturer. The actual behaviour is then evaluated by measuring the frequency responses and sensitivities of 30 smartphones in an anechoic chamber and assessing their SNR. 84 % of the tested models emerged as suitable for measuring the loudspeaker-room impulse response.

Keywords: Loudspeaker–room response equalization, SISO, Minimum-phase filter, Mixed-phase filter, Smartphone microphone, MEMS microphone, Frequency response, SNR.

Artificial Intelligence Disclosure

Artificial Intelligence tools were used during the work on this thesis. Apart from a literature search in AI research agent SciSpace, their use was limited to Microsoft 365 Copilot and Microsoft 365 Copilot Chat under the enterprise data protection (provided by Chalmers). That ensures data privacy, including the exclusion of shared data from further training and security against copyright infringement. These tools were used to

- Search for literature,
  - Example prompt: 'Where can one buy/download the ITU-R recommendations? I am interested in ITU-R BS.1284-2',
- Coding: optimisation, troubleshooting, and finding suitable functions,
  - Example prompt: 'How to sort dataclass from smallest to largest based on attribute x?',
- Table formatting and troubleshooting in LaTeX.

Acknowledgements

First of all, I would like to thank Tom Fletcher from AIAIAI, who proposed the topic of this thesis. It was a pleasure to work with you, and I really hope that my findings come in handy in the future.
Thanks also to my university supervisor, Jens Ahrens. I am so grateful for all the time you spent with me, being available for my questions, and always willing to explain.
I would like to express my gratitude to everyone at the Division of Applied Acoustics at Chalmers. The welcoming atmosphere and all the knowledge you shared with us students – it was an amazing two years. I learned so much, and it was a great fun!
Lastly, my biggest thanks go to my parents. All of this wouldn't be possible without your support. Thanks a ton!

Jan Hanuš, Zlín, January 2026

Contents

List of Acronyms

ADC Analog-to-Digital Converter. 65
AGC Auto Gain Control. 67, 68, 86
AOP Acoustic Overload Point. 64, 65
API Application Programming Interface. 68
ASIC Application-Specific Integrated Circuit. 65, 67

DSP Digital Signal Processing. 3, 78, 79, 83, 86
DTFT Discrete-Time Fourier Transform. 3, 19, 47

ECM Electret Condenser Microphone. xiii, 64, 65
ERB Equivalent Rectangular Bandwidth. 12, 13, 18

FDT Frequency-Dependent Truncation. 34–36, 50
FET Field-Effect Transistor. 65
FFT Fast Fourier Transform. 41
FIR Finite Impulse Response. 43, 46, 85
FRF Frequency Response Function. x, xiii, 17, 66, 68–70, 72, 73, 76, 77, 80–84, I

IEC International Electrotechnical Commission. 68
IIR Infinite Impulse Response. 25, 85, 86
IR Impulse Response. xi, xii, 6, 16, 23–25, 44, 45, 47, 50, 52, 58, 73, 74

JND Just Noticeable Difference. 14

LMS Least Mean Squares. 26, 29
LRIR Loudspeaker-Room Impulse Response. x–xii, 6, 13–20, 23–26, 29–31, 33–35, 40, 43–45, 47–50, 52, 57, 58, 62, 63, 71, 77, 83, 85, 86
LRRC Loudspeaker-Room Response Curve. x–xii, 6, 8, 10, 11, 18, 19, 22, 24, 26, 32–34, 42, 50, 53–56, 59, 60, 63, 85, 86
LRRE Loudspeaker-Room Response Equalization. x, 1, 15, 16, 24, 83, 85, 86

MEMS Microelectromechanical System. xiii, 64–66, 79, 83

PCB Printed Circuit Board. 65–67

RMS Root Mean Square. 19, 30, 41, 70, 73

SNR Signal-to-Noise Ratio. iii, xi, xiv, 16, 30–33, 64, 66, 68, 70–72, 77, 79, 83, 86
SOS Second-Order Sections. 25, 26
SPL Sound Pressure Level. 64

THD Total Harmonic Distortion. 64

List of Figures

List of Tables

Introduction

When listening to music through a loudspeaker system in a room, the perceived sound quality strongly depends not only on the loudspeaker system itself but also on the listening environment. The room's shape, its volume, the materials on the walls, the placement of the speakers, the position of the listener – those are all factors that contribute to the overall experience. It is clear that if the design of the listening space is not done properly, it can have a detrimental effect. Excessive resonances and reflections can result in an undesirable alternation or 'coloration' of the sound. To improve such a situation, one can measure an impulse response in the listening position, and then, based on the measurement, create filters that will be applied to the loudspeaker system, trying to mitigate these negative influences. This technique is referred to as Room Response Equalization [1], [2] or, less frequently but more precisely, as Loudspeaker-Room Response Equalization (LRRE).

The subject of LRRE is nothing new. It has been studied for many decades now (see [3] from 1958, for example). Nowadays, there are many commercial solutions that are available as computer software, standalone units, or fitted into integrated amplifiers, AV receivers, and active loudspeaker systems. Yet this thesis is motivated by a request from the industry. A company named AIAIAI that features a portable stereo set of studio monitors, UNIT-4, in its portfolio, is interested in the possibility of implementing some sort of automatic equalization functionality in their device. Hence, the purpose of the thesis is to explore the topic of LRRE and to design a system suitable for their product that enhances the listening experience. As the Unit-4 monitors are portable, they are intended for transport and use in various spaces. Because of that, it would be very convenient to be able to measure the impulse response, based on which the LRRE is designed, with a smartphone. Thus, a part of this project is dedicated to the microphones used in smartphones and tries to answer the question whether it is possible to make such acoustic measurements with them and if so, to what extent.

The first part of the text is about the LRRE. Initially, the listening room's influence on sound reproduction and relevant parts of psychoacoustics are presented. Then, the theory of LRRE is thoroughly explored, starting with the measurement of the loudspeaker-room impulse response to the final computation of the compensation filter, which takes care of the equalization. The last two sections are dedicated to implementing three different methods and evaluating them.

The second part is devoted to smartphones. This chapter begins with the theory part, discussing the microphone's construction, relevant parameters, and expected performance. Then, the whole measurement procedure for obtaining its sensitivity, frequency response, and signal-to-noise ratio is presented.

Lastly, all results are discussed, and the suitability of using a smartphone as a measurement device for capturing the loudspeaker-room impulse response is evaluated.

1

Loudspeaker-Room Response Equalization

1.1 Necessary Bits of Digital Signal Processing

This section introduces important concepts in DSP that will be used later in the text. It might be a good idea to skip it for now and use it only as a reference.

1.1.1 *Z*-transform

It is a mathematical operation that takes a real or complex discrete signal in the time domain and transforms it to a complex-valued discrete signal in the so-called z-domain. It is defined [4] as

$$X[z] = \sum_{n=-\infty}^{+\infty} x[n] z^{-n}, \tag{1.1}$$

where

$$z = re^{j\omega}.$$

r is a real number and ω is the angular frequency. Complex z can be expressed in polar coordinates, in so-called z-plane: r represents the radius and ω the angle of z. Let's shed some intuitive light on the equation. Z-transform is a generalization of the discrete-time Fourier transform (DTFT), which is described by the same formula but without the term r^{-n}. What DTFT does is it correlates a signal with a complex exponential $e^{-j\omega n} = \cos \omega n - j \sin \omega n$, which results in information about magnitude and phase for a given frequency of the analyzing signal. By correlating the signal also with the real exponential r^{-n}, we can also analyze signals that diverge (with time) and thus assess stability as well as causality of a system. DTFT can be only used for the analysis of causal stable systems.

When it comes to Z-transform, *poles* and *zeros* are important terms to be aware of. Let's assume an arbitrary signal or a system that is represented in the z domain as

$$H(z) = \frac{(1 - 0.5\, z^{-1})(1 - 1.5\, z^{-1})}{(1 - 0.7\, z^{-1})(1 - 1.2\, z^{-1})}. \tag{1.2}$$

Roots of the numerator polynomial ($[0.5, 1.5]$) and denominator polynomial ($[0.7, 1.2]$) are called *zeros* and *poles*, respectively. They are plotted in the z-plane in Fig. 1.1. Poles correspond to peaks and zeros to dips in the magnitude frequency

response. Since the frequency response is evaluated on the unit circle, the distance of the poles and zeros to the unit circle influences the strength of a dip/peak.

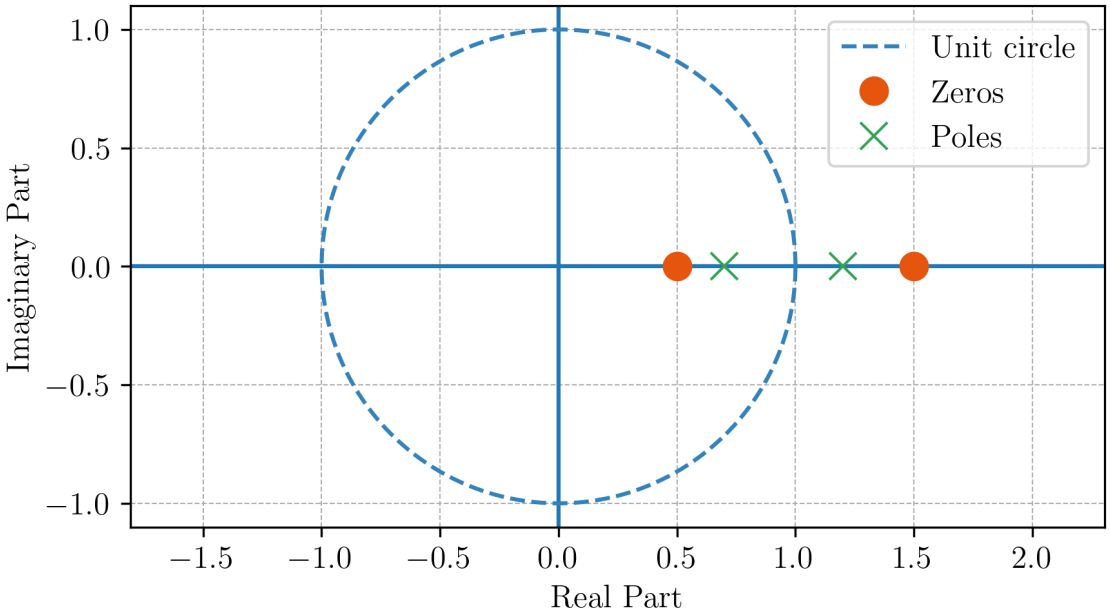


Figure 1.1: Z-plane or pole-zero plot of the system with given transfer function by Eq. (1.2).

1.1.2 Filters

This section considers linear time-invariant (LTI) filters.

1.1.2.1 Causality

'A filter is *causal* if it does not laugh before it's tickled' [5]. In formal language, its output does not depend on any 'future' inputs. For example, $y[n] = x[n+2]$ **is** dependent on future inputs and thus is said to be *non-causal* or *acausal*.

However, a delay can make a non-causal filter causal: $y[n] = x[n+2-D]$. If the delay $D \geq 2$, the filter becomes causal.

1.1.2.2 Stability

A filter is *stable* if its impulse response $h[n]$ 'approaches zero as n goes to infinity' [5]. In the complex plane, this is equivalent to the filter having all its poles inside the unit circle. Any pole outside the unit circle results in 'exponentially increasing component of the impulse response'[1] [5]. A filter that is not stable can cause the output to grow to infinity.

[1]This assumes that the filter is causal. If it is anti-causal, poles inside the unit circle cause instability.

1.1.2.3 Minimum-Phase Filter

A *minimum-phase* filter has 'the least energy delay of all systems with a given magnitude response' [6]. It is sometimes said to have the *fastest decay* [5]. In other words, the accumulated energy from time 0 until given t will be larger for the minimum-phase filter than for any other filter having the same magnitude response. That also means that for a given magnitude response, only one minimum-phase filter exists.

In the complex plane, it has all its poles and zeros inside the unit circle. That implies that both the minimum-phase filter and its inverse are stable and that such a filter is causal by definition. The opposite of the minimum-phase filter is a maximum phase filter that has all the poles and zeros outside the unit circle [5].

1.1.2.4 Linear-Phase filter

A *linear-phase* filter has a symmetric impulse response. It is probably no surprise that its phase is linear. This means that if a signal is filtered with such a filter, every frequency component is delayed by the same amount. That preserves the signal's waveform in the time domain as much as possible. The delay is $(N-1)/2$ samples, where N is the filter length. A linear-phase filter can only exist as an FIR filter. [5]

Any linear-phase filter can be decomposed to a minimum-phase term, maximum-phase term, and a third term containing only zeros on the unit circle [4]. Even though it is sometimes claimed that linear-phase filters are ideal for audio applications as they preserves the signal's waveshape, it does not need to be necesarilly true due to the fact they have a symmetric impulse response and therefore introduce pre-ringing that can be heard. It can be that a filter with a phase response somewhere between linear- and minimum-phase might be superior [5].

1.1.2.5 Mixed-Phase Filter

A *mixed-phase* filter is a filter for which zeros occur both inside and outside of the unit circle [5].

1.1.3 Phase, Phase Delay & Group Delay

A phase response of a filter gives the phase shift in radians or degrees of each sinusoidal component. The phase of a sinusoidal signal is defined as

$$\varphi(\omega) = \omega\, t.$$

We can see that the phase is linearly dependent on angular frequency ω. Thus, we could obtain a quantity that would tell us the time delay in seconds for every sinusoidal component of the input signal. It is called *phase delay* and defined as [5]

$$P(\omega) = -\frac{\varphi(\omega)}{\omega} \quad \text{(s)}. \tag{1.3}$$

The minus makes it a delay.

A *group delay* specifies the delay of narrow-band interval of sinusoidal components around ω. It is defined as [5]

$$D(\omega) = -\frac{d}{d\omega}\varphi(\omega) \quad \text{(s)}. \tag{1.4}$$

It is this quantity that is typically used when one is concerned about how the frequency components are modified/delayed by a filter. [7] claims that one can study acoustic signals by considering them as a set of narrow-band packets. And since an acoustic signal is a band-limited signal, one could describe it as 'a sinusoidal signal with an amplitude envelope variation modulated on it'. In 1989, Marshall Leach, Jr. proposed a quantity called *differential time-delay distortion* which is the difference between the phase delay and the group delay, motivated by the fact that for zero delay introduced by the system/filter, they should be equal [7]. However, I have not found any further information about it in more recent literature or articles, so it was probably not adopted, and the group delay alone remains an indication of any introduced time shifts.

1.2 Room's Influence on Sound Reproduction

Since the objective is to improve our perceived sound quality by compensating not only for the loudspeaker's response but also for the response of a room, it is appropriate to first ask what happens when sound travels in an enclosed space and what we demand from a suitable space for amplified reproduction.

1.2.1 From Sound Source to Listener's Ears

Let's first assume a short impulse that is emitted from a single sound source. It propagates away from the source in various directions. From the listener's point of view, an unaltered pulse from the source arrives at the listener's ears first, a so-called *direct sound*. It is followed by *early reflections*, as the pulse reaches nearby walls or other objects and reflects towards the listening position. The pulse propagates further in the room and reflects many more times until it is fully absorbed by any absorption present in the room. This is being referred to as a *reverberation* or *late reflections* and quantified by the *reverberation time*. All these parts together form a *room impulse response* (sometimes abbreviated as RIR). Fig. 1.2 shows how such IR can look in a schematic way. Its frequency domain equivalent is a *room response curve* (RRC) which is shown in Fig. 1.3. In reality, when the RIR is measured, it contains the IR of the sound source, which influences the direct sound. Therefore, in this work, I will instead use an abbreviation LRIR for *loudspeaker-room impulse response* and LRRC for *loudspeaker-room response curve*, to have this in mind.

Now, let's move from a pulse to a continuous signal. When the source is switched on and a steady state is reached, at frequencies for which multiples of their corresponding half wavelengths fit between walls, floor, and ceiling, so-called *standing waves* or *modes* are formed [8]. These are pressure maxima and minima distributed within the space. It is clear that the room's geometry determines at what frequencies standing waves **can** exist. However, the presence of a standing

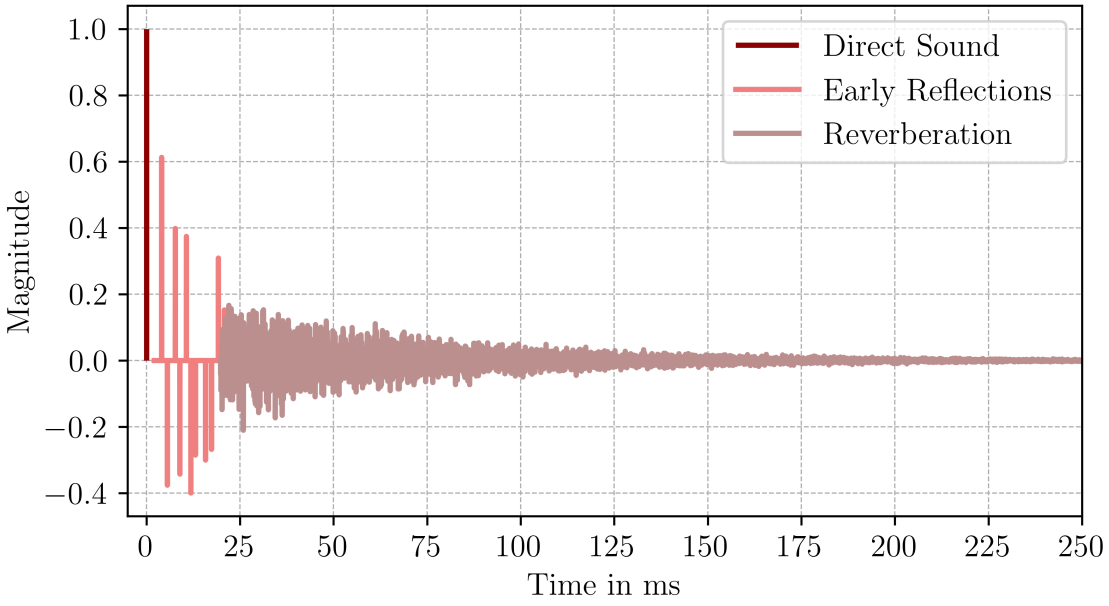


Figure 1.2: Scheme of a room impulse response.

wave (whether it is excited or not) depends on the position of the source within the room, and if we actually hear the standing wave is determined by the listening position.

1.2.2 Early Reflections

When it comes to concert halls' acoustics, the literature typically considers early reflections up to 50 ms as something positive. Firstly, they increase the clarity of music and intelligibility of speech [9], [10]. In this sense, they are considered as a 'support' to the direct sound and prevent the music from beeing 'burried' in the rather long reverberation[2]. Secondly, in spite of the modification mentioned, they can increase a so-called *spaciousness*, subjectively described by Marshall and Barron as the 'sensation of feeling inside the music'. However, it is only reflections from the sides, the *lateral reflections*, that contribute to this sensation [11].

In small listening rooms such as control rooms in recording studios or living rooms, the situation is a little different. Thanks to the room's smaller size, the early reflections are usually considered within a time window of 20 ms after the arrival of the direct sound [12]. While some listening tests showed that the presence of lateral reflections is preferable for 'recreational' listeners, as they made the sound stage 'broader' (also increasing the spaciousness), musicians and sound professionals preferred to limit the level of lateral reflections, as they were 7 times more sensitive to them than the 'ordinary' listeners [11]. The same listening tests also showed that early reflections coming from the front – reflected from the wall behind the speakers – are unwanted because they can distort the sound stage. The term sound stage

[2]This is particularly important for quick passages of a piece or when quieter instruments such as recorder or harpsichord play [10].

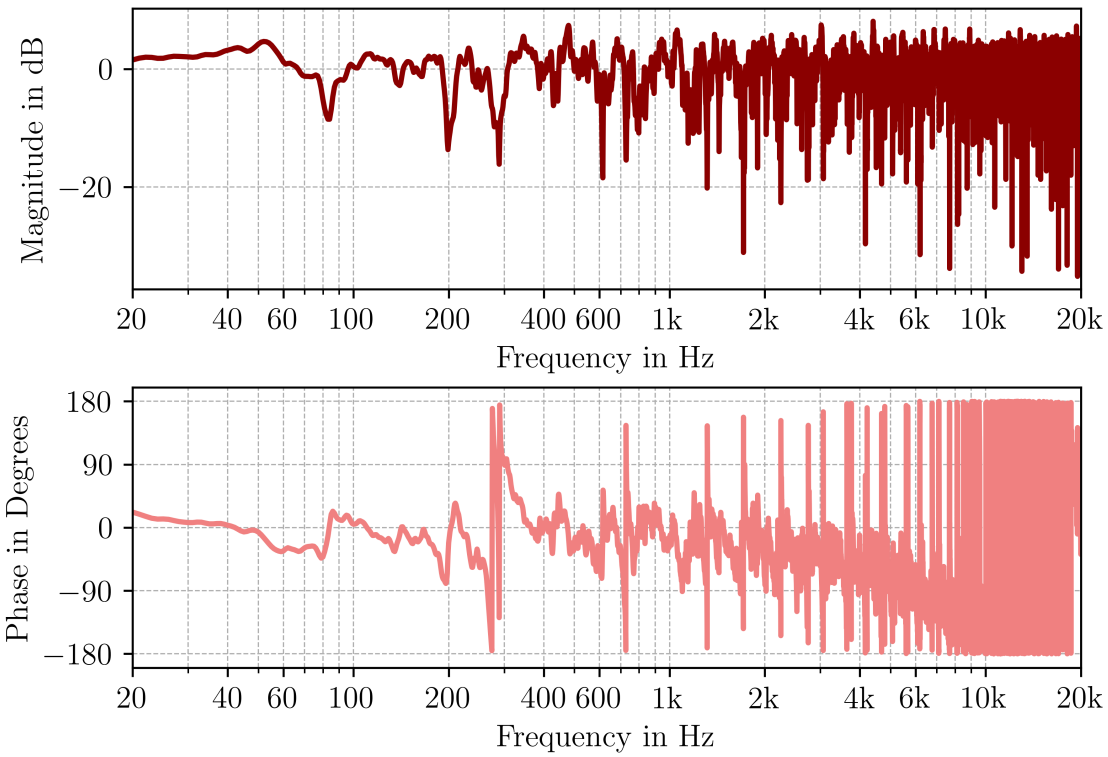


Figure 1.3: Room response curve obtained as $\mathcal{F}\{\text{RIR}\}$.

refers to the width and depth of the perceived sound image and the localization of instruments within it.

The sound stage distortion is also discussed in [12]: Not distinguishing between lateral or non-lateral, for a stereo reproduction, if the early reflections are strong with respect to the direct sound, they will result in the distortion of the perceived sound stage. [8] supports this and also mentions the word 'blurred' when discussing the incorrect location of phantom sources caused by the early reflected sound. It is also the change of *timbre*, a coloration, that is being attributed to the early reflections [13]. This is no surprise because a reflection esentially is a delayed version[3] of the direct sound,and when combined with the direct sound, it results in a modification of both the time structure and the spectral balance of the sound (a timbre). This is clearly visible in the magnitude LRRC as significant notches at frequencies where the reflection delay corresponds to odd multiples of half the signal period (i.e., half the wavelength). This is shown in Fig. 1.4.

That is why the International Communication Union (ITU-R) imposes a demand on the early reflections for a 'reference listening room' with accurate listening conditions: Early reflections reaching listener up to 15 ms should be weaker by 10 dB compared to the direct sound in the frequency range from 1 kHz to 8 kHz [14]. Thus, we should have in mind that it is beneficial to reduce the amount of early reflections when designing the equalization[4].

[3] It also has a lower level and possibly different spectral balance than the direct sound, depending on the absorbing properties of the reflecting surface.

[4] When targeted by passive means, even though it is common to make an extensive reduction by

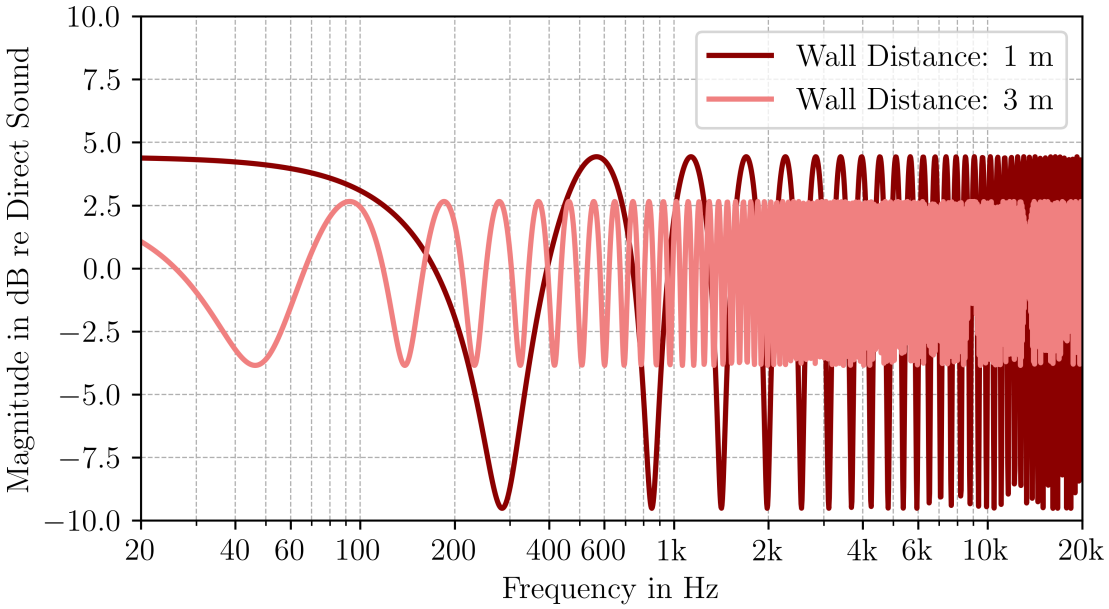


Figure 1.4: The influence of a reflection on the direct sound, when the source-receiver distance is 3 m and the reflection coefficient of the wall $R = 0.8$. Wall distance is the distance of both the source and receiver from the wall. Note: The sound source is assumed to be a monopole point source.

1.2.3 Reverberation

Beranek, in his book *Concert Halls and Opera Houses* [15], refers to reverberation as a key characteristic of a space: 'It is one of the components available to the composer (and the performer) for producing a musical effect'. We can therefore view reverberation as an aesthetic element of the music itself rather than just an acoustic property. This is especially important when we discuss the amplified music. The intended reverberation is already added by the music producer or the mixing engineer onto the record. As a result, the reverberation in listening rooms for sound reproduction no longer serves this purpose. [8] suggests that the reverberation time of such rooms should be shorter than that present on the recording. It should not be increased, especially at lower frequencies; otherwise, the room has a tendency to sound 'boomy'.

The Recommendation ITU-R for listening conditions of control rooms [14] specifies a value of reverberation time

$$T_\mathrm{m} = 0.25\sqrt[3]{\frac{V}{V_0}}, \tag{1.5}$$

where V is the volume of the room and $V_0 = 100\,\mathrm{m}^3$. The value of T_m should be equal to the mean of the measured T over the frequency range from 200 Hz to 4 kHz. A tolerance for the measured T is specified with respect to T_m and is shown in Fig. 1.5. To have some rough idea about what the value of T_m should be for a typical

absorptive material, some listening tests show that diffusion is considered to be superior in helping to decrease the level of early reflections [12].

room in an apartment that could have a ceiling height of 2.5 m and a floor area of 15 m^2, $T_m \approx 0.18$ s. With a floor area of 30 m^2, $T_m \approx 0.23$ s. However, in domestic environments like living rooms, the values are typically significantly higher. In such spaces, it has been observed that reverberation times below 0.4 s are perceived as 'dry', while values above 0.8 s as 'live' [8].

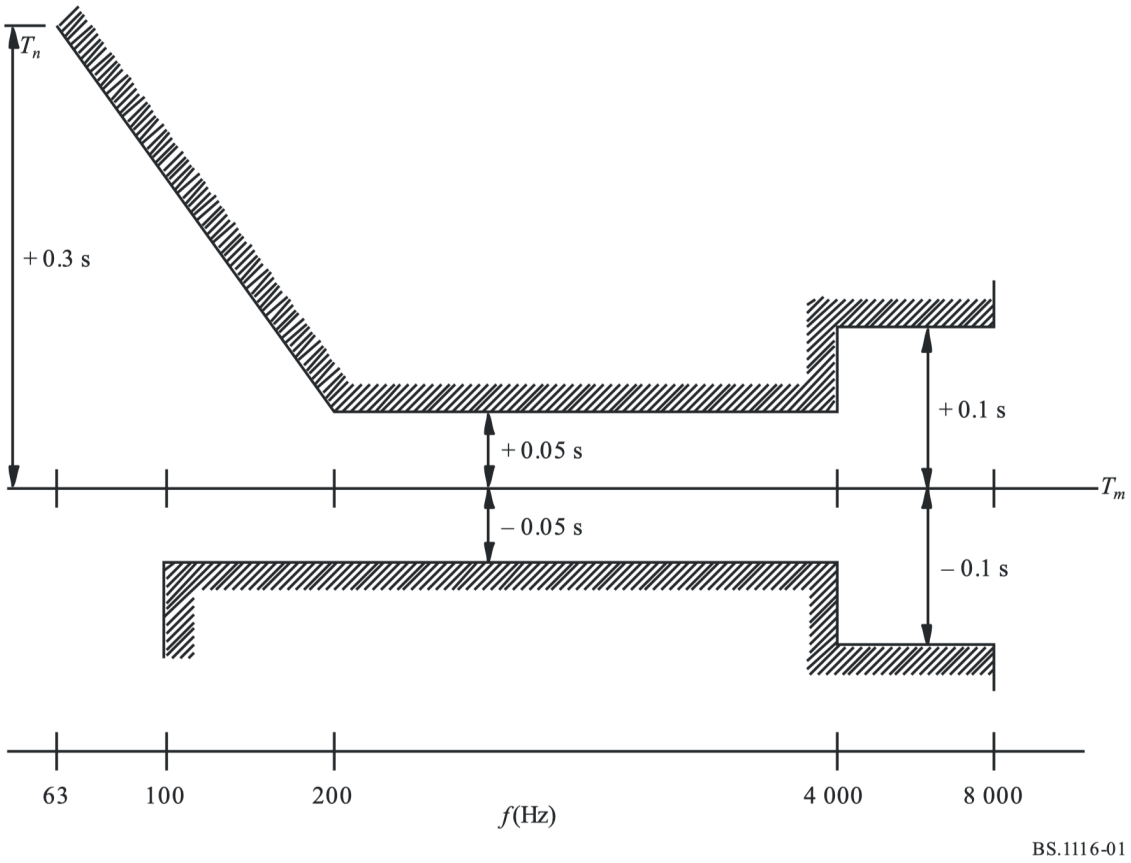


Figure 1.5: Reverberation time tolerance relative to T_m. Adopted from [14].

1.2.4 Room Modes

From the previous text, it is clear that room modes modify the LRRC. The degree of modification depends on both the speaker's and the listener's position. They exist in the whole frequency range of interest (that is, typically from 20 Hz to 20 kHz). However, the modal density rises with the square of the frequency [8]. This means that at low frequencies, there are only a few modes that can be clearly seen in the magnitude LRRC, whereas at high frequencies they are so dense that we cannot distinguish between them and therefore are not so important to us. In practise, a formula from statistical acoustics is used to determine some kind of an estimate of a frequency, up to which it makes sense to track the modes. Up to this frequency, the resonance curves of the modes have the quality factor Q, which describes how wide/narrow the resonance is (see Eq. (1.19)), less than 3 [16]. Its name is the

Schroeder frequency and is given by [16]

$$f_{\text{Schroeder}} \approx 2000\sqrt{\frac{T}{V}}, \tag{1.6}$$

where T and V stand for the reverberation time and the volume of the room, respectively. Again, to have some rough idea, using the same example as before; a room with a ceiling height of 2.5 m, a floor area of 15 m^2 and a reverberation time of 0.5 s, $f_{\text{Schroeder}} \approx 230\,\text{Hz}$. With a floor area of 30 m^2, $f_{\text{Schroeder}} \approx 163\,\text{Hz}$[5]. So now it is evident that when it comes to the room modes, we are concerned with the lowest frequencies.

Room modes can be described by the quality factor Q and their decay time[6] [8]. They contribute the the measured reverberation time of a space. That is, in my opinion, why there are no requirements concerning room modes in the Recommendation ITU-R [14]. [8] stresses that the decay time of room modes should be smaller than the decay time of musical instruments. It also mentions that a listening test was performed where the low decay times of room modes at low frequencies were preferred. Because standing waves at low frequencies have minimum-phase characteristics, reducing their decay times makes the LRRC smoother [11], [17]. This implies that smoothing the LRRC where the standing waves occur will shorten their decay times. To comment on the influence of Q, some listening tests showed that low-Q values that cause broad-band peaks in the magnitude response are easier to recognize than narrow-band peaks with high-Q values [18]. However, other listening tests showed the opposite [19], [20].

1.3 Perception

When it comes to sound reproduction, a key part of what we, as humans, experience is based on how the sound is processed by our auditory system. We might be sensitive to certain aspects, while tend to ignore other. I have already mentioned some subjective preferences earlier in the text, but there is more information that might be useful when designing an equalization system.

1.3.1 Masking

When it comes to psychoacoustics, a phenomenon called *masking* is often discussed. It refers to the fact that some sounds simply cannot be heard in the presence of other sounds due to the way the auditory system operates. There are two types of masking: *simultaneous* and *temporal*. Simultaneous masking is when a sound (a maskee) is masked by another sound (a masker) when both are played at the same time. Temporal masking, on the other hand, is when the maskee is masked by the masker that comes later in time: a *pre-masking*, or by the masker that comes earlier in time: a *post-masking*.

[5] Note that T is usually frequecy-dependent. It is therefore typically calculated for fractions of octave bands.

[6] The definition of a decay time is the same or similar to that of a reverberation time.

For this work, the temporal masking is the more relevant one. When it comes to pre-masking, the masker can mask the preceding sound if it occurs no later than 20 ms afterwards. For post-masking, much longer delays are possible: the maskee can be masked up to 200 ms after the masker. How well the masker can ‚do its job' depends on this time delay, the nature of the sounds, and the level difference between them. Fig. 1.6 shows what level the maskee needs to exceed in order to be heard for different values of time delays when the masker is a uniform noise with the level of 60 dB and the maskee is a sinusoid of 800 Hz. [21]

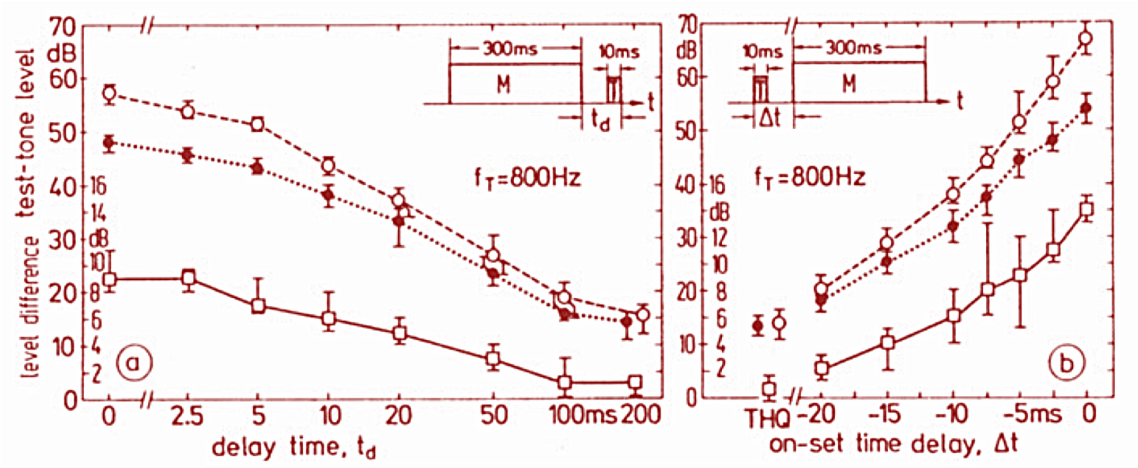


Figure 1.6: Maskee level thresholds when the masker is a 300 ms-long uniform noise with the level of 60 dB and the maskee is a 10 ms-long sinusoid of 800 Hz; post-masking (a), pre-masking (b). The maskee is in both ears in phase (open circles) and out of phase (closed circles). The squares are the median of their difference (the smaller y-axis). Adopted from [21].

1.3.2 Frequency Resolution

The first concept that should be mentioned here is *critical bands*. These are frequency-selective filters that, in parallel, process the sound in our auditory system [8]. One way of measuring the bandwidth of the critical band is to simultaneously play a pure tone and a band-limited noise, and increasing the bandwidth of the noise up to the point when further bandwidth increase has no impact on the masking of the pure tone (the pure tone stays audible at its current level even when the bandwidth is being further increased). This is therefore closely connected to the mentioned simultaneous masking. The concept of critical bands was later slightly improved and mathematically described under the name *equivalent rectangular bandwidth* (ERB). The bandwidths of both are shown as a function of frequency in Fig. 1.7.

Sometimes it is claimed that we 'hear' in critical bands, but that is misleading. The idea behind the critical bands and ERB is that within these bands, the individual frequency components are summed to estimate the overall loudness, define clear boundaries for masking by noise (as discussed), and for frequency separation of two tones in order to be individually identifiable. It **is** possible, however, to hear changes in the frequency response within a single bandwidth. [11]

The quantity that we might be looking for is the *just-noticeable frequency difference*. [21] presents results of listening tests of this quantity, and it was shown that this JND can be as high as $20\%f$, approx. a third-octave band, at low frequencies (125 Hz) when the test tone is short (10 ms) but as low as $0.2\%f$, approx. a 1/35-octave band, at mid frequencies (1 kHz) when the test tone is rather long (500 ms). But does this imply that with the same level of resolution, we can also recognize the spectral features? I couldn't find any more information regarding this.

As much as we may wish otherwise (I truly wished it myself), it has to be concluded that it is not really possible to find a single curve that would precisely describe the frequency resolution we, humans, have. However, [11] suggests that when it comes to measurements of LRIR, the coloration effects as discussed are likely to be less audible if they fall within the bandwidth of the critical band/ERB.

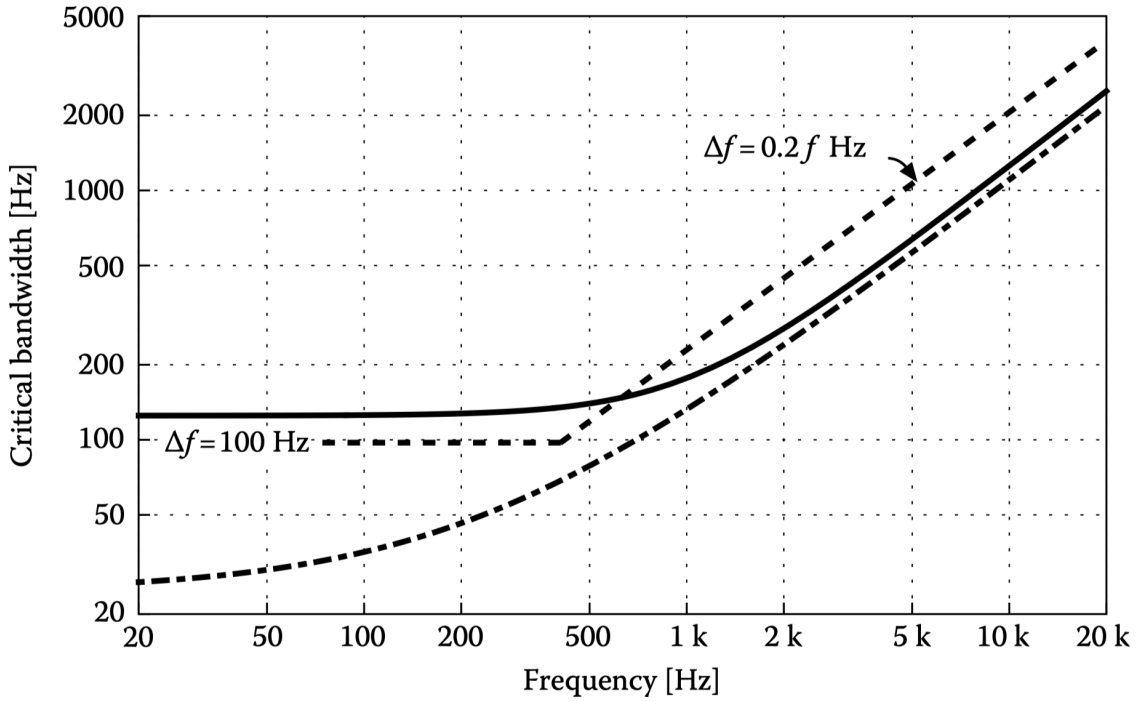


Figure 1.7: The bandwidth of critical bands (solid), ERB (dotdashed), and asymptote that is constant at the lowest frequencies and then follows a third-octave bandwidth (dashed). Adopted from [8].

