***Toningenieur Project* seminar paper**

Objective audio quality assessment of MEMS microphones

David Neussl, BSc.

Supervision:
Univ.Prof. Dipl.-Ing. Dr.techn. Alois Sontacchi (KUG: IEM)
Daniel Neumaier, MSc. (Infineon Technologies GmbH)

in co-operation with

Abstract

In the development of audio components such as microphones and loudspeakers, the measures that are used to quantify these products are in most cases directly translated from the field of electrical engineering and general signal processing. This is of course a valid approach, especially since the ranges of these parameters have not been exhausted for most of the time. But with the growing expertise of OEMs producing better and better components, traditional measures lose their significance regarding whether or not those differences are still perceptible.

The target of this project is to analyse the audibility of a specific group of signals and the amount of disturbance they generate if audible. The test signals are mad up of a changing noise floor that is accompanied by a transient peak, which for the ease of reading will be called glitch in the following.

While there are approaches towards developing such methods, that take the human perception into account, those algorithms are not always that well understood or developed for a very specific purpose (for example the audio quality of speech codecs). To understand the audibility of transient glitches with a change in the noise level and the influence of such artifacts on the *disturbance*, these parameters evaluate glitches with regards to their influence on the overall influence on perceived quality rather than focus on the disturbance of a glitch. In other words, with those measures, a short glitch is interpreted as a short dip in quality, which is no problem if the rest of the signal quality is great. In order to acquire a deeper understanding of the relationship between such glitches and the human perception, a listening test was performed. Subsequently the data was then analysed with different known methods, where the parameter *Loudness* proved as the most versatile and reliable measure to base statements regarding *audibility* and *disturbance* on.

While the prediction of the audibility of an artefact is rather reliable across different groups and ages, from what could be gathered during the listening tests, predictions regarding disturbance are of less significance. This may be due to a real difference in what is perceived as *annoying* but may also result from the listening test methodology. Still it is possible to determine loudness ranges wherein rather reliable statements can be made. In the end, loudness seems to be a very capable measure to to indicate perceptible relevant audible disturbances and will be investigated further for the applicability to similar problems regarding the perception of sound.

Contents

List of Figures

1 Introduction

The development of microphones largely a discussion of topics from the field of electro-acoustical engineering. Certainly acoustic engineering is essential to finalize any such product, but a large part of the process is performed mainly with electrical engineering in mind, often separated from the field of acoustics and audio. This is even more the case with MEMS (micro-electro-mechanical systems) microphones, where an additional challenge is to make microphones as small as possible, while performing at similar levels as *traditional microphones*, regarding sound pressure and quality. This generally involves the discussion of parameters such as THD (Total Harmonic Distortion) and dynamic range which are well established measures for audio components and can be found in data sheets, ranging from microphones to loudspeakers or headphones and audio amplifiers.

At the current state, the improvement of one over the other microphone is often very differential, such that a difference is often only detectable in the measured quantity that does not correlate with an audible difference. Even more so, the difference between the microphones of two different vendors or even between generations of the same product have very often shifted from improving a certain measure to the introduction of additional features like shifting the operating point to improve performance in certain ranges. These new features often introduce problems that can not be measured and compared with traditional means, but require different measures and approaches.

Therefore, the objective of this project is to deepen the understanding in the field of perceived audio quality. This project in particular will focus on the perceptibility and disturbance of transient artifacts that are accompanied by a change in the ground noise level, called glitches in short. Further, the research will discuss how to derive a measurable quantity that sufficiently describes the influence on perceived audio quality of these artifacts. Ideally, this measure will then provide information about the audibility of *new* (as in different implementations) glitches as well as the disturbance, provided they are audible. Further studies may then evaluate the usability of this measure on other transient artifacts stemming from a different source than what is described in this work. It is also possible that the found method is applicable to completely different sources of signal degradations, such as distortion and may be researched in the context of rather different problems regarding audibility and quality degradations.

2 Audio quality

Before discussing audio quality, it is necessary to define the term audio signal in the context of this work. It is important to note, that audio signals constitute a sub group of the more general group of (analogue or digital) signals (note that this definition is sufficient for this work, but excludes the definition of signal for everything beyond the engineering framework). This group also contains very different signals like ECG (Electrocardiography) readings or vibrational measurements for example. For this large group of signals, there are certain methods that can be applied, independent of the nature of the signal. Time-frequency analysis for example is not only used in the context of audio, but also a tool for the "investigation of cardiac abnormalities" like heart blocks [18] or "to extract machinery health information contained in non-stationary signals" [5]. These are just two examples, where methods of general signal analysis are applied to a specific subgroup as well. This is visualised in figure 2.1 (right) by the interleaving parts of the circles that represent signal specific methods.

Beyond such *general* signal analysis methods, there often are methods that are more attuned to a specific field of signals and applications. For the field of audio, this may be special transformations like the *LUFS* (Loudness Units relative to Full Scale) which includes the human on the perceiving side and is used in audio production and broadcasting, as mentioned in this article by Hugh Robjohns [16]. Such very domain specific methods are what is implied by the areas of the circles, that do not interleave with others as they are intended for use only in a very specific context.

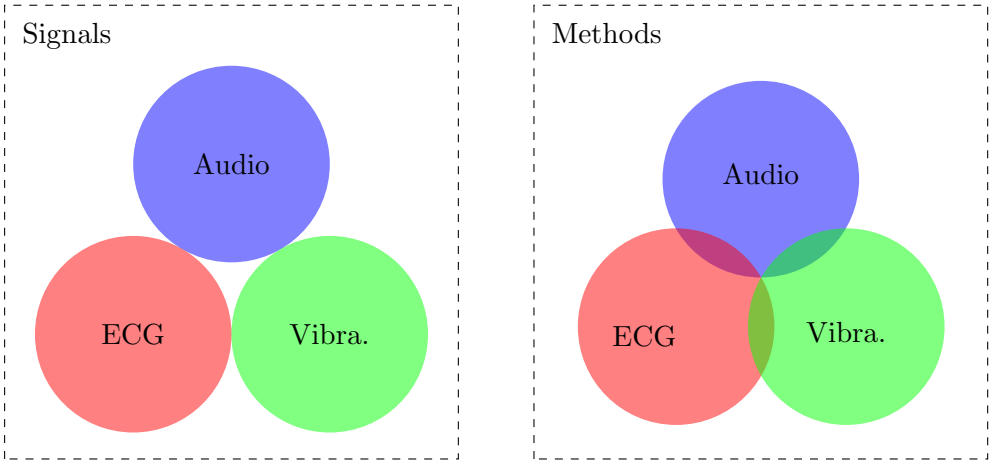


Figure 2.1: Signal spaces, where signals (left) can be clearly assigned to different sub categories such as audio, vibrational measurements or ECG (beyond many others) and processing methods (right), that can be exclusive to a sub domain (non interleaving segments) or used across different sub categories of signals (interleaving segments)

2.1 Objective and subjective

In general, the description of anything can be done in an objective (*this box is 1m wide*) or subjective (*this box is big*) way. While both statements describe the size of an object, only the objective description will produce repeatable results that are independent of the person that is involved in this assessment as it relies on standardised methods. To achieve a similar reproducibility with the subjective assessment, firstly it is necessary to utilise persons that are capable of reliably performing said estimation. Secondly it is also necessary to provide an *anchor* or *boundaries*, in order to contextualise the assessment. Note also, that the objective assessment only determines a specific parameter (in this example the *width* of the box) while the subjective description is of a more *holistic* natures (a very wide but but narrow box may still be small while its width is large).

The same principle is applicable to the space of signals, as there are many well established methods to objectively describe the quality or degradation of a signal. These are often parameters such as the *total harmonic distortion* (THD), *signal-to-noise ration* (SNR) or the *bit depth* (number of bits per sample). For many applications, such parameters are very useful ways to extract information about either the signal itself or the processing system (i.e. a sensor, an amplifier, etc.). Yet these measures only describe one dimension of the signal, while the investigation of *perceived* audio quality requires a broader and more subjective method. To perform this investigation, in a first step a listening test will be necessary, where an anchor or boundaries are provided as described for the box example. This method will also involve some anchors or boundaries (as described before) to provide context to the subjects of a listening test. With this data, a measure can be derived, that at the least reproduces these results and ideally will also prove as a predictive tool for new data.

2.2 Objective audio quality

To find a measure, that describes the influence of an interference, artefact or any other general alteration of a signal, the approach that is followed for this project is to combine standardised, measurable quantities like the signal level and combine them with information about how humans perceive and process auditive information. This allows to categorise influences in to categories with differences that are measurable and perceptible, measurable but not perceptible and immeasurable but perceptible, where each category has different implications on perception.

2.2.1 Measurable and perceptible

With artefacts, that are perceptible as well as measurable, it is generally very well possible to estimate how much of the artefact is perceived. The challenge in this case is to find the correct mapping and interpretation between measurand and perception.