1.3.3 Group Delay

How sensitive are we to the modification of group delay of a signal? [22] performed listening tests where the subjects listened to impulse responses of various speaker systems convolved with pink noise. The group delay of the result was artificially modified, and different versions were compared using a paired-comparison listening test. It was shown that within the frequency range from 300 Hz to 1 kHz, group delay variations up to 1 ms are inaudible.

Another study, which examined the effect of phase modifications on perception [23], found that the auditory system is sensitive to modifications of complex har-

monic signals comprising harmonics with certain fixed relations, which is the case for music. The listening tests were conducted using synthetic signals of this nature. It was concluded that phase modifications at a certain frequency can influence the perception of frequencies up to one octave lower and higher. Above 800 Hz, the differences in phase spectrum are inaudible.

1.3.4 Magnitude

When it comes to the just noticeable difference (JND) in magnitude, 1 dB is generally considered to be the value of change we can just perceive. [10] mentions that for sounds exceeding 40 dBSPL with frequencies above 100 Hz, the JND is less than 1 dB. The most sensitive frequencies for detecting the magnitude change are within the range of 1 kHz and 4 kHz, where the JND is just 0.25 dB for sounds with the level greater than 60 dBSPL.

1.4 Theoretical Basis of LRRE

From the section 1.2, we know that the listening room changes the sound we perceive, and in order to have a good or accurate listening experience, we impose certain demands on the early reflections, the excitation of room modes, and the reverberation. Traditionally, this is solved by strategically placing absorbers and diffusers within the space. But what can we do instead with a signal processing?

If we were to mimic the behaviour of the passive treatment, we would need to control the whole sound field in the room. And this is possible to do. One measures the LRIR in n positions with m loudspeakers distributed within the room. This results in $n \cdot m$ LRIRs needing to be compensated by filters applied to the individual loudspeakers, thereby controlling the sound field. This is possible up to a certain frequency which is determined by the number of n and m, the size of the room, and how large the equalized area should be. This is referred to as the MIMO approach – Multiple-Input (m) Multiple-Output (n). [24], [25]

Designing a system with the MIMO approach is a very complex task, and it is not further discussed in this work. Instead, we will look at what is possible to do by treating individual loudspeakers seperatly, equalizing one by one. The general idea is as follows: If we know the LRIR $h_{\mathrm{LRIR}}(t)$, we could design a compensation filter $h_{\mathrm{comp}}(t)$ which is essentially an inverse of $h_{\mathrm{LRIR}}(t)$. As a result, the influence of $h_{\mathrm{LRIR}}(t)$ would be canceled out.

In the time domain:	in the frequency domain:
$h_{\mathrm{LRIR}}(t) * h_{\mathrm{comp}}(t) = h_{\mathrm{target}}(t),$	$H_{\mathrm{LRRC}}(f)\, H_{\mathrm{comp}}(f) = H_{\mathrm{target}}(f).$

$H_{\mathrm{target}}(f)$ is defined by the user. The simplest case is 0 dB at all frequencies making $H_{\mathrm{comp}}(f)$ an inverse of $H_{\mathrm{LRRC}}(f)$. In time domain, $h_{\mathrm{target}}(t)$ should be as

close to a delta fuction $\delta(t)$ as possible; $\int_{-\infty}^{\infty} \delta(t)\, dt = 1$. The compensated response in time and frequency domain (or the target function) is shown in Fig. 1.8. It is evident that such a filter cannot modify the sound field, so it would decrease the level of early reflections or reverberation. However, it can alter the spectral balance and time structure of the direct sound, potentially limiting the coloration or sound stage distortion discussed in Sec. 1.2.2 (be it correcting for the imperfect response of the loudspeaker itself or trying to mitigate the room's influence).

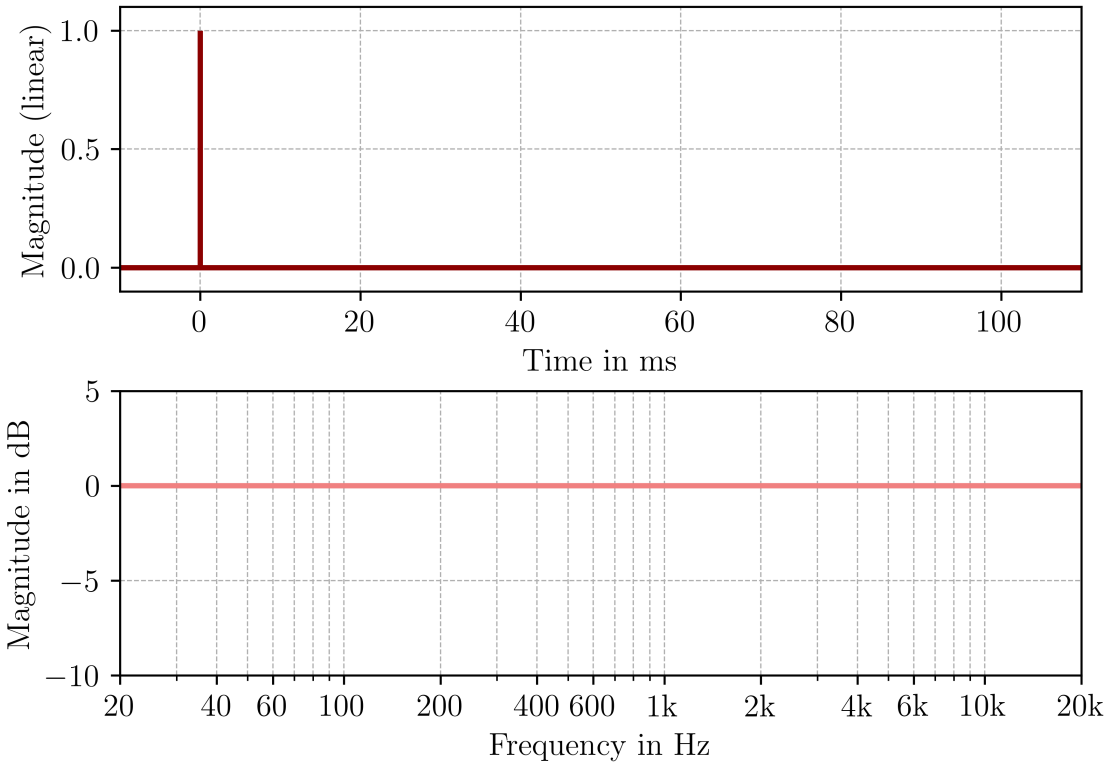


Figure 1.8: Ideal resulting loudspeaker-room response after compensation in the time and the frequency domain.

Therefore, the idea is to focus on the **equalization of the early part of the LRIR**. That is in accordance with [1], [11], which presents that in small listening rooms, how the reproduced music will sound is dominated by direct sound and early reflections. Another motivation is that the human ear perceives sound by integrating individual sound events within a time window of a length up to 60 ms [26].

1.4.1 General Procedure of LRRE

In the beginning, I would like to acknowledge a review paper *Room Response Equalization–A Review* from 2017 [1], where the authors managed to create an extensive overview of the existing methods for the design of LRRE. It was a great resource for the following text. More methods and approaches can be found there if one is interested.

Fig. 1.9 show a block diagram of the LRRE including all the steps that need to be performed. First, the LRIR has to be measured in the listening position. This response is then modified or *pre-processed*. If more than one measurement of LRIR were done, a *prototype LRIR* is created from them. These multiple measurements are performed in the considered 'listening area', each in a slightly different position. Then, a *compensation filter* is computed based on the pre-processed (prototype) LRIR.

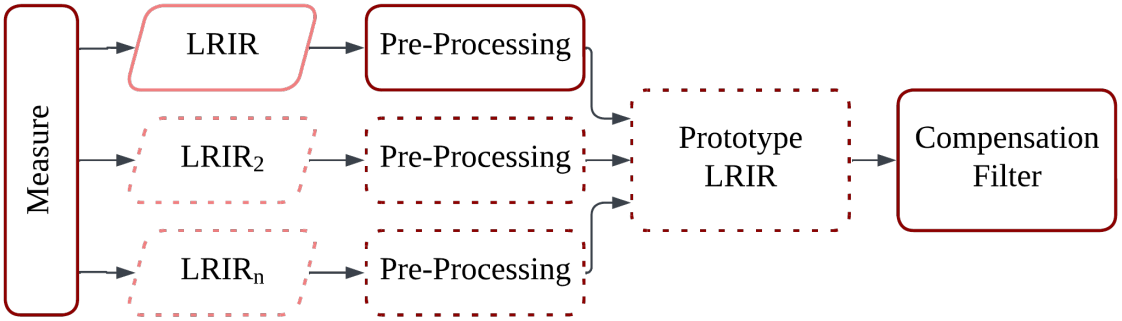


Figure 1.9: Block diagram of LRRE. Dotted blocks are optional.

1.4.2 Measurement of LRIR

First, the LRIR has to be measured in the listening position. There are several methods for obtaining an IR when the excitation source is a loudspeaker. The three most common ones are sine sweep, H_x-estimator, and MLS (maximum-length sequence). LMS is known for producing distortion artifacts due to the inherent nonlinearities of the measurement system [27]. The H_x-estimator is based on averaging, so it requires quite a long measurement time and also some involvement of the user to set the averaging appropriately. Therefore, for our application, the winner is the sine sweep technique, which will be discussed in detail.

1.4.2.1 Exponential Sine Sweep

In this method, the excitation signal is a sine wave that has a varying frequency over time. This is done in a desired frequency range, typically starting from the lowest to the highest frequency of interest. The word exponential means the following: The instantaneous sweep is the slowest at the lowest frequencies and the fastest at the highest frequencies – it increases exponentially with time. This improves the SNR at low frequencies. Apart from increasing the SNR of the resulting IR by increasing the level of the excitation signal, the increase is also possible by making it longer: Every doubling of the duration of the sweep increases the SNR by 3 dB [28].

The impulse response is obtained by a convolution of the measured output with the (time domain) spectrum inverse. This is known as deconvolution[7]:

$$h_{\mathrm{LRIR}} = y * x^{-1}. \tag{1.7}$$

[7]When a circular convolution is used, it is equivalent to taking the inverse Fourier transform of the quotient between output and input spectra.

x and y stand for the input and output time signals, respectively. However, there is one problem. In the frequency domain, that would mean dividing $\mathcal{F}\{y\} = Y$ by $\mathcal{F}\{x\} = X$. If we then get values of X close to zero (typically at frequencies outside the measured frequency range or at low frequencies where the loudspeaker's output is weak), X^{-1}, and therefore the response would approach infinity. This can be overcome by a so-called packing filter that calculates a regularized inverse in the frequency domain (shown in Fig. 1.10): A small frequency-dependent regularization parameter is used, so that the inversion is applied only in a specified frequency range [29]:

$$X(f)^{-1} = \frac{X^*(f)}{X^*(f)X(f) + \epsilon(f)}. \tag{1.8}$$

$\epsilon(f)$ is the regularization parameter and * denotes a complex conjugate. Eq. (1.8) refers to a specific solution developed by Ole Kirkeby. The inverse Fourier transform of such defined X^{-1} can be then safely use as $x^{-1} = \mathcal{F}^{-1}\{X^{-1}\}$ in Eq. (1.7).

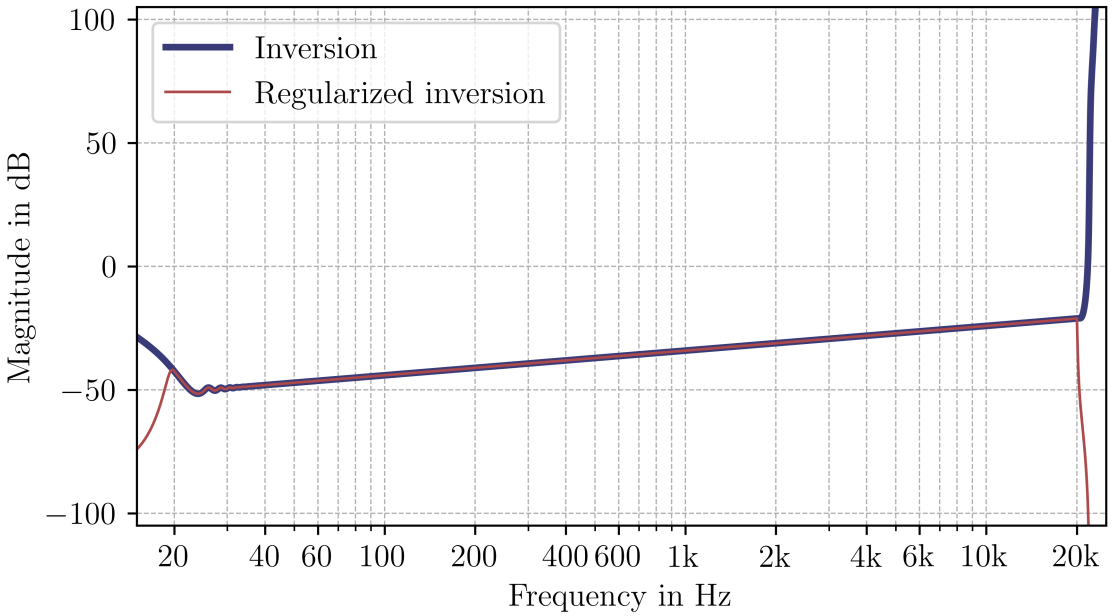


Figure 1.10: Magnitude of the normal and regularized inversion of the FRF of the excitation signal.

1.4.3 Pre-Processing

This section gathers various modifications that might be applied to the measured LRIR to improve the behavior of the compensation filter.

1.4.3.1 Initial Truncation

We said that the focus of the equalization is the direct sound and early reflections, in other words, the early part of the LRIR. So, the response needs to be truncated such that the late reflections are cut out. Based on what has been said in Sec. 1.2.2 this could be up to 20 ms.

If the late part is simply removed, this, in theory, corresponds to multiplication of the LRIR (which is a time signal) with a rectangular window. Multiplication in the time domain corresponds to convolution in the frequency domain, and because the spectrum of the rectangular window is a sinc function, this produces a distortion of amplitude of the spectrum of the truncated LRIR. To reduce this effect, smooth windows (or half windows if it is necessary to truncate only the beginning or the end of the signal) such as *Hanning* window are used instead.

1.4.3.2 Smoothing

The idea behind smoothing is to 'even out' the response in the frequency domain, so that the compensation filter, which is essentially an inverse of the smoothed LRIR, does not overcompensate. First, the response's fluctuations in the frequency domain might be much finer than we can perceive, and second, even a small change in the measurement position can change the position of the peaks and notches. So, limiting those by smoothing makes the compensation filter more robust [1].

It can be implemented both in the time and frequency domain [1], [30]. The definition in the frequency domain is [30]

$$H_{\mathrm{sm}}[k] = \sum_{i=k-m(k)}^{k+m(k)} H[i]\, W_{\mathrm{sm}}\Big[i - k + m(k)\Big]. \tag{1.9}$$

$H_{\mathrm{sm}}[k]$ is a spectral component of the smoothed spectrum H_{sm}, H is the LRRC and W_{sm} is a smoothing window with variable length depending on the frequency (m is frequency-dependent). This equation essentially says that every spectral component $H_{\mathrm{sm}}[k]$ is a sum of the product of the window W_{sm} centered at k and H at corresponding indices i.

This smoothing process can be done for

- a complex H (a so-called *complex smoothing*),
- a magnitude and phase of H separately and then cobined as $H_{\mathrm{sm}} = |H|_{\mathrm{sm}} \arg(H)_{\mathrm{sm}}$.

Of course, if one is only interested in the magnitude spectrum, only the magnitude can be smoothed. The window length is usually a fraction-octave band (typically 1/3, 1/6, or 1/12), or another frequency-dependent window, such as the ERB. If the fraction-octave bands are used, once can find the frequency band limits based on the center frequency f_c as

$$f_{\mathrm{lower}} = f_c\, 2^{-\frac{1}{2n}}, \qquad f_{\mathrm{upper}} = f_c\, 2^{\frac{1}{2n}},$$

where n specifies the fraction ($n = 3$ corresponds to 1/3, etc.).

1.4.3.3 Frequency-Dependent Truncation

In [31], this method is used to remove reflections from semianechoic impulse responses, but I have found that a few commercial software packages and the famous Room EQ Wizard also use it for our purpose. The idea is as follows: Instead of computing the frequency response from the full length of the impulse response, the

impulse response is truncated based on which spectral component is being computed. It can be formulated using the Discrete-Time Fourier Transform (DTFT) as [31]

$$H_{\mathrm{FDT}}[k] = \sum_{n=0}^{K-1} W_k[n]\, h[n]\, e^{-\frac{2\pi i n k}{K}}. \tag{1.10}$$

$H_{\mathrm{FDT}}[k]$ is a spectral component of the frequency-dependent windowed spectrum H_{FDT}, h is the LRIR, W_{k} is a truncation window with variable length depending on the frequency index k, n is an index specifying samples of the LRIR and corresponding truncation window, and K is the length of LRIR. It can also be formulated using the Short-Time Fourier Transform formulation of the DTFT [31].

In [31], the length of W_{k} is adjusted manually for relevant frequency bins k, tweaking it such that the present reflection(s) in the measured impulse response are removed. For our purpose, the length of W_{k} could be formulated as multiples of a period $T = 1/f$.

1.4.3.4 Other Considerations

A normalization of the measured LRRC has to be considered before calculating the compensation filter. Because the compensation filter is an inverse to the measured LRRC, we should normalize the measured LRRC to be close to 0 dB. One can do this by calculating the mean or RMS of the magnitude spectrum over a certain frequency range and normalizing it to 0 dB. For the mean:

$$H_{\mathrm{norm}} = \frac{H}{\frac{1}{n}\sum_{k=m}^{m+n} |H[k]|}, \tag{1.11}$$

where k represent a frequency bin and n, $m+n$ coresponding frequency range limits chosen by the user. If other value than 0 dB is required, one can multiply H_{norm} by g (in linear scale) or by specifying it in dB as G:

$$g = 10^{G/20}.$$

Another thing to consider is limiting the compensation frequency range: Outside this range, we would like not to compensate anything – leave the compensation filter at 0 dB. At the same time, the transition should be smooth between the band where the compensation happens and outside this band. One could achieve it by following the formula:

$$H_{\mathrm{limited}} = H + W\,(H - g). \tag{1.12}$$

W is a weighting function consisting of a flat passband with tapers (e.g., Hanning) on both ends, and g is the (linear) value at which the taper ends will settle. It could be given by G using the formula above. Unfortunately, this can be used only for the magnitude. When considering the complex spectrum, I was not able to find or come up with any formula or approach that yielded satisfactory results.

Lastly, one should consider the audio equipment's frequency response (if it is not flat, which is typically not the case). When it comes to the microphone, one should compensate for its response so that the correction reflects the reality. When it comes

to the loudspeaker, you might argue that since we want to correct the loudspeaker-room impulse response, it does not make sense to correct for it, and that is true. However, its limitation will always be the low-frequency roll-off; therefore, we should compensate for this with a high-pass filter so the compensation filter does not force the loudspeaker to reproduce something it is not physically able to.

1.4.4 Prototype LRIR

If the LRIR is measured only in one position, its pre-processed version is used for the computation of the compensation filter, and therefore no prototype is needed. This approach is called SISO (Single-Input Single-Output). If more than one measurement is performed, we need to find a way to determine a representative response, a *prototype*, that will later be used for the compensation. Since the compensation filter is based on multiple measurement positions (multiple outputs), this approach is called SIMO (Single-Input Multiple-Output). By doing so, we not only make the compensation more robust but also enlarge the equalization zone [1].

There are different approaches to determining the prototype response. In [32], a mean average, a median, an RMS-average, a min-max prototype, and a fuzzy c-means prototype were compared. The min-max prototype is derived by minimizing the maximum error between the computed prototype and the measured responses. The fuzzy c-means algorithm extracts 'the common patterns of the room magnitude responses by means of c centroids' [32]. It was shown that all tested methods performed similarly, but the mean average yielded the best results. Some of those methods have been implemented in commercial products [33].

However, [6] argues that a correction that is good for the mean response doesn't necessarily need to be good for any of the measured responses. What's more, if the compensation filter is of the mixed-phase type (see the next section), one has to be careful about the positions of zeros (dips in the frequency domain). Fig. 1.11 shows zeros of 18 different loudspeaker-response transfer functions measured in different positions along a line. We can see that there are three zeros outside the unit circle (excess-phase zeros) – one around 135 Hz (left picture), another around 185 Hz, and the third just above 200 Hz (right picture). The one on the left is fairly stable, while the others move significantly with changes in position. According to [6], [34], it is safe to invert only fairly stable zeros outside the unit circle. The bigger the move, the higher the pre-ringing that the compensation filter will cause in the listening positions. To decide whether a zero should be inverted or not is based on the allowed pre-ringing error [34].

1.4.5 Compensation filter

To be able to compute the compensation filter, two things are needed: the pre-processed (prototype) LRIR and a target filter/response. The target filter is the desired filter we would like to obtain when the compensation filter is applied to the LRIR.

There are two approaches to the compensation filter: a minimum-phase and a mixed-phase. The former only corrects for the magnitude response, while the latter

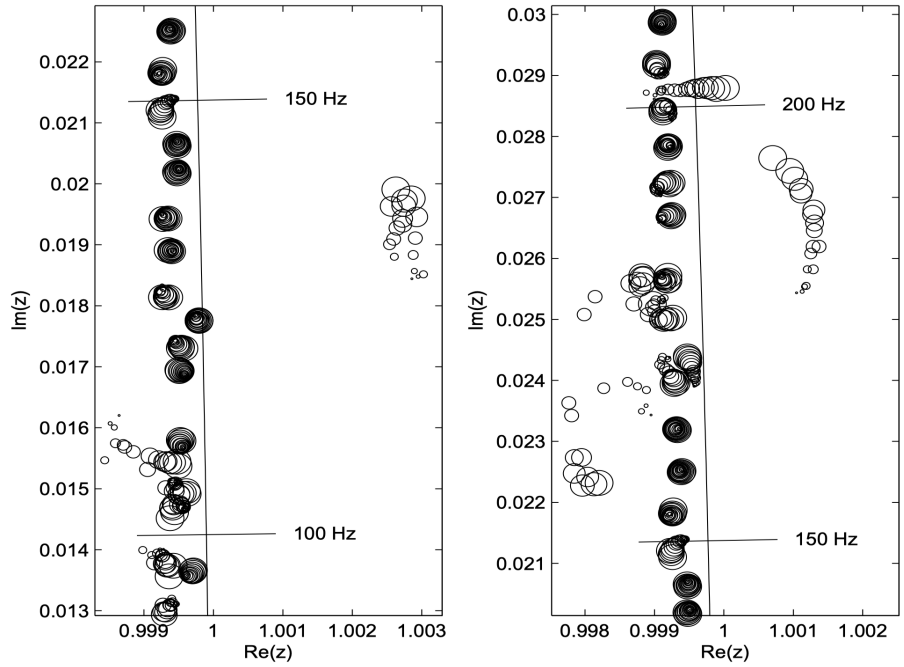


Figure 1.11: Zoomed segments of the complex plane where zeros (circles) of 18 different loudspeaker-response transfer functions measured in different positions along a line (2.5 cm separation), where different radii distinguish various measurement positions. Adopted from [34].

compensates for both the magnitude and the phase.

1.4.5.1 Target Response

Sometimes also referred to as the target curve, it is the response we target; the resulting response after the compensation. Before, we said that we aim for $\delta(t)$ in the time domain. In frequency domain, we might expect to desire a flat magnitude response. However, discussions on the internet suggest that if the target curve is flat, the sound is perceived as too 'bright'. It seems that things are more complicated.

The Recommendation ITU-R for listening conditions of control rooms [14] specifies relative limits for the magnitude frequency response when the system is excited by a pink noise. The graph is shown in Fig. 1.12. The broadband magnitude spectrum of pink noise decreases by 3 dB per octave, so we need to account for this if we want to use it as an indication of the magnitude target response. Adding those −3 dB, it roughly suggests a decrease of the target response by 3 dB per octave.

[35] explores this in great detail and discusses the reasons behind this tilt. Even though the on-axis magnitude response of a loudspeaker is flat, typically, the off-axis responses are tilted: the larger the angle, the steeper the tilt. When a curve for the sound power is calculated (that would correspond to averaging all the radiation angles), the tilt is somewhere between −2 and −2.5 dB per octave. That is quite close to those −3 dB.

However, it is even more complicated. The limits given in Fig. 1.12 are for

a steady-state (the response is calculated from a long measurement of noise). For our scenario, the situation is different because in the compensation, we assume only the direct sound and early reflections. So, how is it for our case, then? [35] argues that the target response should be flat for the direct sound. But what about those early reflections we take into account? When the average magnitude response is calculated in an anechoic environment for the pressure of angles ±30 degrees (roughly assuming angles for the early reflections), the average tilt is about −1 dB per octave. So, this could be our desired magnitude target curve. It also seems to agree with some commercial software solutions. It might seem that 1 dB is negligible but when we take into account the full audio spectrum from 20 Hz to 20 kHz, that is roughly 10 octaves – 10 dB difference between the lowest and highest frequencies. Tolerance limits in Fig. 1.12 are manually modified according to our assumptions and shown in Fig. 1.13.

So far, we have only discussed the magnitude. What about the phase, though? [33] mentions that the target filter should be ideally close to a unit pulse; concentrated in time. Therefore, the response could be a minimum-phase filter.

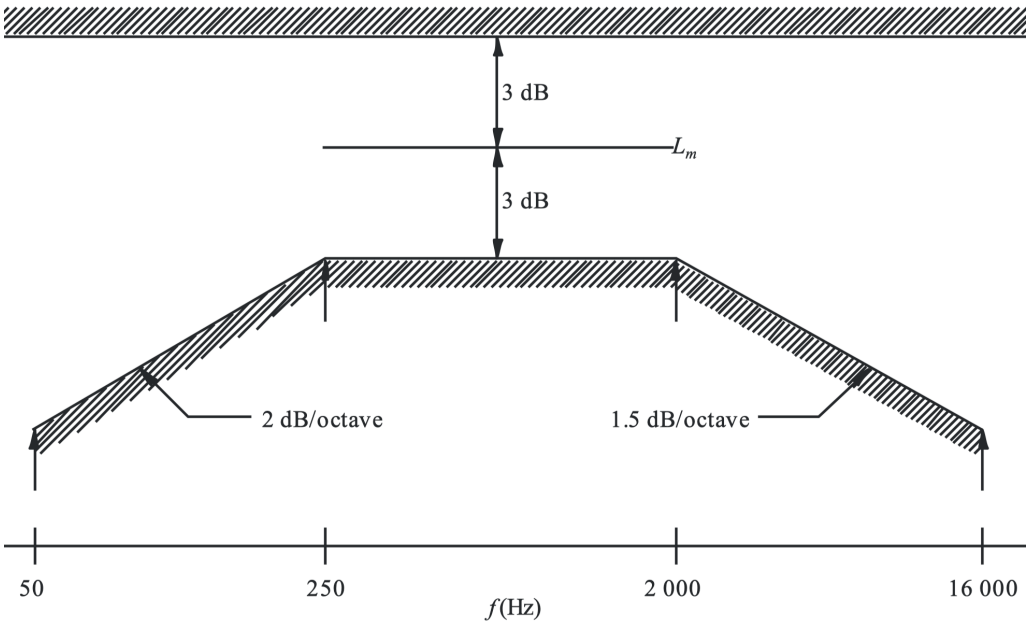


Figure 1.12: Tolerance limits for operational room response curve when excited by a pink noise; the Recommendation ITU-R. Adopted from [14].

1.4.5.2 Minimum-Phase Approach

This is the easier approach to implement and is probably also the more common one. We said in Sec. 1.1.2.3 that only one minimum-phase filter exists for a given magnitude response. That means that we can only work with the magnitude response in the whole process and in the end we just add the phase response, which is unique under the assumption that the required compensation filter is minimum-phase.

This is based on the idea that a mixed-phase response $H_{\mathrm{LRRC}}[z]$ (yes, the LRRC is a mixed-phase filter) can be factored out into a minimum-phase term $H_{\mathrm{LRRC,MP}}[z]$

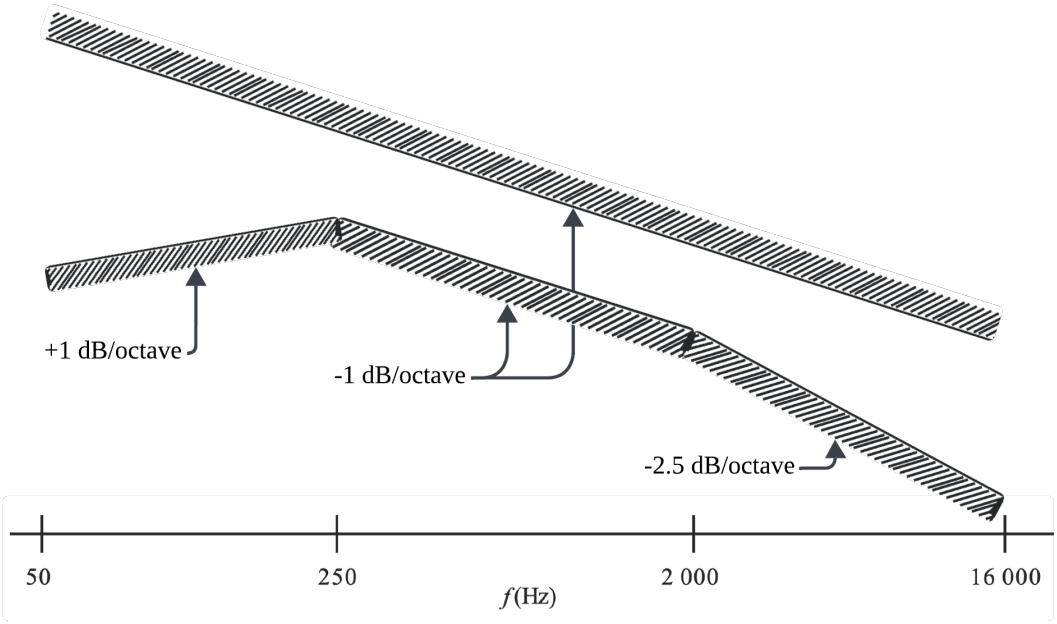


Figure 1.13: Possible target curve limits based on the tolerance limits for the room response curve in Fig. 1.12 and our assumptions about the direct sound and early reflections.

and a causal stable allpass filter $H_{\mathrm{AP}}[z]$ [1], [5]:

$$H_{\mathrm{LRRC}}[z] = H_{\mathrm{LLRC,MP}}[z]\, H_{\mathrm{AP}}[z]. \tag{1.13}$$

Thus, we correct only for the minimum-phase part $H_{\mathrm{LLRC,MP}}[z]$. $H_{\mathrm{AP}}[z]$ is sometimes referred to as the *excess phase*. Fig. 1.14 shows the original and minimum-phase version of the measured LRIR. To obtain the $H_{\mathrm{LLRC,MP}}[z]$ from $H_{\mathrm{LRRC}}[z]$, all zeros outside the unit are has to be reflected inside [5].

There are many advantages to this approach. First, it is simpler to work just with the magnitude response than with the full response that assumes also the phase response and the time domain. The minimum-phase filter has the least possible delay/latency, which can be beneficial for many applications (for the implementation in our portable monitors, it definitely is). Thanks to that, the compensated LRIR ($h_{\mathrm{LRIR}} * h_{\mathrm{comp}}$) has only post-ringing[8] and no pre-ringing. Pre-ringing is an undesired artifact as it has to be way shorter than post-ringing in order not to be heard (see Sec. 1.3.1). Fig. 1.15 shows an impulse response with pre- and post-ringing. Lastly, there are no stability issues, as the minimum-phase filter is stable by definition.

But there is one big disadvantage to this approach. As we already said, the room impulse responses are in general mixed-phase filters. Some frequency regions are minimum-phase (usually the low frequency range), but others are not [36]. To correct such a filter, in theory, one also needs a mixed-phase filter. If a minimum-

[8]Post-ringing are components in the IR that are present after the main peak, while pre-ringing is a name for such components present before the main peak [33].

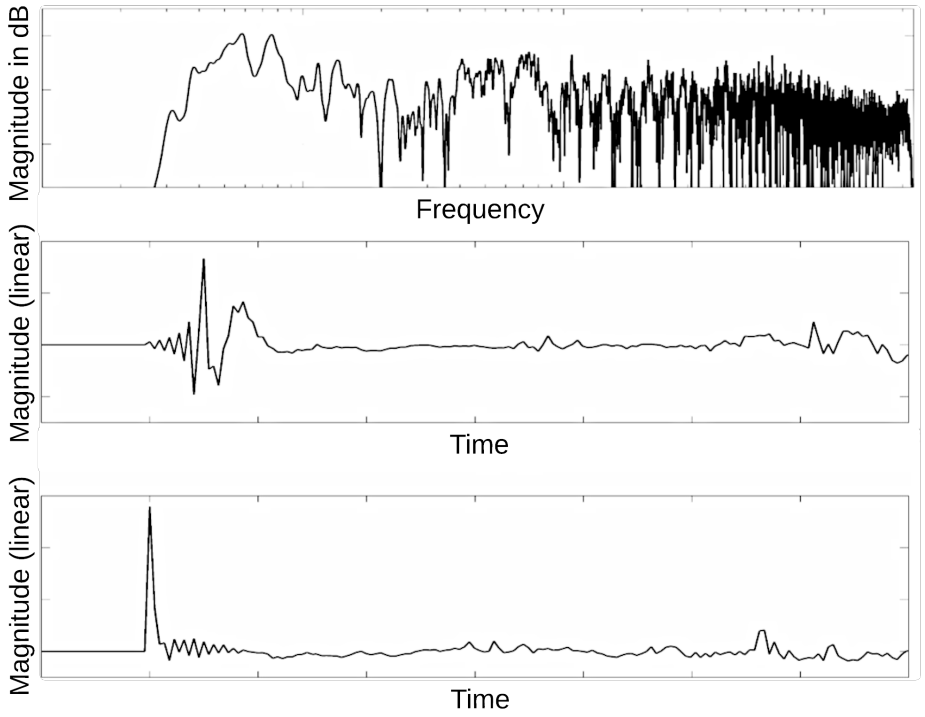


Figure 1.14: Magnitude of LRRC (top), corresponding LRIR (middle), and its minimum-phase version (bottom). Adapted from [33].

phase equalizer is used instead, it can 'inject' energy at the wrong times – the allpass filter $H_{\mathrm{AP}}[z]$ from Eq. (1.13) that is left out after the compensation can have much worse IR than the complete system [6]. For example, if a dip is caused by an early reflection, such a dip can't be simply boosted by increasing the level at that frequency, neglecting the phase response. If the loudspeaker is boosted at that frequency, the direct sound will be boosted, as well as the reflection that is out of phase, and the resulting boost will probably be way smaller than expected. Despite this, minimum-phase equalizers have been widely used as compensation filters for the LRRE, also as commercial products.

There are several techniques how one can implement the minimum-phase compensation filter. The first and obvious one is to take the inverse of the magnitude LRRC:

$$|H_{\mathrm{comp}}| = \left| \frac{H_{\mathrm{targ}}}{H_{\mathrm{LRRC}}} \right|, \tag{1.14}$$

and then calculate the minimum-phase filter based on this magnitude response. This is equivalent to calculating the minimum-phase version of LRRC and then taking its inverse. That is called the frequency-domain deconvolution.

Other methods have been proposed to implement the minimum-phase compensation filter. For example, LPC analysis, where the LRRC is modelled with a minimum-phase all-pole filter (it models spectral peaks, but cannot precisely capture notches), or homomorphic filtering, which is, however, subordinate to LPC, as the filter has to be very long and it is highly sensitive to source/receiver position.

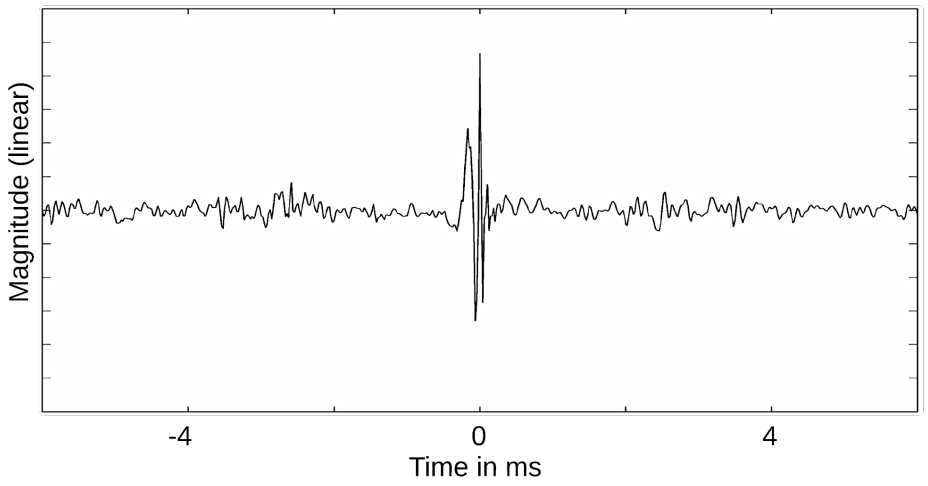


Figure 1.15: IR with pre-ringing (to the left) and post-ringing (to the right). Adapted from [33].

Details can be found in [1].

Another interesting realization uses second-order sections (SOS) (=biquad filters), which are minimum-phase and have infinite impulse response (IIR). SOS peak filters are connected in series such that their sum results in having $|H_{\text{comp}}|$ as in Eq. (1.14). A big advantage of this approach is that it is very easy to implement, because such filters are typically present in the DSP chips of active loudspeakers and pre-/integrated amplifiers. Another advantage is that one has quite a big control over the correction process (i.e., not correcting for small deviations). [37], [38]

1.4.5.3 Mixed-Phase Approach

To use a mixed-phase filter for correcting the mixed-phase LRIR is the 'correct' way as it takes into account the phase as well. However, an inverse filter (the compensation filter) of a mixed-phase filter (with poles outside the unit circle) must be non-causal to ensure stability. This makes sense: Let's use the example with the early reflection that causes a dip in the magnitude response. If we were to correct the direct sound such that the early reflection contribution is cancelled out by the compensation filter, we would have to send a pulse slightly before the main pulse because the reflected sound travels a longer path than the direct sound. It is not possible to do that if the main pulse starts at $t = 0$. Therefore, we need to introduce a delay [1]. Then the compensation filter becomes causal.

The disadvantage of the mixed-phase compensation filter is that it can introduce not only post-ringing but also pre-ringing in the compensated response. It happens when the compensation filter is 'less than perfect'. If not masked, pre-ringing is easily noticeable because they don't occur naturally in the physical systems' impulse responses. [33]

The pre-ringing can be different with a slight position change, and it can be limited by correcting only the 'fairly stable' zeros based on the pre-ringing error criterion as discussed before. Another disadvantage is the introduced delay if the

latency of the playback system plays a role.