Psychoacoustic models as described in [20] can be applied to estimate and derive a measure, that describes the human perception of a certain signal type. One example for this equal loudness contours [8] that describe the frequency sensitivity of the human ear and which absolute SPL levels are perceived as equal in level. For this approach it is necessary to fully understand how these models are developed, as the results can be very specific to only a certain type of stimulus (i.e. only small band white noise).

A different approach is to first perform a hearing trial and then derive a measure from the results of the hearing trial. One advantage of this method is, that the estimation can be fit to a specific behaviour/result directly, without applying psychoacoustic models, that are often not specifically made for the use case and therefore only sub optimal. If this method is chosen, it is necessary to consider the risks of hearing trials, that may yield biased or overfitted results if not performed correctly.

2.2.2 Measurable but imperceptible

If a quantity is measurable, but from psychoacoustic models or research results we know that humans can not perceive this, this quantity is very easy to handle, as it will not influence the final result.

An example for this are frequencies above 20kHz , which are inaudible to humans ([17]), even though they can be perceived by many animals (cats can perceive frequencies up to 80kHz for example [6]).

2.2.3 Immeasurable but perceptible

This group of artifacts is very special, as at first it may seem negligible or not worth the time to discuss. Contrary though, this aspect should be discussed, especially in the context of listening tests, as such influences are often not inherent to the audio signal but to the circumstances. *Quality* is an attribute merged from different sensory impressions, where the actual performance can be overshadowed by good or bad product design and have a bigger influence on buying decisions (compare to [14]). In the context of audio devices, which is already a very general term, *look and feel* (vision and haptics) are very important factors as well, beside the actual audio quality. As this is not the main topic of this research, artefacts of this group will not be further discussed, unless the results of the hearing trial explicitly show effects that can be assigned to this group.

As the artifacts that are to be examined are measurable and perceptible, the chosen approach is to perform a hearing trial of which the results can be used to examine existing measures or derive one that can accurately describe the results given known and potentially new artifacts.

3 Existing approaches

The challenge with objectively measuring audio quality is one that is present from the beginnings of telecommunications and therefore there is a large number of standards available today to deal with this problem (in this context). With the existence of such many standards, algorithms were implemented and tested with a small number of subjects in order to establish their usability for the given problem of predicting the quality of the given samples. In addition to these parametric approaches, the usability of the *A-weighted RMS* was evaluated.

One difficulty that arose for every measure tested in this stage was that there no ground truth available, in order to properly verify the results from any of the parameters. This is one more indication to perform a hearing trial, that would result in data than can be used as ground truth for further investigation.

3.1 PAQM (Perceptual Audio Quality Measure

This parameter is the oldest parameter that was examined over the course of this project and already uses a framework that is described in [20] and applied with different adaptations to other parameters.

As with all parameters that are discussed here, the PAQM is a *double ended* approach, that always compares a test signal against a test signal. In general, the reference signal would be a *perfect reference* containing no impairments at all to rate the overall quality of the test signal. This also means, that the computation can only give relative results, that do not yield information about absolute audibility thresholds. Additionaly, in order to investigate only a certain aspect, it is necessary to have a reference signal that contains only those disturbances under investigation, increasing the complexity of this approach.

Finally, this parameter can be very sensitive to noise, which is a problem in the investigated scenarios where the additive noise and the glitch can not be separated and only be evaluated together. As noise is an essential part of this analysis problem, this parameter is not sufficient for this purpose.

3.2 PEAQ, PESQ (Perceptual Evaluation of Audio/Speech Quality)

One of the more recent and prominent examples for audio quality measures is a standard by the ITU (*International Telecommunications Union*) called the *Perceptual Evaluation of Speech Quality* [12]. This standard is already an improvement over early approaches to assess the objective audio quality of speech signals in telecommunications but was adapted to use more broad band signals, as telecommunication works with sampling frequencies of 8kHz. As by now the PESQ itself is already more than 20 years old, with 2010 the *POLQA* (*Perceptual Objective Listening Quality Assessment*, [13]) started to supersede the *PESQ*. One big motivation for this renewal of the PESQ was the use of VoIP (Voice over Internet-Protocol) telephony again

extends the frequency range (beyond other improvements) and introduces new challenges to assess objective audio quality.

Similar to most approaches to estimate overall quality, the PEAQ/PESQ is also *double ended* and requires a reference according to the test subject, introducing the same complexities as mentioned with the PAQM. In addition to psychoacoustic models, this parameters also applies artificial neural networks to predict the perceived audio quality.

In 2020 a performance analysis of the PEAQ was done, that comes to the conclusion, that the PEAQ would probably benefit especially from an "improved cognitive model" [2, p.1] to make it more generally applicable to a greater range of disturbance artifacts. This is in accord with findings from testing this method before the hearing trial, as performance was varying heavily with the type of artifact that was applied to a test signal. The estimation results from this parameter were compared against the subjective ratings of a few testers, but was not satisfying enough to proceed with the PEAQ.

3.3 PEMO-Q (Perception-Model based Quality

Similar to the parameters discussed before, the PEMO-Q is again a double-ended method that evaluates the quality of a test signal given a high quality reference signal [7]. Opposing to the approach of comprising psychoacoustic models and neural networks, PEMO-Q tries to estimate the quality only applying known psychoacoustic models.

Similar to the previous methods, for test signals that are very noisy or dominated by noise, the PEMO-Q does not produce reliable results. This results was observed mostly when the noise in the test and reference signal were different realisations of the same noise (i.e. the noise parameters are the same but not each sample in the noise). For this reason, this parameter could not be used for this specific application.

3.4 A-weighted RMS (Rooot Mean Square

A very different approach to what was discussed in the preceding paragraphs is the application of the A-weighted RMS value. The first difference lies in its single ended approach as the result is the RMS value of a filtered signal, which can be computed with a moving window to get instantaneous values instead of one measure for the complete signal. This already reduces the complexity, as no special reference signal is required which by extend also implies that this measure is not sensitive to noise in the same way, the parameters discussed before, are.

While it makes sense to apply the A-filter [8], it is also not completely unexpected to see that the application of such a simplified *psychoacoustic model* is not enough for a satisfying prediction. Still it proves as a first proof of concept, whether or not psychoacoustic modeling may be a step in the right direction.

This parameter will be revisited during the evaluation of the listening test data (chapter 5 in the context of deriving a parameter with sufficient predictive capability.

4 Listening Test

As the previously discussed parameters did not yield satisfactory results without further information, it was decided to perform a hearing trial. This chapter describes the targets of the hearing trial as well as the modalities and execution of the trial.

4.1 Listening test objectives

4.1.1 Audibility threshold of glitches

One of the main goals of this listening test is to have explicit information about audibility thresholds of different switching signals. This will help understand at which levels these transient disturbers become audible and ideally aid in deriving a measure, that can predict the audibility of different glitches. As an additional gain, the different thresholds should serve as a show case, whether or not different switching settings can demonstrably reduce the audibility of these glitches.

4.1.2 Disturbance/annoyance of glitches

For ranges above the audibility threshold it is also interesting to examine how the experiment subjects will rate the annoyance of audible glitches. Furthermore, a *nice-to-have* information is whether and how the perceived disturbance due to the glitches increases with an increasing level. This could later on be used to allow for a trade off between artifacts that are audible but still tolerable/acceptable (i.e. not annoying).

4.1.3 Specification

Once the previously discussed targets are achieved, the last goal is to derive a generally applicable specification that can be used during the evaluation of existing and development of new switching behaviours. For this, the results from the two previous targets (audibility threshold and disturbance) will have to be converted to different measures which will be discussed in more detail in chapter 5.1, especially subsection 5.4.

4.2 Listening test mode

In order to yield the results as defined in section 4.1, further requirements to the listening test have to be considered:

- Close proximity to perception threshold (audibility threshold)
- Audibility of a minor impairment (audibility threshold of glitches)
- Information about disturbance in audible range

For this specific task no standardised trial strategy was available that would cover all of the above described requirements sufficiently. Therefore, it was decided to investigate the usability of two standardised trial modes that each fulfil parts of the requirements and then perform the listening test in a combined mode.

4.2.1 AB-X trial

To address the first two requirements of detecting the audibility of small impairments/impairment close to the perception threshold, the AB-X trial is often recommended as it is very sensitive in the "detection of small impairments" [11, p. 5]. In this form of trial, the subjects are provided with a reference file (X) and are asked to listen to the audio samples behind A and B and afterwards decide which of the two samples differs the most from X (triple-stimulus with hidden reference). Using this method therefore provides verifiable answers whether or not a difference was detectable or not. In addition this method can include a query of disturbance in the form of a five-grade impairment scale as provided in [11, p. 5] ranging from *Imperceptible* to *Annoying*.