There are three implementation techniques of the mixed-phase compensation filter that are widely discussed in the literature [1]: a frequency-domain deconvolution, a least mean squares (LMS) method, and the minimum-phase approach in combination with an all-pass filter.

The frequency-domain deconvolution was already briefly discussed. The compensation filter is calculated as follows:

$$H_{\text{comp}} = \frac{H_{\text{targ}}}{H_{\text{LRRC}}}. \tag{1.15}$$

Before doing so, the modeling delay has to be added to the LRIR. If the LRRC is not adequately pre-processed, such that its inversion could result in excessive gains, a deconvolution with regularization can be used (see Eq. (1.8), for which H_{LRRC} would be X).

LMS is a well-known optimization algorithm that is used to find a desired filter based on the least mean square of the error, the difference between the desired and the actual signal. For this application, the suitable version is the x-filtered LMS algorithm [39] or even better, the x-filtered NLMS, where the step size adapts to the input's power, which tends to converge faster (N stands for normalized). This algorithm can also be used for joint optimization: The individual measured and pre-processed LRIRs can be input directly without constructiong the prototype response. Details can be found in [39].

The last technique assumes the minimum-phase/allpass filter factorization from Eq. (1.13). If we compensate not only for the minimum-phase term but also for the allpass term, we end up with correcting with a mixed-phase compensation filter. This technique was successfully used in [34].

1.4.5.4 Stereo Reproduction: Timing is Important

If two sound sources that are equally far away from a listener play the same sound at the same output level, the listener will perceive a so-called *phantom* (virtual) source exactly in between those speakers. If we move one speaker so that the two distances to the listener differ, the phantom source will move. That means if one listens to a stereo system, and the distance of both speakers to the listener is not the same, the perceived sound stage will shift.

The music we listen to was mixed in a professional studio where the monitoring system was probably perfectly aligned. We should therefore also aim to achieve perfect distance alignment of our speakers. HiFi enthusiasts often use a tape measure for this purpose; the core audiophiles even opt for laser measurement tools. However, it is possible to use an allpass filter to introduce a time delay to the speaker that is closer than the other one.

1.4.5.5 SOS Cascade Approach: Full Theory

As mentioned before, in this technique, the compensation filter consists of a cascade of individual SOS, IIR peak filters $H_{\text{peak},1}$, $H_{\text{peak},2}$, ..., $H_{\text{peak},n}$:

$$H_{\text{comp}} = H_{\text{peak},1}\, H_{\text{peak},2} \ldots H_{\text{peak},n}. \tag{1.16}$$

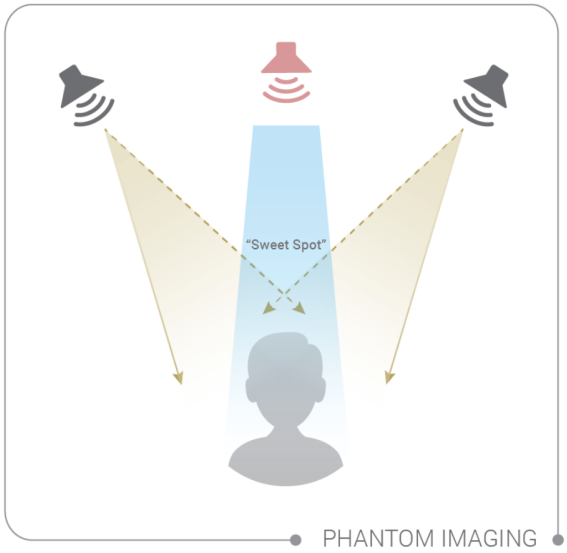


Figure 1.16: Stereo sound reproduction layout, the red speaker represents a phantom source. Adopted from [40].

H_{comp} is found by minimizing a mean error e between the desired target response and the filtered loudspeaker-room response. In dB [37]:

$$e_{\text{dB}} = H_{\text{target,dB}} - (H_{\text{LRRC,dB}} + H_{\text{comp,dB}}). \tag{1.17}$$

The whole concept can be split into three parts: the division of the frequency spectrum into bands, the initial parameter guess for each band, and optimization.

Part 1: The magnitude response $H_{\text{LRRC,dB}}$ is split into bands. Each band is reserved for one peak filter. According to [37], [38], the bands are split based on zero-crossings of the magnitude response $H_{\text{LRRC,dB}}$. This can work well, but it can happen that the response within two zero-crossings has significant peaks or dips and could benefit from more than just one peak filter. That is why I propose checking the slope of the response, too. First, the zero-crossings are determined (by a sign change between two consecutive values). Then, the mean response value between consecutive zero-crossings is computed. If positive, after every combination of monotonically increasing and decreasing trend, a new crossing is marked. If the mean is negative, the order of the trends is swapped: monotonically decreasing is followed by the increasing one, and a new crossing is marked. Lastly, the areas above/below the target curve are calculated and sorted in descending order (larger area = higher priority). By this, one can limit the number of filters used (the hardware might allow only for a limited number N) and ensure that the filters will correct the N most significant bands.

Part 2: Initial parameters of the peak filters are guessed. These are

- central frequency f_c: a geometric mean between crossings,
- gain G_{dB} calculated as

$$G_{\text{dB}} = -H_{\text{LRRC,dB}}(f_c) + H_{\text{target,dB}}(f_c), \tag{1.18}$$

- and quality factor Q calculated as

$$Q = \frac{f_c}{B_{\pm 3\ \mathrm{dB}}}, \tag{1.19}$$

where $B_{\pm 3\ \mathrm{dB}}$ stands for bandwidth – the difference between the upper and lower frequencies at which the response's magnitude reduces (for peaks) or increases (for dips) by 3 dB from the peak/dip value given by G_{dB}. The higher the Q, the narrower the filter is.

Part 3: [37] suggests a direct (random) search method. After the initial guess of a band, the three parameters are randomly varied by up to 5 %. Then, e_{dB} from Eq. (1.17) is evaluated. If it is smaller, the new values are adopted; otherwise, they are discarded. After 100 to 200 iterations, the results will be very close to the global minimum [37].

According to [37], this optimization is performed sequentially, band by band (starting with the largest band area). The problem with such an approach is that the interaction between the individual filters is not fully accounted for. That's why it is suggested to do a post-optimization: one or more rounds of the optimization process are added, but now with bands ordered according to frequency.

[38] points out that the risk of the proposed direct search method is that if the initial guess is not close to the values that minimize the error, there is a potential risk of falling onto a local minimum. Instead, the Rosenbrock method is suggested. It keeps track of past search results and adapts the search direction.

At this stage, we have datermined N filters uniquely specified by 3 parameters: f_c, G_{dB} and Q. From these parameters, 6 coefficients: a_0, a_1 and a_2, determining the poles of the filter, and b_0, b_1 and b_2, determining the zeros of the filter, can be calculated by analytical formulas [41]:

$$\begin{aligned} a_0 &= 1 + \tfrac{\alpha}{A} & \qquad b_0 &= 1 + \alpha A \\ a_1 &= -2\cos(\omega_0) & \qquad b_1 &= -2\cos(\omega_0) \\ a_2 &= 1 - \tfrac{\alpha}{A} & \qquad b_2 &= 1 - \alpha A \end{aligned}$$

where

$$A = 10^{G_{\pm 3\mathrm{dB}}/40}, \qquad \omega_0 = 2\pi\frac{f_c}{f_s}, \qquad \alpha = \frac{\sin(\omega_0)}{Q}.$$

f_s stands for the sampling frequency. These can be directly plugged into a difference equation which computes the current output sample (sample on the filter output) based on past and present input samples and past output samples. For our biquad filter, the difference equation is

$$y[n] = \frac{b_0 x[n] + b_1 x[n-1] + b_2 x[n-2] - a_1 y[n-1] - a_2 y[n-2]}{a_0}. \tag{1.20}$$

Note: In practise, when these coefficients are required, it is usually assumed that they are normalized by a_0: $a_{0,\mathrm{norm}} = a_0/a_0 = 1$, $a_{1,\mathrm{norm}} = a_1/a_0$ and so on.

1.4.5.6 *X*-Filtered NLMS Approach: Full Theory

The x-filtered normalized least mean squares algorithm is an instantaneous gradient descent method, in which the desired filter is adapted only based on the error at the current time/iteration. It is an adjusted version of the LMS algorithm that was invented by B. Widrow and T. Hoff in 1960.

A block diagram of the algorithm is shown in Fig. 1.17: The delayed input signal $x[n]$ (typically noise) denoted as $d[n]$ (desired signal) is compared with the input signal $x[n]$ that is filtered by the LRIR and the current estimate of the compensation filter h_{comp}. It is denoted $\hat{d}[n]$ as the estimation of the desired signal. The compensation filter is updated based on the instantaneous value of the error $e[n]$, which is the difference between $d[n]$ and $\hat{d}[n]$ [39]:

$$\begin{aligned} e[n] &= d[n] - \hat{d}[n] \\ &= d[n] - \mathbf{x}_{\text{filt}}^T[n]\, \mathbf{h}_{\text{comp}}[n], \end{aligned} \tag{1.21}$$

where $\mathbf{x}_{\text{filt}} = x * h_{\text{LRIR}}$, thus the name x-filtered. $\mathbf{x}_{\text{filt}}^T[n]$ is a row vector (as T stands for a transpose) in the n-th iteration:

$$\mathbf{x}_{\text{filt}}^T[n] = \left[x_{\text{filt}}[0], ..., x_{\text{filt}}[n]\right],$$

and $\mathbf{h}_{\text{comp}}[n]$ is a column vector in the n-th iteration of chosen lenght N. Note: $\mathbf{x}_{\text{filt}}^T[n]$ is zero-padded to match the lenght N.

$$d[n] = (x * h_{\text{targ}})[n - \Delta], \tag{1.22}$$

where Δ denotes number of samples.

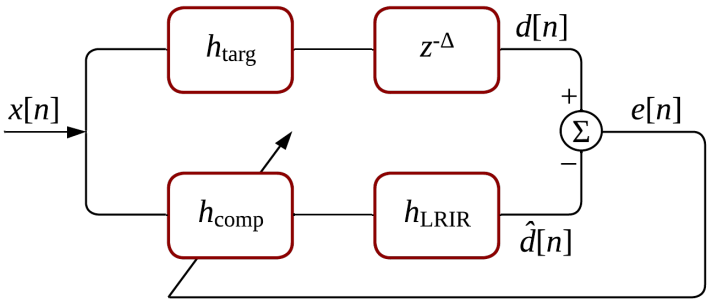


Figure 1.17: Block diagram of x-filtered (N)LMS algorithm. Inspired by [39].

$\mathbf{h}_{\text{comp}}$ is updated in every iteration as

$$\mathbf{h}_{\text{comp}}[n+1] = \mathbf{h}_{\text{comp}}[n] + \mu\, e[n]^*\, \mathbf{x}_{\text{filt}}[n], \tag{1.23}$$

where * denotes a complex conjugate and

$$\mu = \frac{\mu_{\text{scale}}}{\mathbf{x}_{\text{filt}}^T[n]\, \mathbf{x}_{\text{filt}}[n] + \delta}. \tag{1.24}$$

μ_{scale} in typically chosen within the range $(0, 1]$. δ is a small regularization factor to prevent division by 0. N in NLMS stands for *normalized* and it denotes that the

step size μ is scaled by the instantaneous input power ($\mathbf{x}_{\text{filt}}^T[n]\,\mathbf{x}_{\text{filt}}[n]$), which ensures stability and faster convergence. The delay of Δ samples is intoroduced to a) match the possible delay of the initial peak in h_{LRIR} and to b) make the h_{comp} causal as already discussed before[9].

1.5 Implementation

The implementation was done in Python (v. 3.13.2) using the Miniconda distribution on MacOS. The main libraries used were NumPy (v. 2.2.4), SciPy (v. 1.15.2), and Matplotlib (v. 3.10.1). For measurements, the library PyFar (v. 0.7.2) was used primarily for convenience (signal definition, manipulation, and plotting).

I have implemented three approaches that are fully explained later on in this section. But first, I would like to present details about the LRIR measurement and some other findings and practical limitations.

1.5.1 Measurement of LRIRs

As mentioned before, the exponential sine sweep was used as the excitation signal for the loudspeaker system. The measurement procedure was created and executed using Python. *Focusrite Scarlett 4i4 3rd Gen* was used as the audio interface, and *Superlux ECM999* as the measurement microphone. A loopback cable was used for recording the excitation signal. This recorded signal was then used as the input signal for the impulse response computation. By doing that, the round-trip latency of the measurement chain is cancelled out.

1.5.1.1 Signal-to-Noise Ratio

It would be practical to come up with a metric that would provide us with information about the signal-to-noise (floor) ratio so we could decide if the measurement is 'good enough' and therefore can be used for further processing, or that we should increase the level or the length of the excitation signal. Because of that, as part of the measurement, I have also recorded 5 s of background noise to obtain information about the noise floor.

My initial idea was to compare the magnitude spectra of the sweep and background noise (possibly their smoothed versions) and then test the smallest possible SNR (their difference in dB) at a particular frequency over a frequency range of interest that would give reasonable results. This turned out not to be a good idea because when the measurement position was exactly at a pressure node of a room mode at that frequency, even for very long sweeps with a high amplification level, the SNR would be around 0 dB. Thus, it could not serve as any indicator at all.

Therefore, I decided to use a different metric: I defined the SNR as the ratio of the RMS values of the sweep signal and background noise:

[9] It is not that if the delay is not introduced, h_{comp} would be acausal; the acausal part that is before the main peak would be cut and the compensation would not be good.

$$\mathrm{SNR} = 20 \log \left(\frac{\sqrt{\frac{1}{n} \sum_{i=1}^{n} s_{\mathrm{sweep},i}^2}}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} s_{\mathrm{noise},i}^2}} \right). \tag{1.25}$$

Even though with this method one does not know exactly what the SNR is at a particular frequency, I believe it could be used as a relevant parameter.

Fig. 1.18 compares magnitude spectra of measured LRIR by using 10 s exponential sweep length and varying sweep signal level, resulting in different SNR values. We can see that above a certain SNR, the curves stop changing. That should be a clear indication that the response is valid. We can see that in this case, the SNR is roughly 9 dB or higher. Fig. 1.19 shows exactly the same, but when the curves are smoothed in 1/3-octave bands. If we consider the minimum-phase compensation approach where we care only about the smoothed magnitude, this curve might be enough to consider – neglecting all the small deviations in the measured response. We can see that even for smaller SNR values, the results are good. It is surprising for me that above 100 Hz, even for the two lowest SNRs where the sweep signal was practically buried in the background noise, the smoothed curve is very good.

It has to be said that the SNR values are only orientational. When testing this, I was increasing the sweep level by 3 dB for each subsequent measurement (not all of them are plotted), but the resulting SNR didn't closely match the expected increase. It could even happen that 3 dB higher excitation level would result in no SNR improvement at all. It could be due to the variable background noise level and the too-short recording length of the background noise.

1.5.2 Practical Limitation

It turned out there are some practical limitations that make some parts of the theory impossible to use.

1.5.2.1 Initial Truncation

We said earlier that for small listening rooms, it is common to consider early reflections up to 20 ms. However, it appeared that truncating the LRIR after 20 ms is problematic. Fig. 1.20 shows the truncated impulse response for two different truncation lengths and Fig. 1.21 their corresponding magnitude spectra. We can see that at mid and high frequencies, the 20 ms version behaves as expected – limiting the depth of peaks and dips as the late reflections are removed. However, at least up to 400 Hz, the response does not follow the trend of the untruncated (raw) one. It is probably because the window is too short relative to the periods T of low frequencies. If we consider our LF limit to be 40 Hz, then the 20 ms window corresponds only to $0.8T$. By trial and error, I have come up with the shortest window that seems to capture low frequencies well, and that is 80 ms. That corresponds to $3.2T$. The issue now is that the truncated response is not much different from the untruncated one, which makes this step questionable. On the other hand,

a shorter impulse response will save computation time even with the consideration of zero-padding (next paragraph).

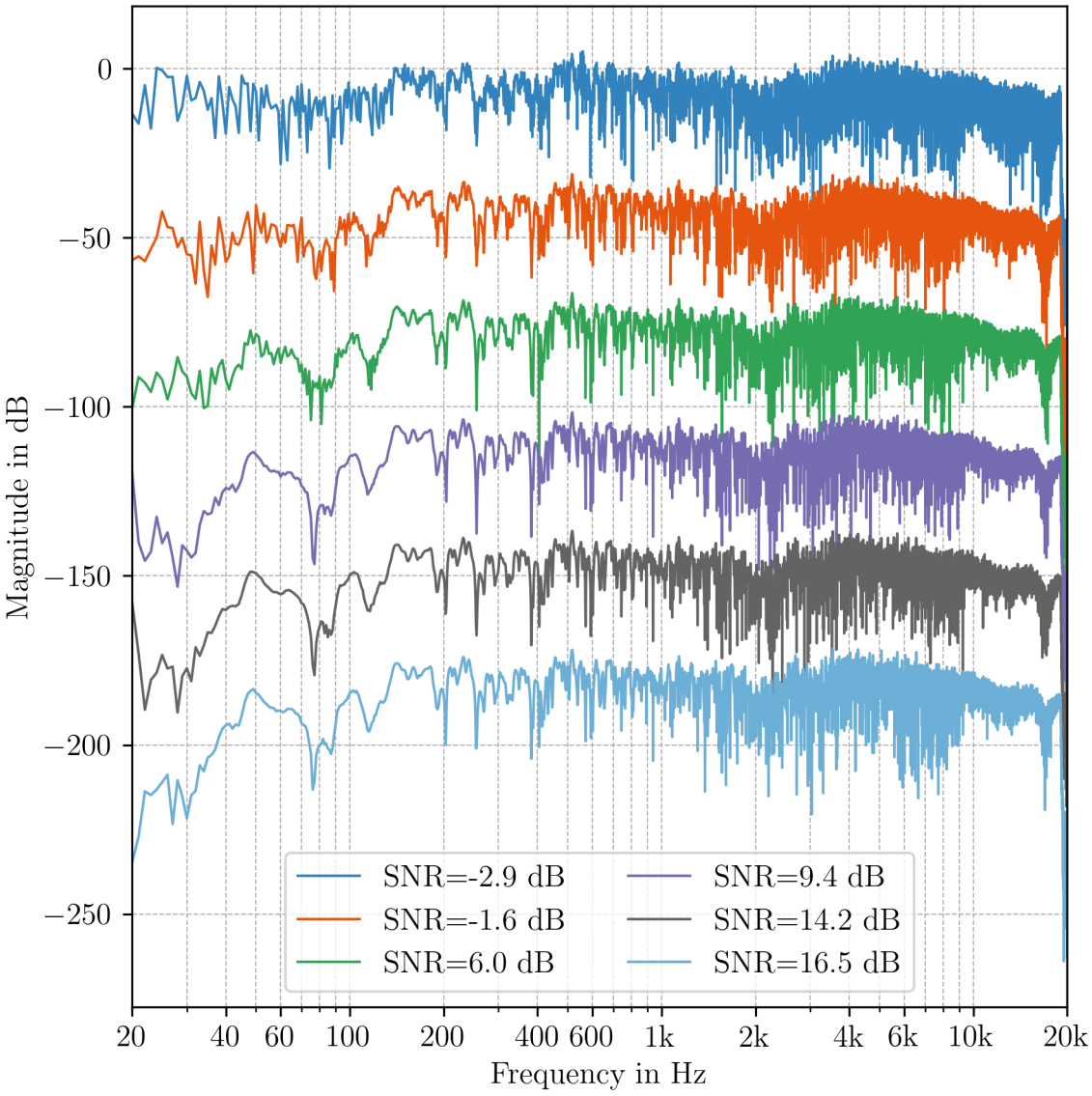


Figure 1.18: LRRC from impluse response measurements with different values of SNR.

After the truncation, one has to zero-pad the impulse response to obtain the desired frequency resolution, which is dependent on the impulse response length. The frequency step Δf is calculated as

$$\Delta f = \frac{f_s}{N}, \tag{1.26}$$

where f_s and N are the sampling frequency and length of the impulse response (or its double-sided spectrum), respectively. Even for the 80 ms version, Δf would be 12.5 Hz, which is insufficient at the lowest frequencies. Δf for the shown responses is 2 Hz, which corresponds to the length of 0.5 s.

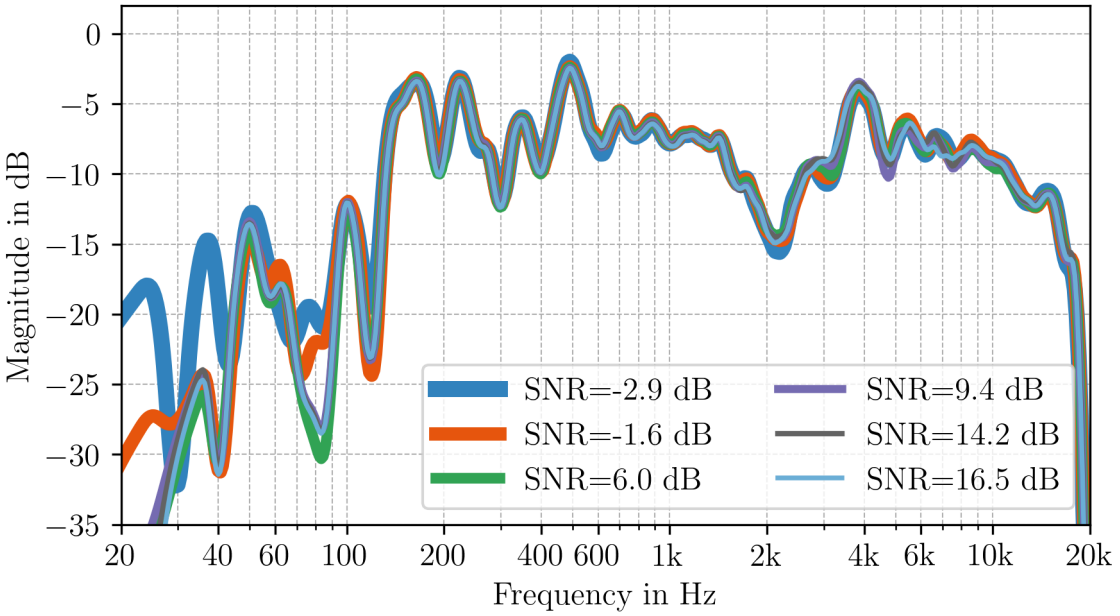


Figure 1.19: LRRC from impluse response measurements with different values of SNR, 1/3-octave smoothing.

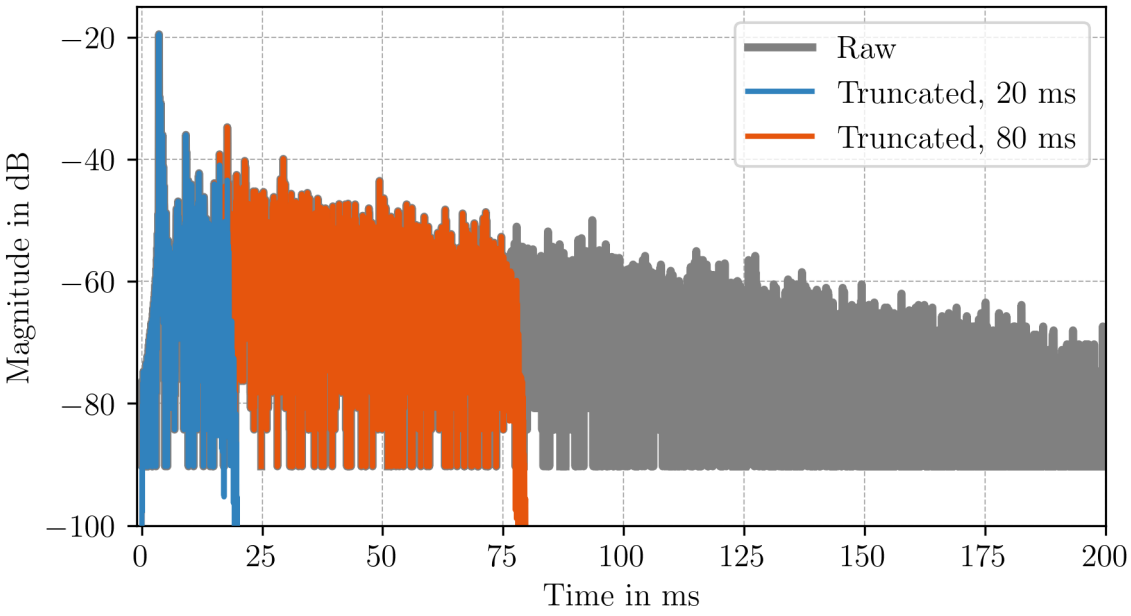


Figure 1.20: Different truncation lengths of the LRIR.

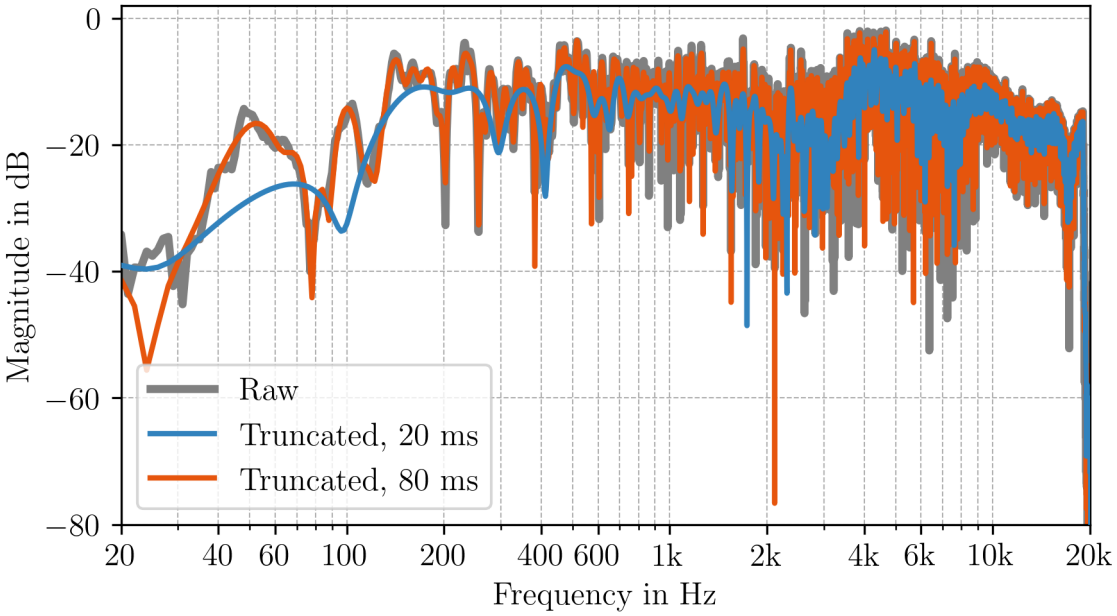


Figure 1.21: Magnitude spectra of the LRIR of different truncation lenghts.

1.5.2.2 Methods for 'Evening Out' the LRRC

In the theory part, we mentioned three methods: complex smoothing, separate magnitude and phase smoothing, and frequency-dependent truncation (FDT). It turned out that some of the methods work better and some worse.

Fig. 1.22 shows a comparison of a magnitude LRRC and its phase delay (which is essentially just a scaled phase response for easier analysis, see Sec. 1.1.3) between the measured (only truncated) response and the mentioned methods. Smoothing was done in 1/3-octave bands and FDT with window lengths corresponding to $4T$, which was experimentally chosen as it followed the trend of the 1/3-octave smoothing the closest.

If we focus on the magnitude response, from the lowest frequencies up to approx. 600 Hz, all methods perform very similar. The only difference is that the separate magnitude smoothing reduces the depth of the notches. At higher frequencies, the situation is different. The separate magnitude smoothing appears to best approximate the measured response. This makes sense, as it is essentially an average of the measured response. The FDT one behaves very similarly apart from the frequency range of 2 to 4 kHz. The reason is: The magnitude of the FDT response is dependent on the sound energy that falls within the specific time window of a particular frequency bin. For illustration, at 3 kHz, the window length is 1.3 ms (assuming the time interval of $4T$) and the reflected sound at this frequency is reflected not as strongly to the measurement position as other frequencies.

Let's look at the magnitude of the complex smoothed response. We can see that, as the frequency increases, the curve drops significantly. [42] mentions this problem and outlines the cause: the real and imaginary parts of the response oscillate, and when the window function stretches over one or multiples of phase cycles of the oscillations, thanks to the summation in the Eq. (1.9), they will cancel out. Thus,

[42] suggests the separate magnitude and phase smoothing instead.

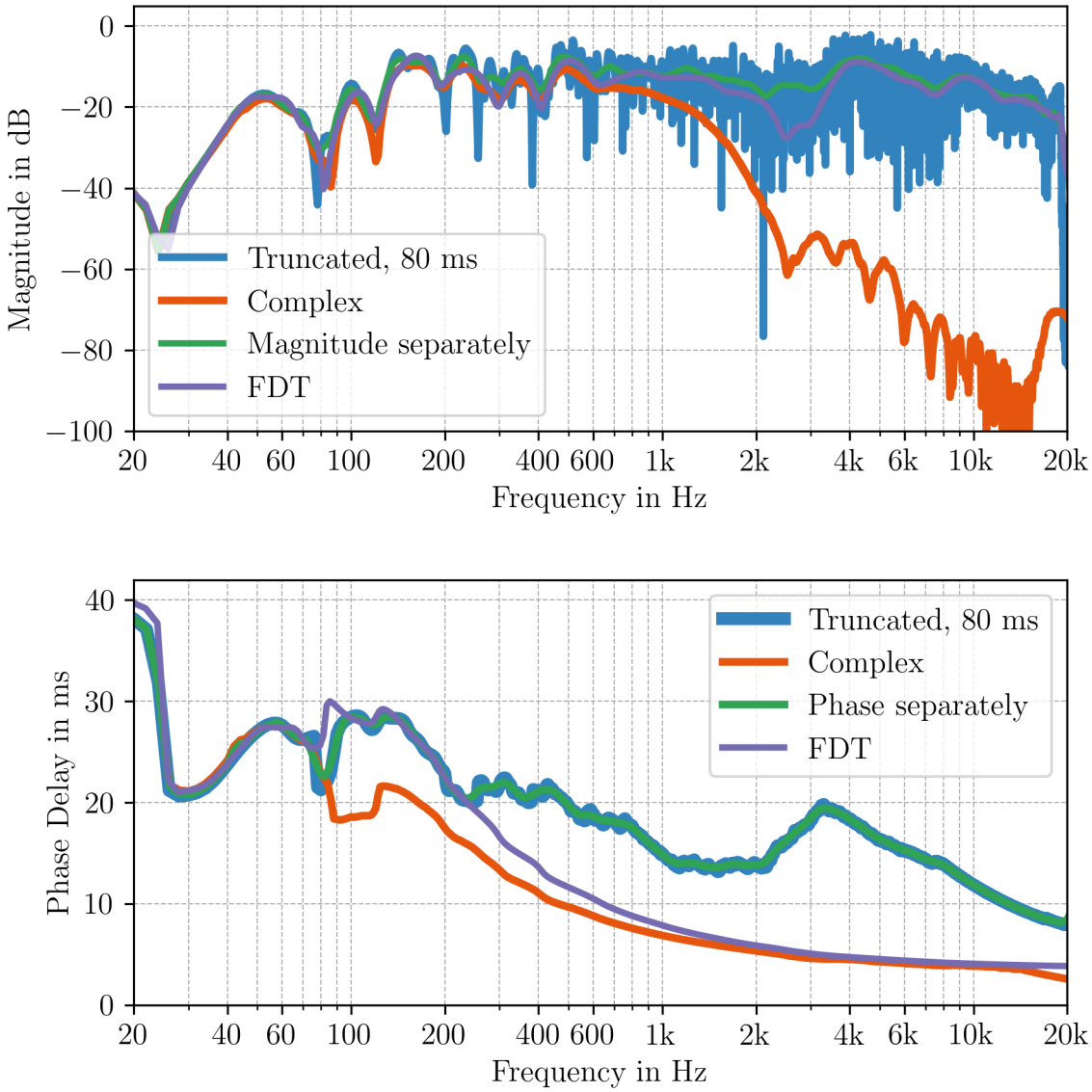


Figure 1.22: Comparison of magnitude response and phase delay of measured LRIR with its smoothed or frequency-windowed versions.

Both for magnitude and phase delay plots, this method seems promising. However, the situation changes when we look at the impulse response (Fig. 1.23) computed from the combined, separately smoothed magnitude and phase response. It looks very different from the measured one. The same is true for the impulse response of the complex smoothed spectrum. I listened to music convolved with these responses, and it confirmed the visual representation: for both of them, the output signal was very strange and different from the output of the measured impulse response. The only exception was the third method, FDT, where its impulse response both looks and sounds reasonable.

The takeaway is that if we want to perform the minimum-phase equalization, we can choose either separate magnitude smoothing (since magnitude is the only thing

we care about) or the frequency-dependent truncation. However, for the mixed-phase approach, only FDT can be used.

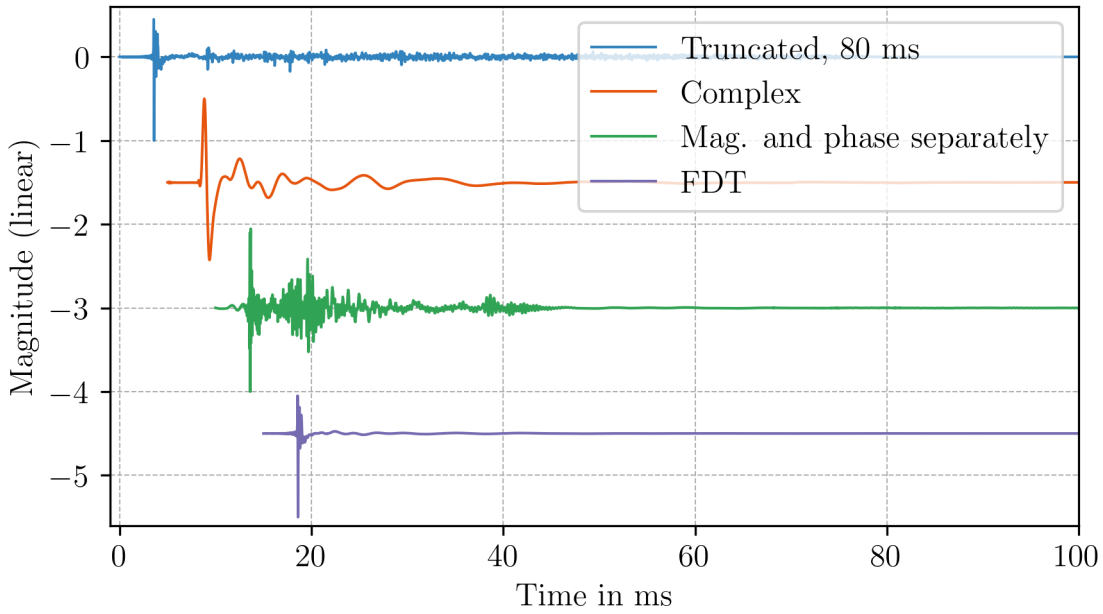


Figure 1.23: Comparison of measured with its smoothed or frequency-windowed versions. Responses are horizontally and vertically shifted for better visibility.

1.5.3 Prototype Response

I performed measurements at 13 positions on a grid of 40 cm, assuming a quite generous equalization zone for a single listener. Fig 1.24 illustrates all the measuring positions by dots, and the corresponding magnitude responses (1/3-octave smoothing) are shown in Fig 1.25. The responses are plotted in groups and spaced apart for better visibility. The purple curves are all measurements within 20 cm grid and will be used for further prototype calculations.

Based on the theory, I decided to perform a mean average of the measured responses. We can see that, at least up to 500 Hz, the average could perform well because the peaks and dips are well aligned. At higher frequencies, the situation is different. The blue curve in Fig 1.26 shows the mean average performed. Since above 1 kHz the peaks and dips in the individual measurements can be at different frequencies, the resulting peaks and dips of the mean don't necessarily have to physically exist.

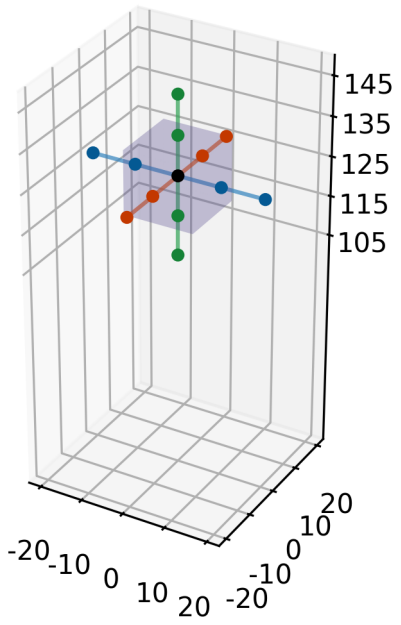


Figure 1.24: Visualization of the measurement positions. The unit is cm.

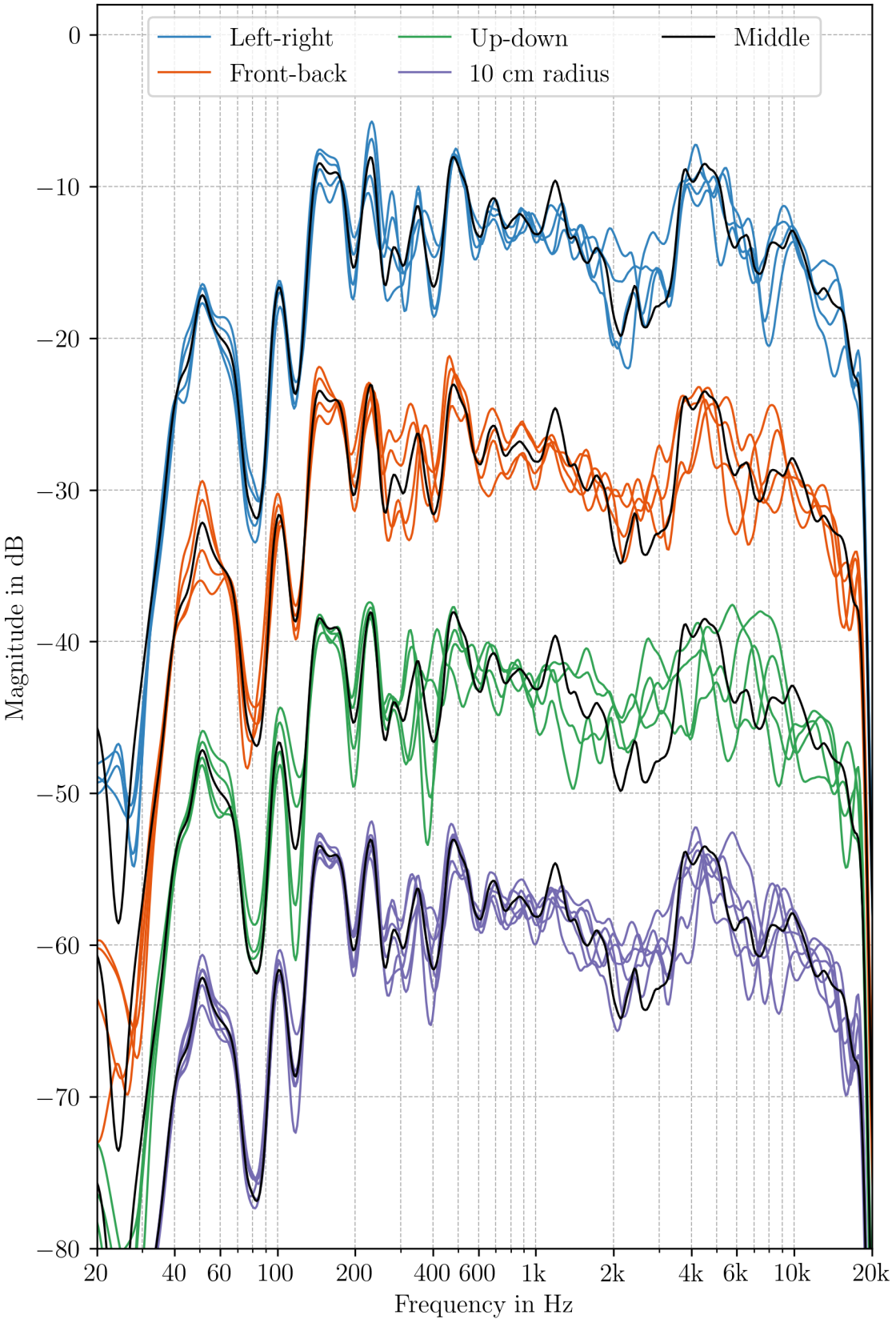


Figure 1.25: Magnitude responses of 13 measurement positions on 40 cm grid shown in Fig. 1.24, 1/3-octave smoothing. Responses are vertically shifted for better visibility.