This method is not an ideal solution though, mainly for the reason, that it would already require a sufficient estimate of the audibility threshold in order to properly set up the hearing experiment. Only with knowledge about the lowest detectable glitch level, all test runs could be set up in an efficient manner that prevents the playback at levels where the glitches are clearly inaudible. In addition to that, the method described in [11] is rather time consuming if repeated for samples with very little relative differences (for example level differences of 1 dB). This would needlessly exhaust the trial subjects and could potentially distort the results of the listening test.

4.2.2 Up-Down trial

In order to address the problems that arose during the discussion of the AB-X trial, a method of listening tests is discussed that focuses on estimating a threshold.

The large group of Up-Down design (UDD) trials provides a solution to this problem, as it was developed with just that target in mind. Even though it is often applied in clinical research regarding the efficiency of a drug, they can also be applied to the problem presented here, as they are a well established and understood method of finding or estimating thresholds in the context of audio [1].

With the goal of finding an audibility threshold, this method will start at a level that is well within the audible range. From there, the level will be decreased if a stimulus can be detected

and increased if said stimulus was not detected. In order to improve reliability and accuracy of this method, additional mechanisms can be applied:

1. 1 up, 2 down: In order to decrease the level, the detection of the glitch has to be confirmed at the same level two times in a row. This helps make sure, that the participants can still safely detect the glitch and the glitch is therefore still above the audibility threshold.

2. Decreasing step size: once the playback levels falls below the threshold for the first time (i.e. after not detecting the glitch for the first time), it can be reasonable to decrease the step size, as the real threshold will very likely be somewhere between the *no detection* level and a previous level. While it is possible to further decrease the step size, for the purposes of this test, the step sizes used are 2 dB in the beginning and 1 dB after the first non-detection.

In contrast to the AB-X trial method, this trial does not provide information about the disturbance of the audible glitches. Furthermore, in order to verify the correctness of the answers, some form of reference must be given which is already inherent to the AB-X method.

4.2.3 Combined mode

To overcome the shortcomings of both trial methods, one approach could be to simply first perform a 1 up 2 down trial to estimate the threshold and then perform a limited number of AB-X trials where the levels for each sample are set by the results of the corresponding Up-Down trial. This method would require the test subjects to perform two different and strenuous tasks, that could lead to premature exhaustion and thereby distort the results.

Therefore, it was decided to operate on a testing method, that incorporates parts of both methods and is similar to the *Two-alternatives forced choice* test, which was already described in [4, p.242 and following][1]. This seems to be a more reasonable choice both with regards to the task complexity and the quality and information content of the results:

1. Instead of three samples, the participants are only presented with two samples, of which one only contains constant noise that is equivalent to one of the two noise levels contained in the artifacted samples. The other contains glitches and noise level changes and is the sample that is to be detected. The participants are informed about this and are presented with an exemplary audio sample before the listening test begins. The decision to only using two samples is also justified by the fact that the reference would be the same for all samples, as it is always the same noise signal as well as the fact that by intention, the first few iterations, the audio samples will be played back at levels well above any audibility threshold.

2. After listening to the samples, the participants must decide whether the glitches and/or noise changes were heard in sample A or B. This is a forced choice and there is no possibility to give an answer similar to *no difference detected*, as this will be handled in the next step. The participants were informed about this circumstance beforehand.

[1] This test describes a method to detect the just noticeable difference between the shadows cast by two differently illuminated lamps

3. After listening to the samples and deciding on one of the two, the participants must rate the disturbance of the detected artifacts on a 5 step scale from *Perceptible, but not annoying* to *slightly annoying* and *annoying* with intermittent levels between those three. If no glitch could be detected, there is also a 6th option named *imperceptible*.

4. After these tasks are performed the playback level is adapted according to the answer behaviour and the query is repeated until a sufficient number of playback levels was performed.

The decision whether to de- or increase the level is made according to the rules of *Up-Down* trials, where a *non-detection* is constituted by either

- Choosing the reference sample with a disturbance rating other than *Imperceptible* (Wrong choice)
- Choosing any sample with a disturbance rating of *Imperceptible* (Inaudible)

Sample type \ Disturbance quality	Imperceptible	Perceptible
Reference sample	Non-detection	Non-detection
Test sample	Non-detection	Detection

Table 4.1: Visualisation of detection and non-detection

To terminate the trial for a given sample, the sign of the level change (in- or decrease) had to be reversed 4 times, after which the threshold is assumed to be approached sufficiently.

4.3 Listening test implementation

4.3.1 Sample selection

Since there are multiple switching behaviours, the first step of sample selection was to decide on settings that would be investigated. For this, the choice fell to on two different switching behaviours that would produce rather different results, named *Switch 1* and *Switch 2* in the following. For a proper representation of the range of possible audio artifacts an analysis of the peak heights was performed on 2 minute long recordings (see figure 4.1, left plots). These recordings contain segments with varying switching frequencies to take into account segments of many consecutive switches but also segments with no or just a few switches. The analysis parameter of this peak analysis is the peak height, of which the peak height distributions are shown in figure 4.1 (right plots). With this distribution, peak values were chosen to represent the higher end of typical values (*avg*) as well as extreme peaks (*high*). In addition to that, the peaks were chosen, such that they would be intercomparable to make the interpretation and comparison of the listening test results easier.

As a fifth sample, an idealised switch was added, that was synthesised by blending the noise levels in a way that does not produce any glitch at all. This should serve as an anchor in the listening test results as well as a benchmark sample for what could be achieved with perfect switching behaviour.

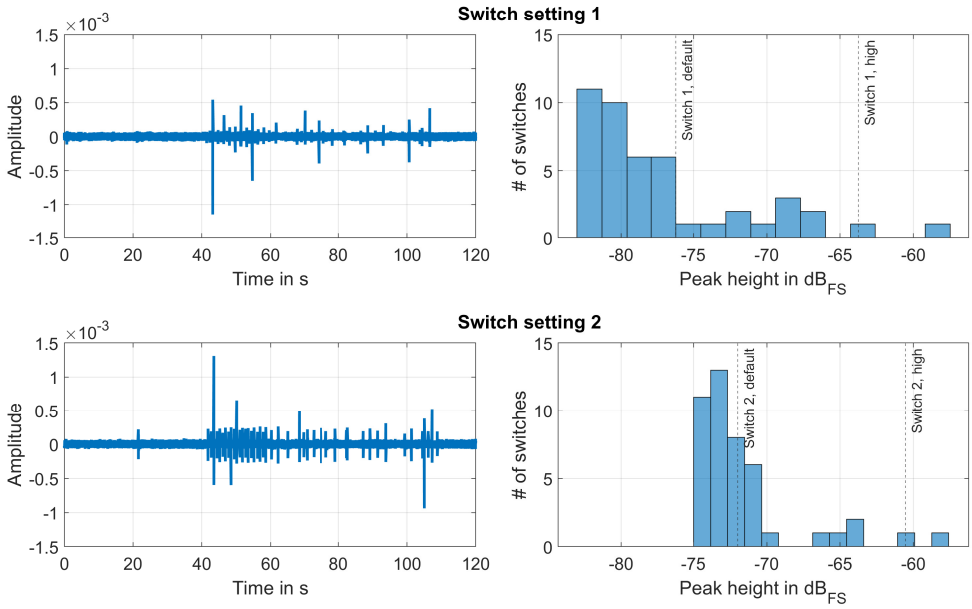


Figure 4.1: Peak distribution of *Switch 1* and *Switch 2* setting. The peak heights for the selected signals from *Switch Setting 1* are $Pk_{avg} = -76.2\ dB_{FS}$ and $Pk_{high} = -63.8\ dB_{FS}$, for *Switch Setting 2* they are $Pk_{avg} = -72\ dB_{FS}$ and $Pk_{high} = -60.5\ dB_{FS}$. These peaks are selected to represent comparable peak heights that are chosen at the higher end of typical values (*avg*) as well as extremes (*high*).

4.3.2 Listening test panel

As the listening test was performed in two different locations, the panels also varied with both trials. The two panels were made up of 10 and 11 people, which is just enough to gather reliable information, as the participants of both panels provide a certain expertise [11, p.4].

Expert listener panel

In the first trial, that was carried out on university premises, the trial subjects were drawn from an *expert listener panel* that is employed for listening tests in many student performed listening tests as well as commercial ones. This panel consists of people of whom it is known that they have no hearing impairments (i.e. normal hearing) and who participate in listening tests regularly. In addition, these participants all have a certain background in world of acoustics..

Infineon expert panel

The second trial was performed on a panel of Infineon employees whose work is closely related to the project on which the listening test was performed on. Therefore also this panel can be handled as *specialised expert assessors* [10] whose results can as well be treated as very reliable even from smaller panels.