That's why I propose further evening out of the magnitude response by widening the smoothing window as a function of frequency. The result of the smoothing by using a 'high-frequency scaling' of the window is the orange curve. We can see that it works well: The curve captures the overall trend of the response but neglects the small, possibly not physical, fluctuations.

The formulas for calculating the band-limiting frequencies of the fractional octave bands are changed by a factor k to

$$f_{\text{lower}} = f_c\, 2^{-\frac{k}{2n}}, \qquad f_{\text{upper}} = f_c\, 2^{\frac{k}{2n}},$$

where

$$k = 1 + s_{\text{scale}}\,(f_c - f_{\text{tresh}})\,0.001. \tag{1.27}$$

f_{tresh} is the frequency chosen by the user, which specifies the frequency threshold above which the scaling should be applied, and s_{scale} is a scaling factor specified by the user, determining how much wider the band will be. The number 0.001 is just a scaling factor for convenience, so s_{scale} can be in a reasonable range (approx. integers).

The green curve shows when this scaling is applied to the smoothing of the midpoint measurement. We can see that the results are very similar except between 2 to 3 kHz. Therefore, an option might be to leave out the prototype completely, and measure only at one position. Not only is this way more convenient for the user, but it also saves computing power (mainly on the smoothing procedure of multiple responses). Performance-wise, something in between is to first compute the mean of the measured magnitude responses, and then perform the smoothing (which I have also tested).

So far, we have only mentioned the magnitude smoothing. However, if we want to use the mixed-phase approach, to 'even out' the response, frequency-dependent windowing has to be used, which outputs an impulse response; thus, we need to average the corresponding complex spectra.

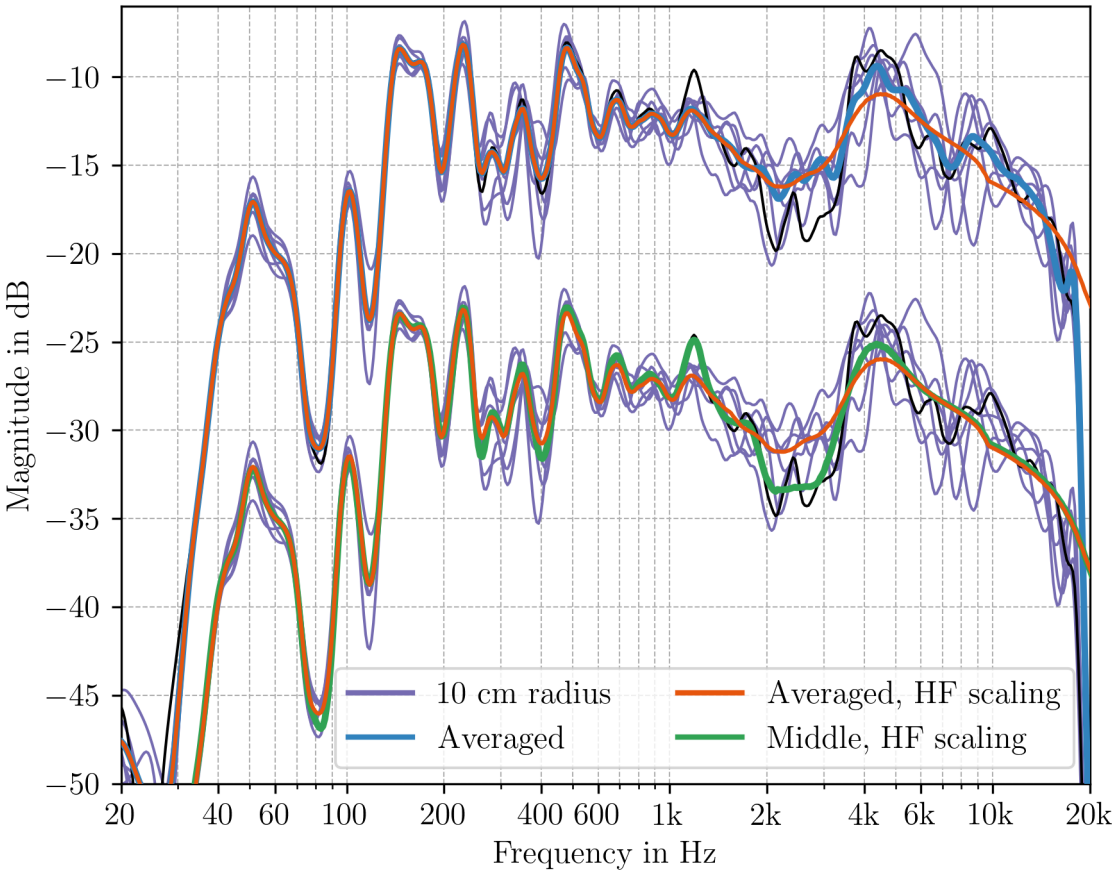


Figure 1.26: Comparison of averaged and high-frequency scaled magnitude responses with the measurements on 20 cm grid, 1/3-octave smoothing. Responses are vertically shifted for better visibility.

Fig. 1.27 compares averaging complex spectra together with separate magnitude and phase smoothing. Based on the magnitude and corresponding impulse responses, we can conclude that averaging complex spectra produces artifacts, but we can use the separate magnitude and phase smoothing. Listening to music convolved with these impulse responses confirmed the findings from the plots. Moreover, it is in agreement with [42].

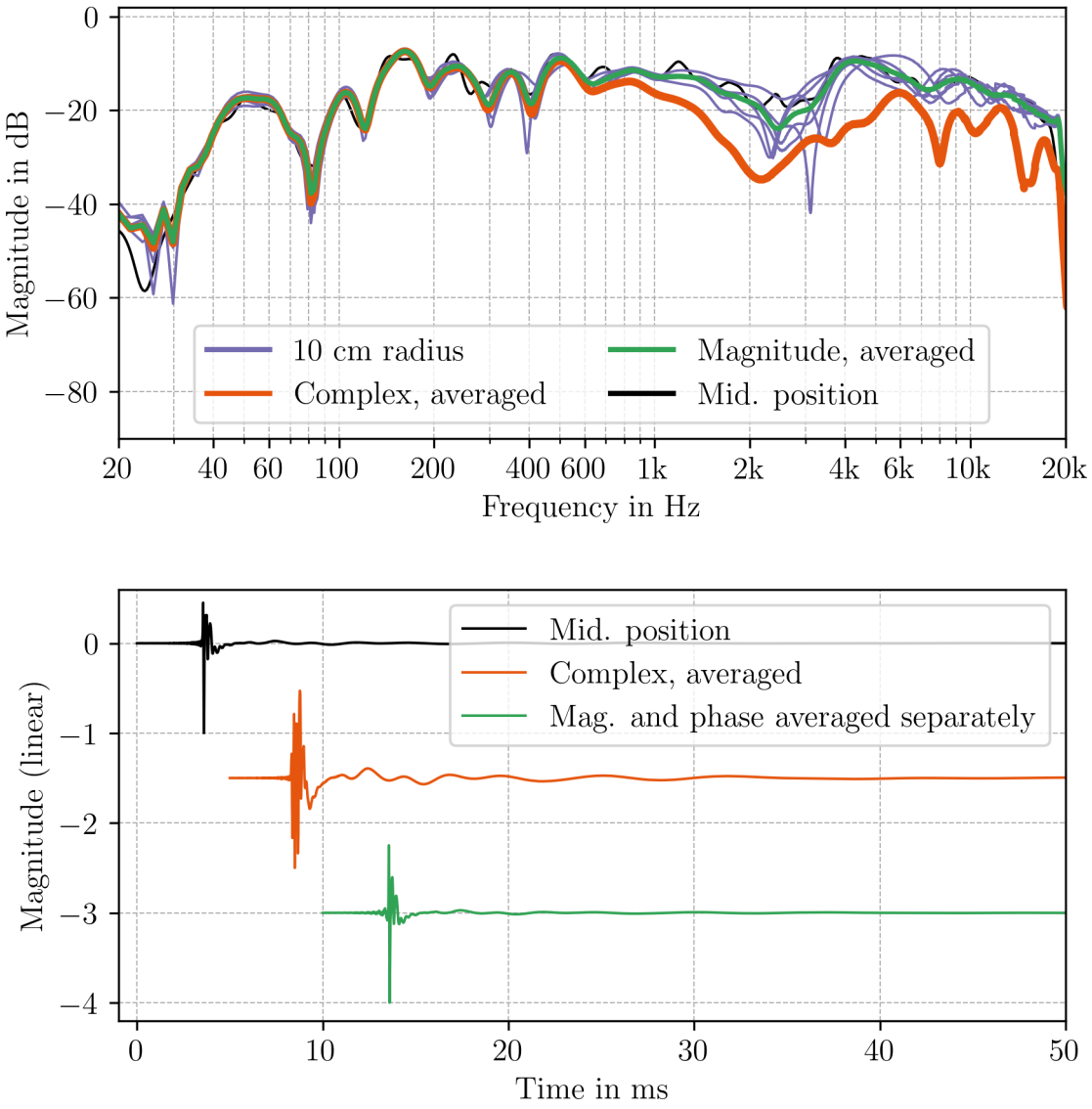


Figure 1.27: Comparison of averaging complex spectra (here when FDT was applied beforehand), magnitude spectra, and corresponding impulse responses. Impulse responses are horizontally and vertically shifted for better visibility.

1.5.4 Approach 1: Minimum-Phase FIR Compensation Filter

As the first approach, I have implemented the simplest method based on the Eq. (1.14). Fig. 1.28 illustrates the whole procedure after the LRIR measurement. Here, the procedure is limited to only one speaker/channel and one measurement point (no prototype response is calculated).

Initial Truncation & **Zero-padding**: The measured LRIR is truncated the first 80 ms after the initial peak and afterwards zero-padded to obtain sufficient frequency resolution, which is specified by the user as discussed in Sec. 1.5.2.1. The truncation is performed by a multiplication between the impulse response and

a weighting function consisting of a flat passband with Hanning taper at the beginning (128 samples) – assuming that the peak of the impulse response is never at index 0 as there is always some propagation delay till the signal from the loudspeaker reaches the microphone – and at the end (256 samples).

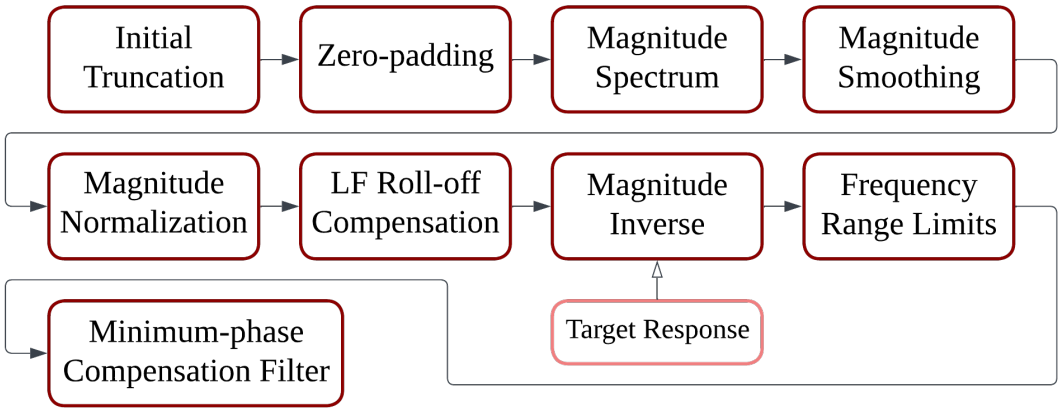


Figure 1.28: Detailed block diagram of the procedure of the minimum-phase inversion.

Magnitude Spectrum & **Magnitude Smoothing**: Single-sided magnitude spectrum is obtained by performing FFT. Afterwards, it is smoothed within the chosen fractional-octave bands. Optionally, the smoothing can be made more robust at high frequencies by widening the smoothing window according to Eq. (1.27). For the results shown in Fig. 1.26, one has to choose

$$s_{\text{scale}} = 8.$$

Magnitude Normalization: The magnitude spectrum is normalized in chosen frequency range to a specified dB value G. The normalization method can be either RMS or mean.

Low-frequency Roll-off Compensation: The frequency response of the *AIAIAI Unit 4* monitor was measured in the anechoic chamber, and based on that measurement, a Butterworth high-pass filter was manually fitted to approximate its low-frequency roll-off. Both are shown in Fig. 1.29. Its parameters are cut-off frequency f_c and order n:

$$f_c = 65\,\text{Hz}, \qquad n = 4.$$

The 4th order specifies the roll-off of 24 dB per octave. This matches nicely with the theory because it corresponds to the slope of a bass-reflex (= vented, ported) loudspeaker system, which is the case for the *Unit 4* monitor.

The Magnitude response of the Butterworth HP filter can be calculated analytically as

$$|H_{\text{HP}}[k]| = \frac{1}{\sqrt{1 + (f[k]/f_s)^{2n}}}, \tag{1.28}$$

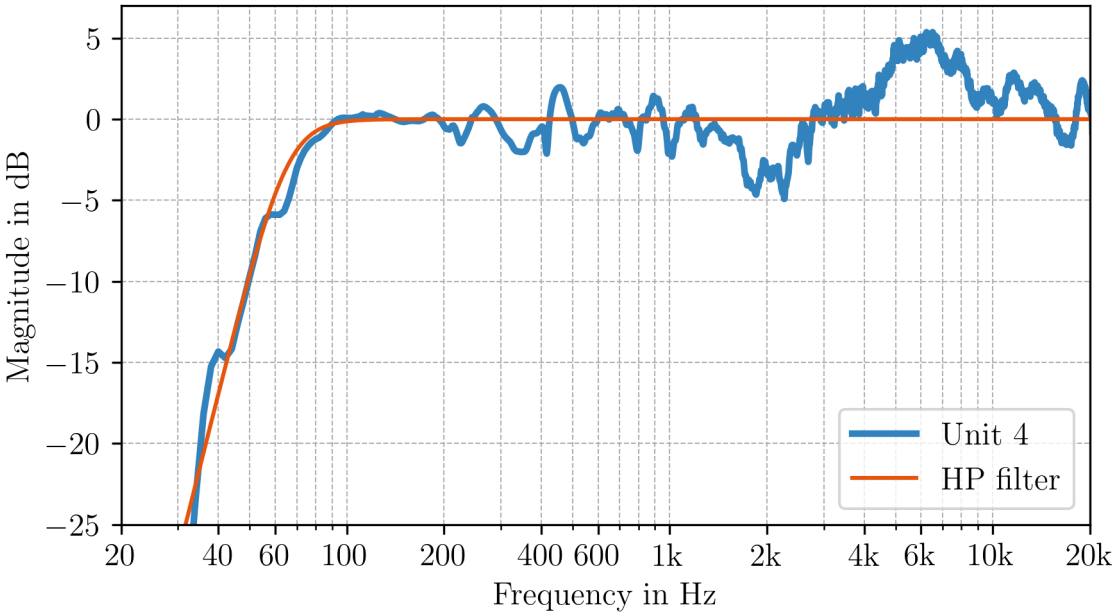


Figure 1.29: Magnitude frequency response of the *AIAIAI Unit 4* monitor and fitted HP filter.

where $f[k]$ is frequency at k frequency bin and f_s is the sampling frequency. The normalized smoothed LRRC is compensated with this response as

$$|H_{\text{LRRC, comp}}| = \frac{|H_{\text{LRRC}}|}{|H_{\text{HP}}|}. \tag{1.29}$$

Magnitude Inverse & **Frequency Range Limits**: The target magnitude response H_{targ} is created as a linearly decreasing function with specified slope in dB per octave, and the compensation filter's magnitude response is calculated as

$$|H_{\text{comp}}| = \frac{|H_{\text{targ}}|}{|H_{\text{LRRC, comp}}|}. \tag{1.30}$$

Then, the $|H_{\text{comp}}|$ is limited outside the user's specified frequency range of interest using Eq. (1.12).

Minimum-phase Compensation Filter: Minimum-phase compensation filter h_{comp} is computed using a cepstrum of a spectrum: Every zero outside the unit circle (a so-called maximum-phase zero) corresponds to an acausal exponential, and every zero inside the unit circle (a so-called minimum-phase zero) corresponds to a causal exponential in the corresponding cepstrum. So, the cepstrum is computed from the given spectrum, acausal exponentials are converted to causal exponentials, and the corresponding spectrum becomes minimum-phase [5].

1.5.4.1 Results

Fig. 1.31 - Fig. 1.34 show results measured in my 13 m^2 dorm room when compensation within the range of 40 Hz and 15 kHz was considered. Let's first look at

Fig. 1.31 showing the magnitude spectra when 1/3-octave smoothing is used. The green curve is the measured response (grey) convolved with the compensation filter, and the purple one is also a measured response, but this time, with the filter applied to the speaker before the measurement was taken (more accurate).

Looking at the green curve, it seems that the compensation improved the situation, as the response looks flatter than the grey one. However, when the measurement was repeated with the compensation filter, the results were slightly worse. The biggest difference is at the highest frequencies, and also, we did not get rid of the dip at 80 Hz. What might be the cause of it? In general, it could be either that the speaker or microphone position is at or near a mode node, or that it is caused by early reflection cancellation. After measuring the positions, it turned out to be the second reason. As discussed in theory, one should be careful about boosting such regions with a minimum-phase filter. When we consider that a 20 dB boost corresponds to 10 times the original amplitude – which is a lot! – It would be beneficial to be able to limit the correction to a certain dB ceiling, as it is safer not to compensate so strongly under these circumstances; unfortunately, with this method, it is not possible.

When listening to music with the compensation filter on, the LF boost was indeed too high, and the compensation filter would benefit from limiting it. When the same compensation was computed but limiting the correction's low frequency limit to 150 Hz, the subjective impression was better. Theoretically, the LF correction limit could be set by the user based on trial and error, but I don't consider this a suitable solution for limiting compensation boosts at low frequencies.

Fig. 1.33 shows spectra related to the same compensation but now with 1/6-octave smoothing. We can see that the results are a little different, but we can hardly say they are better. I believe the same is true if we compare the resulting compensated impulse responses shown in Fig. 1.31 and Fig. 1.33 – the compensated LRIR when the 1/6-octave smoothing is used is not closer to our desired delta function in any way than the compensation using the 1/3-octave smoothing. One aspect, where the finer smoothing is actually worse, is the group delay. For the minimum-phase filter, the phase response is linked with a particular magnitude response. If the magnitude response is less smooth, as with the 1/6-octave smoothing, the group delay will be less smooth as well, potentially increasing the risk of audible artifacts (see Fig. 1.34 for comparison).

In this approach, the compensation filter h_{comp} is an FIR filter. This filter has to have many taps to provide a reasonable frequency step Δf at the lowest frequencies (with increasing frequency, this is less and less relevant as the frequency step is linear)[10]. For $f_s = 48\,\text{kHz}$ and $\Delta f = 4$, it is 12000 taps (used for the 1/3-octave smoothing; for the 1/6-octave, twice that was used).

[10]This can be overcome if the whole processing is done in a *warped* domain. Details can be found in [1].

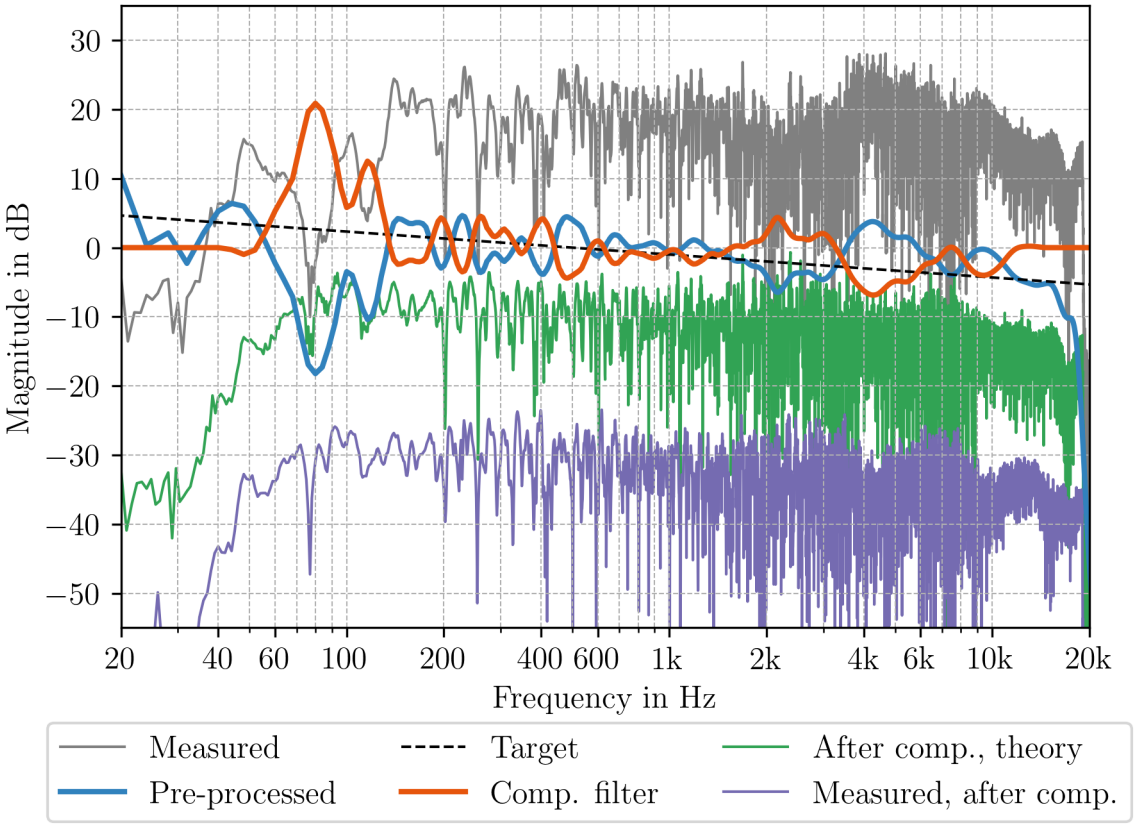


Figure 1.30: Magnitude spectrum of compensation results using 1/3-octave smoothing. 'Full' spectra are shifted for better visibility.

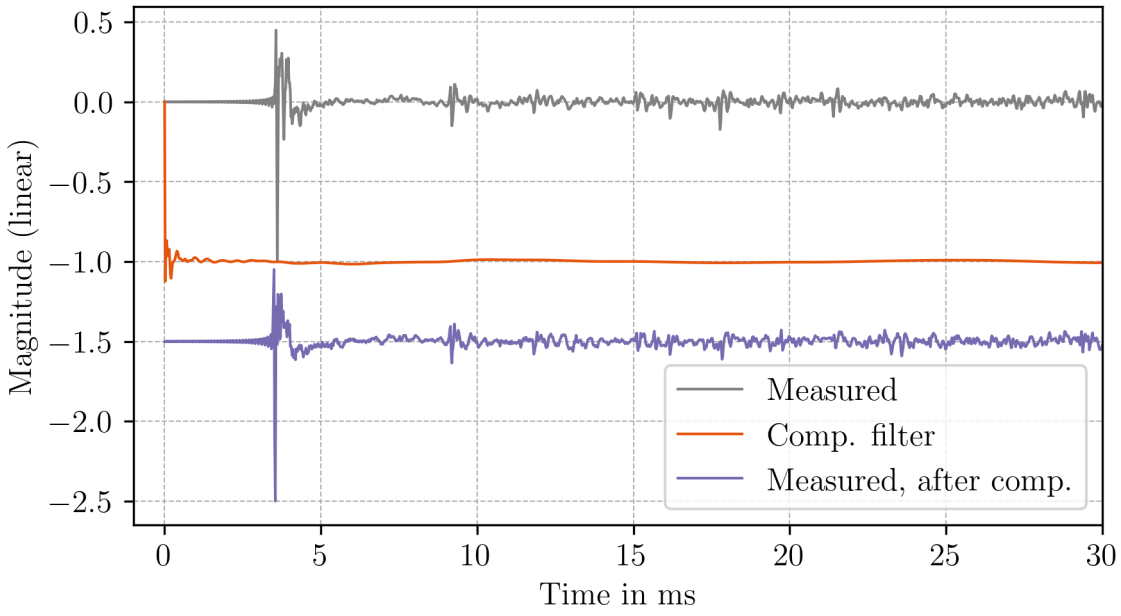


Figure 1.31: LRIR before and after compensation and IR of the compensation filter h_{comp} using 1/3-octave smoothing. Responses are vertically shifted for better visibility.

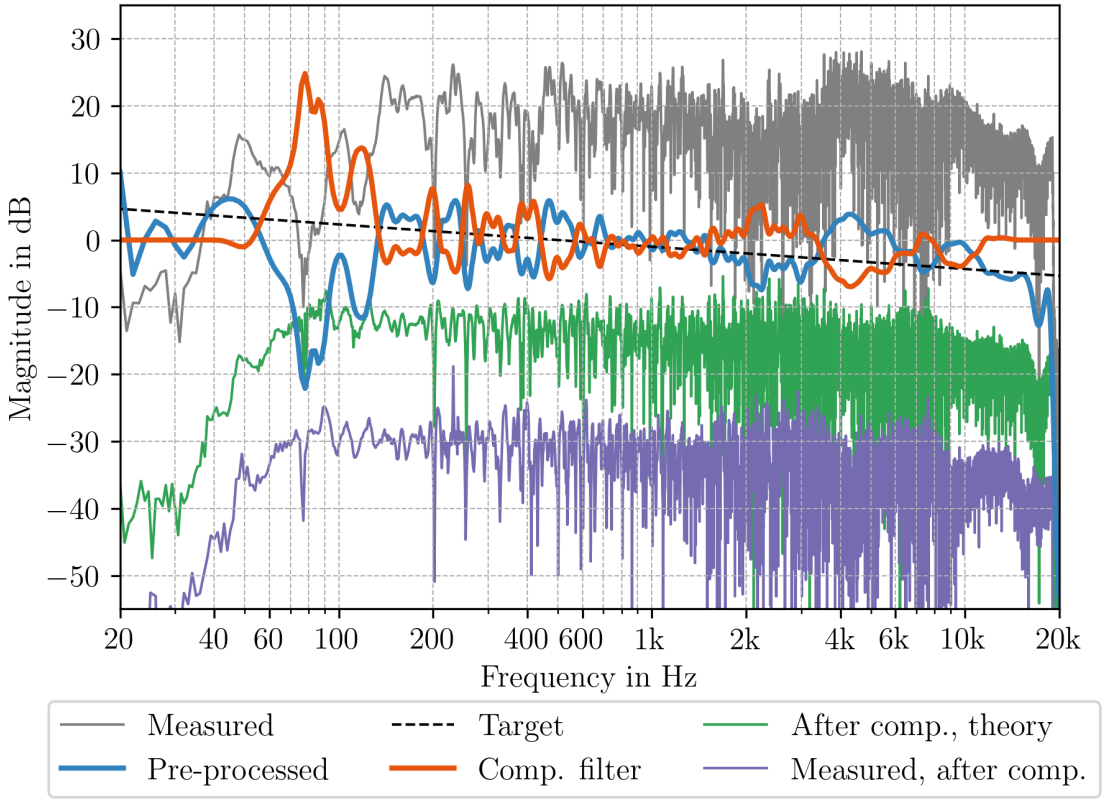


Figure 1.32: Magnitude Spectrum of compensation results using 1/6-octave smoothing. 'Full' spectra are shifted for better visibility.

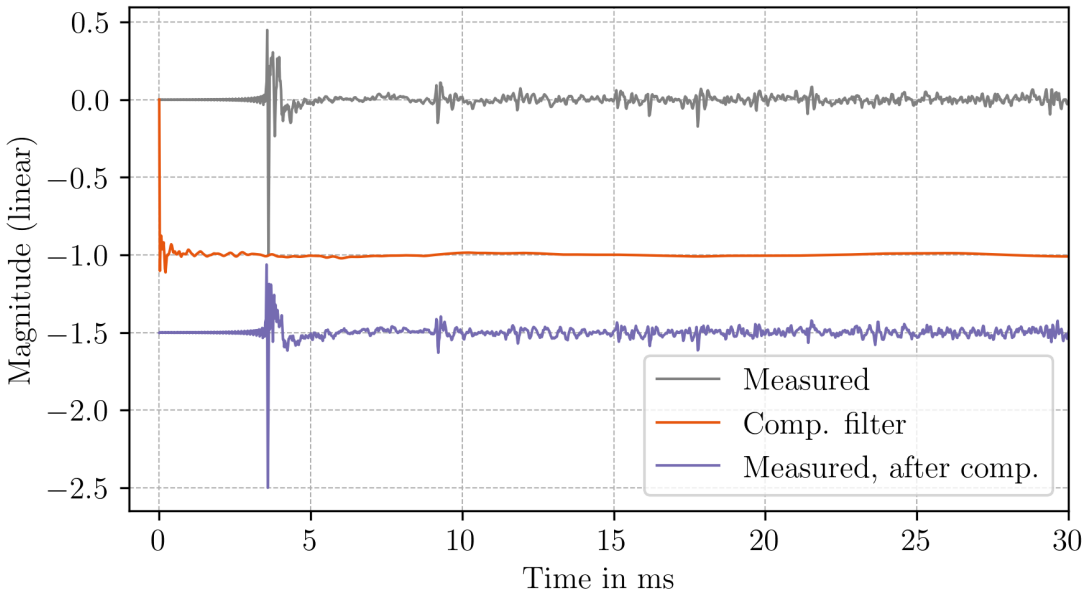


Figure 1.33: LRIR before and after compensation and IR of the compensation filter h_{comp} using 1/6-octave smoothing. Responses are vertically shifted for better visibility.

1.5.4.2 Evaluation

This approach is certainly promising for implementation based on the presented results. Its main limitations are the inability to limit the 'amount' of correction applied and the fact that the compensation is implemented as a long FIR filter, which can be computationally demanding for hardware implementation.

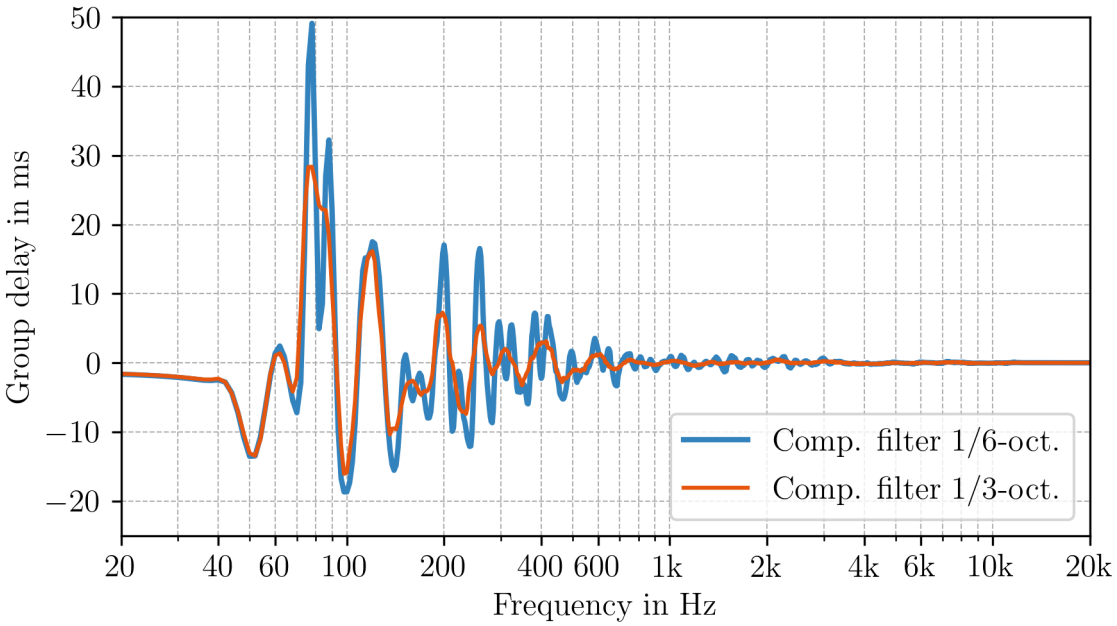


Figure 1.34: Comparison of group delay of the compensation filters when different smoothing resolution is used.

1.5.5 Approach 2: Mixed-Phase Compensation FIR Filter

For the mixed-phase compensation filter, I have decided to implement it using the x-filtered NLMS algorithm. Fig. 1.35 shows the individual steps of the procedure. As for the first approach, the following description involves no prototype response – it assumes only a single speaker/channel and one measurement point.

Figure 1.35: Detailed block diagram of the procedure of the x-filtered NLMS inversion.

Frequency-dependent Truncation: The frequency-dependent truncation is

used here as it is the only discussed method that works for 'evening out' the complex spectrum. In this step, the initial truncation is also hidden. The user inputs

- the lowest assumed frequency f_{lowest},
- the multiples of period T n_T, based on which the window truncation length will be determined for each frequency bin,
- and frequency step Δf.

First, the LRIR is initially truncated either based on Δf or specified f_{lowest} together with n_T (the IR lenght $n_h = n_T * f_s / f_{\text{lowest}}$), whichever outputs a higher number. As I already mentioned, for every frequency bin, a different window truncation length W_k is calculated as $T \cdot n_T$ and the DTFT is computed using the Eq. (1.10). At the end, the obtained spectrum is converted back to the time domain using the inverse Fourier transform.

Magnitude normalization: The spectrum is normalized based on its magnitude in the exact same way as for the first approach.

X-filtered NLMS: It is implemented exactly as shown in the block diagram in Fig. 1.17, implementing Eq. (1.21) - Eq. (1.24). The user has to input the length of the input signal x, the initial delay Δ of the compensation filter, the compensation filter length, and the step-size μ_{scale}. x is generated as a sequence of normally distributed samples, approximating white Gaussian noise.

1.5.5.1 Testing

First, I tested different values of the input parameters to see how they affect the compensation filter. The length of the compensation filter was set to the same length as the (frequency-dependent) truncated LRIR denoted as *Pre-processed*.

Fig. 1.36 shows the effect of the input signal length, which is expressed as a multiple of the compensation filter length. We can see that every increase in length results in a better compensation filter. The biggest differences are at the lowest frequencies, at the sharp resonance at 80 Hz, and in the frequency range between 2 kHz and 3 kHz.

How the step-size influences the resulting compensation filter is shown in Fig. 1.36. Here, the differences are not so big, although the best results are obtained with the biggest step-size.

Fig. 1.38 shows the magnitude and corresponding impulse responses for different delays Δ. Here, the delay determines the filter's lower frequency limit. For example, from the highest frequencies down to 600 Hz, the magnitude response of the filter with the delay of 5 ms looks almost the same as the one with the delay of 100 ms. However, below that frequency, the responses differ. Only the filter with the longest delay, 100 ms, works at the lowest frequencies. It is linked to the acausality of the mixed-phase compensation filter: we need the delay to capture what precedes the main peak, as it is part of the filter as well. As shown when discussing truncation, low frequencies require longer time windows, thus a longer delay for an accurate compensation filter at low frequencies is needed.

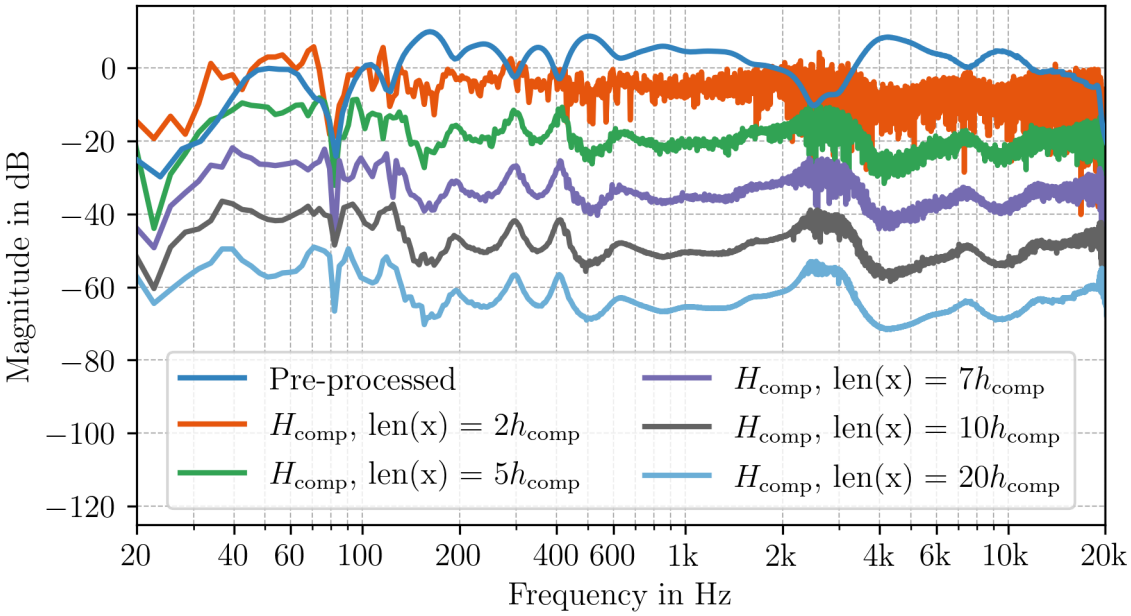


Figure 1.36: Magnitude response of the truncated LRIR (Pre-processed) together with a comparison of the magnitude responses of compensation filters computed using different input signal lengths. $\Delta = 100\,\text{ms}$, $\mu_{\text{scale}} = 0.9$. Curves are vertically shifted for better visibility.

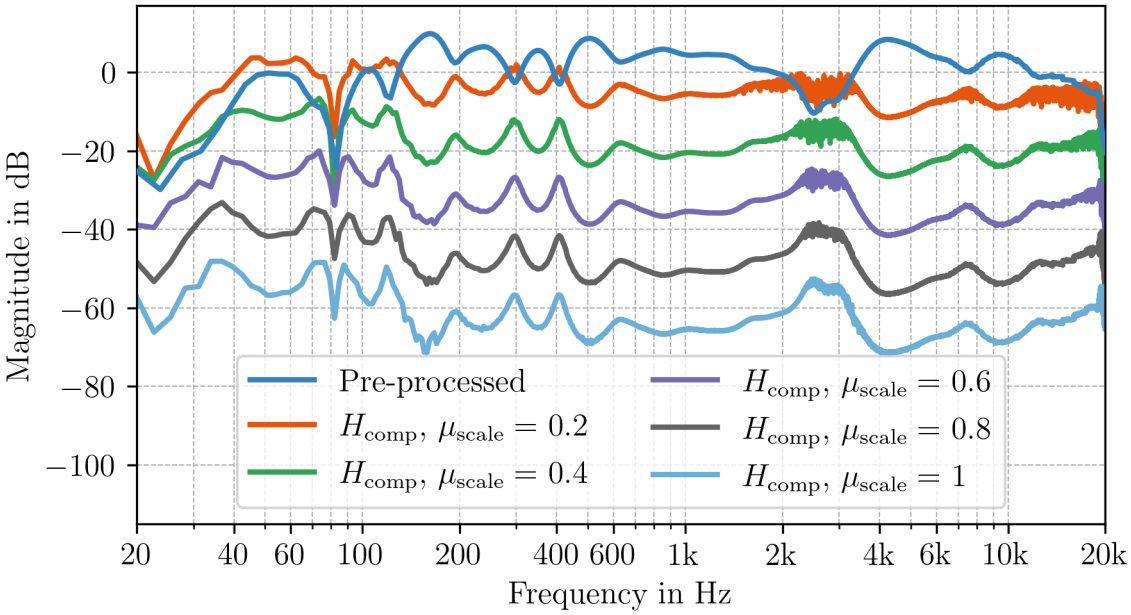


Figure 1.37: Magnitude response of the truncated LRIR (Pre-processed) together with a comparison of the magnitude responses of compensation filters computed using different step-size by changing μ_{scale}. $\Delta = 100\,\text{ms}$, $\text{len}(x) = 20h_{\text{comp}}$. Curves are vertically shifted for better visibility.

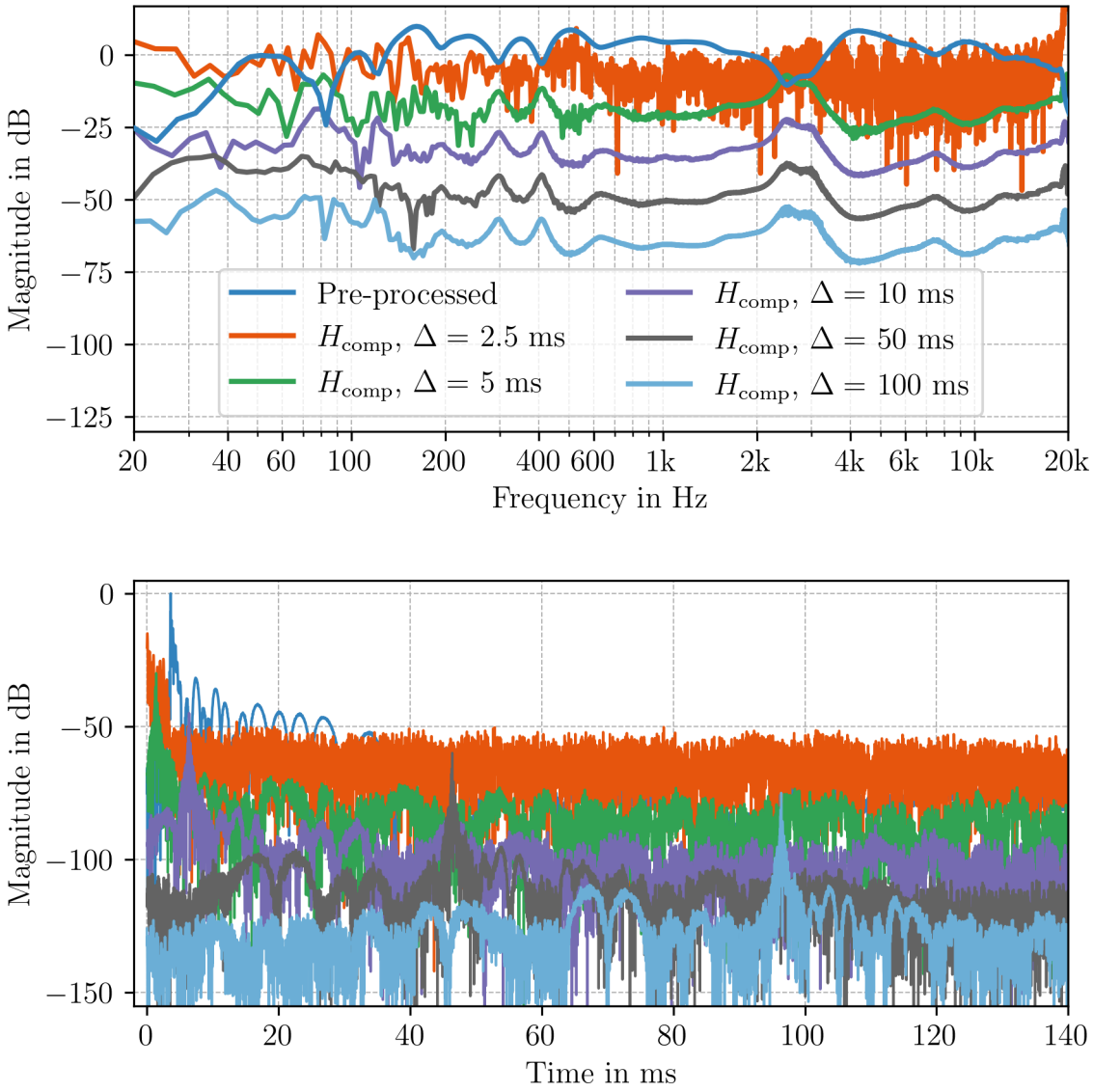


Figure 1.38: Magnitude responses and corresponding impulse responses of the truncated LRIR (Pre-processed) and compensation filters computed using different delays Δ. $\mu_{\text{scale}} = 0.9$, $\text{len}(x) = 20h_{\text{comp}}$. Curves are vertically shifted for better visibility.

1.5.5.2 Results

Based on findings from testing described in the previous section, the compensation filter is obtained with these parameters:

- $\text{len}(h_{\text{comp}}) = \text{len}(h_{\text{FDT}})$,
- $\text{len}(x) = 40h_{\text{comp}}$,
- $\Delta = 100\,\text{ms}$,

and for the parameters of the frequency-dependent truncation

- $n_T = 4$,
- $f_{\text{lowest}} = 40\,\text{Hz}$,
- $\Delta f = 4\,\text{Hz}$.

Fig. 1.39 shows the magnitude spectra of the measured, pre-processed (FDT), compensated response as well as the magnitude response of the compensation filter. Looking at the compensation filter's curve, we can see that even with input length being 40 times the length of h_{comp}, the expected resonance at 80 Hz is still not captured. However, we can also see it as an advantage as we would rather not like to compensate for such deep antiresonances. We can conclude that, for both the calculated and measured LRRC, after compensation, it is flatter than without compensation, except in the region between 2 kHz and 3 kHz. This is caused by the FDT pre-processing method: In the FDT blue curve, which corresponds to the truncated measured LRIR, we can see a significant dip even though it is not really present in the grey curve. It is influenced by the parameter n_T, which is in this case 4. That means that the time window for the Discrete Fourier transform computation at 2.5 kHz is 1.6 ms, starting from the main peak of the measured IR. So, up to this time, there was a lower amount of the early reflections reflected in the measurement position around this particular frequency. This can also be seen in Fig. 1.22 in the difference of the FDT and magnitude smoothing in this particular frequency range. The question is now which method better captures what we perceive, because that is time-dependent, and therefore, whether assessing the results by showing the full compensated spectrum has any meaning at all. We said we need at least 80 ms for the initial truncation to capture the low-frequency content, and that the ear integrates sound events up to 60 ms, which is 20 ms less. Purely based on the listening comparison (FDT versus magnitude smoothing), I can say that the peak between 2 kHz and 3 kHz caused by the FDT actually sounds as an amplification in that region and therefore does not sound right.

Fig. 1.40 shows the measured LRIR before and after compensation together with the IR of the compensation filter. We can clearly see that the artifacts coming just after the initial peak are reduced. In my opinion, these are still the direct sound (an imperfection of the loudspeaker system), very early reflections from the table where the speakers are located, or a combination of both. However, the early reflections, which occur later (9 ms and beyond), are not really compensated. Moreover, significant post-ringing, as discussed in Sec. 1.4.5.3, is present, too. As for the previous approach, the resulting compensation filter is an FIR filter, coming with the downsides already mentioned.

1.5.5.3 Evaluation

The parameters of the compensation filter could probably be 'tweaked' so that the results obtained would be slightly better. However, I decided not to spend time on that because of certain aspects that make this approach unsuitable for our implementation purpose. The biggest disadvantage is the delay of at least tens of milliseconds for the initial peak needed so that the compensation filter 'behaves' also at the low-frequency region. This is way too big for implementation in studio monitors, where musicians or producers need almost instant response from a microphone or an instrument when monitoring while singing/playing.

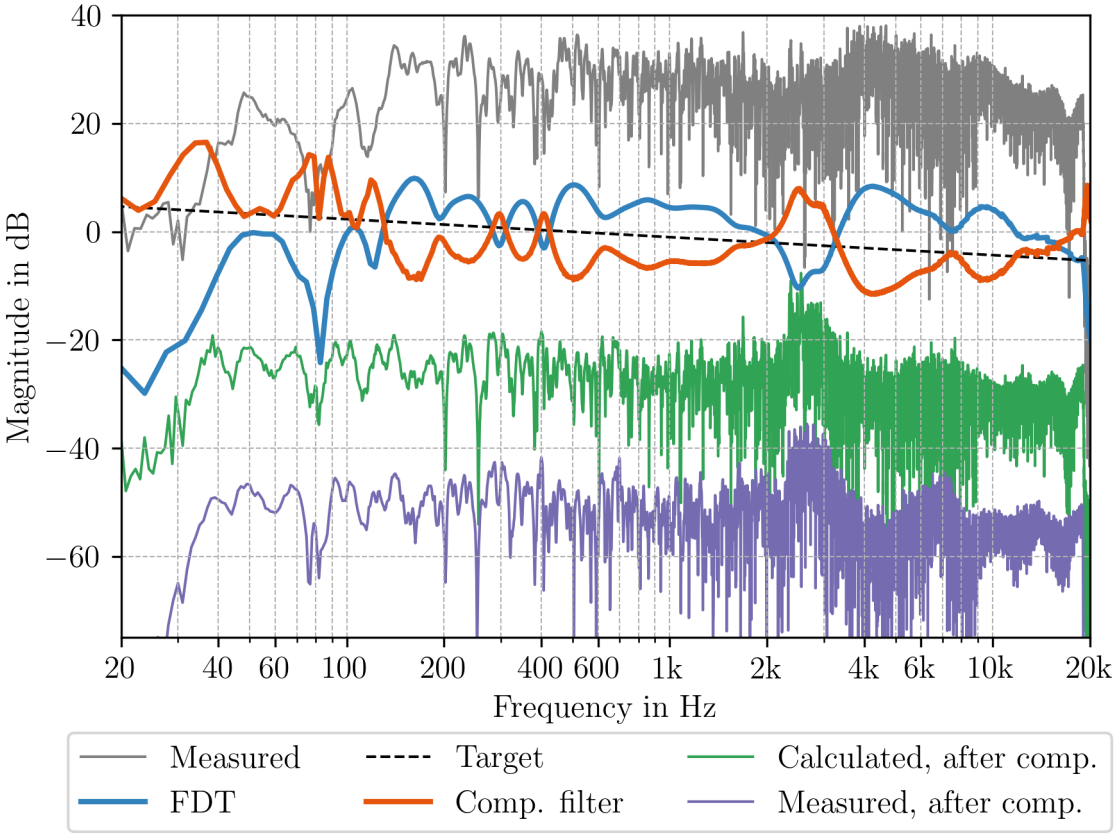


Figure 1.39: Magnitude spectrum of compensation results obtained using the x-filtered NLMS algorithm. 'Full' spectra are vertically shifted for better visibility.

Another limitation of this approach is its computational demands. For the chosen parameters corresponding to the shown results, the computation of the compensation filter took more than a minute for my laptop with a 4-core Intel Core i5 processor and 16 GB of RAM. That is probably already way too long for the user to wait, not to mention that, when implemented on a mobile device as intended, the computational time would most likely increase further.

Lastly, a further limitation is the impossibility to limit the frequency range of compensation, as the spectrum to be manipulated is complex, and such a formula as shown in Sec. 1.4.3.4 and used in the previous approach cannot be used.

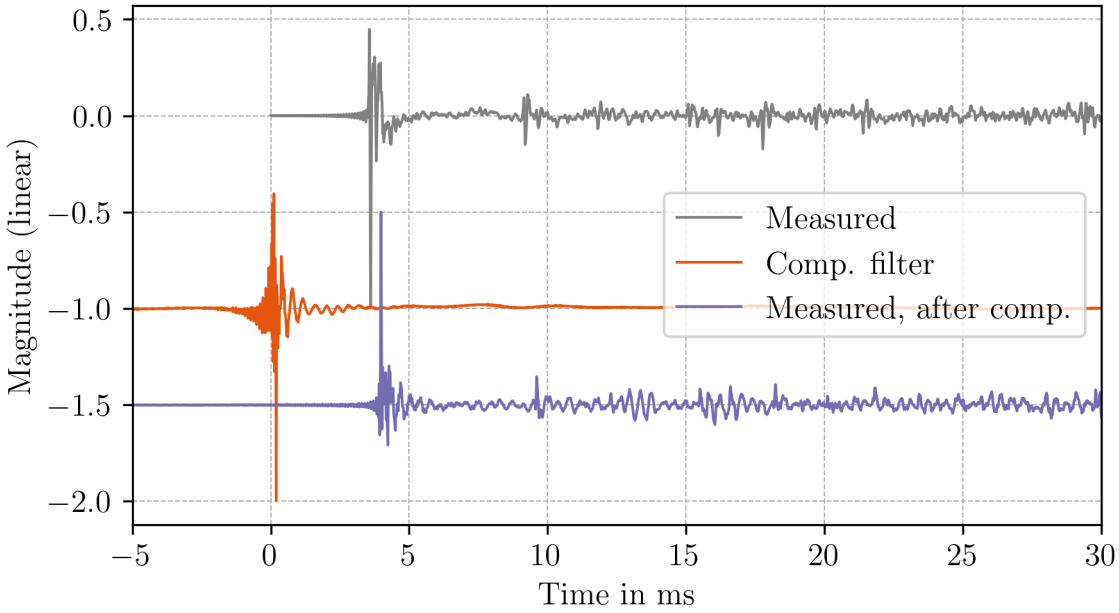


Figure 1.40: LRIR before and after compensation and IR of the compensation filter h_{comp} obtained using the x-filtered NLMS algorithm. The orange and purple curves are shifted by approximately 100 ms for better visibility.

1.5.6 Approach 3: Minimum-Phase IIR Compensation Filter Using a SOS Cascade

Fig. 1.41 shows all the individual steps of the implementation of this method. The first three blocks, and the **Magnitude Smoothing**, **Magnitude Normalization**, **LF Roll-off Compensation** are identical to those already described in 'approach 1'. The new ones are:

Logarithmic Interpolation: In this, third approach, there is no step of going from the frequency domain back to the time domain – the resulting filter will be a combination of IIR filters characterized by parameters. Therefore, we can save computational effort by resampling the linear magnitude spectrum grid to a logarithmic one. I chose the grid to be 30 points per octave. This ensures about 1 Hz step at the lowest frequencies, which is fine enough. This reduces the number of data points to less than 2 % of the linearly spaced grid with the same degree of low-frequency resolution.

Prototype Response: This block is dotted as it is an optional step if more than one LRIR is measured. Then, they are averaged as described in Sec. 1.5.3. The contribution of this step from the subjective point of view will be discussed in the following section.

Frequency Range Limits: This one works slightly differently from the one explained in the previous implementation. Outside the correction band, the compensation is not left at g as was shown in Eq (1.12). Instead, H_{limited} has to follow the target function so Eq. (1.17) works:

$$H_{\text{limited}} = H + W\,(H - H_{\text{targ}}). \tag{1.31}$$

Split Into Frequency Bands: This is already described in 'part 1' of the Sec. 1.4.5.5. However, even though [37] suggests finding frequencies where pre-processed LRRC crosses zero, this is only valid when the target function is zero (0 dB in the whole frequency range). In general, we need to find 'target response-crossings' instead – frequencies, where the pre-processed LRRC crosses the target function (see Fig. 1.42). Since the bands are sorted by area, the user specifies which bands will be considered for the compensation filter by specifying the number of peak filters used, N. The user can also specify the minimum area considered – any band with an area below that threshold won't be considered for compensation.

Initial Parameter Guess & Compensation optimization: This was also described before in 'part 2' and 'part 3' of the Sec. 1.4.5.5. Eq. (1.17) - Eq. (1.20) are implemented. The random search method is used, and by experimenting, the change of the parameters is chosen as follows:

- Q: up to $\pm 5\,\%$,
- f_c: up to $\pm 5\,\%$, but only for the post-optimalization. For the first optimization round, the initial-guessed value is used,
- G_{dB}: up to $\pm 0.2\,\mathrm{dB}$.

When it comes to the optimization, [37] suggests taking one band at a time, first making an initial guess of the parameters and then performing parameter optimization. Optionally, the optimization can be performed one or more times (post-optimization). However, there are more strategies: One can first make an initial guess for all the bands, and then either optimize them iteratively (one band at a time) or simultaneously in parallel. Which method performs best will be discussed in a later section.

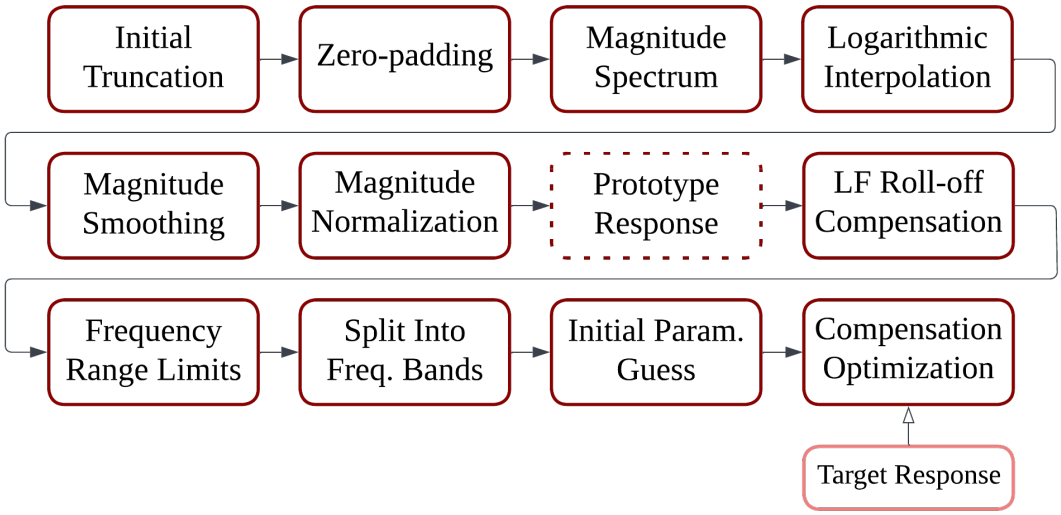


Figure 1.41: Detailed block diagram of the procedure of the minimum-phase IIR compensation filter.

1.5.6.1 Testing

Fig. 1.43 shows the mean error in dB according to Eq. (1.17) for all the different initial guess & optimization strategies. The smoothing window for this example was chosen to be a 1/3-octave band. 100 iterations were chosen for every band. The tested response was split into 24 bands, resulting in 2400 iterations in total. We can see that there is no point in increasing this number, as further iterations would most likely not result in lower error. 'init. guess' stands for making an initial guess for all the bands first before the iterative optimization starts, 'post-optimization' refers to performing one more round of optimization (so two optimization rounds in total). In this case, the number of iterations per band is 50 in the first optimization round and another 50 in the second. For the total number of iterations, the second round starts after 1200 iterations. The effect can be seen in the graph for the purple and dark grey curves, resulting in a sudden steeper error decrease.

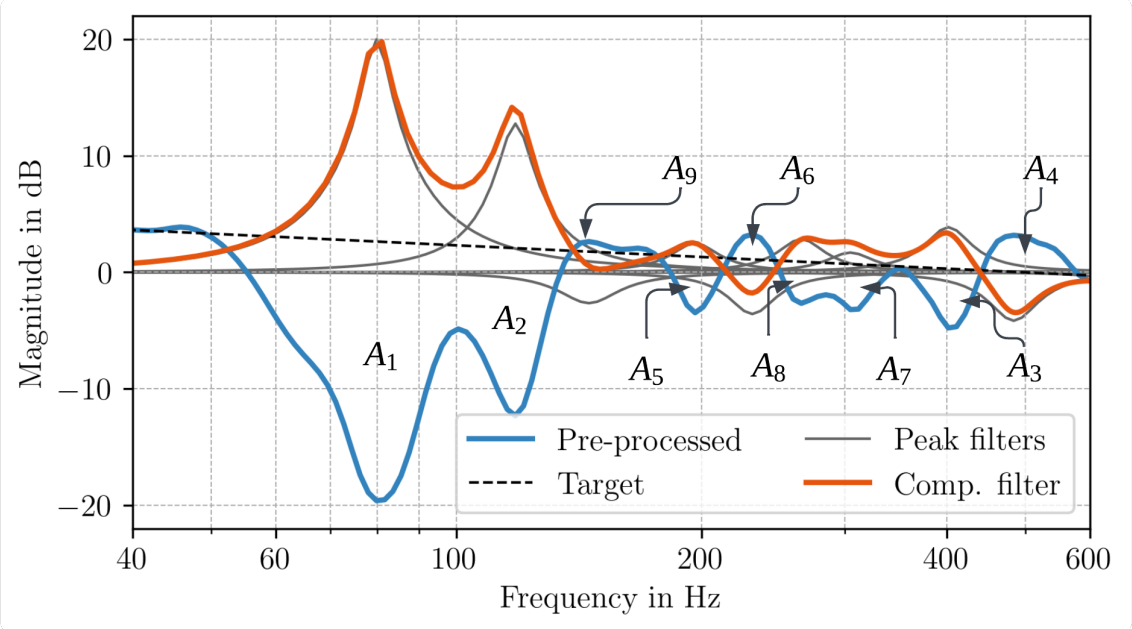


Figure 1.42: Identified areas denoted A_x, already optimized peak filters, and their sum (in dB) as the resulting compensation filter.

To compare the results, we can say that the resulting error is quite similar for all the methods. The biggest difference is how fast the error decreases, i.e., how fast the parameters of the peak filters that form the compensation filter converge. We can see that the initial guess for all the bands immediately brings the error down by a significant amount. Then, the iterative band compensation (the green and dark grey curves) outperforms the parallel approach. We can conclude that in this case, the best method is the 'Iterative with init. guess'. However, the presented results varies slightly for different LRRC analyzed. We should therefore also consider the 'Iterative with init. guess and post-optimization', because in theory this is the more robust method, being able to optimize also the center frequencies of the peak filters f_c, which is not a good idea to do for the first optimization round as it might shift

f_c away from its actual position dut to the fact that the other (interfering) bands are not yet optimized. Both, however, offer the possibility of reducing the number of iterations, as they converge quickly.

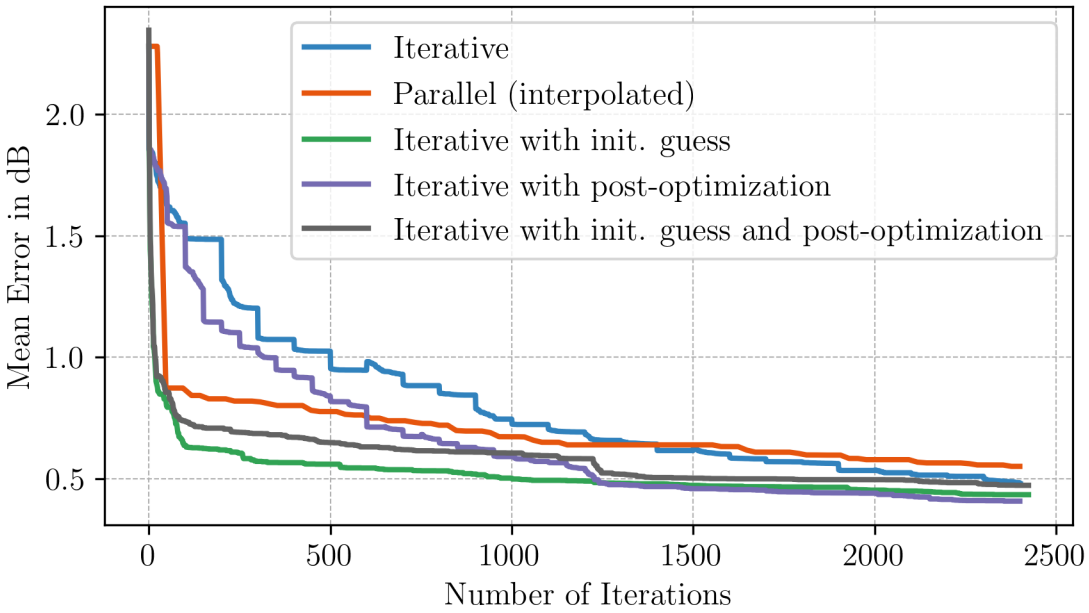


Figure 1.43: Comparison of the mean error between different initial guess & optimization strategies.

Fig. 1.44 shows the magnitude spectra of the compensation filter (the last two mentioned techniques) and the pre-processed LRRC. The thin curves are the actual compensation filters; the thick ones are when they are plotted with a minus sign, so one can see how well they approximate the pre-processed LRRC. For this purpose, the target response was set 0 dB at all frequencies, and the low-frequency roll-off compensation was neglected. We can see that despite the fact that the errors are slightly different, they both perform very similarly.

1.5.6.2 Results

For the presented results, the 'Iterative with init. guess and post-optimization' strategy was chosen with 100 iterations per band. The following parameters for the compensation were used:

- frequency range of compensation from 40 Hz to 15 kHz,
- 1/3-octave magnitude smoothing,
- unrestricted number of bands N,
- minimal area of 0.1 dB · octave of the peak filter to be considered for compensation,
- magnitude target response with a -1 dB/octave tilt,
- limits of the gain G_{dB} of the peak filters [min,max] = $[-12, +6]$,
- limits of the quality factor Q of the peak filters [min,max] = [0.01,6].

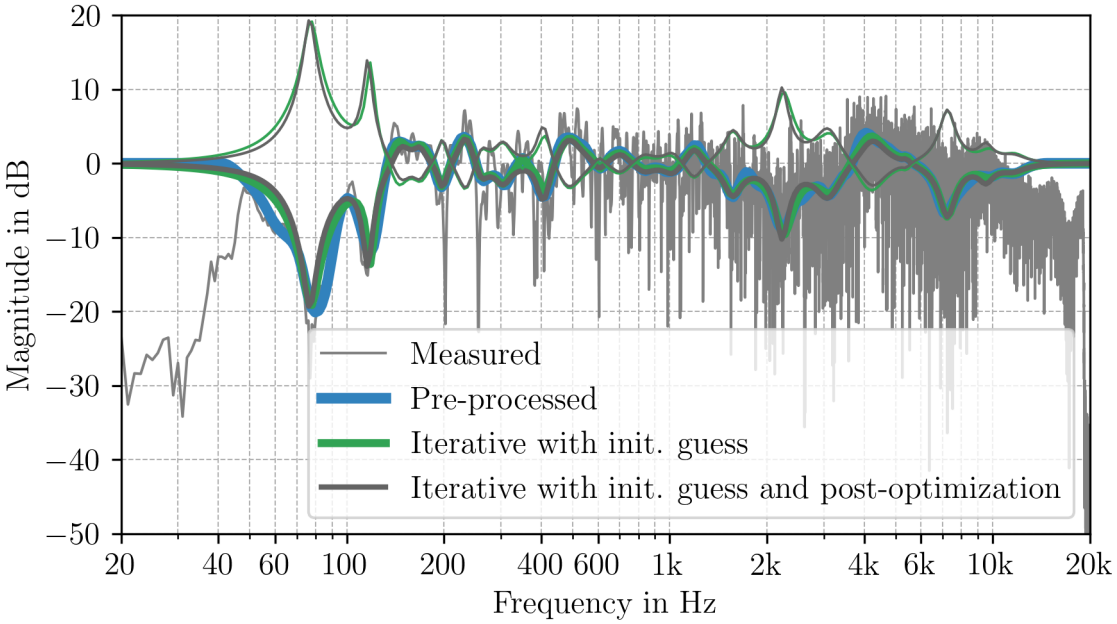


Figure 1.44: Magnitude spectra of the compensation filters of the two best candidates of the initial guess & optimization strategies, together with their inverse to see how well they match with the pre-processed LRRC. Note: $H_{\text{targ, dB}} = 0$, LF roll-off compensation is neglected.

Fig. 1.45 shows the magnitude spectra. For both the calculated and measured compensated spectra, we can see that the situation has improved compared to before compensation. The peak filters would provide a good approximation of the 'perfect' inverse presented in the first approach if the peak filters' parameters weren't limited. This is, however, the big advantage of this method – that one can specify bounds to G_{dB} and Q of the peak filters and thus not 'overdo' it with the compensation. This is especially helpful in areas where large dips occur, in this example just below and above 100 Hz.

Fig. 1.46 shows the corresponding impulse responses. Even here, we can see a positive impact of the compensation filter. The earliest artifacts coming just after the initial peak are reduced, and some of the early reflections coming later in time are slightly reduced as well. Furthermore, it seems that no significant post-ringing is introduced (the situation is way better than for the mixed-phase filter shown in the previous approach).

1.5.6.3 Evaluation

This approach is promising for our implementation goal. The compensation results seem satisfactory, and there is a big benefit of being able to control the degree of compensation, so extensive deviations in the measured response, especially dips, are not fully compensated. Furthermore, the compensation filter, an SOS IIR filter cascade, is ideal for hardware implementation.

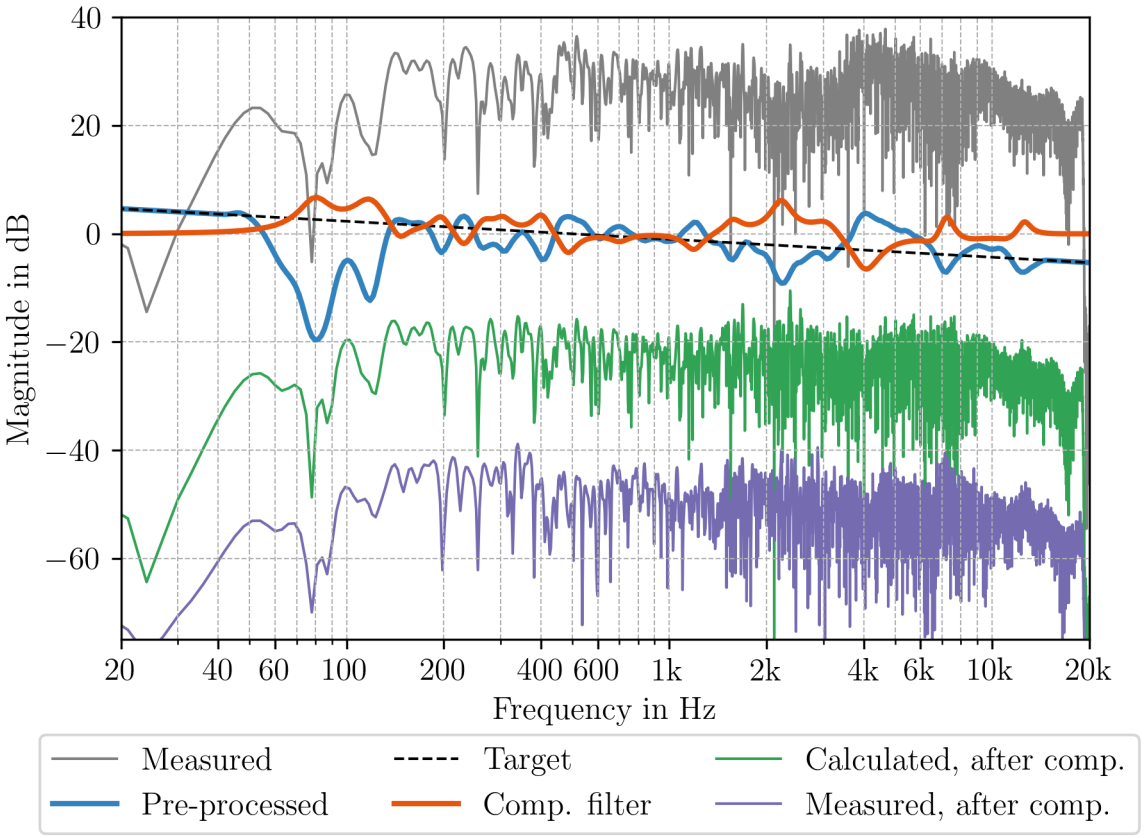


Figure 1.45: Magnitude spectrum of compensation results obtained using the SOS cascade. 'Full' spectra are vertically shifted for better visibility.

1.5.7 Stereo Considerations

When a stereo reproduction is considered, the entire procedure is performed twice: LRIR is measured, and the compensation filter is computed for both speakers. Moreover, considering what was said about the importance of timing in Sec. 1.4.5.4, the measured LRIRs of left and right speakers are compared and their relative shift in samples is determined based on the location of the highest (initial) peaks in the impulse responses. Then, the leading speaker is delayed by the number of samples obtained. Fig. 1.47 and Fig. 1.48 show the compensation done for both left and right channels.

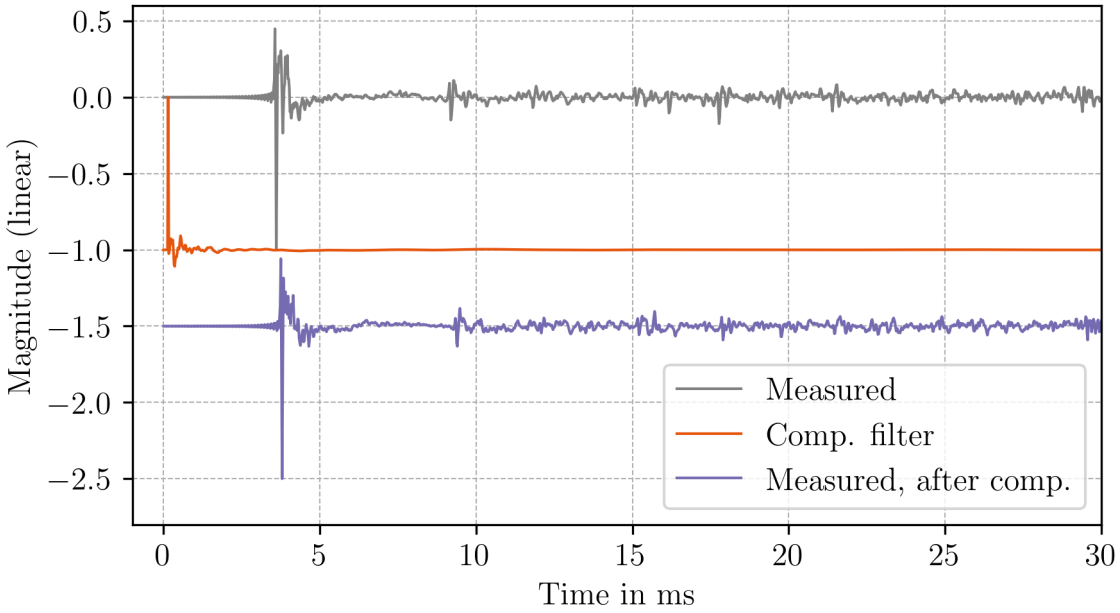


Figure 1.46: LRIR before and after compensation and IR of the compensation filter h_{comp} obtained using the SOS cascade.

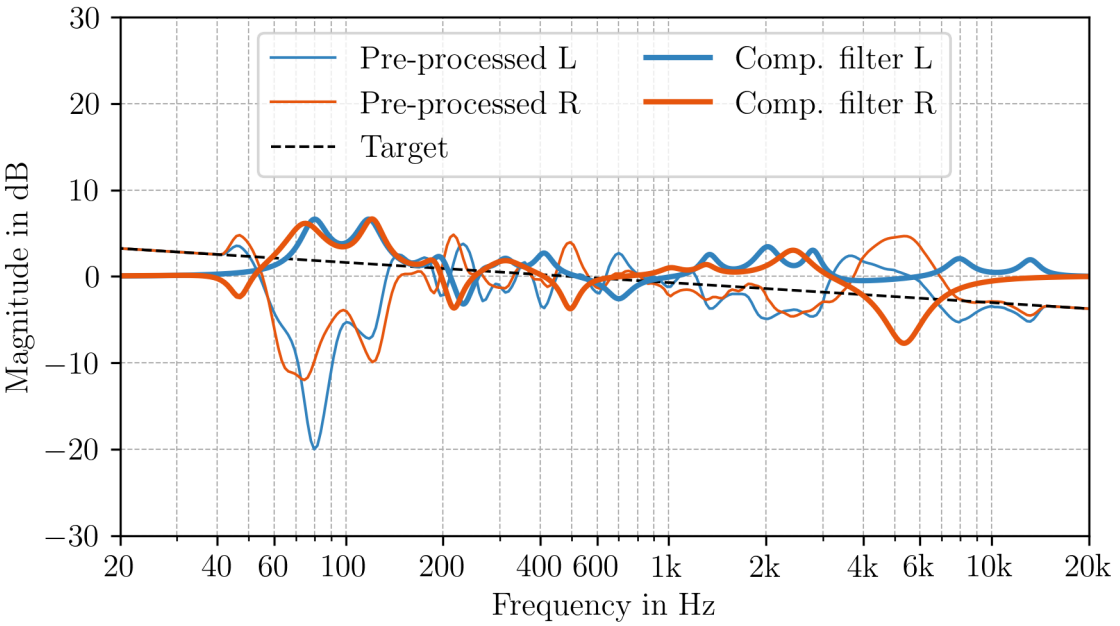


Figure 1.47: Magnitude spectra of pre-processed LRIRs and compensation filters for left and right channels.

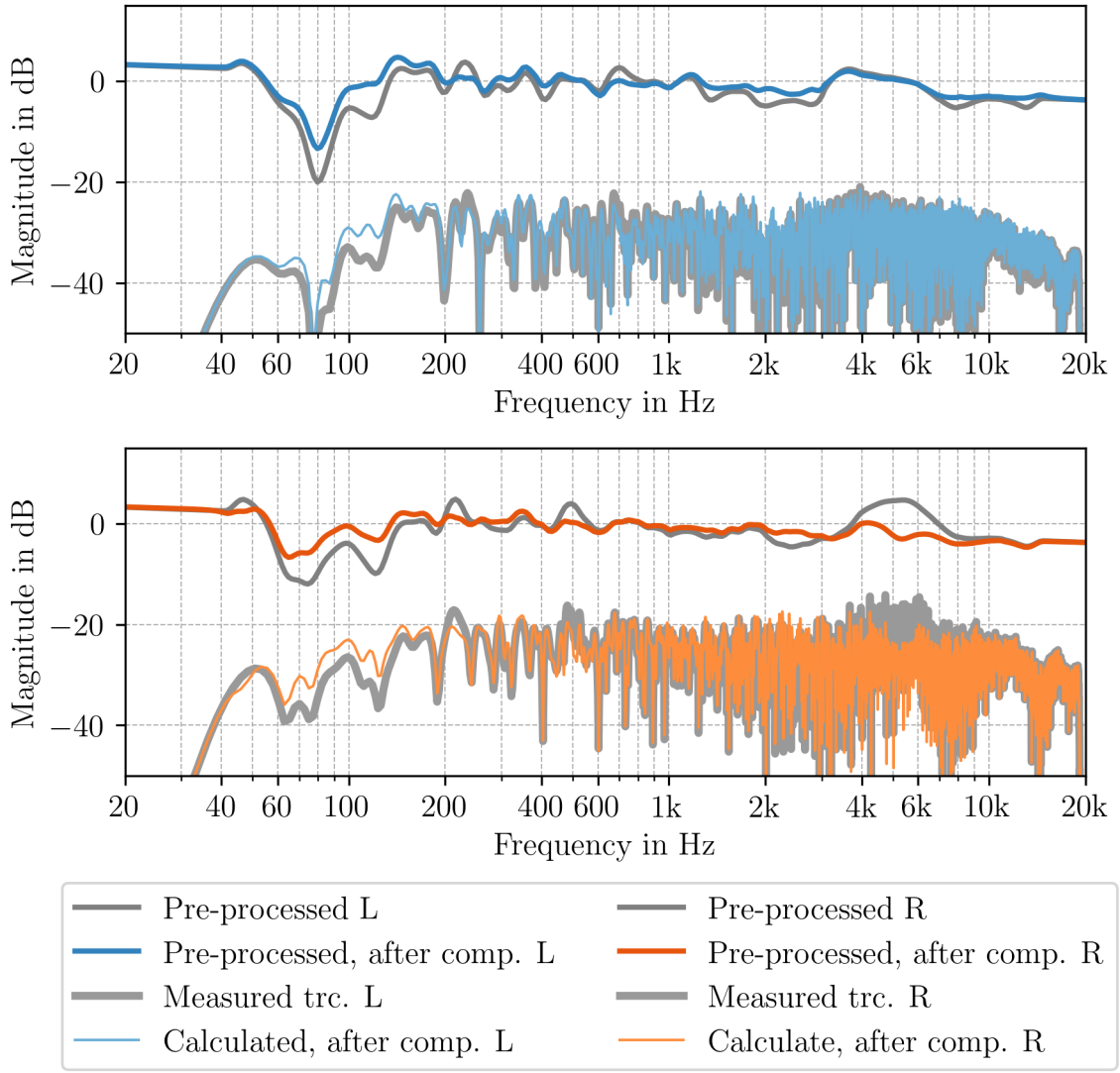


Figure 1.48: Magnitude spectra of compensation results: The compensation filter is applied to the pre-processed LRRC and initially truncated measured LRRC for left and right channels.