Combining both trial runs, a total participant count of 22 specialised listeners is achieved which should be sufficient to lead to significant results.

4.3.3 Software

As discussed in section 4.2 *Listening test mode*, the chosen method for this trial is not fully standardised and therefore there is no software package that can be used straight away. For this reason, the decision was made to implement the required functionality using the *MATLAB* framework. As at the point of implementation it is possible that further listening tests will be performed, the listening tests software framework was implemented in such a way, that it could be easily adapted and expanded for different trials as well as different trial methods.

To fulfill this requirement, the framework is setup such that the listening test is always controlled by a *main* application that handles the listening test as configured. This configuration allows for different trial modes by implementing each as a separate application that is called by *main*.

Configuration

As one requirement to this framework is the modularity, the operation mode can be configured using a *MATLAB* file that contains information necessary for the calibration of the listening test (mainly conversion factors) but also the modalities of each run (trial mode, audio files, playback gains).

Main

The *main* window of the listening test simultaneously interprets the information contained in the configuration file and coordinates the trial accordingly. This includes presenting only the relevant windows as well as handling and storing the trial data once the trial is finished.

Thresholds

The *thresholds* window is a sub program of *main* and is used to determine the audibility thresholds of each participant. In the listening test for this project this was performed to track the (set) threshold of each participant. For this, the participants were asked to in- and decrease the levels of two *wobbletones* [2] to their respective hearing thresholds. This part of the trial is not verifiable, but as it was only little additional effort, it was performed for the following reasons.

- Investigate the correlation of hearing threshold and audibility thresholds (i.e. listening test results)
- Set the initial playback levels for the following trial runs to guarantee that the start for each trial is set in the audible range.
- Potentially detect and exclude subjects with inadequate hearing ability.

ABUpDown (2 alternatives forced-choice with 1-up 2-down, *2AFC-1U2D*)

This subprogram implements the required logic from *ABX* and *UpDown* trials as discussed and derived in subsection 4.2.3 *Combined mode*.

[2] Frequency modulates pure tones with center frequencies 1 kHz and 10 kHz, respectively

This user interface informs the participants about the progress of the whole trial (performed number of repetitions and number of executed trials, *Sample Trial*) and provides additional information to perform the trial (description of process and required steps, *Instructional Text*).

The lower three segments of the GUI are for actual interaction with the particiants.

1. The left controls audio playback, where randomly one of the two buttons will trigger the known noise reference and the other button triggers the test signal. This stage is only for listening and trying to detect the sample that contains the glitches and requires the user to make a choice (*"Sample A"* or *"Sample B"*).

2. In the center panel, the user is asked to rate the disturbance of the detected glitch. Here they can select *Imperceptible* if they could not detect the glitch/a difference or any of the other values if they think the correct sample was selected in the previous step. Depending on the choice, the level for the next round will be in- or decreased (see table 4.1). This answering behaviour will also affect the evaluation as the number of *false positives* (i.e. wrong sample and disturbance value other than *Imperceptible*) will be investigated.

3. If the previous steps are performed in this order, the *"Ok"* button in the right panel will be enabled, allowing the participants to proceed to the next step which is either another repetition with the same sample set or the beginning of the next set. In addition, the users can provide additional feedback or commentary in case that anything out of the expected happened during the trial or if any unexpected artifacts occurred. This step is optional though and does not affect the state of the *"Ok"* button.

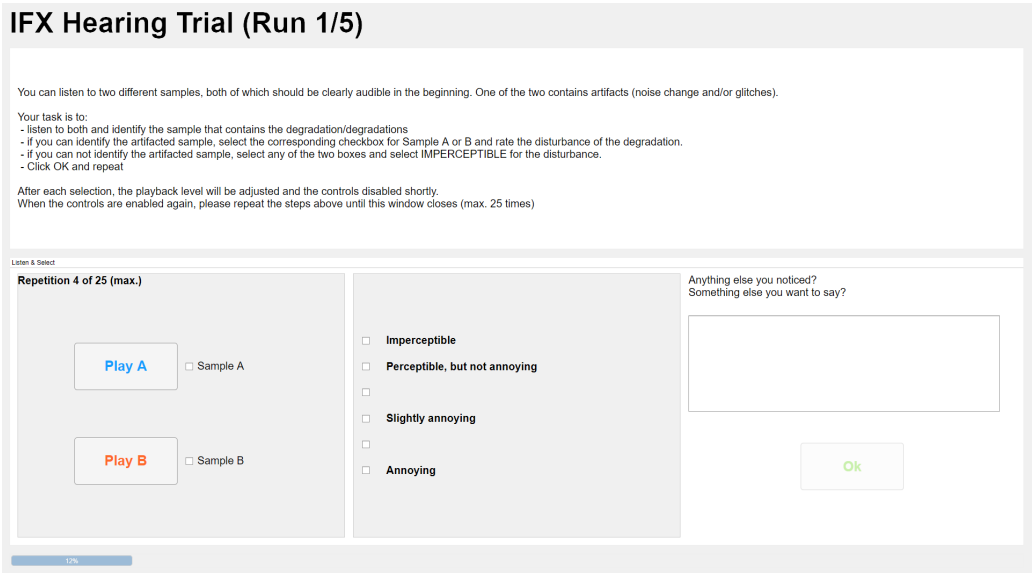


Figure 4.2: GUI for *2AFC-1U2D* test that shows the current run (top, *1/5*), Instructions for the listening test and the interactional area (three segments at the bottom). In addition, the information about the progress in the current run is displayed at the bottom, to provide visual feedback after each iteration.

4.3.4 Setup

Hardware

With one of the objectives being the determination of the audibility threshold of small impairments, it was very quickly decided to perform the listening test in an acoustically treated room via headphone playback in order to prevent any external disturbances (as good as possible). For the audio hardware, the following products were chosen/used:

- PC with Windows 10
- Audio interface: Focusrite Scarlett 2i2 (3rd generation)
- Headphones: Beyerdynamic DT-770 Pro (250Ω)

In order to compensate the influence of the frequency response of the headphones, the software audio equalizer *SoundID Reference* by *Sonarworks*[3] was used. This equalizer performs a predistortion of the audio signal in order to provide an overall system frequency response as *flat* as possible. For this model of headphones, *SoundID Reference* provides a pre made calibration profile (see Figure 4.3). To perform this trial it is also necessary to run *MATLAB* in the version R2020a or higher, as some functionality is only available from this release on.

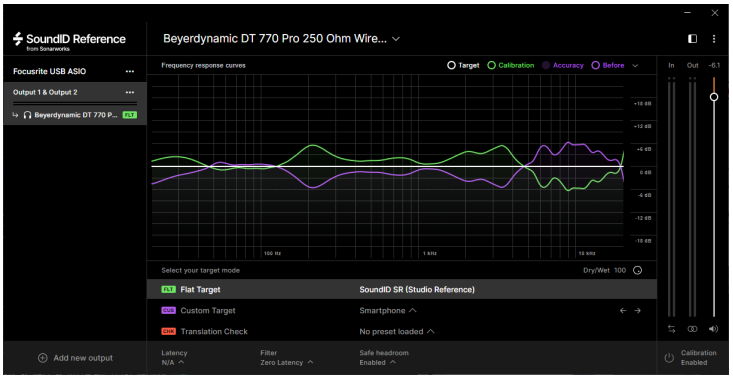


Figure 4.3: *SoundID Reference* setup for listening test. The *Calibration* (green) shows the frequency response of the headphones without correction, *Before* (purple) presents the equalizer that is applied and *Target* (white) is the target frequency response, which is flat but could be adapted to simulate different environments/situations (playback in car, via smartphone, different headphone model, ...)

Calibration of listening test setup

As the metrics we want to derive from this listening test will finally be in absolute values (dB_{SPL} or similar) it is necessary to calibrate the listening test setup before performing the trials. For this reason, a calibrated measurement setup is required, which in this case is the Br+el&Kjaer artificial head provided by the *IEM*[4]. With this setup, it is possible to measure the absolute level of a sound played back via headphones, which is necessary for this trial as it will also be performed using headphones.

[3] https://www.sonarworks.com/soundid-reference
[4] https://iem.kug.ac.at/institut-fuer-elektronische-musik-und-akustik-iem.html

For the calibration, a *MATLAB*-script was used, that performs the following iterative calibration method (see also figure 4.4):

- To begin the calibration, a target level (L_T) must be specified (i.e. 94 dB_{SPL}) as well as some arbitrary initial playback amplitude (A) for the measurement signal (X, in this case a 1kHz sine).
- The signal is played back and measured with a calibrated measurement setup to determine the level of the sine L_M.
- Using the target level (L_T in dB_{SPL}), a conversion factor (C) is computed, by which the signal amplitude is scaled: $A = A \cdot 10^{C/20}$.
- With this re-calibrated amplitude, the signal is again scaled to the new amplitude A and the measurement is repeated
- Until a deviation from the target level that is less than 0.05 dB is achieved, this process is repeated.
- Once a sufficiently exact calibration is achieved, the signal amplitude is A is equivalent to the sound pressure amplitude of the signal. By scaling this value by the target level, the digital equivalent D_0 to 0 dB_{SPL} can be derived: $D_0 = A \cdot 10^{-L_T/20}$.