1.6 Subjective Evaluation

Subjectively, the results of the equalization were tested as follows: four reference tracks with different musical properties were chosen. First, they were listened to through Beyerdynamic DT 770 PRO headphones, assuming this to be the reference, and then, the tracks were listened to in the room with and without the equalization applied through the stereo pair of AIAIAI's UNIT-4 monitors. The reference tracks were

- Pack'd My Bags, Chaka Khan (Funk This, 2007),
  - Heavily compressed drums in such a way that it is very rich for transients, deep, and punchy bass,
- Summer 1982, Chris Commisso (Funk Punk, 2025),

  - In my opinion, a well-balanced mix when it comes to the spectral balance across low, mid, and high frequencies,
- Norwegian Wood, Jacob Collier[11] (The Light For Days, 2025),
  - Pronounced acoustic guitar in the mid-range and clear vocals that sit in the middle of the stereo image,
- Vltava (The Moldau) from Má vlast, Bedřich Smetana (Czech Philharmonic, Jiří Bělohlávek, 2018),
  - a symphonic poem, smooth, sustained passages with limited percussiveness.

1.6.1 Comparison Between Approaches

From the results presented in the previous section, the clear winner is the third approach, where the compensation filter comprises a cascade of SOS IIR filters. The difference is nothing short of subtle: Both approaches using the FIR filter are not really usable in real life thanks to their inability to control the degree of the compensation. It is quite common that, especially in the low-frequency region, there are deep, broad dips in the measured LRRC (as is also the case in my room, where the main testing was done). If the compensation is not limited, it may result in boosts as high as 20 dB that perceptually do more harm than good. That is why, for further perceptual evaluation, only the third approach mentioned is considered.

1.6.2 Minimum-Phase IIR Compensation Filter

Firstly, I would like to mention that for whatever testing done in mono (using only one speaker), the difference between the equalized and unequalized versions was very subtle. Therefore, all the following comments are based on testing in a stereo configuration. Let's start with a discussion about the choice of different compensation parameters that significantly influence the outcome.

Smoothing: I have tested 1/3-octave and 1/6-octave smoothing and perceptually, the 1/3-octave gave better results.

Magnitude target response tilt: This has quite some influence on the spectral balance of the compensated results. Subjectively, good values were in the interval of $[-0.5, -1.0]$ dB/octave. The higher the (absolute) value, the bassier the sound, with less pronounced highs.

Upper limit of the gain $G_{\mathbf{dB}}$: This parameter is crucial for limiting the maximum gain boost of the individual peak filters. 6 dB is on the safer side.

Upper limit of the quality factor Q: This parameter sets a bound to how narrow the resonance of the peak filters can be. Our perception is less sensitive to narrow notches than boosts [1], so we should be careful about compensating them. $Q = 6$ is also on the safer side.

Frequency range of compensation: This is relevant for two reasons. Firstly, the frequency range of our loudspeaker system that is able to reproduce the sound above a certain frequency. Below, the output power is negligible, and therefore, nothing should be compensated below that threshold. For our loudspeakers, it is

[11] Originally by the Beatles

set to 40 Hz[12]. The second reason might be that we are limited by the measurement device (i.e., the smartphone's microphone), which might deviate from an ideally flat frequency response: If we don't have the calibration data for the exact model for the whole frequency range, we can assume its reliability only in a limited frequency range (discussed later).

(a) (b)

Figure 1.49: POV from the listening position and room where the main testing was done.

1.6.2.1 Equalization On Versus Equalization Off

In total, I have listened to the outcome in five different rooms of smaller to mid-size (approximately from 10 m^2 to 20 m^2). Three of them were very reverberant (being empty or with very little furniture with any absorptive properties), and the other two were more of a normally equipped bedrooms or living rooms with an average expected acoustic qualities.

It is possible to say that the result strongly depends on the acoustic conditions present in the room: If the room is very reverberant, the outcome of the compensation is hardly noticeable – the late reflections are simply too strong. Moreover, for highly undamped modes in the listening room that result in high-gain peaks, even though the compensation filter deploys quite aggressive peak filters to bring those resonances down, it is not enough, and the subjective perception is still unsatisfactory. However, if the room's acoustics are decent (or at least not very bad), the compensation does make quite a noticeable difference.

The main testing was conducted in my single-room apartment, as shown in Fig. 1.49. The corresponding uncompensated and compensated responses are in Fig. 1.48 and the filters used in Fig. 1.47. As seen in the plots, the uncompensated version was quite short on the low end, which was also the subjective perception. The compensation filter helped with this. In addition, the upper mid-range didn't sound quite right, and the compensated version was subjectively much better. Moreover, the sound stage sounded slightly wider. Another listening position, as close as

[12]Note that according to Eq. (1.31), the half window Hanning taper already reduces the amount of compensation from half octave above (i.e., approximately 57 Hz).

possible to the back wall, was also tested. The sound's coloration was significant, and it was quite short on the high frequencies. The computed compensation filters helped with this and significantly improved the perceived image.

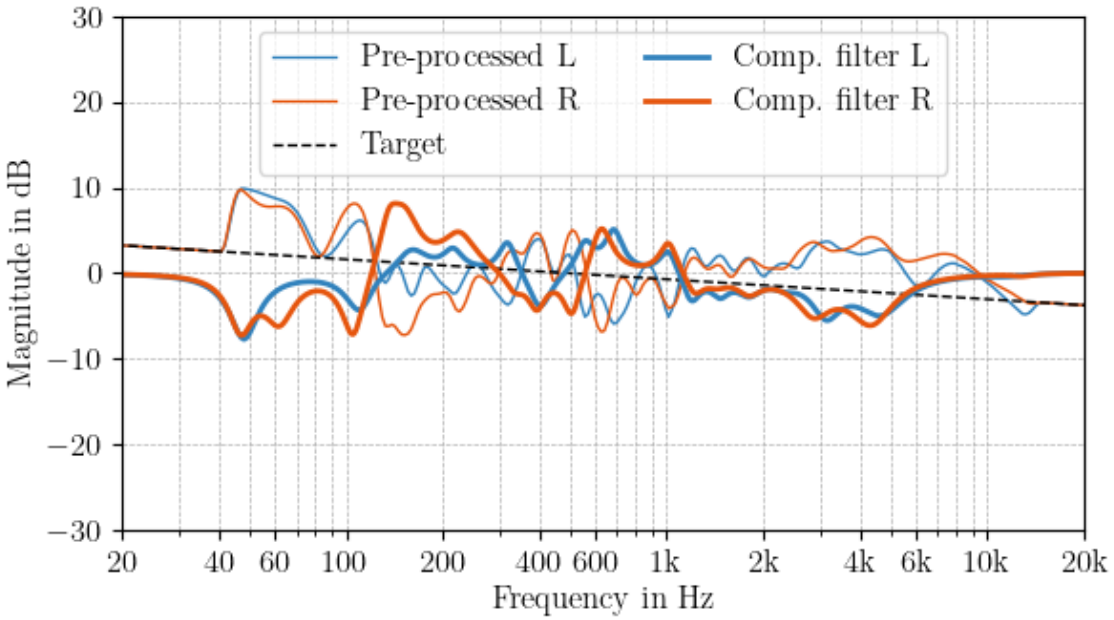


Figure 1.50: Magnitude spectra of pre-processed LRIRs and compensation filters for left and right channels; the listening lab at Chalmers.

For the other test room with reasonable acoustics (the listening lab at Chalmers), the results are shown in Fig. 1.50 and Fig. 1.51. The situation here was quite different: the low frequencies were boosted by the room, so the compensation filter made cuts to compensate. While this is objectively reasonable, subjectively, I would prefer the uncompensated version (at least in the low-frequency range). Finally, in this case, the sound stage widened, too.

Regarding the low-frequency observations, it could be beneficial to provide users with an option to shape the target response interactively, allowing them to influence the compensation filter's outcome, even if it might not yield the most accurate results. When it comes to the increased width of the sound stage, despite the fact that it could be subjectively evaluated as superior, as far as a 'faithful reproduction' is concerned, this is considered an unwanted modification.

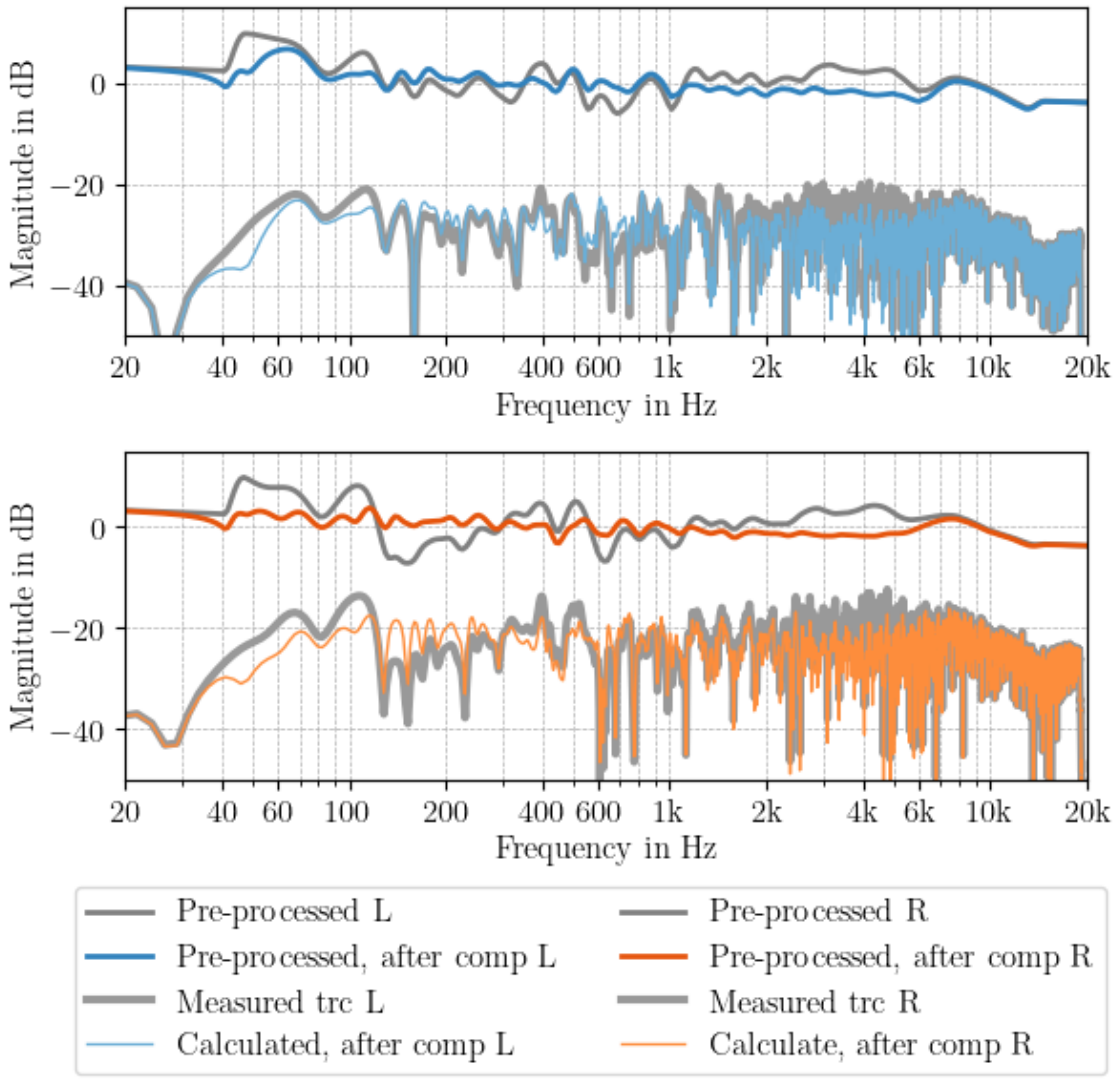


Figure 1.51: Magnitude spectra of compensation results: The compensation filter is applied to the pre-processed LRRC and initially truncated measured LRRC for left and right channels; the listening lab at Chalmers.

The cause is likely that a different compensation filter is computed for each loudspeaker. [43] presents a design framework for compensation filters that account for channel similarity and discusses the motivation behind it: When it comes to a sound field created by two or more loudspeakers, phantom source localization is based on inter-channel intensity and phase differences. This means that if the two LRIR (L speaker-listening position and R speaker-listening position) differ, the sound stage has to be altered. One could argue that with our compensation, we make them more similar. That is true, but only for the magnitude response. When it comes to the phase, since we are using the minimum-phase compensation filters here, not only do we not correct for the phase, but we also slightly modify the original phase by the compensation filter. I believe that this might be the reason why the sound stage appears wider after compensation.

2

Smartphone Microphones

2.1 Construction, Parameters & Expected Behaviour

According to [44] from 2019, microphones in nowadays cell phones can be either of electret condenser type (ECM) or microelectromechanical system (MEMS). While cheaper mobile devices are equipped with ECMs, MEMS microphones are found in more expensive models. [45] from 2020 suggests that ECMs are found rather in featurephones (also known as brick phones) while MEMS microphones are fitted both in featurephones and smartphones. Yet, ECM can be found in smartphones, too, but it is rare: 'For mid-to-high-end smartphones, the penetration rate of MEMS microphones is even more than 90 %' [45]. Therefore, I discuss MEMS microphones in much greater detail in the following text.

Apart from the microphone's size, key parameters are [45]: sensitivity, signal-to-noise ratio (SNR), and acoustic overload point (AOP). A sensitivity of an analog microphone tells us how many (milli)volts we get at the output when the microphone is exposed to the acoustic pressure of 1 Pa and the unit is (m)V/Pa. Regardless of whether a particular smartphone has an analog or digital microphone, we can only access the converted digital data stream. Then, it is expressed as the digital value per Pa, typically with unit FS/Pa with FS referring to *full scale*, when the maximum possible amplitude is scaled to ± 1. It is common to express it in decibels as $20 \log(S)$ in dBFS with S referring to sensitivity. SNR describes the inherent noise of the microphone and it is usually A-weighted and expressed in dBA. For the best microphones, it ranges between 66 dBA and 69 dBA [45]. AOP is the maximum sound pressure level (SPL) a microphone can sense. That is usually defined via total harmonic distortion (THD), another microphone's parameter relating generated harmonics at its output to the fundamental frequency, typically at four or 10 % [45]. There is one key parameter for our application that hasn't been mentioned yet: the frequency response. Its magnitude should be reasonably flat, and the phase as linear as possible (the group delay should be constant).

2.1.1 Electret Condenser Microphone

ECM has a capacitive working principle. A common construction type is that the back electrode is fixed and the front electrode moves due to the sensed sound pressure. The distance between the two electrodes varies, and as a result, the capacitance and the voltage across the electrodes alter accordingly [16]. The back electrode (or

back plate) is a thin perforated metal coated with electret[1] and thus charged. The electrodes are connected to a discrete preamplifier – a field-effect transistor (FET), assembled onto a printed circuit board (PCB) and encapsulated in a metal can. Fig. 2.1 shows its cross-sectional sketch and how such a device looks like in real life.

(a) (b)

Figure 2.1: Cross-sectional sketch of ECM (a) and a real model – type LF-M6027-O from Ariose Electronics (b). Adopted from [45], [46].

2.1.2 Microelectromechanical System Microphone

Over the last 20 years, the demand for MEMS microphones has rapidly grown – from 250 million units sold in 2006 [47] to 7.1 billion in 2021 [48]. And it has a simple explanation: The MEMS microphone has many advantages over the ECM. The unquestionable foremost benefit is its smaller size – in modern smartphones it has a volume of about 5 mm^3, at least ten times smaller than a typical ECM with the same electro-acoustical performance. Its size and surface mountability[2] make it a better candidate for mounting into today's slim smartphones. It also has higher stability in varying temperatures and humidity. The first iPhone fitted with MEMS microphone was iPhone 4, released in 2010 [47]. [44], [45]

MEMS microphones can have several working principles. The most common, the 'industry standard', is the capacitive one, the same as for the ECM. The simplest version comprises the movable membrane and the rigid perforated back plate shown in Fig. 2.2. The difference is that there is no electret material, and the polarization voltage is provided by a so-called application-specific integrated circuit (ASIC) that also contains a preamplifier. In the hi-end models, an analogue-to-digital converter (ADC) is integrated as well (then we talk about a digital MEMS microphone) [44]. The microphone and ASIC are glued onto a PCB and covered by a metal lid. The configuration with one membrane and one back plate is called *single-ended*. Two superior designs exist, improving the nonlinear distortion and AOP. The first one consists of a single membrane between two back plates and two ASICs, each connected to one of the back plates – a *true-differential* configuration. The other one, being also a true-differential configuration, but named as *sealed dual membrane*,

[1] Electret is a dielectric with embedded permanent charge.

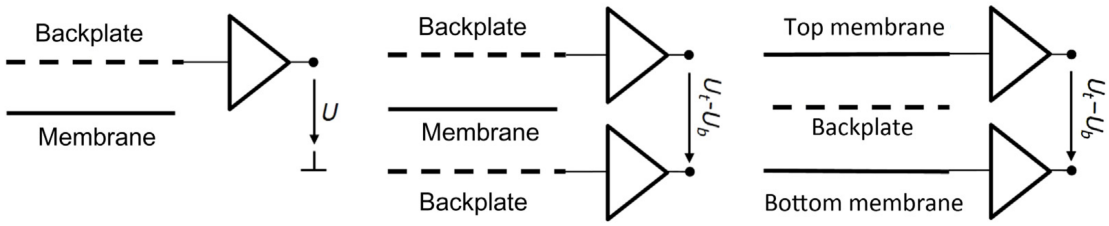
[2] MEMS microphone can be surface mounted onto a printed circuit board (PCB) as opposed to the ECM, which is mounted 'through-hole'.

has two membranes and only one back plate. This construction offers the highest SNR. All three are shown in Fig. 2.3. The SNR is also proportional to the capacitor's size (the membrane together with the back plate). For the increase in SNR, the usual approach is to deploy multiple capacitors simultaneously (in either of the configurations). [45]

Other commercially available transducer principles are piezoelectric and optical. The piezoelectric one does not require the back plate, reducing the manufacturing cost. However, the SNR is limited to 62 dBA, making it unsuitable for top-end smartphone models. On the contrary, the optical principle is rather more sophisticated and expensive, but can offer high SNR. [45]

(a) (b)

Figure 2.2: Cross-sectional sketch of MEMS (a) and a real model – type IMP23ABSU from STMicroelectronics (b). Adopted from [45], [49].



(a) (b) (c)

Figure 2.3: Three configurations of capacitive MEMS microphone: *single-ended* (a), *true-differential* (b) and *sealed dual membrane* (c). Adopted from [45].

What kind of magnitude of the frequency response function (FRF) can we expect from such a microphone? That is what Fig. 2.4 shows. The graph was created using a complex lumped-element model of the most common capacitive principle MEMS microphone [50]. Let's go through the most important elements that we can directly see in the response: The vibrating diaphragm has its mass and stiffness that together form a resonant system. That's the reason for the first resonance with the peak just below 30 kHz[3]. The frequency of the peak is also influenced by the cavity behind the back plate acting as a stiffness. How damped the resonance is is determined by the holes in the back plate [51]. Then, we can see

[3]In this scenario, the opening in the PCB is not only one but four holes that move the resonance up in frequency.

a roll-off at the lowest frequencies. That is due to small ventilation holes that are in the membrane (not shown in Fig. 2.2a) [50]. Lastly, the second resonance with the peak below 200 kHz is caused by the opening in PCB (sound port) and the cavity in front of the membrane: The volume of air in the opening acts as a mass, and the volume of air in the cavity as a stiffness.

In this example, the magnitude response is very flat in the audible range, but in reality, the resonances can be shifted lower (for example, due to only one opening in PCB). That's when additional signal processing that can be implemented within ASIC comes into play. In order to suppress the resonance, a digital second-order low-pass filter can be used [52].

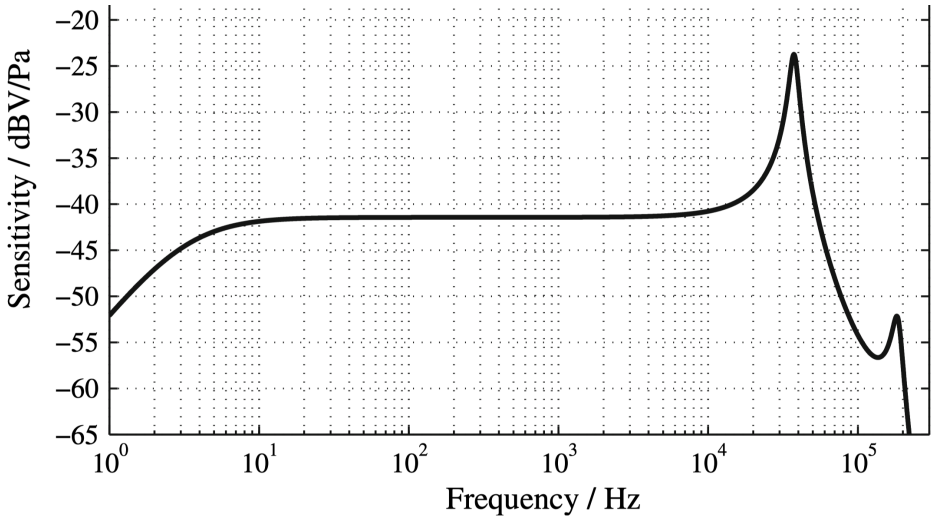


Figure 2.4: Magnitude frequency response of a MEMS microphone. Adopted from [50].

Fig. 2.5 shows the expected group delay of the MEMS microphone. The peak above 20 kHz corresponds to the resonance of the diaphragm. It should match the first resonance from the magnitude response plot. Unfortunately, it does not, because the figures are taken from different sources, which most likely assume different model parameters. Since the vertical axis is logarithmic and a time delay of a fraction of a millisecond is unnoticeable, we can conclude that the MEMS microphone's group delay is expected to be very flat.

2.1.3 Digital Signal Processing

I already mentioned that smartphones can use a filter to suppress a resonance. This can be seen as a processing that has a positive impact on acoustic measurements. However, there is usually way more signal processing going on that should improve the acquisition of speech signals. Low-pass filter for eliminating the low-frequency rumble due to handling or wind, or a so-called auto-gain control (AGC) that serves for maintaining the same amplitude at the output, can be mentioned as examples. [54]

It is clear that such signal manipulation is rather unfavourable for measurements, as the measured data doesn't necessarily have to reflect reality anymore.

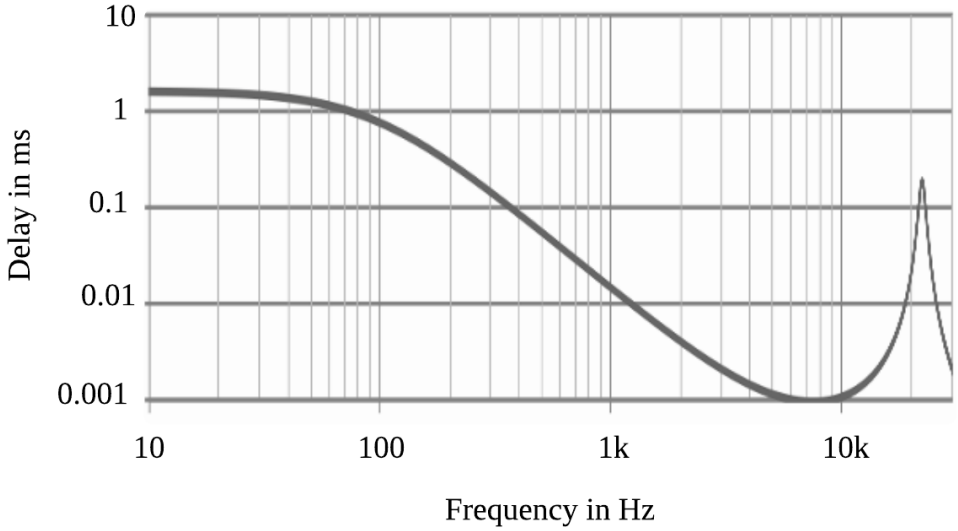


Figure 2.5: Group delay of a MEMS microphone. Adapted from [53].

Thus, it is important that the operating system of the device can offer such an application programming interface (API) that allows the developers of third-party apps to disable such processing. Fortunately, this is the case. For Apple devices running on iOS, from iOS version 6 released in 2012, it is possible to opt for a so-called *measurement* in the *AVAudioSession.Mode* class that disables AGC and high-pass filter with the cut-off frequency of 200 Hz [54]. When it comes to Android, a similar functionality is hidden under the name *UNPROCESSED* of the *MediaRecorder.AudioSource* class under the *android.media* API available from Android 7.0 released in 2016. In the API documentation, one can find that this mode can be accessed 'if available, behaves like *DEFAULT* otherwise' [55]. I discussed with two app developers what this actually means, and the answer is that the manufacturer of the device can decide whether they make it available or not. To make it worse, it seems that this information is nowhere to be found. The only way to find out is to ask for such output and see whether one gets an output signal or not.

2.2 Measurements

The goal is to obtain a frequency-dependent sensitivity, and SNR. Equivalent to frequency-dependent sensitivity is measuring the sensitivity only at one specific frequency – in this case 1 kHz was chosen – and then measuring a relative ('uncalibrated') FRF. I decided to use an indirect measurement method that uses a calibrated reference microphone: The sensitivity (in our case, the uncalibrated values of FRF) of the unknown microphone can be determined when both microphones are exposed to the same free-field pressure. In other words, the same signal is measured by both microphones – either simultaneously or sequentially. There is an IEC standard number 61094-8 [28] that provides details about the measurement procedure when both devices are conventional analog microphones. Even though it was not

possible to meet the standard fully, I followed it where possible, trying to obtain the most accurate results. I performed the measurements in the anechoic chamber at the Division of Applied Acoustics, Chalmers. It has a volume of 800 m^3 and its official operation frequency range is between 75 Hz and 10 kHz where its absorption exceeds 99 %.

2.2.1 Theory: Sensitivity at 1 kHz

As already mentioned before, the sensitivity of a smartphone's microphone is expressed as the digital output value per 1 Pa. In other words, it is a ratio between the digital output value, $d_{\mathrm{SMART,rms}}$, and sensed acoustic pressure $p_{\mathrm{SMART,rms}}$:

$$S_{\mathrm{SMART}} = \frac{d_{\mathrm{SMART,rms}}}{p_{\mathrm{SMART,rms}}} \quad (\mathrm{FS/Pa}). \tag{2.1}$$

The idea is to excite the loudspeaker with a 1 kHz harmonic signal and record that onto the smartphone. The problem is that we don't know what the input pressure $p_{\mathrm{SMART,rms}}$ is. That is why we need the reference microphone. If we use the calibrator that outputs 1 $\mathrm{Pa_{rms}}$, we can get the overall sensitivity of the reference microphone together with pre-amplifier and further amplification and AD conversion in the audio interface:

$$S_{\mathrm{REF}} = \frac{d_{\mathrm{cal,rms}}}{p_{\mathrm{cal,rms}}} \quad (\mathrm{FS/Pa}). \tag{2.2}$$

$d_{\mathrm{cal,rms}}$ is the digital output recorded when the microphone is put into the calibrator and senses $p_{\mathrm{cal,rms}} = 1\,\mathrm{Pa_{rms}}$. Subscript $_{\mathrm{cal}}$ stands for calibrator. Having S_{REF}, we can express the pressure sensed by the reference microphone:

$$p_{\mathrm{REF,rms}} = \frac{d_{\mathrm{REF,rms}}}{S_{\mathrm{REF}}} \quad (\mathrm{Pa}). \tag{2.3}$$

As both microphones are placed at the exact same location,

$$p_{\mathrm{SMART,rms}} = p_{\mathrm{REF,rms}}, \tag{2.4}$$

and we have the final equation for the smartphone's microphone sensitivity:

$$S_{\mathrm{SMART}} = \frac{d_{\mathrm{SMART,rms}}}{d_{\mathrm{REF,rms}}} S_{\mathrm{REF}} \quad (\mathrm{FS/Pa}) \tag{2.5}$$

which is expressed in decibels as

$$S_{\mathrm{SMART,dB}} = 20\log(S_{\mathrm{SMART}}) \quad (\mathrm{dBFS}). \tag{2.6}$$

2.2.2 Theory: Frequency Response/Impulse Response

Traditionally, to obtain FRF of a microphone, one needs an input (test signal) and an output recorded by the microphone. In our scenario, the 'output' is the output recorded by the smartphone, and the 'input' is the output recorded by the reference microphone. This complies with the standard [28] and by doing so, one eliminates

the influence of the reference loudspeaker system and reflections in the chamber (more about it later).

There are several methods for obtaining the FRF of a microphone. I decided to test two that are widely used and choose the one giving better results: the *Exponential Sine Sweep*, which was already fully explained in Sec. 1.4.2.1, and the H_1 *Estimator*.

2.2.2.1 H_1 Estimator

This method uses random noise as a test signal. I used a pink noise whose magnitude in broadband spectrum decreases with 3 dB per octave to improve SNR at low frequencies (the same broadband spectrum is obtained with the exponential sine sweep). The FRF is defined as

$$H(f) = \frac{S_{yx}(f)}{S_{xx}(f)}, \tag{2.7}$$

where S_{yx} and S_{xx} denote the cross-spectrum between input and output the and auto-spectrum out the input, respectively. Eq (2.7) is implemented as follows: Both input and output signals are split into blocks. Then, for every block S_{yx}, S_{xx}, and therefore also $H(f)$ are calculated. The resulting $H(f)$ is calculated as an average of the individual transfer functions. By this technique, we suppress an additive noise that is present during the measurement (apparent in the output signal) compared to calculating the transfer function from the whole output and input signals.

Important parameters are the number of blocks from which we calculate the average, and the number of samples within each block. The more averages, the better estimate of $H(f)$ and the higher block length, the better frequency resolution of the estimate (determines the length of the impulse/frequency response), which is important especially at low frequencies. These together dictate what should be the length of the measurement signal. The quantity that describes how accurate the transfer function is is called coherence function, and it is defined as

$$\gamma^2_{xy}(f) = \frac{|S_{yx}(f)|^2}{S_{xx}(f)\, S_{yy}(f)}, \tag{2.8}$$

where $S_{yy}(f)$ denotes the autospectrum of the output. $\gamma^2_{xy}(f)$ ranges from 0 to 1 with 1 indicating a good measurement – the spectra of the input and output are completely coherent. A value between 0 and 1 indicates partly coherent spectra of the input and output (additive noise is still present).

2.2.3 Theory: Signal-To-Noise Ratio

SNR of a microphone is defined as a difference between the RMS pressure of 1 Pa (94 dBSPL) and the level of A-weighted RMS signal the microphone's self-noise:

$$\mathrm{SNR}_{\mathrm{smart}} = 94\,\mathrm{dBSPL} - 20\log(s_{\mathrm{A\text{-}weigh.,rms}}), \tag{2.9}$$

where $s_{\mathrm{A\text{-}weigh.,rms}}$ is the A-weighted RMS signal of the microphone's self-noise. Having the sensitivity S_{smart} and recorded digital signal of silence (= the self-noise)

$d_{\text{A-weigh.,rms}}$, we can use Eq. (2.3) and rewrite Eq. (2.9) as

$$\text{SNR} = 94\,\text{dBSPL} - 20\log\left(\frac{d_{\text{A-weigh.,rms}}}{S_{\text{smart}}}\right) \quad (\text{dBA}). \tag{2.10}$$

Let's determine the minimal smartphone's SNR, that is required so that the measurement of the LRIR is valid. We want the level of total noise $L_{\text{noise,tot}}$ to be smaller than the level of the test sweep signal L_{sweep} made smaller by the required SNR of the measurement SNR_{req}:

$$L_{\text{noise,tot}} < L_{\text{sweep}} - \text{SNR}_{\text{req}}. \tag{2.11}$$

$L_{\text{noise,tot}}$ can be expressed as a sum of the self-noise of the microphone and the background noise in the room. Using Eq. (2.9), we can write

$$L_{\text{noise,tot}} = 10\log\left(10^{L_{\text{bgn,room}}/10} + 10^{(94\,\text{dBSPL}-\text{SNR}_{\text{smart}})/10}\right). \tag{2.12}$$

Substituting Eq. (2.12) into Eq. (2.11) and solving for $\text{SNR}_{\text{smart}}$:

$$\text{SNR}_{\text{smart}} > 94\,\text{dBSPL} - 10\log\left(10^{(L_{\text{sweep}}-\text{SNR}_{\text{req}})/10} - 10^{L_{\text{bgn,room}}/10}\right). \tag{2.13}$$

According to my measurements,

- $L_{\text{bgn,room}} = 35\,\text{dBA}$,
  - for the background noise level in my test room,
- $L_{\text{sweep}} = 65\,\text{dBA}$,
  - for the conservative level that the AIAIAI's Unit 4 monitors can produce in the listening position in my test room, in the whole frequency range,
- $\text{SNR}_{\text{req}} = 10\,\text{dB}$
  - for the sufficient SNR, according to the results presented in Fig. 1.18.

Plugging these values into Eq. (2.2.3),

$$\text{SNR}_{\text{smart}} > 39\,\text{dBA}. \tag{2.14}$$

2.2.4 Equipment

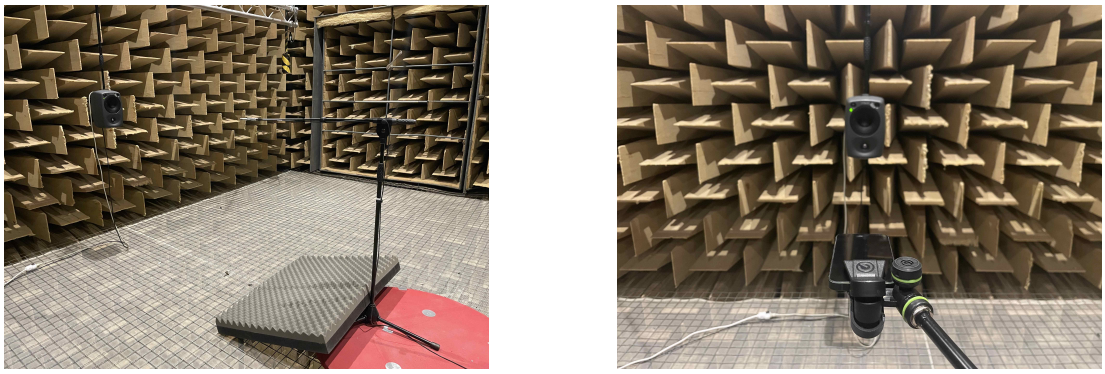
For the measurements, the following equipment was used:

- Audio Interface: Steinberg UR22mkII,
- Power supply: GRAS 12AA,
- Pre-amplifier: Brüel & Kjær Type 2669-C,
- Free field microphone cartridge: Brüel & Kjær 4950,
- Sound source: Genelec 8030A,
- Sound level calibrator: Brüel & Kjær Type 4231.

When it comes to the capture of audio onto smartphones, two apps were used: *Stereo Sound Recorder* by HARDCODED JOY S.R.L. on Android and *SoundMeter X* by Faber Acoustical, LLC. Both of them support the measurement mode or unprocessed audio stream, but for the *Stereo Sound Recorder* (as it is for all the other existing recorder apps for Android), they are programmed with a fallback option – if the uprocessed output is not available, then the default (processed) one is output. Therefore, whether a given Android device supports the unprocessed audio stream cannot be properly assessed. All recordings are done with a sampling frequency of 48 kHz and 24-bit depth.

2.2.5 Setup

Fig. 2.6 shows the measurement setup. Both the reference and the smartphone's microphones were placed at the exact same position and measured sequentially. As the microphone stand was placed on a reflective surface (the red circle), I placed a porous absorber on the edge, trying to reduce reflections. Unfortunately, this was not their only source – on one spot in the ceiling of the chamber, a few wedges of the porous absorbers were missing, causing the reflected wave to reach the microphone. Therefore, I experimented with the distance between the loudspeaker system and the microphone that had a certain influence on the outcome, and finally settled on 120 cm. Fig. 2.7 illustrates how the spectrum of the recorded signal onto the reference microphone looks. There is a clear comb filter from 200 Hz onwards with the biggest difference about ±2 dB. Of course, the fluctuation can also be caused to a certain degree by the response of the loudspeaker system. Especially at frequencies below 200 Hz, we can see limited output from the system that decreases significantly below 55 Hz.



(a) (b)

Figure 2.6: Setup of the smartphones' microphones measurement.

2.2.6 Initial Testing

According to the standard [28], 'The sound source shall not produce distortion components that may generate a significant response from the microphone under test and/or reference microphone at frequencies other than the test frequency'. Therefore, I measured the response of the reference loudspeaker system with the reference microphone at several gains to see what gain is the best compromise between good SNR and no significant harmonic distortion.

As a next step I did measurements both with the reference microphone and the smartphone to be able to extract our interest – the FRF of the smartphone's microphone – and tested various lengths of the test signals. For the sweep excitation, I tested lengths from 2 s to 16 s, and the obtained FRF was almost invariant for different measurement lengths. This indicates that we can trust the obtained response. I decided to measure with the length of 10 s to obtain a good SNR.

When it comes to the H_1 estimator, it turned out that this measurement method did not work when obtaining the frequency response of the smartphone's microphone

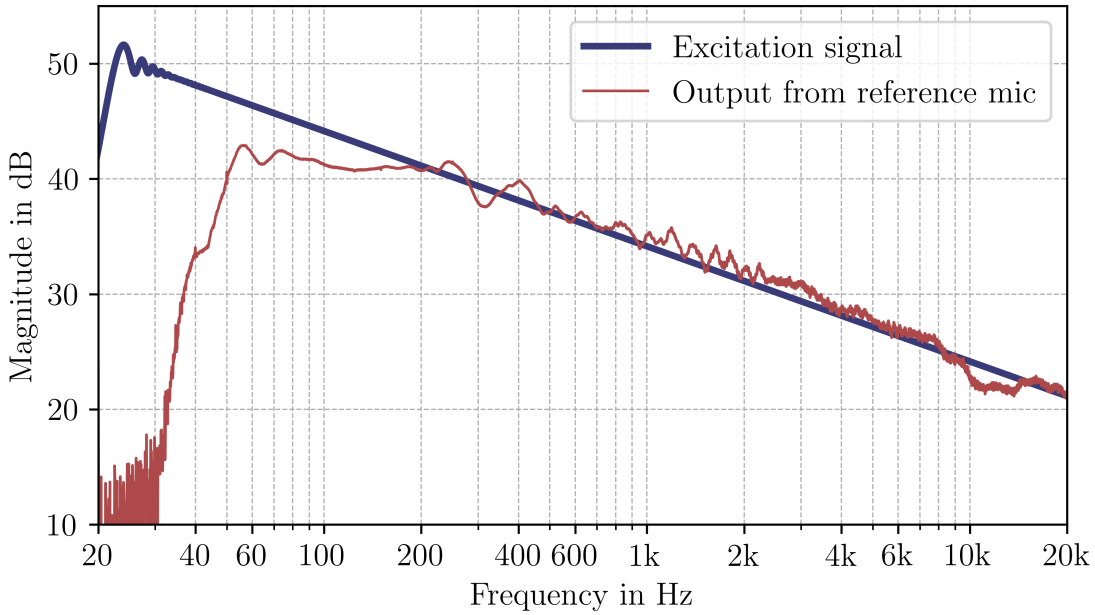


Figure 2.7: Magnitude response of the sweep excitation and recorded signal onto the reference microphone, overlayed.

using the output of the reference microphone as the input. Even with long measurement signals (two minutes), ensuring significant averaging and long block lengths, the values of the coherence function were very low in almost the whole frequency range, and the frequency response didn't match the one obtained through the sweep method. Unfortunately, I have no explanation for such behaviour. Fig. 2.8 shows this on the computed magnitude frequency response of iPhone 12 mini.

2.2.7 Procedure

For the actual measurements, once for the reference microphone and then for every smartphone I recorded

- 10 s of 1 kHz sine wave:
  - The RMS value is calculated from it and used for obtaining the sensitivity.
- Three times 10 s of exponential sine sweep, each with 5 dB step in gain:
  - This is used to obtain the FRF/IR.
  - The IR is cut to 300 ms length. As this can create sharp edges at the lowest frequencies and therefore change the response, the IR is filtered by an 8th order high-pass filter with a cut-off frequency of 30 Hz before cutting.
  - The FRF is calculated three times (for each measurement with different gain) to test whether it has any influence on the FRF.
- 30 s of pink noise:
  - As I showed previously, the H_1 estimator doesn't work as expected, so this is recorded just to possibly see if there is some processing happening (attenuation, filtering, etc.).

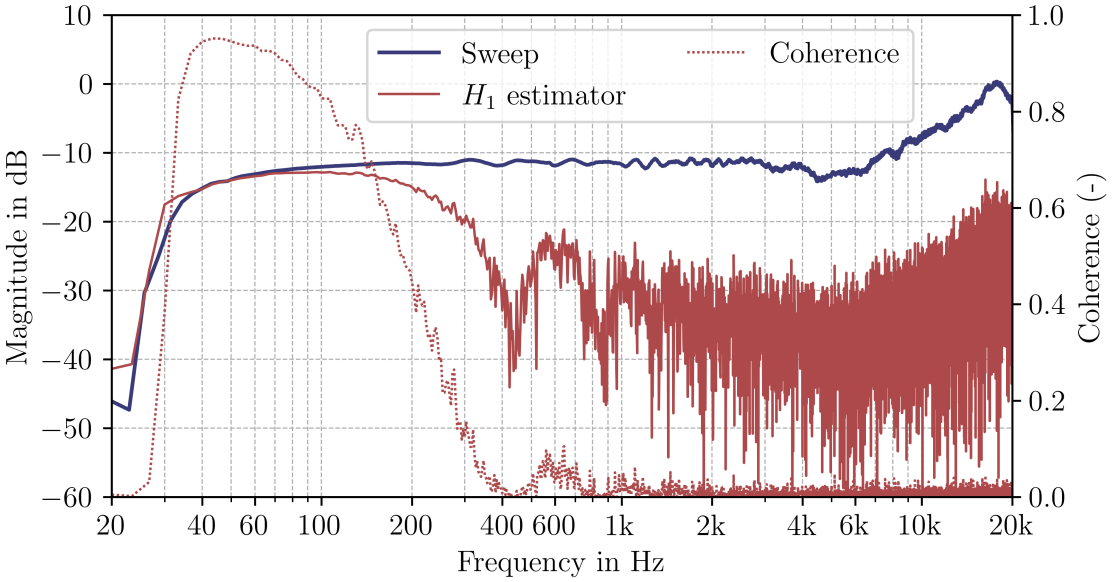


Figure 2.8: Magnitude response of iPhone 12 mini both from the sweep excitation and H_1 estimate (of 192 blocks with the length of 300 ms – 30 s of noise with 50% block overlap) together with its coherence.

Fig. 2.9 shows a comparison between the raw and processed impulse response and its corresponding magnitude frequency response, together with different levels of gain of the sweep excitation. In the impulse response plot, we can see quite a significant peak about 30 ms after the main one. One could think that this is the main reflection causing the comb filter seen in Fig 2.7. However, I've cut it from the IR and it had no influence on the ripple, so it is not the cause. In the magnitude response we can see how noisy the response is at the lowest frequencies. I suggest this is mainly due to the very low output of the loudspeaker system (see Fig. 2.7), and therefore I decided to consider valid results of the measurements only above 40 Hz. To align the recordings from a smartphone and the reference microphone, a cross-correlation is used[4].

2.2.7.1 Problems With the Processing

As already mentioned in section 2.2.2, calculating the impulse response of the smartphone microphone using the output of the reference microphone signal as an input, one in theory eliminates the influence of the loudspeaker system and response of the room. Let's look at it in more detail. If an excitation signal s_{in} is recorded in the anechoic chamber onto a smartphone microphone, it can be expressed as follows:

$$s_{\text{out,smart}} = s_{\text{in}} * h_{\text{speaker}} * h_{\text{room}} * h_{\text{smart}}, \tag{2.15}$$

where $s_{\text{out,smart}}$ stands for the recorded output signal and h_{speaker}, h_{room} and h_{smart} for the impulse response of the loudspeaker system, the impulse response of the anechoic chamber (possible reflections) and the impulse response of the smartphone

[4]That worked in 80 % of cases – otherwise the signals were manually aligned.

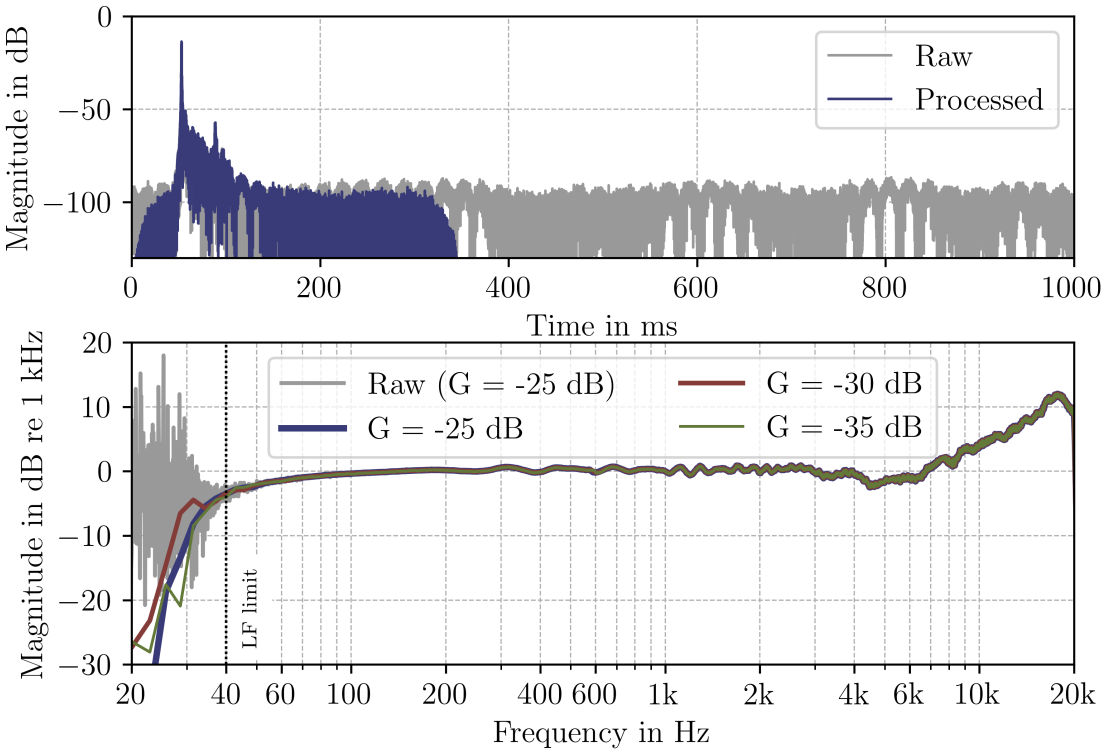


Figure 2.9: Comparison of raw and processed impulse response and its corresponding magnitude response for different values of gain (G) of iPhone 12 mini.

microphone, respectively. In a similar way, we can express the output recorded onto the reference microphone $s_{\text{out,ref}}$. If we assume that its impulse response is 1[5],

$$s_{\text{out,ref}} = s_{\text{in}} * h_{\text{speaker}} * h_{\text{room}}. \tag{2.16}$$

Now we can substitute s_{in} from Eq. (2.16) into Eq. (2.15):

$$s_{\text{out,smart}} = s_{\text{out,ref}} * h_{\text{smart}}, \tag{2.17}$$

and obtain the desired impulse response of the smartphone microphone:

$$h_{\text{smart}} = s_{\text{out,smart}} * s_{\text{out,ref}}^{-1}. \tag{2.18}$$

Fig. 2.10 compares this to simply convolving the output with the excitation signal. We can see that it has certain influence, but there is still some ripple left.

As I already mentioned, the reference microphone signal was recorded only once at the beginning of all the measurements. Unfortunately, as the measurements of the smartphones were ongoing for more than a week, for part of the measurements, the measurement conditions were slightly changed: the grey porous absorber on the ground (see Fig. 2.6a) was slightly shifted, which resulted in changing the impulse response h_{room}. We can then rewrite Eq. (2.15) as

$$s_{\text{out,smart}} = s_{\text{in}} * h_{\text{speaker}} * h_{\text{room 2}} * h_{\text{smart}}, \tag{2.19}$$

[5]The response of the microphone is so close to it that we can do that.

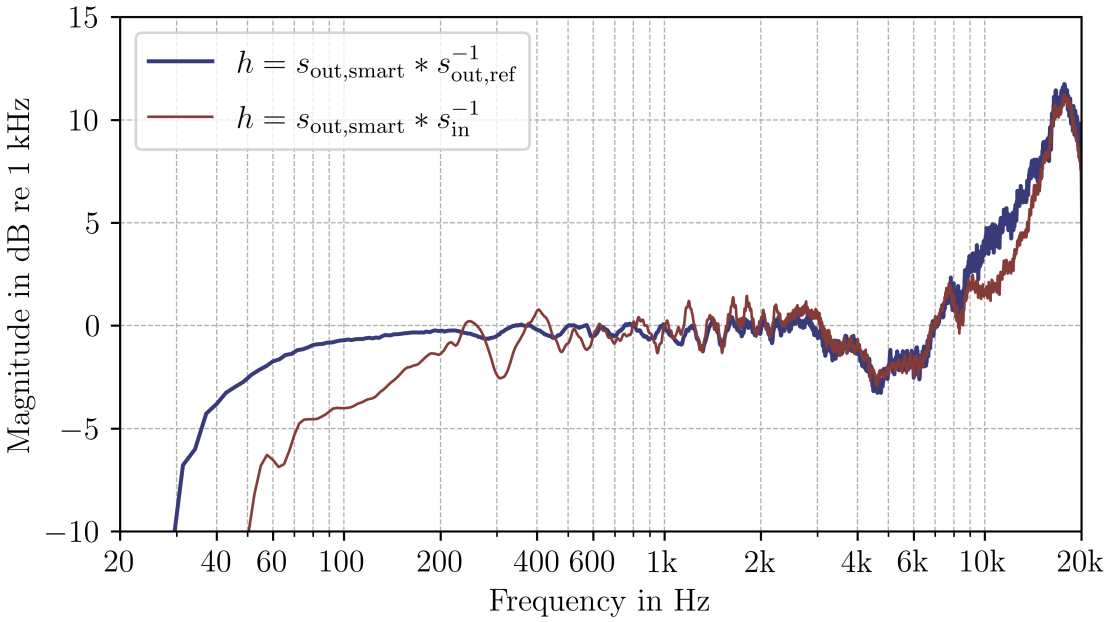


Figure 2.10: Comparison of magnitude FRF of iPhone 12 mini calculated using the output of the reference microphone signal $s_{\text{out,ref}}$ and excitation signal s_{in} as the input.

with $h_{\text{room 2}}$ representing the impulse response of the room when the conditions have changed. If we substitute s_{in} from Eq. (2.16) into Eq. (2.19), we get

$$s_{\text{out,smart}} = s_{\text{out,ref}} * h_{\text{smart}} * h^{-1}_{\text{room}} * h_{\text{room 2}} * h_{\text{smart}}. \tag{2.20}$$

This will result in a modified smartphone impulse response

$$h_{\text{smart,mod}} = h_{\text{smart}} * h^{-1}_{\text{room}} * h_{\text{room 2}} = s_{\text{out,smart}} * s^{-1}_{\text{out,ref}}. \tag{2.21}$$

Fortunately, I recorded the same phone twice under both these conditions, and thus I could evaluate the impulse response 'of the condition change':

$$h_{\text{change}} = h_{\text{room 2}} * h^{-1}_{\text{room}} = h_{\text{smart,mod}} * h^{-1}_{\text{smart}}, \tag{2.22}$$

and express the correct impulse response as

$$h_{\text{smart}} = h_{\text{smart,mod}} * h^{-1}_{\text{change}}. \tag{2.23}$$

Eq. (2.21) was used for all the phones from a certain measurement day when the porous absorber was shifted (from certain day, all the responses were more 'wiggly'). Magnitude FRF of all impulse responses from Eq. (2.21)-Eq. (2.23) are shown in Fig. 2.11.

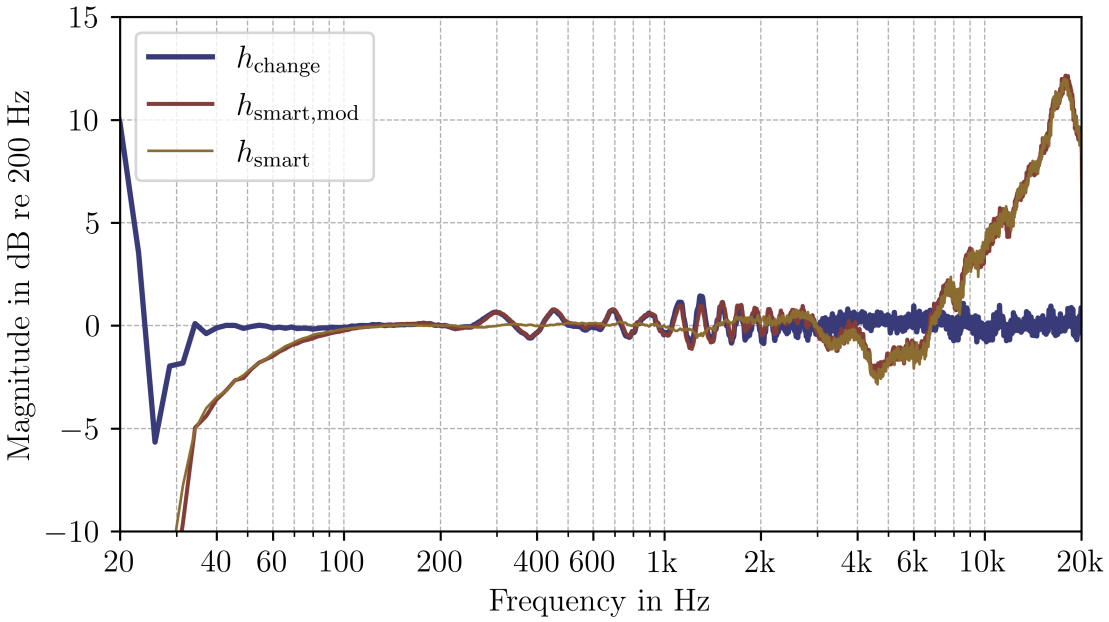


Figure 2.11: Magnitude FRF of h_{change}, $h_{\text{smart,mod}}$ and h_{smart}.

2.2.8 Results

In total, 30 devices were measured: 16 iPhones (11 different models) and 14 devices running Android (14 different models). Tab. 2.1 shows all the measured sensitivities, where green and red colours mark the highest and lowest sensitivities measured, respectively. We can see that the spread of iPhones is ±2 dB, which is very small. For Android devices, the range is within ±19 dB. The highest number in total has Samsung Galaxy S23 with −13.7 dBFS, while the lowest one is −51.5 dBFS for Motorola Edge 50 Pro. The value for Samsung Galaxy A52 5G is missing because during the measurement, both the 1 kHz sine wave as well as the pink noise got attenuated after a few seconds, signifying that the *UNPROCESSED* output option couldn't be accessed, and some sort of noise suppression was engaged, making it impossible to do such measurements.

Sensitivity is not so important when it comes to the performance of the microphone. However, thanks to sensitivity, we can determine the SNR, which is shown in Tab. 2.2. Earlier, we derived a minimum desired SNR value for the measurement device: 39 dBA. All measured phones have SNR values that exceed this limit, making them, in theory, suitable for the LRIR measurements from the perspective of the self-noise. However, there are big differences between the models. All iPhone models exceed the SNR of 63 dBA, and the best-performing model, iPhone 16 Pro, has a value of 72.1 dBA. For the measured devices using Android, the span is higher than 20 dB with the lowest value of 42.6 dBA for the Motorola Edge 50 Pro, and the highest value of 63.7 dBA for the Samsung Galaxy S23.

Magnitude frequency responses af all the individual models are given in Appendix A. Fig. 2.12 shows the mean and spread of all the iPhone devices. We can see that the mean stays within ±5 dB limits. The highest dispersion is ±20 dB, but that is true only for the highest frequencies in the audible range (and it is mainly

Table 2.1: Measured sensitivities.

Model (iPhone)	$S_{\mathbf{dB}}$ **in dBFS**	**Model (Android)**	$S_{\mathbf{dB}}$ **in dBFS**
iPhone 7	-23.8	Fairphone 3 Plus	-37.4
iPhone 11	-22.9	Google Pixel 7a	-35.4
iPhone 12 (I)	-22.2	Huawei Mate 20 Pro	-23.6
iPhone 12 (II)	-23.7	Huawei P30 (ELE-L29)	-34.6
iPhone 12 mini (I)	-26.0	Motorola Edge 50 Pro	-51.5
iPhone 12 mini (II)	-25.7	Motorola Moto G8 Power	-23.0
iPhone 12 Pro Max	-22.5	OnePlus 9	-50.2
iPhone 13	-26.2	Samsung Galaxy A8 (2018)	-30.2
iPhone 13 mini (I)	-22.7	Samsung Galaxy A34 5G	-43.6
iPhone 13 mini (II)	-25.4	Samsung Galaxy A52 5G	–
iPhone 13 mini (III)	-24.5	Samsung Galaxy A53	-18.1
iPhone 14 (I)	-22.8	Samsung Galaxy S8	-38.9
iPhone 14 (II)	-23.0	Samsung Galaxy S23	-13.7
iPhone 15	-24.3	Xiaomi Redmi 9	-33.1
iPhone 16 Pro	-23.4		
iPhone SE (2022)	-25.6		

caused by one model, which deviates from the trend of all the other devices). Towards lower frequencies, the responses become more and more consistent, and below 1 kHz, the deviation is only up to ±1 dB. There is no device whose response was 'ugly', significantly deviated from the others.

For Android devices, the situation is different: 11 of the 14 measured devices had 'reasonable' responses. Their magnitudes are plotted in Fig. 2.14. We can see that the mean is within ±11 dB – worse than for iPhones. However, when it comes to the highest spread, it is only ±13 dB, which is significantly lower than for iPhones. Towards lower frequencies, the trend is similar to that of iPhones, but the responses are less coherent in comparison, especially below 200 Hz. At the lowest valid measured frequency, 40 Hz, the deviation is bigger than ±10 dB, whereas for iPhones, it is up to ±0.5 dB.

The corresponding group delays of the so far mentioned devices are shown in Fig. 2.17 and Fig. 2.15, respectively, for iPhones and Androids. In both cases, the performance is the same: Smooth curve with no present spikes, and 0 ms delay down to 150 Hz. Then, it climbs up to 20 ms at the lowest valid measured frequency. This is way higher than what was presented in the theory part. It is probably due to the internal DSP, most likely a high-pass filter.

Three of the measured devices using Android had inferior results. These were: Huawei P30, Motorola Edge 50 Pro, and OnePlus 9. The magnitude responses are shown in Fig. 2.16 and the corresponding group delays in Fig. 2.16. If we look at the magnitude plot, we can say that the curve is anything but flat. Surprisingly,

Table 2.2: SNR of all measured devices.

Model (iPhone)	SNR in dBA	Model (Android)	SNR in dBA
iPhone 7	63.5	Fairphone 3 Plus	61.1
iPhone 11	66.7	Google Pixel 7a	63.6
iPhone 12 (I)	64.5	Huawei Mate 20 Pro	58.5
iPhone 12 (II)	63.8	Huawei P30 (ELE-L29)	45.2
iPhone 12 mini (I)	63.6	Motorola Edge 50 Pro	42.6
iPhone 12 mini (II)	63.3	Motorola Moto G8 Power	62.6
iPhone 12 Pro Max	64.4	OnePlus 9	48.9
iPhone 13	66.7	Samsung Galaxy A8 (2018)	61.7
iPhone 13 mini (I)	67.0	Samsung Galaxy A34 5G	57.0
iPhone 13 mini (II)	66.4	Samsung Galaxy A52 5G	–
iPhone 13 mini (III)	67.2	Samsung Galaxy A53	62.6
iPhone 14 (I)	67.8	Samsung Galaxy S8	58.8
iPhone 14 (II)	67.1	Samsung Galaxy S23	63.7
iPhone 15	67.4	Xiaomi Redmi 9	57.9
iPhone 16 Pro	72.1		
iPhone SE (2022)	63.3		

all of the devices have the same trend: strong dips around 1 kHz, 3 kHz, 5 kHz, and 11 kHz. As expected, we can see spikes in the group delay plot at the corresponding frequencies. I believe that it is not likely that the MEMS chip would have such a bad performance. It contradicts the expected behaviour based on the chip's physical properties. Therefore, in my opinion, it is the internal DSP that is most likely engaged, which causes that behaviour.

2.2.8.1 Deviations Among One Model

I received three devices of the model iPhone 13 mini, so we can look at how different their frequency responses are. Their average and spread are shown in Fig. 2.18. We can see that they are very coherent, deviating by up to ±1 dB in the low-frequency region and by ±2 dB at the dip above 11 kHz.

2.2.8.2 Influence of the Case and Directivity

Fig. 2.19 shows how both the presence of a phone case and holding the smartphone at an angle influence the magnitude frequency response. We can see that on-axis measurement versus holding the phone at a 45 ° angle makes almost no difference to the results. However, when the case is on (in this instance, the case was quite thick – it created a few mm waveguide for the sound before reaching the sound inlet), above 3 kHz, the response starts to deviate because the resonance being originally just below 20 kHz moves lower to approximately 11 kHz.

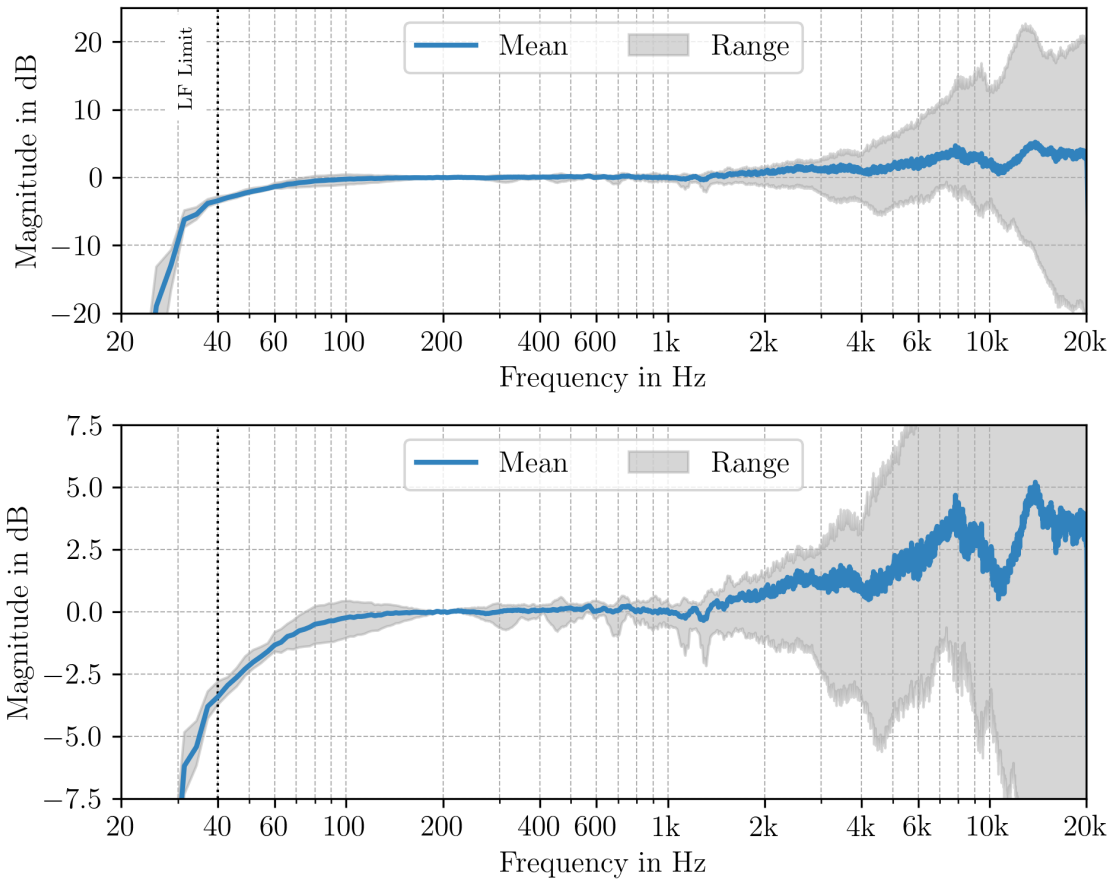


Figure 2.12: Magnitude FRF of all measured iPhones. Normalized to the value at 200 Hz.

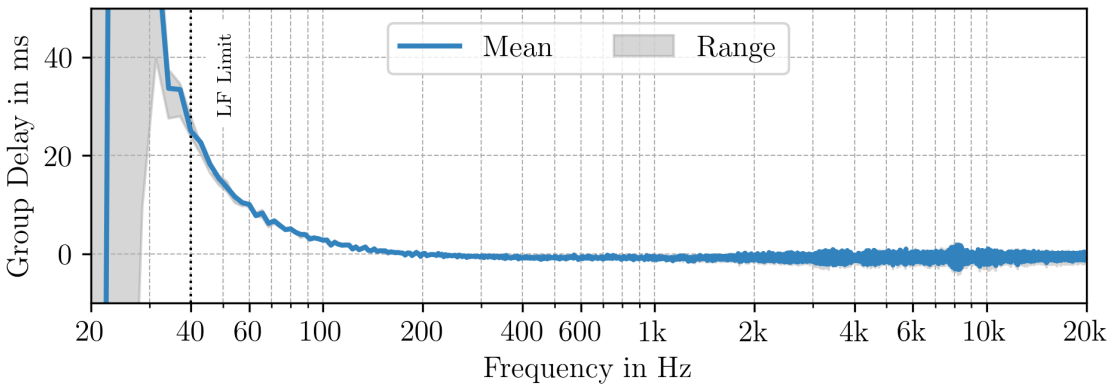


Figure 2.13: Group delay of all measured iPhones.

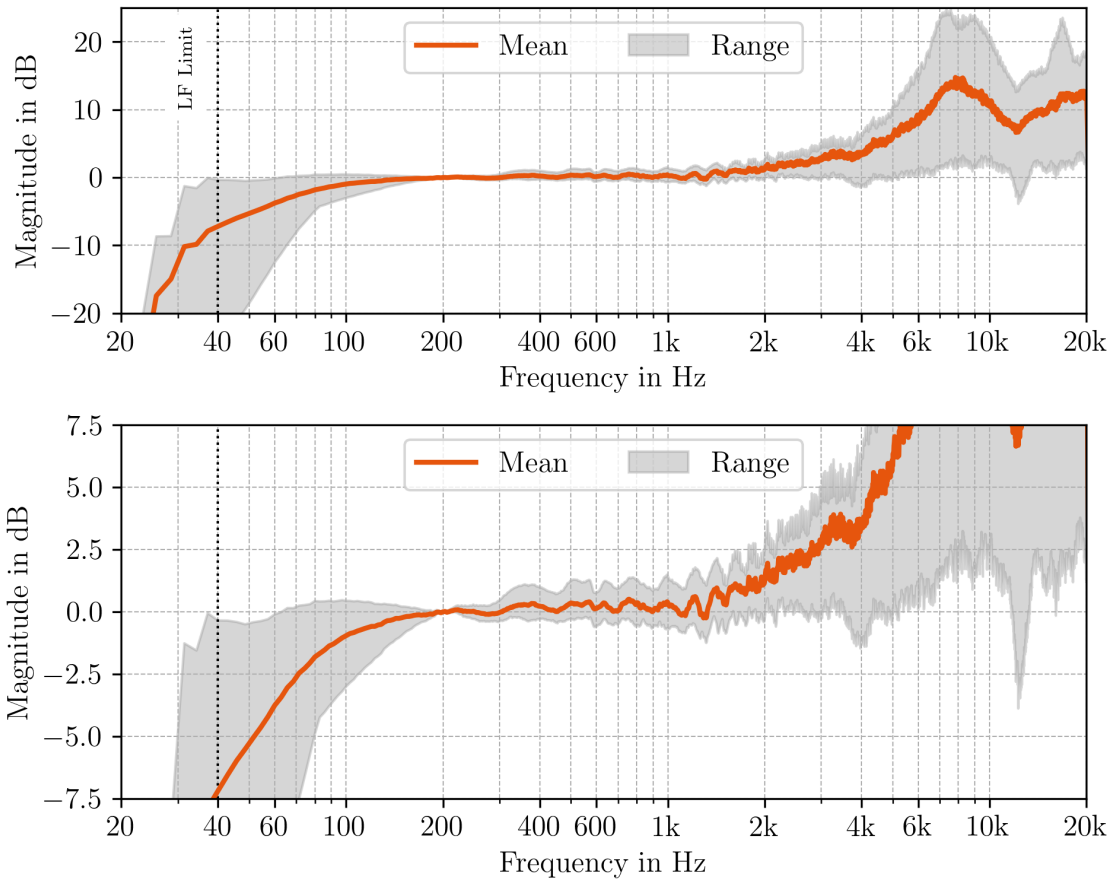


Figure 2.14: Magnitude FRF of all measured Androids except for three inferior ones. Normalized to the value at 200 Hz.

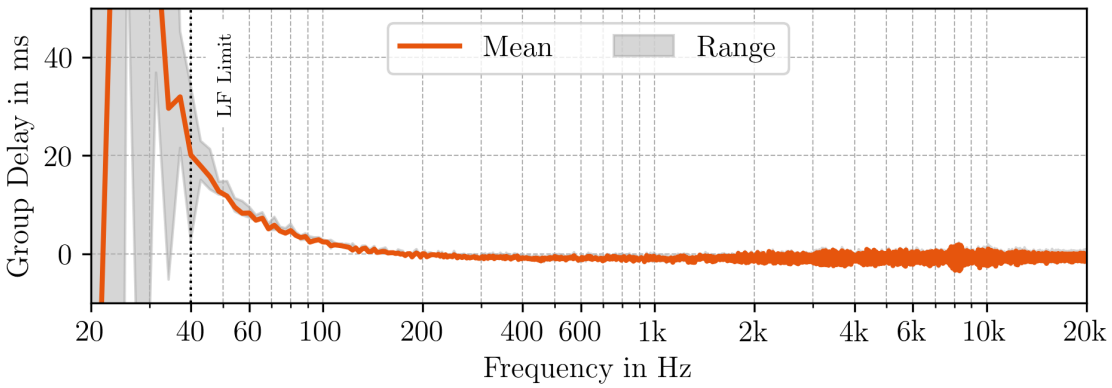


Figure 2.15: Group delay of all measured Androids except 3 bad ones.

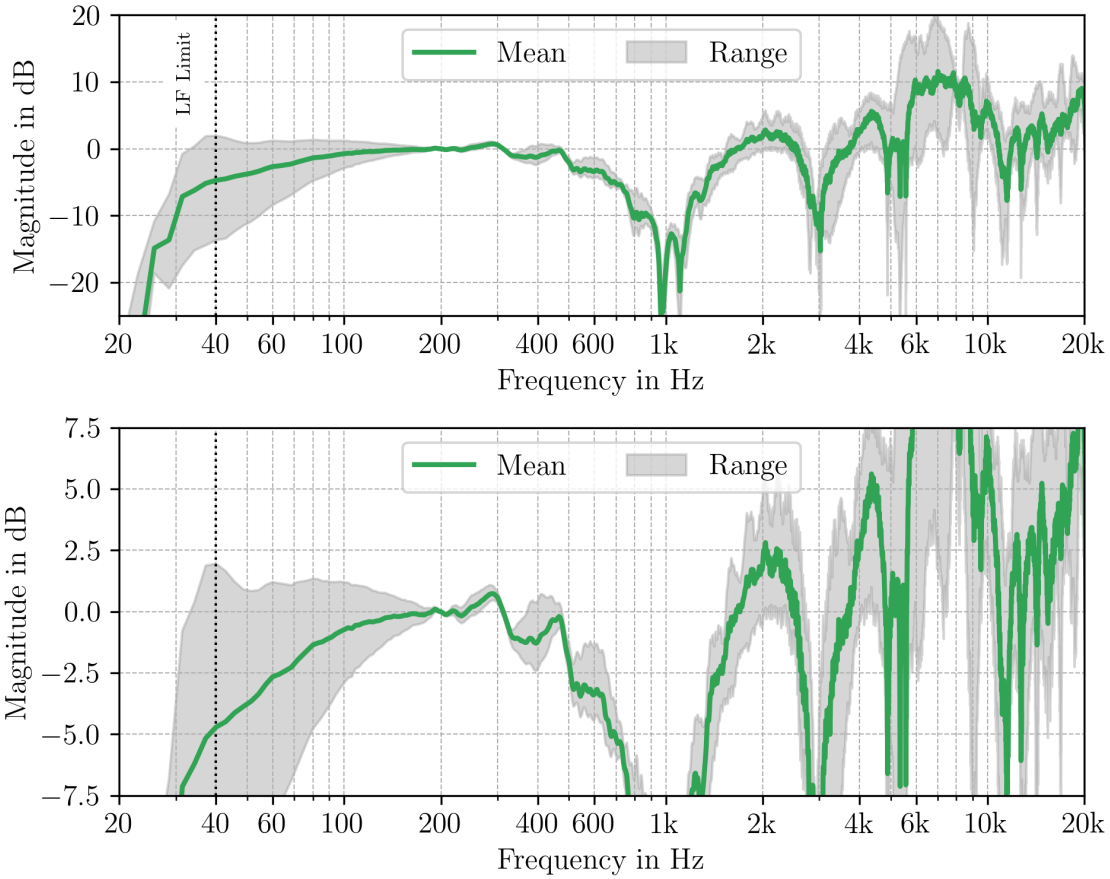


Figure 2.16: Inferior Magnitude FRF of three models: Huawei P30, Motorola Edge 50 Pro, and OnePlus 9. Normalized to the value at 200 Hz.

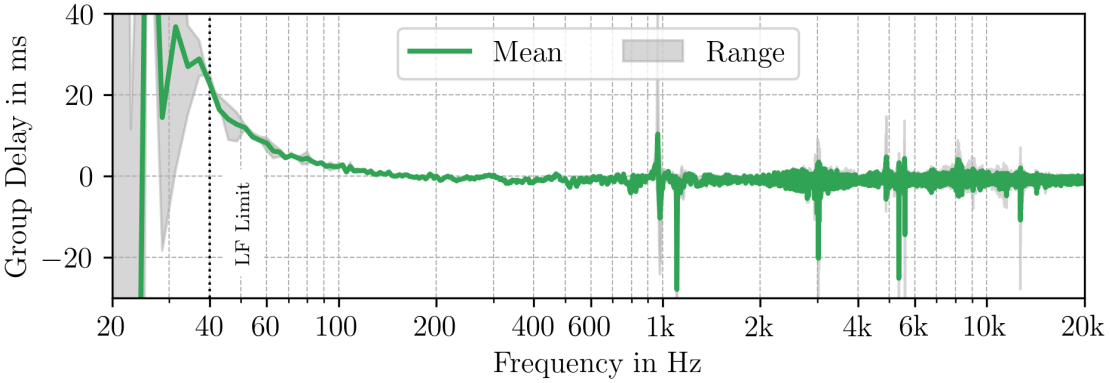


Figure 2.17: Inferior group delay of three models: Huawei P30, Motorola Edge 50 Pro, and OnePlus 9.

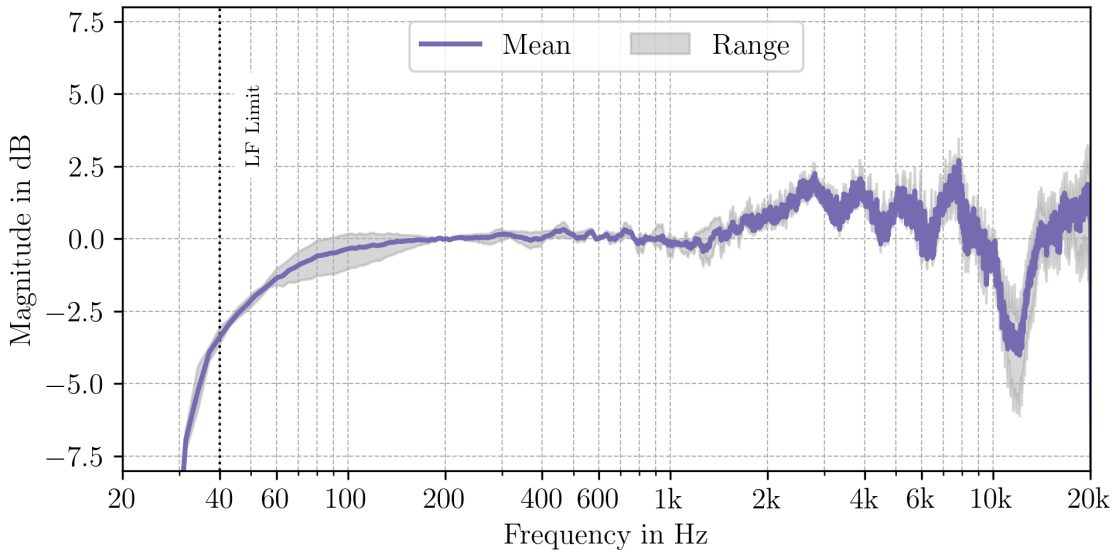


Figure 2.18: Magnitude FRF of three devices of the same model: iPhone 13 mini. Normalized to the value at 200 Hz.

2.2.9 Using Smartphone's Microphone for LRIR Measurements

Based on the presented data, it appears that smartphones can generally be used to measure the LRIR. SNR is not the limiting factor. Just as a side note, I have also found a measurement microphone manufactured just for this exact purpose, which uses the MEMS capsule as well [56].