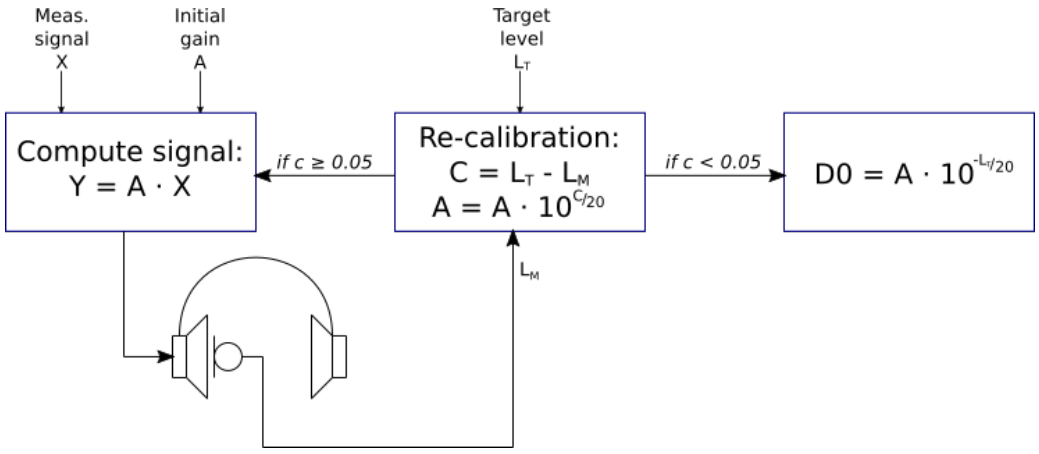


Figure 4.4: Flowchart, that shows the general steps that are performed to calibrate the hearing trial setup on absolute sound pressure levels

To control for headphone placement, this measurement was repeated 10 times for each headphone, where for each measurement, the headphone was removed and then re seated to the artificial head. The result of this calibration is shown in figure 4.5, where each data point is one repetition of the measurement. As the used audio interface does not allow for the output gain to be set in software, the gain potentiometer was permanently fixed to an amplification of approximately 60%.

Trial 1

The first trial with the *expert listener panel* (section 4.3.2) was held in Graz in the *Akustiklabor*, which is a room, that is specially treated and often used for measurements and listening experiments similar to this one. A distinctive feature for this location is the *room-in-room* construction

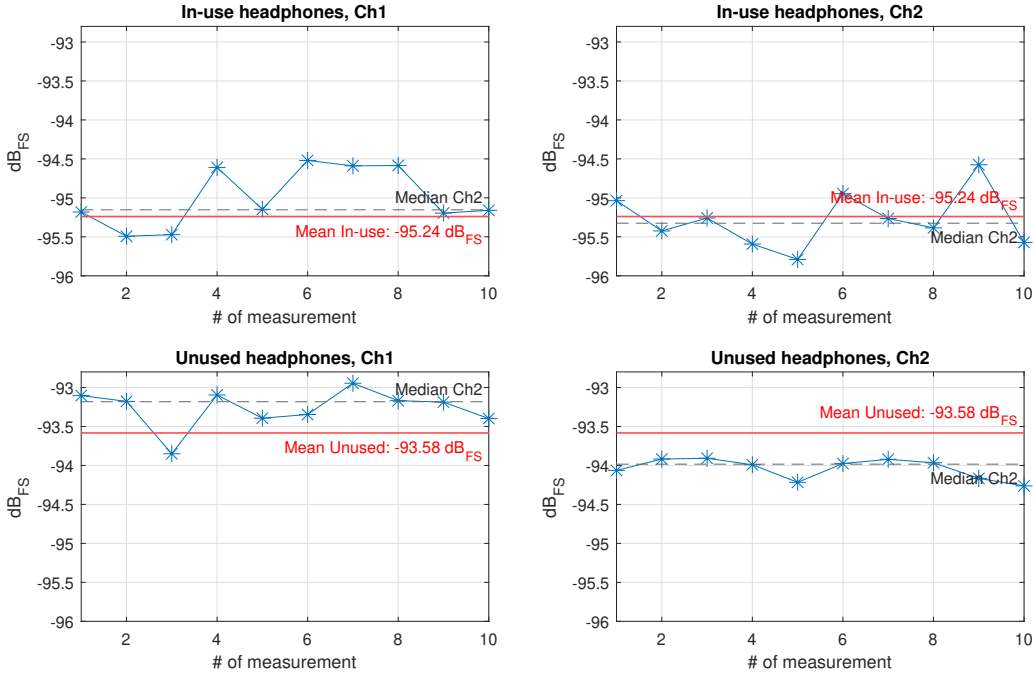


Figure 4.5: Calibration results for two headphones, where each data point shows the measured sound pressure level in dB_{FS} that is equivalent of 0 dB_{SPL}. For the listening test, the median result of all measurements for the *In-use* headphones were used, the second pair of headphones (*Unused*) was measured as a backup for the listening test.

Figure 4.6: Calibration setup using artificial head

to isolate the inside from external disturbances, beside the lining of walls, doors, windows and other reflective surfaces with acoustic foams (as can be seen in figure 4.7a).

Trial 2

For the second trial on Infineon premises a similar space was used, be it of much smaller volume but with more focus on canceling reflections from inside. The *anechoic chamber* in which the second trial was performed is much like the *Akustiklabor*, built applying the *room-in-room* technique to provide acoustic decoupling from the environment, and lined with acoustic foams to reduce reflections inside. Even though this chamber is much smaller in size, as it is normally used for measurements, it is still big enough for this trial, as the setup is rather minimal (see figure 4.7b).

(a) Setup in *Akustiklabor* (*IEM/KUG)*, Graz

(b) Setup in *anechoic chamber* (*IFX)*, Villach

Figure 4.7: Listening test setups in Graz and Villach

5 Evaluation of results

5.1 Data review

To begin the process of evaluating the hearing trial results, the first step was to review the raw data, in order to gain a better understanding of the collected information. In the simplest form, this data is available as information about the playback RMS and peak levels in dB_{SPL}, which are plotted against the number of iterations in figure 5.1. Note, that the data will often be discussed separately for peak and RMS values as those values are drastically different due to the nature of the signals. Short transient glitches do not have a big impact on the overall RMS value, which is dominated by the noise, that is same (level, distribution) for every sample. In contrast, the peak level is defined by the maximum height of the highest glitch and (more or less) independent of the noise floor.

In a first step, the data is colour separated (Blue: *Switch 1, average peak*, Orange: *Switch 1, high peak*, Yellow: *Switch 2, average peak*, Purple: *Switch 2, high peak*, Green: *Glitchless switch*) as well as participant separated using marker styles (one marker per person) to make the following plots easier to understand, highlighting different *clusters* (especially in the RMS plot). In addition, the marker coding allows to find corresponding trajectories for the different trial runs of specific participants, while the identity of each person is still protected. With this it is possible to scan for potential *outliers* or participants that need further investigation. This may for example be necessary, if the results for every trial are too high or low with respect to the residual data. Looking at the right plot in 5.1 for example, we see that participant $\triangle$ had a problem with *Switch 2, high peak* (purple), but performed average in every other test, which is why the data was kept. Figure 5.1 is already adjusted for two participants whose results are too divergent.

To further improve this data display, the individual trajectories were aligned with respect to the playback levels. As mentioned in section 4.3.3 *Thresholds*, the initial playback levels may be different for each participant, due to the levels set in the beginning of the trial. For this reason, as a result of this alignment, the iteration number on the x axes in figure 5.1 do begin at negative values. Each trajectory was shifted to the left (i.e. negative iterations), in order to compensate for a higher initial playback level. This step drastically improves the visibility of clusters for each sample and suggests the computation of individual thresholds for each trajectory (i.e. one threshold per participant and sample), to further analyse and visualize the data.

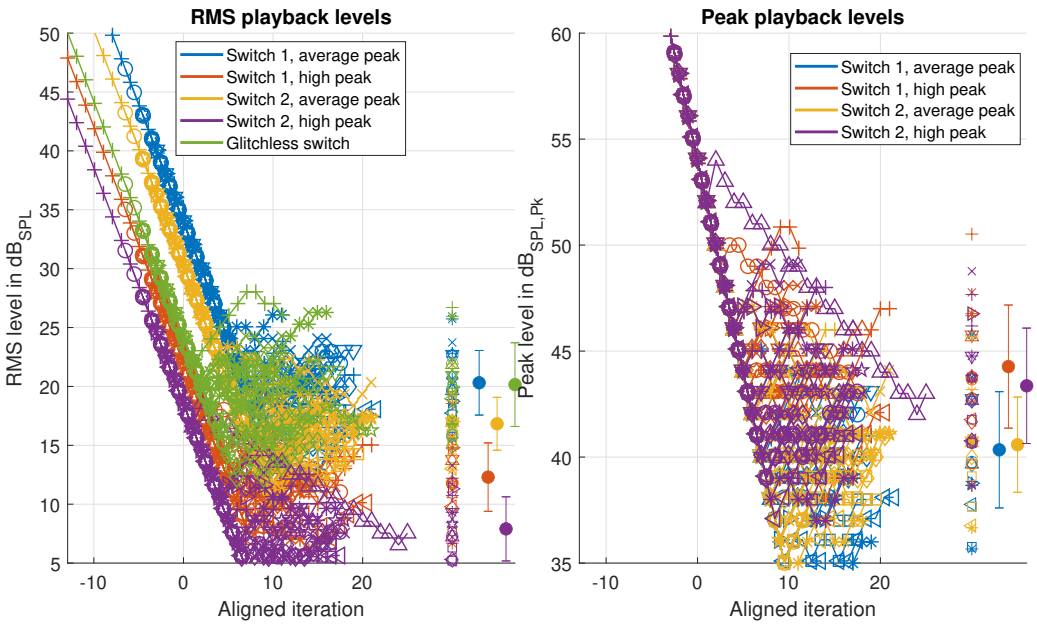


Figure 5.1: Colour (sample) and marker (participant) coded playback level trajectories. Left: RMS, Right: Peak. The trajectories are aligned with respect to the initial playback levels to improve the visibility of clusters. In addition the individual (per sample and participant) and grouped (per sample) thresholds are marked at the unlabelled end of the axes.

In addition to investigating the data regarding playback levels and thresholds, it is reasonable to also look at gathered meta data. In this case, this was done to evaluate whether or not the initial gain may be correlated to the final thresholds, as such correlation would have great impact on statements that can be derived in the next steps of the evaluation.

Figure 5.2 shows the number of iterations, the final thresholds and the number of participants per initial gain level. Assuming, that a higher initial playback level would be correlated to higher final thresholds, this plot would have to show the same or very similar numbers of iterations for each listener (within the limit of 25 iterations at most). Contrarily though, this plot shows a general trend towards a higher number of iterations with an increasing initial gain. This suggests, that participants with high initial gains require more iterations to approach a low threshold than those that started at lower playback levels to begin with. This is also in consistency with what can be seen in the center plot that shows no clear relation between initial gain and the final threshold.

The lower plot shows how many participants started at which initial gains, with an average increase of approximately 10 dB. This increase can be explained by the average participant age of 32 years, where already slight effects of presbycusis (especially at high frequencies) is anticipated (as shown in [15, p.60], dissertation focusing on the effects of presbycusis).

For these reasons, it was necessary to show that even if participants did not properly perform the assignment of the calibration, it does not influence the final results, while still making certain, that the trials will begin above the audibility threshold for every participant.

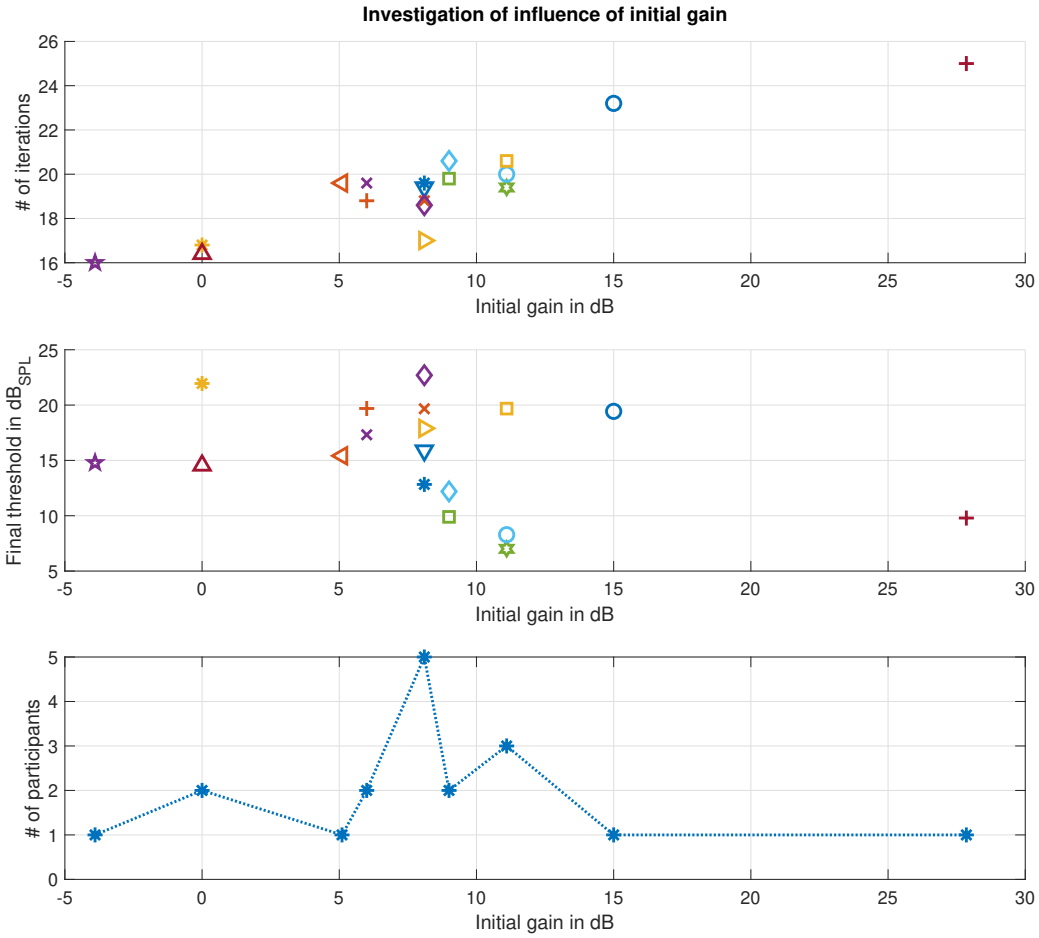


Figure 5.2: To investigate the correlation between initial gain and final thresholds and trial length, the corresponding values were plotted against each other.
In the top figure we see the average number of iterations required by each participants, that shows a correlation between higher number of iterations with a higher initial gain. This is not only expected but also a good indication that the final threshold is less correlated with the initial gain. Naturally, if a participant starts at a higher gain, they will require more iterations to approach the same low threshold as a different participant starting already closer to said threshold.
This is also supported by the center plot that does not show the same correlation between final thresholds and the initial gains. With this information, a potential connection between initial gain an final threshold can be dismissed.
The lowest plot shows the number of participants per initial gain level.

5.2 Audibility thresholds

To derive meaningful audibility thresholds from the data, evaluation methods from *Up-Down* tests are utilised. Since the test for each sample will run until 4 inversions occurred (see subsection 4.2.3), it can be assumed that the final levels are already sufficiently close to a threshold. For this reason, the individual thresholds were computed from the average of the last three values in each trajectory. This method mitigates a bias that may be introduced due to the higher starting playback level for some participants as discussed above in figure 5.2. From these individual thresholds a second set was computed where the average threshold for each sample (i.e. grouped by colour) was computed.

The results of these computations are also included in figure 5.1, positioned on the upper (unlabelled) ends of the x axes respectively. For a more conclusive discussion of these thresholds, especially with respect to the differences in RMS and peak level, the thresholds are displayed in figure 5.3, where the peak and RMS threshold values are plotted against each other. Doing so grants further insight into the connection between those values and how audibility changes, with respect to RMS as well as peak level.

One outlying result is the data for *Glitchless switch* (green), for which the data is located at the lower end of the ordinate around 20 $dB_{SPL,RMS}$ (abscissa) of this plot. This is because the peak is displayed by the ordinate as the glitchless sample per definition does not have a peak. Even though a maximum value could be derived for this signal, this value does not carry any information and only leads to confusion in understanding the results. For this reason, a secondary y axis was introduced on the right side, that displays zero for the peak of the glitchless sample.

This already gives insight to a first learning regarding the audibility due to the noise change. Since *Glitchless switch* is the idealised case where no glitch is present in the signal, the audibility threshold is only due to the change in the noise level. This also means that this is already a *hard limit* regarding the audibility, that is completely independent of any glitch.