Out of 25 measured models (30 devices in total), only four were completely unsuitable:

- Three for strongly fluctuating and therefore unsuitable frequency response:
  - Huawei P30,
  - Motorola Edge 50 Pro,
  - OnePlus 9,
- And one for clearly engaged DSP that prevents any kind of acoutic mesurements:
  - Samsung Galaxy A52 5G.

That makes 84 % of the tested dataset usable.

iPhone models benefit from the certainty of disabling the default DSP and the frequency response coherence across models up to 1.5 kHz. Apart from being able to use the models that were measured in the anechoic chamber and therefore one has their calibration data, it seems that it is also safe to use an average calibration data if the LRRE is limited up to 1.5 kHz.

The main disadvantage of devices running Android is that there is no guarantee that the default DSP can be disabled for a particular model. Because of that and the lack of coherence of the frequency responses, only models for which the calibration data exist can be safely used for measurements.

If a minimum-phase compensation filter is used, it is sufficient to compensate

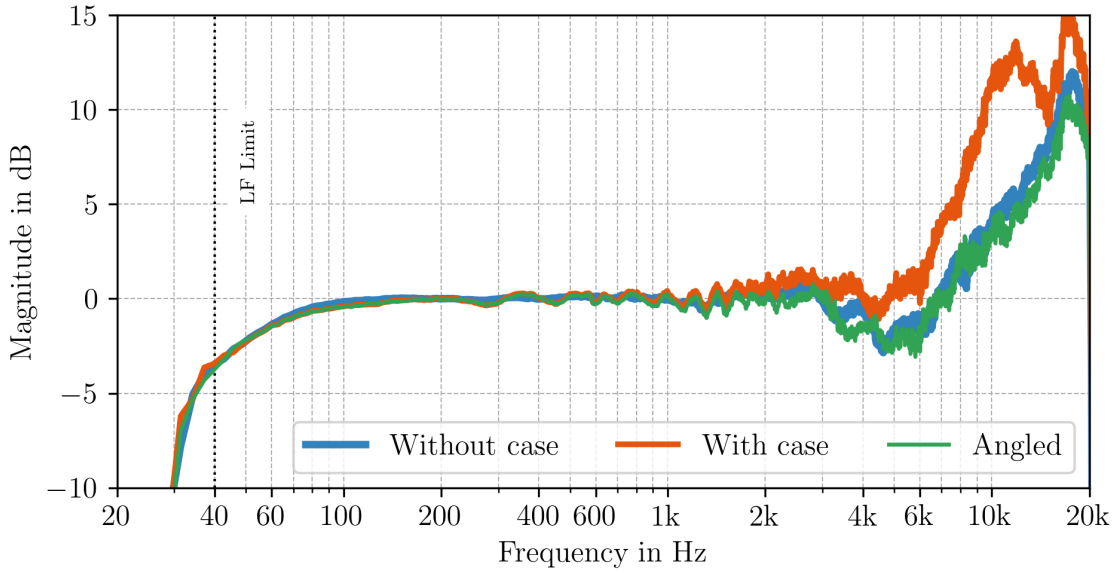


Figure 2.19: Magnitude FRF of iPhone 12 mini when measured without case on-axis, with case on-axis, and without case under 45 ° angle. Normalized to the value at 200 Hz.

only for the magnitude of the smartphone's frequency response. However, if a mixed-phase compensation is of interest, one should also compensate for the phase response due to the group delay rise at the lowest frequencies.

When it comes to the performance of the measurement, the user doesn't have to pay much attention to directing the sound inlet of the microphone at the speaker, as it has no influence. However, if a full range equalization is to be done, the user should be advised to do the measurement without the phone case, as it can influence the results in the high-frequency range.

3
Conclusion

The goal of this thesis was twofold: One task was to explore the topic of loudspeaker-room response equalization (LRRE) and find and implement a relevant method that could be later fitted into a pair of compact studio monitors. Three complete approaches were developed:

- Minimum-phase compensation using an FIR filter,
- Mixed-phase compensation using an FIR filter,
- And minimum-phase compensation using a cascade of IIR biquad peak filters.

The other task was to assess whether today's smartphones can be used as an acoustic measurement device, so one could measure the impulse response of a room, which is the first step in performing the LRRE.

3.1 Loudspeaker-Room Response Equalization

First, I would like to present a few remarks that emerged from the implementation work. When it comes to working with the early part of the LRIR, the theory says that one should work with the first 20 ms, considering correcting only for the direct sound and early reflections. It was shown that this is too short in comparison to the period of the lowest frequencies of interest. Experimentally, the shortest window possible to capture frequencies down to 40 Hz is 80 ms, which corresponds to $3.2T$ at 40 Hz. When it comes to smoothing the measured (complex frequency) responses, it is impossible to use either complex smoothing or separate magnitude and phase smoothing – both methods produce incorrect impulse responses. The only way to smooth the measured responses while accounting for the phase response is frequency-dependent truncation. If the phase response is not of interest, smoothing the magnitude response alone is sufficient.

13 LRIRs were measured on 40 cm grid to assess the sensitivity of the LRRC on the position change. For 1/3-octave smoothing, the magnitude responses are very coherent up to 500 Hz. This means that if the LRIR is measured only in one position, the computed compensation filter will be valid on a grid size up to $\lambda/2$, or in other words, within the radius of $\lambda/4$. For higher frequencies (smaller wavelengths), the best option is to use an average of the measured LRIRs with a frequency-dependent smoothing window width that extends with frequency and therefore neglects small fluctuations that differ across the grid and captures only a broad trend of the spectral balance. However, subjectively, it is sufficient to measure only at the mid position and use the scaling of the smoothing window.

To summarize the outcome of the three developed compensation approaches,

the minimum-phase compensation using a cascade of IIR biquad peak filters performs the best and at the same time is the most suitable option for the application purpose. The main limitation of the other two approaches is the inability to limit the amount or degree of compensation applied, which is mainly a problem when the measured LRRC has extensive dips that the compensation filter then inappropriately compensates. Another important limitation of the mixed-phase approach using the x-filtered NLMS algorithm is the fact that in order to have a reasonably working filter at low frequencies, a delay of tens of millisecods has to be introduced to the compensation filter, making the compensated signal way too delayed for implementation in studio monitors, where musicians or producers need almost instant response while monitoring or recording a performance.

To comment on the LRRE performance, objectively, it was shown to improve the spectral balance of the LRRC. Subjectively, its performance depends on the acoustic environment: If the listening space is too reverberant, the compensation has a negligible effect. However, for typical home-environment rooms with average acoustics, the compensation filter makes a noticeable difference.

In my subjective opinion, the spectral balance of test tracks while using the compensation filters is more even. The problem is that it can happen that certain 'bass boost' by modal peaks in the room might be subjectively preferable and thus counterproductive to compensate. That is why in possible future work, I propose making the target response customizable, allowing users to make certain modifications to their liking. One negative aspect of the compensated responses was found: The sound stage widens. The reason is most likely due to the fact that each speaker in the stereo configuration uses its own compensation filter. The problem with this approach is that under certain conditions, the compensation filters, since they are minimum-phase, can broaden the inter-channel phase differences, which are important to be as small as possible for correct source localization.

3.2 Smartphone Microphones

As far as using a smartphone for the LRIR measurement, it was shown that this is generally possible. The SNR was not the limiting factor. The span across the measured devices ranged from 42.6 dBA to 72.1 dBA while it was shown that the SNR of 39 dBA is sufficient. Four of the 25 measured models performed poorly, making them unusable for measurements. That means that 84 % of the tested smartphone models are suitable. The inferior performance of the four mentioned devices was caused either by strongly fluctuating frequency response (three devices) or by AGC during the measurement (one device). Both reasons are most likely caused by the default DSP, which, in these cases, could not be disabled.

If the LRRE is to be performed over the full frequency range, calibration data (the frequency response) for the given smartphone model must be available, regardless of the brand or operating system. However, when it comes to iPhones, thanks to the coherence of the frequency responses from the lowest frequencies up to 1.5 kHz, which is less than ± 1.5 dB, if the equalization were limited to this frequency range, an average calibration data could be used instead. Another advantage of Apple devices is their inherent ability to disable the default DSP of the smartphone. For

devices running Android, this is only possible if a manufacturer allows it, and this information is nowhere to be found prior to testing it.

For minimum-phase compensation, it is sufficient to have the magnitude-frequency response calibration data. However, thanks to a group delay rise at the lowest frequencies for all tested devices (20 ms at 40 Hz), the phase response has to be calibrated as well if a mixed-phase filter were used for the compensation. Lastly, the presence of a phone case can have an effect on the measurement at the highest frequencies, so if the full-range equalization is of interest, the user should be asked to remove the case prior to performing the measurement.

Bibliography

[1] S. Cecchi et al., "Room response equalization–a review," *Applied Sciences*, Dec. 2017. DOI: https://doi.org/10.3390/app8010016.

[2] J. Brooks-Park and S. van de Par, "Reverberant sound field equalisation for an enhanced stereo playback experience," in *Forum Acusticum 2023*, Turin, Italy, 2023. DOI: https://www.doi.org/10.61782/fa.2023.0407.

[3] W. Rudmose, "Equalization of sound systems," *Noise Control*, 1958. DOI: https://doi.org/10.1121/1.2369323.

[4] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd. Prentice Hall, 2009, ISBN: 9780131988422.

[5] J. O. Smith, *Introduction to Digital Filters with Audio Applications* . http://www.w3k.org/books/: W3K Publishing, 2007, ISBN: 978-0-9745607-1-7.

[6] M. Johansson, "On room correction and equalization of sound systems," Dirac Research AB, Tech. Rep., 2021, Accessed: 2025-01-29. [Online]. Available: https://www.dirac.com/wp-content/uploads/2021/09/On-equalization-filters.pdf.

[7] J. W. Marshall Leach, "The differential time-delay distortion and differential phase-shift distortion as measures of phase linearity," *Journal of the Audio Engineering Society*, vol. 37, no. 9, pp. 708–723, Sep. 1989.

[8] M. Kleiner and J. Tichy, *Acoustics of small rooms*. CRC Press, 2014.

[9] A. Gade, "Acoustics in halls for speech and music," in *Springer Handbook of Acoustics*, ser. Springer Handbooks, T. Rossing, Ed., New York, NY: Springer, 2007. DOI: https://doi.org/10.1007/978-0-387-30425-0_9.

[10] M. Long, *Architectural Acoustics*. Academic Press, 2014, ISBN: 978-0123982650.

[11] F. E. Toole, *Sound Reproduction: Loudspeakers and Rooms*, 2nd. Oxford, UK: Focal Press, 2008, ISBN: 978-0-240-52009-4.

[12] A. Walker, "Room acoustics for multichannel listening: Early reflection control," in *22nd AES Conference: Illusions in Sound*, 2007. [Online]. Available: https://secure.aes.org/forum/pubs/conferences/?elib=17282.

[13] P. Zahorik, "Spatial hearing in rooms and effects of reverberation," in *Binaural Hearing*, ser. Springer Handbook of Auditory Research, R. Y. Litovsky *et al.*, Eds., vol. 73, Springer, 2021. DOI: https://doi.org/10.1007/978-3-030-57100-9_9.

[14] *Methods for the subjective assessment of small impairments in audio systems*, BS Series – Broadcasting service (sound), Geneva: International Telecommunication Union, Feb. 2015. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1116-3-201502-I/en.

[15] L. Beranek, *Concert Halls and Opera Houses: Music, Acoustics, and Architecture*, 2nd. New York, NY: Springer-Verlag, 2004.

[16] M. Kleiner, *Acoustics and audio technology*, 3rd ed. Fort Lauderdale, FL, U.S.A.: J. Ross Publishing, 2012.

[17] R. Wilson, "Can we get the bass right?" In *21st AES Conference: Audio at Home*, 2006. [Online]. Available: https://secure.aes.org/forum/pubs/conferences/?elib=17270.

[18] F. E. Toole and S. E. Olive, "The modification of timbre by resonances: Perception and measurement," *Journal of the Audio Engineering Society*, vol. 36, no. 3, pp. 122–142, 1988.

[19] S. E. Olive *et al.*, "The detection thresholds of resonances at low frequencies," *Journal of the Audio Engineering Society*, vol. 45, no. 3, pp. 116–128, Mar. 1997.

[20] B. M. Fazenda *et al.*, "Difference limen for the Q-factor of room modes," in *Proceedings of the 115th Audio Engineering Society Convention*, Audio Engineering Society, New York, USA, 2003.

[21] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models* (Springer Series in Information Sciences), 3rd. Berlin, Germany: Springer, 2007, vol. 22, ISBN: 978-3-540-23159-2.

[22] J. Liski *et al.*, "Audibility of loudspeaker group-delay characteristics," in *Proc. 144th Audio Engineering Society International Convention*, Audio Engineering Society, Milan, Italy, May 2018, pp. 879–888. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=19404.

[23] M.-V. Laitinen *et al.*, "Sensitivity of human hearing to changes in phase spectrum," *Journal of the Audio Engineering Society*, vol. 61, no. 11, pp. 860–877, Nov. 2013. [Online]. Available: https://www.researchgate.net/publication/286939870_Sensitivity_of_Human_Hearing_to_Changes_in_Phase_Spectrum.

[24] L.-J. Brännmark and A. Ahlén, "Multichannel room correction with focus control," *Journal of the Audio Engineering Society*, vol. 63, pp. 21–30, 1/2 Jan. 2015.

[25] Dirac, "Dirac live: A technical overview," Dirac Research AB, White Paper: Technical Overview, 2025, Accessed: 2025-09-01. [Online]. Available: https://www.dirac.com/wp-content/uploads/2024/06/Dirac-Live-a-technical-overview-white-paper.pdf.

[26] B. C. J. Moore *et al.*, "The shape of the ear's temporal window," *The Journal of the Acoustical Society of America*, vol. 83, no. 3, pp. 1102–1116, 1988.

[27] G.-B. Stan et al, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, pp. 249–262, 2002.

[28] *Measurement microphones – part 8: Methods for determining the free-field sensitivity of working standard microphones by comparison,* IEC 61094-8:2012, Standard, Geneva, Switzerland, Sep. 2012.

[29] A. Farina, "Advancements in impulse response measurements by sine sweeps," *Journal of The Audio Engineering Society*, 2007. [Online]. Available: `https://api.semanticscholar.org/CorpusID:108967814`.

[30] P. Hatziantoniou and J. Mourjopoulos, "Generalized fractional-octave smoothing of audio and acoustic responses," *Journal of the Audio Engineering Society*, vol. 48, pp. 259–280, 2000.

[31] F. Denk *et al.*, "Removing reflections in semianechoic impulse responses by frequency-dependent truncation," *Journal of the Audio Engineering Society*, vol. 66, no. 3, pp. 146–153, 2018. DOI: `10.17743/jaes.2018.0002`.

[32] S. Cecchi et al., "Evaluation of a multipoint equalization system based on impulse response prototype extraction," *Journal of the Audio Engineering Society*, vol. 59, pp. 110–123, 2011.

[33] L.-J. Brännmark and M. Sternad, "Controlling the impulse responses and the spatial variability in digital loudspeaker-room correction," in *Proceedings of the 2015 International Symposium on ElectroAcoustic Technologies (ISEAT)*, Dirac Research AB and Uppsala University, Uppsala, Sweden, 2015.

[34] L.-J. Brännmark and A. Ahlen, "Robust loudspeaker equalization based on position-independent excess phase modeling," in *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA: IEEE, 2008, pp. 385–388. DOI: `10.1109/ICASSP.2008.4517627`.

[35] F. E. Toole, "The measurement and calibration of sound reproducing systems," *Journal of the Audio Engineering Society*, vol. 63, no. 7/8, pp. 512–544, Jul. 2015. DOI: `10.17743/jaes.2015.0064`.

[36] *Minimum phase*, `https://www.roomeqwizard.com/help/help_en-GB/html/minimumphase.html`, Room EQ Wizard Help, Accessed: 2025-10-29.

[37] G. Ramos and J. J. López, "Filter design method for loudspeaker equalization based on iir parametric filters," *The Journal of the Acoustical Society of America*, vol. 54, pp. 1162–1178, 2006.

[38] H. Behrends *et al.*, "Automatic equalization of flat tv speakers using parametric IIR filters," in *Proceedings of the 126th Audio Engineering Society Convention*, Convention Paper 7802, Audio Engineering Society, Munich, Germany, May 2009.

[39] S. J. Elliott and P. A. Nelson, "Multiple-point equalization in a room using adaptive digital filters," *Journal of the Audio Engineering Society*, vol. 37, pp. 899–907, 1989.

[40] Trinnov Audio. "It's about time #2: Speaker time alignment." [Online; accessed 31-October-2025], Trinnov Audio. (Mar. 2024), [Online]. Available: https://www.trinnov.com/en/blog/posts/its-about-time-2-speaker-time-alignment/?utm_source=chatgpt.com.

[41] R. Bristow-Johnson. "Audio EQ cookbook." W3C Working Group Note, 08 June 2021. (2021), [Online]. Available: https://www.w3.org/TR/audio-eq-cookbook/.

[42] J. Panzer and L. Ferekidis, "The use of continuous phase for interpolation, smoothing and forming mean values of complex frequency response curves," in *116th Convention of the Audio Engineering Society*, Convention Paper 6005, Audio Engineering Society, Berlin, Germany, May 2004.

[43] A. Bahne *et al.*, "Symmetric loudspeaker-room equalization utilizing a pairwise channel similarity criterion," *IEEE Transactions on Signal Processing*, vol. 61, no. 24, pp. 6276–6290, 2013.

[44] L. Beranek and T. Mellow, *Acoustics: Sound Fields, Transducers and Vibration*, 2nd ed. London, United Kingdom: Elsevier, 2019.

[45] M. Fueldner, "Chapter 48 - microphones," in *Handbook of Silicon Based MEMS Materials and Technologies (Third Edition)*, ser. Micro and Nano Technologies, M. Tilli *et al.*, Eds., Third Edition, Elsevier, 2020, pp. 937–948, ISBN: 978-0-12-817786-0. DOI: https://doi.org/10.1016/B978-0-12-817786-0.00048-7.

[46] Ariose Electronics. "LF-M6027-O electret condenser microphone." Accessed: May 18, 2025. (), [Online]. Available: https://www.ario.com.tw/en/product-380800/Electret-Condenser-Microphone-LF-M6027-O-series.html.

[47] J. Czarny, "Conception, fabrication and characterization of a MEMS microphone," Master's thesis, INSA de Lyon, Lyon, France, 2015. [Online]. Available: https://theses.hal.science/tel-01247487v1.

[48] A. Kumar et al., "Recent development and futuristic applications of MEMS based piezoelectric microphones," *Sensors and Actuators: A. Physical*, Sep. 2022. DOI: https://doi.org/10.1016/j.sna.2022.113887.

[49] STMicroelectronics. "IMP23ABSU MEMS microphone." Accessed: May 18, 2025. (), [Online]. Available: https://cz.mouser.com/new/stmicroelectronics/stm-imp23absu-mems-microphone/.

[50] M. Winter *et al.*, "Influence of a chip scale package on the frequency response of a MEMS microphone," *Microsystem Technologies*, May 2010. DOI: https://doi.org/10.1007/s00542-009-0994-z.

[51] P. Sjösten. "Electroacoustics." (), [Online]. Available: https://www.ta.chalmers.se/content/protected/courses/ele/Electroacoustics_Lecture_Notes_Per.pdf.

[52] L. Sant *et al.*, "MEMS microphones: Concept and design for mobile applications," in *Low-Power Analog Techniques, Sensors for Mobile Devices, and Energy Efficient Amplifiers: Advances in Analog Circuit Design 2018*, K. A. A. Makinwa *et al.*, Eds. Cham: Springer International Publishing, 2019, pp. 155–174. DOI: 10.1007/978-3-319-97870-3_8.

[53] Knowles Electronics, LLC, "Frequency response and latency of MEMS microphones: Theory and practice," Knowles Electronics, LLC, Application Note AN22, 2017, Accessed: 2025-06-15. [Online]. Available: https://www.knowles.com/docs/default-source/default-document-library/frequency-response-and-latency-of-mems-microphones---theory-and-practice.pdf?sfvrsn=4.

[54] B. M. Faber, "Acoustical measurements with smartphones: Possibilities and limitations," *Acoustics Today*, Summer 2017. [Online]. Available: https://acousticstoday.org/wp-content/uploads/2017/05/Faber.pdf.

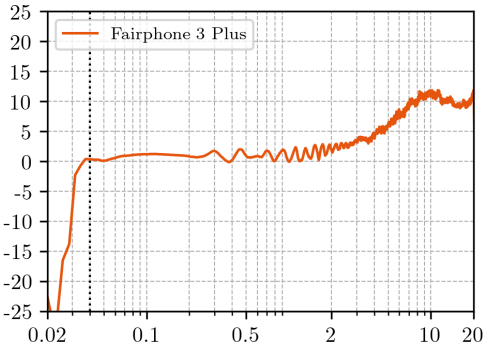
[55] Android Developers. "Mediarecorder.audiosource.unprocessed." Accessed: 2025-06-17. (), [Online]. Available: https://developer.android.com/reference/android/media/MediaRecorder.AudioSource#UNPROCESSED.

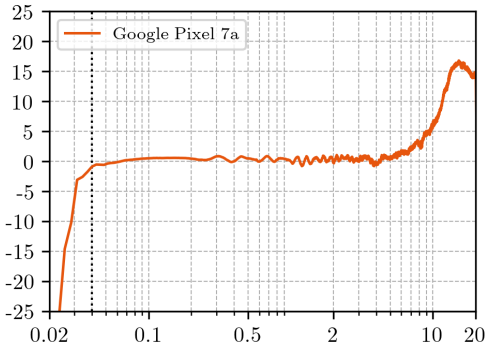
[56] Sonarworks Inc., *SoundID reference measurement microphone*, Online, Without Software — Sonarworks Store, 2025. [Online]. Available: https://store.sonarworks.com/products/soundid-reference-measurement-microphone.

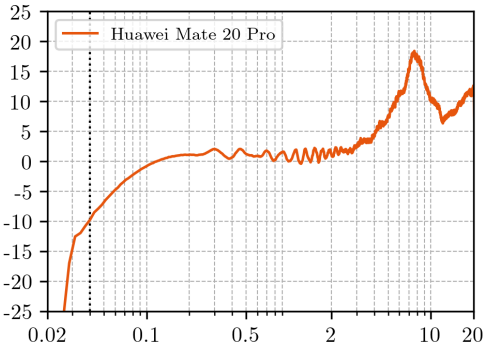
A

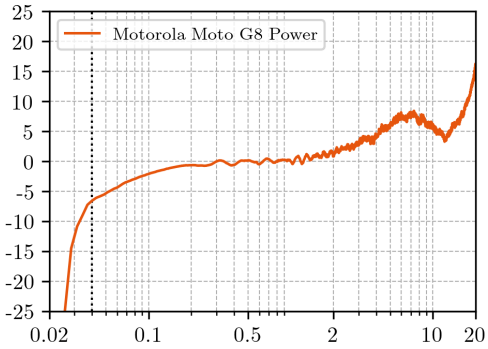
Magnitude Frequency Responses of Smartphones

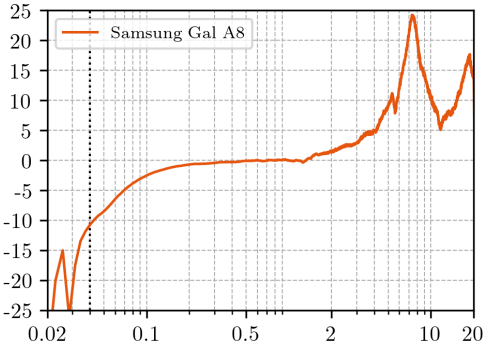
Below are shown the magnitude FRFs of 30 measured smartphones. Vertical axis is magnitude in dB normalized to 1 kHz, horizontal axis is frequency in kHz. Vertical line at 40 Hz is the low frequency limit of the measurement.

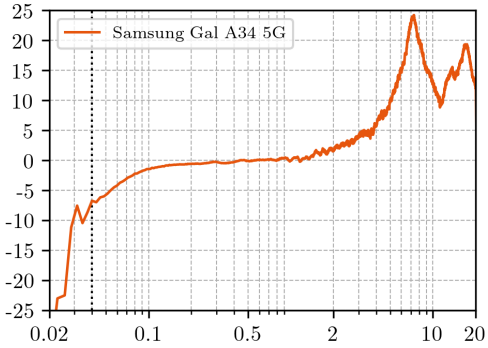


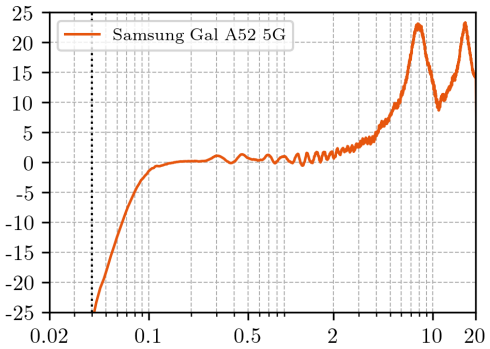


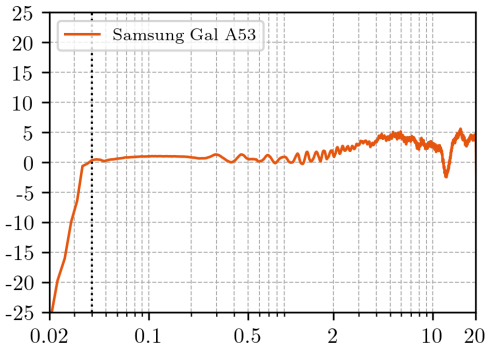


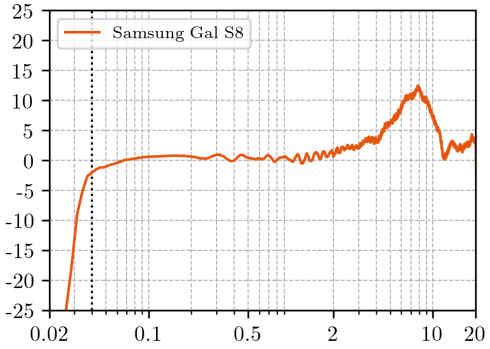


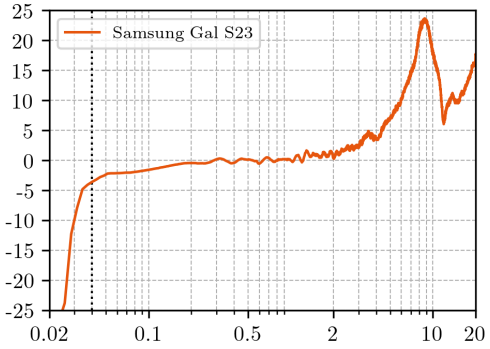


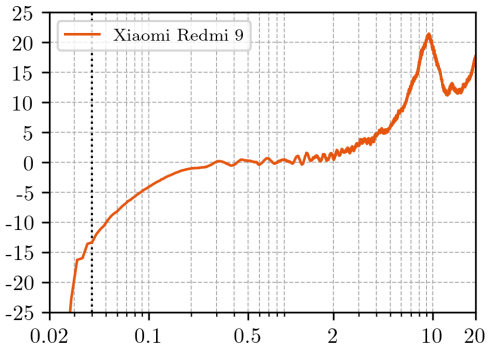


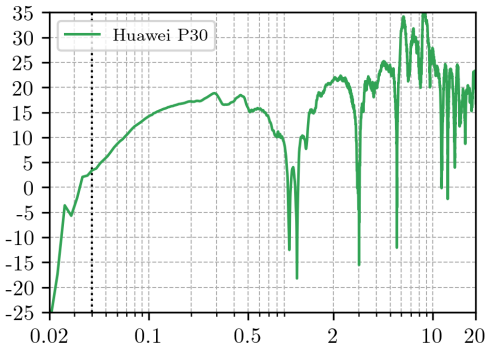

Samsung Gal A52 5G
25 20 15 10 5 0 -5 -10 -15 -20 -25
0.02 0.1 0.5 2 10 20

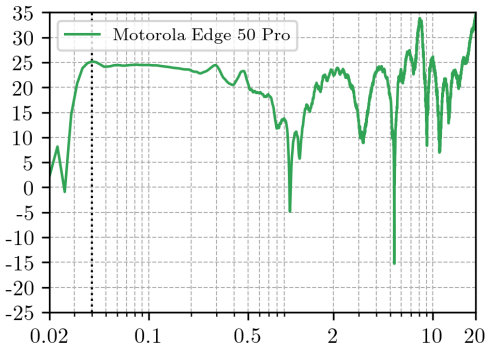

Samsung Gal A53
25 20 15 10 5 0 -5 -10 -15 -20 -25
0.02 0.1 0.5 2 10 20

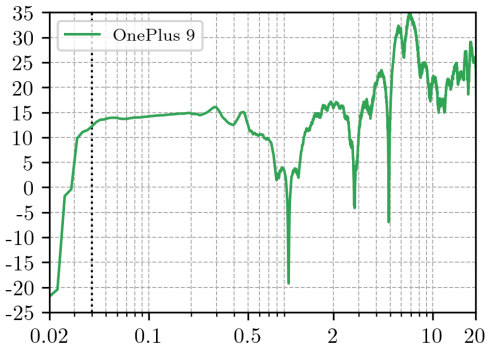

Samsung Gal S8
25 20 15 10 5 0 -5 -10 -15 -20 -25
0.02 0.1 0.5 2 10 20

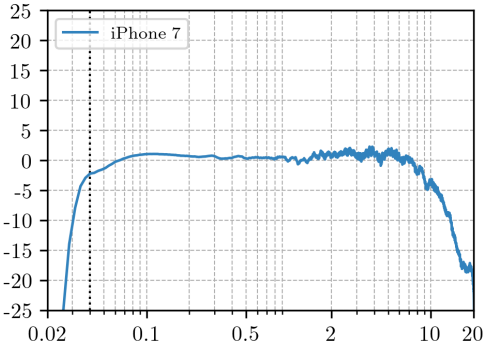

Samsung Gal S23
25 20 15 10 5 0 -5 -10 -15 -20 -25
0.02 0.1 0.5 2 10 20

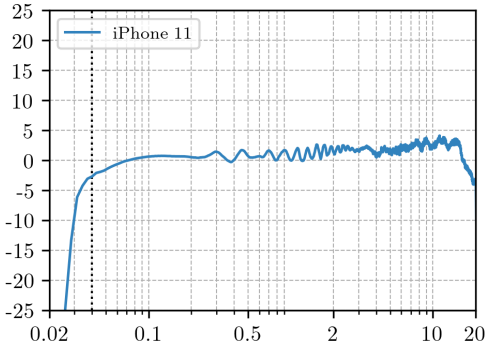

Xiaomi Redmi 9
25 20 15 10 5 0 -5 -10 -15 -20 -25
0.02 0.1 0.5 2 10 20

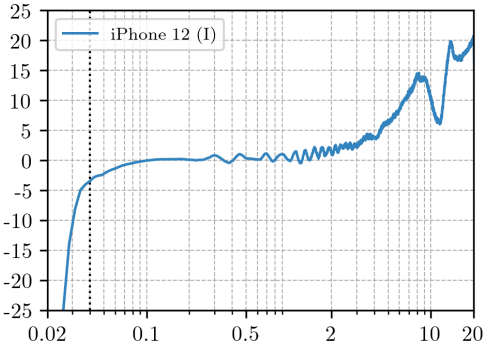

Huawei P30
35 30 25 20 15 10 5 0 -5 -10 -15 -20 -25
0.02 0.1 0.5 2 10 20

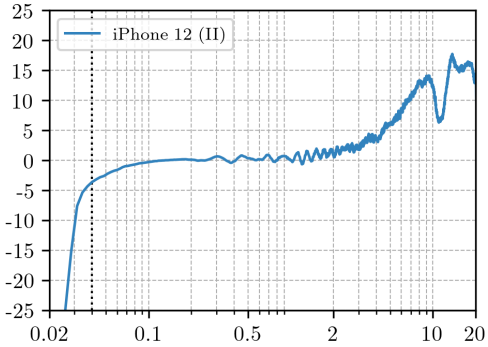

Motorola Edge 50 Pro
35 30 25 20 15 10 5 0 -5 -10 -15 -20 -25
0.02 0.1 0.5 2 10 20

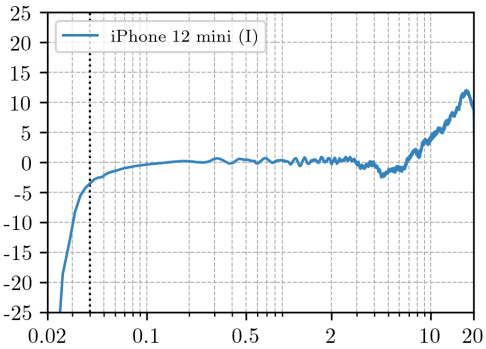

OnePlus 9
35 30 25 20 15 10 5 0 -5 -10 -15 -20 -25
0.02 0.1 0.5 2 10 20

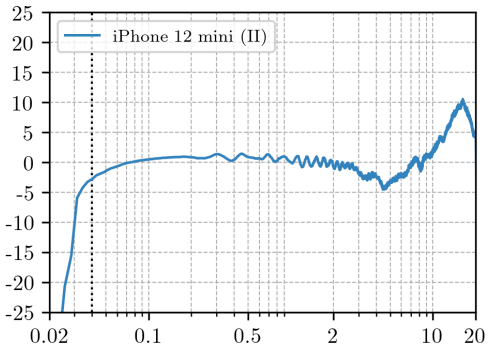

iPhone 7
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

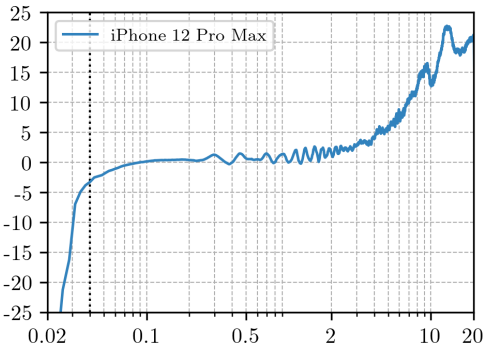

iPhone 11
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

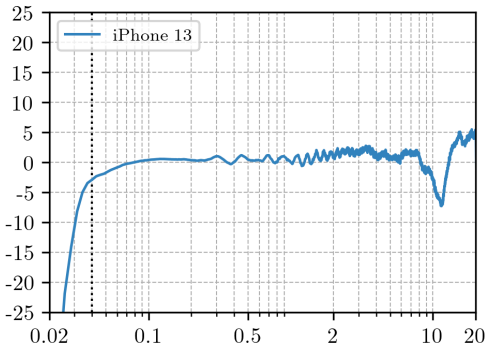

iPhone 12 (I)
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

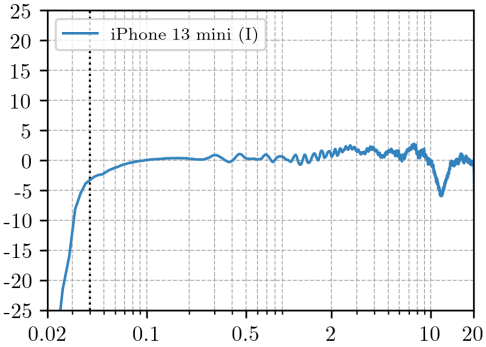

iPhone 12 (II)
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

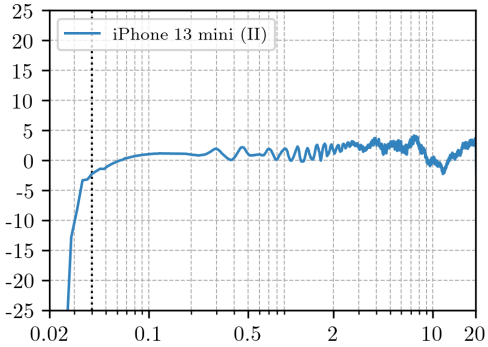

iPhone 12 mini (I)
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

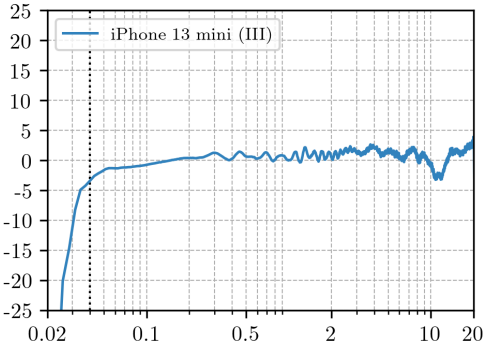

iPhone 12 mini (II)
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

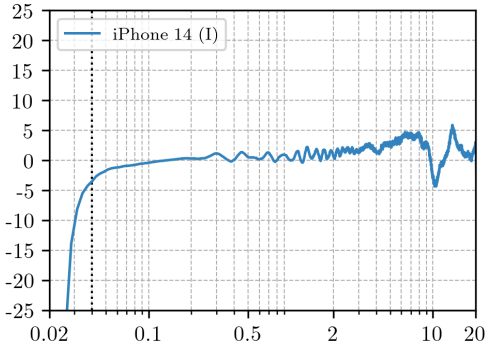

iPhone 12 Pro Max
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

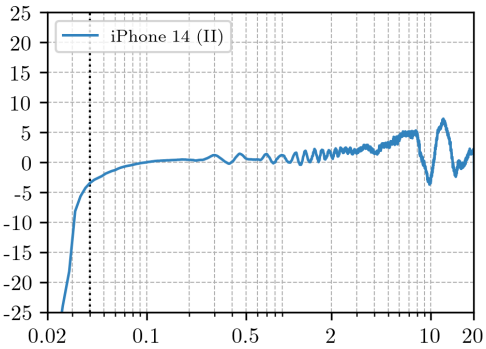

iPhone 13
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

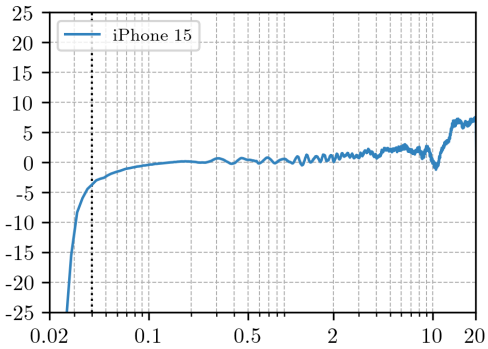

iPhone 13 mini (I)
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

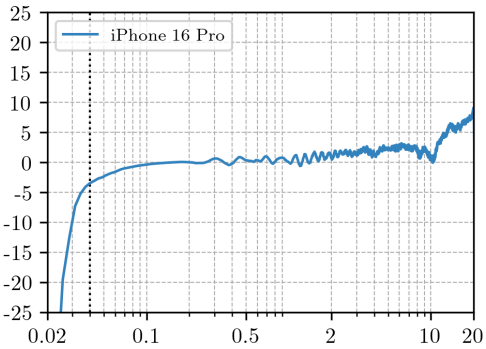

iPhone 13 mini (II)
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

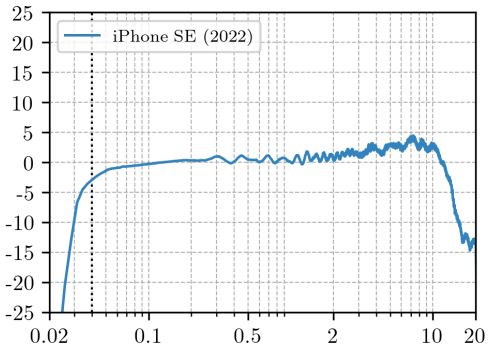

iPhone 13 mini (III)
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20


iPhone 14 (I)
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20


iPhone 14 (II)
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20


iPhone 15
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20


iPhone 16 Pro
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20


iPhone SE (2022)
25
20
15
10
5
0
-5
-10
-15
-20
-25
0.02
0.1
0.5
2
10
20

DEPARTMENT OF ARCHITECTURE AND CIVIL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se