With this limit in mind, the result for *Switch 1, average peak* (blue) stands out, as the threshold on the RMS axis is approximately the same (with respect to variance), suggesting that for low levels, *Switch 1, average peak* is already a rather optimised switching setting. At higher levels, this glitch may become more audible, which should be visible once the disturbance values are analysed (higher disturbance at lower RMS playback levels).

Furthermore, one can inspect the differences between the *Switch 1* and *Switch 2* by grouping the corresponding thresholds (blue and orange, yellow and purple) and comparing them side by side. This shows, that in both cases on average the *Switch 1* produces glitches, that allow for higher peaks before they become audible. For those specific settings one could also derive an area from the specific thresholds (below threshold peak levels and below 20 $dB_{SPL,RMS}$), that can help make a prediction whether or not glitches will be audible or not, by positioning them according to the two values on the axes.

While this analysis shows, that using different switching settings can have a positive influence on the audibility of said glitches, it is also noteworthy, that the simple analysis of the peak and/or RMS value of a signal is not predictive. In order to achieve this goal, a more complex and sophisticated approach will be necessary, that also includes knowledge about the human perception.

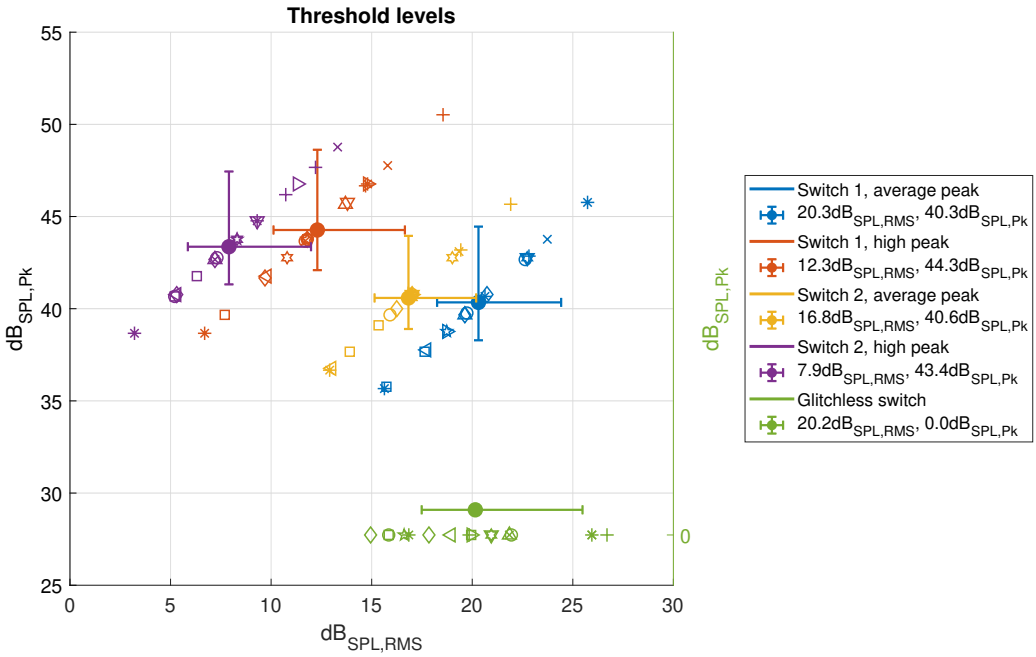


Figure 5.3: Display of threshold values regarding RMS (abscissa) and peak (ordinate) values respectively, where *Glitchless* has its own peak-axis as the peak is not informative in this sample. The whiskers show the confidence interval around the mean threshold, assuming a normal distribution.
Overall, there is an improvement visible for *Switch 1* (blue and orange) over *Switch 2* (yellow and purple).

To derive a measure that holds said predictive capabilities, the tested signals were investigated at their respective threshold levels as determined by the listening test, which can be seen in figure 5.4. Note that these signals were already chosen in such a way that they are representative for typical and high peak values for both settings, as explained in subsection 4.3.1 (Sample selection, p. 14). The first column shows the time signals for both switches with *average peak*, while the second columns shows the *high peak* signals. In order to make the comparison of the signals easier, the time and magnitude axes are aligned.

This presentation highlights the difference in structure between the *high* and *average* signals and their respective peak and RMS levels. While the RMS in both *average peak* signals (left) is about 10 *dB* higher than their *high peak* counterparts, this inequality is inverted for their peak height. The following tendency can be concluded:

- A high peak requires the noise floor to be lower (i.e. a lower RMS value)
- A low peak allows for the noise floor to be higher (i.e. a higher RMS value)

This realisation further confirms the assumption, that a simple peak analysis on the time signal will not be sufficient, but a more involved method is required, that may consider the total energy and even some psychoacoustic influences.

5.2.1 A-weighted RMS

From the results discussed above, the presumption, that to detect audibility, some kind of energy based measure is required. One such energy based measure was already discussed on the first

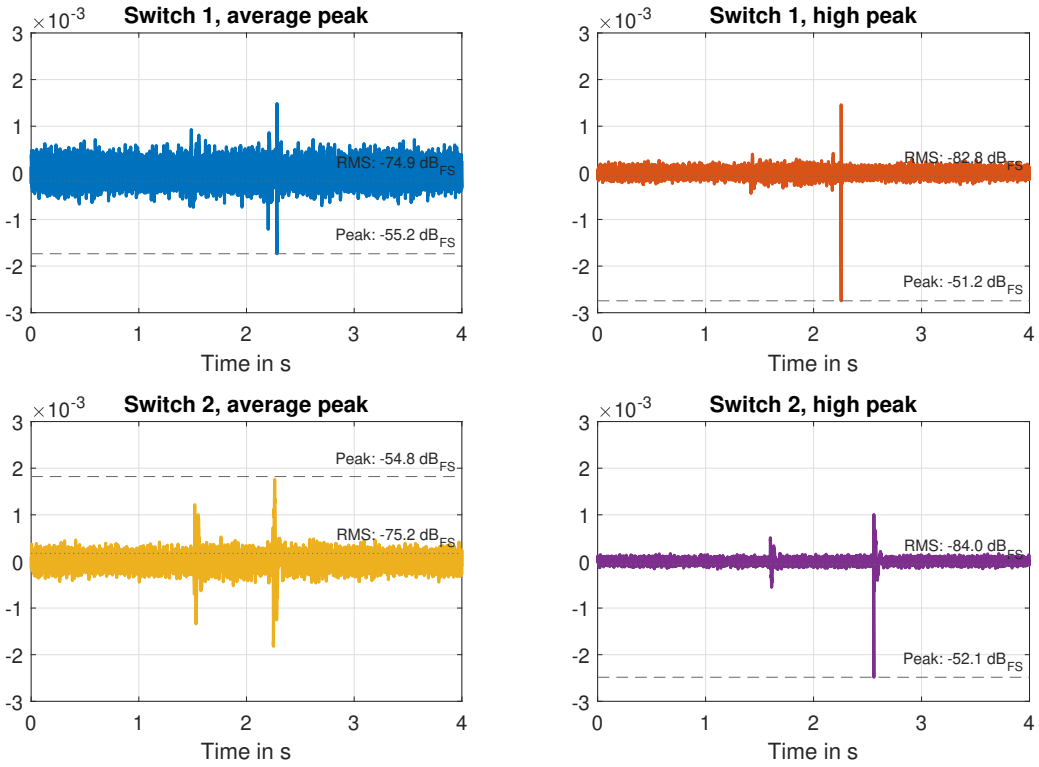


Figure 5.4: This plot compares all test signals at their respective audibility thresholds in the time domain. One can see that the noise floor as well as the glitch height are very different from one another. In addition, the differences are "inverted", where the noise is higher in *average peak* but the glitch is lower

pages of this work in section 3.4 *A-weighted RMS (Rooot Mean Square* with the assumption, that the A-weighted RMS should already improve some aspects of the the detection problems, but probably is not sophisticated enough to yield satisfying results. To validate this, all signals were filtered with an A-filter[1] at their respective audibility thresholds, determined from the listening test in order to compare the general signal levels but especially the peak levels of the highest peaks.

Figure 5.5 shows the transition from analysing the A-filtered signal (top plot) to analysing the RMS value of said signal. While there may be a small improvement over an analysis of the filtered time signal, the spread of the peak levels is still too large ($\approx 15\ dB$) and does not allow reliable prediction of audibility (the peak of *Switch 2, high peak* is below the pure noise in *Glitchless switch*).

5.2.2 Loudness

A parameter, that increases the complexity towards a more sophisticated psychoacoustic modeling and that is mainly focused on estimating the truly perceived level of an audio signal is *Loudness*. This scale was first defined by Stanley Smith Stevens in 1936 [19] with the idea to find a measure that maps an audible event to a *measurable sensation*. In 1991 Eberhard Zwicker developed a procedure to compute the loudness of signals which was also standardised in the DIN norm *DIN 45631:2010-03* [3] and the more widely used ISO 532-1 [9]. With the release *R2020a*, *MATLAB* introduced new functions which also includes several functions regarding

[1] With the *weightingFilter* function implemented by *MATLAB*

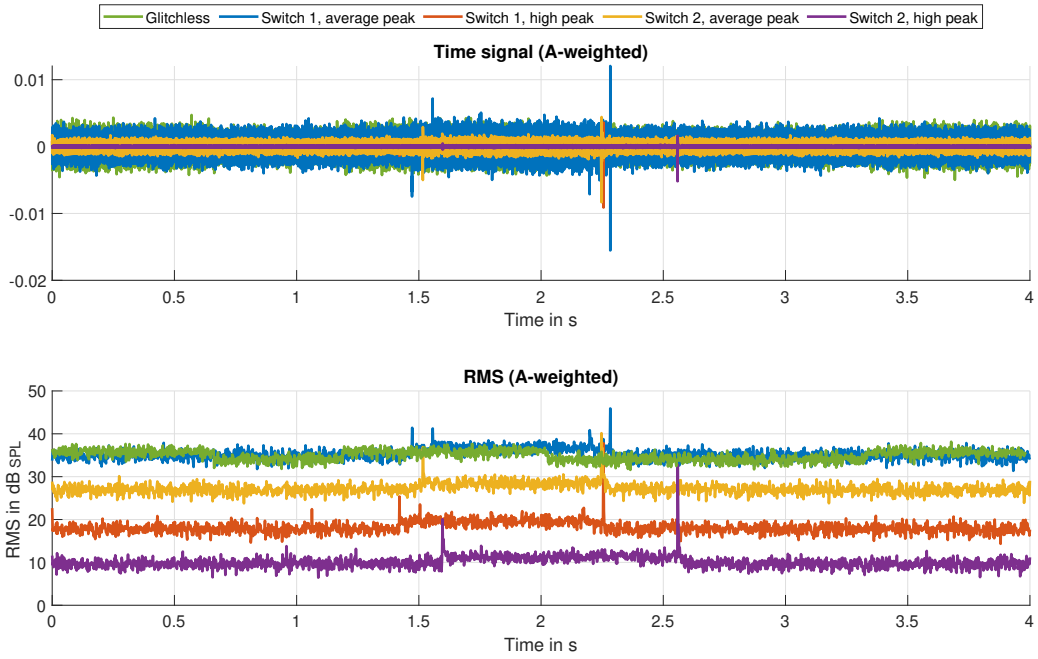


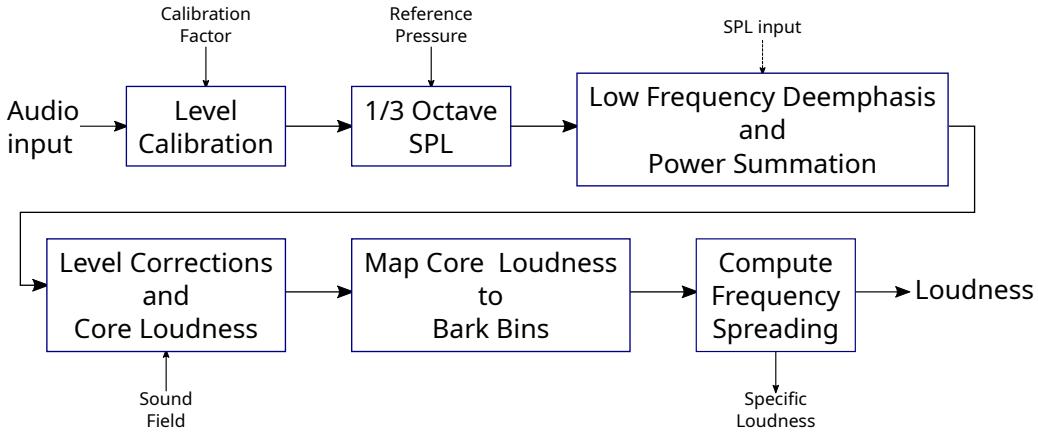
Figure 5.5: The transition from A-filtered time signals (top) and the A-weighted RMS (bottom) shows a slight improvement in the peak separation, but is not unique enough to derive a good threshold.

acoustic perception and thereby the function *acousticLoudness*[2]. The computations that are performed in this function are based on the standards defined in [9] and would also allow the use of ISO 532-2 (*Moore-Glasberg method*) which only allows the computation for time invariant signals.

Before being able to apply this function to the signals we want to analyse, a conversion must be performed in order to utilise *acousticLoudness* correctly. This conversion is necessary, since the test signals are generated by microphones without the input of a signal with a known level in dB_{SPL}. But with knowledge about the microphone sensitivity it is possible to map the signal from the full scale domain to sound pressure levels (see section 4.3.4).

A second setting that must be configured is the adjustment for the *soundfield* (see figure 5.6,

[2]https://www.mathworks.com/help/audio/ref/acousticloudness.html



This figure is recreated according to:
`https://de.mathworks.com/help/audio/ref/acousticloudness_5321_stationary_diagram.png`, 14.02.2022

Figure 5.6: Signal path, as described in the *MATLAB* Documentation for *acousticLoudness*

4th block). As stated in [9], depending on the sound field of the playback situation, certain levels must be attenuated, which is different for free and diffuse fields. As the *Beyerdynamic DT-770* generate a diffuse sound field at the ear, this must be considered, by changing this setting.

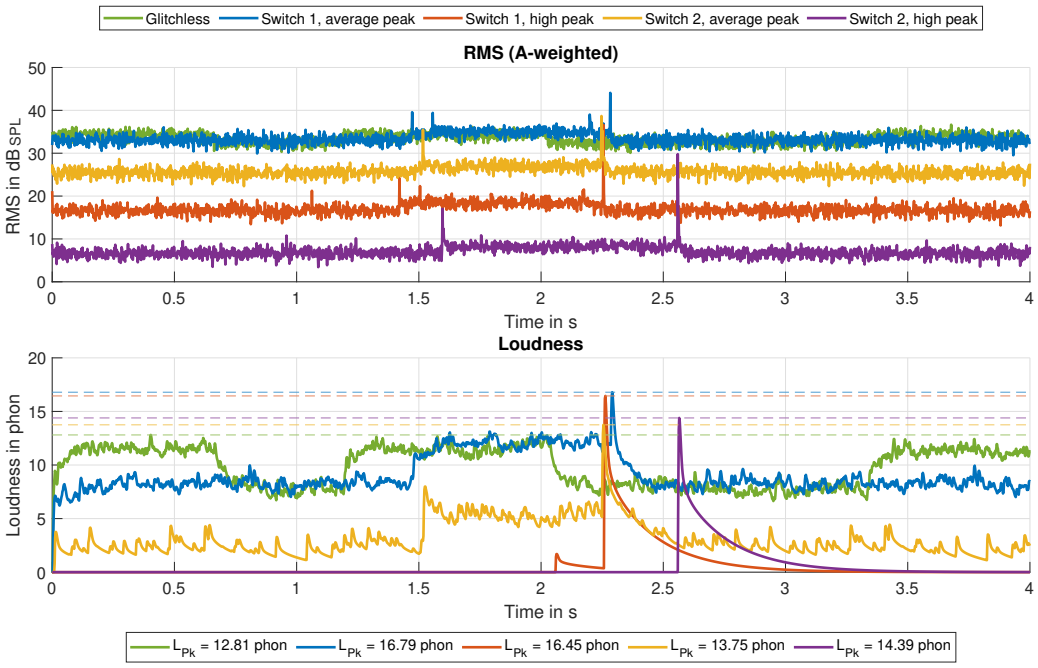


Figure 5.7: Comparison of a-weighted RMS and loudness for the analysed signals at their respective audibility thresholds (from tests). The peaks only show a spread of $\Delta L \approx 3phon$ and by that a vast improvement over the audibility prediction with a-weighted RMS. In addition, a coarse threshold of $L_{thr} \approx 13phon$ can be derived at this stage.

Figure 5.7 again shows the analysis of the tested signals at their respective thresholds as determined by the listening test. While the top plot again shows the A-weighted RMS of the tested signals (same as in figure 5.5, lower plot), here we also see the loudness analysis of those same signals at the bottom. In contrast to the RMS analysis, this transformation now shows an immense improvement regarding the spread of the peaks, reducing it from $\approx$ 15 dB to $\approx$ 3 *phon*. In addition, this measure is less sensitive to the different switching settings making this parameter a usable measure to investigate and compare the signals of this listening test regarding their audibility but may also prove useful in the next steps of analysis, where the disturbance of audible signals is of concern.

5.3 Disturbance in audible range

Since the *Loudness* measure works sufficiently well for predicting the audibility of the glitches and noise changes, the disturbance in the audible range is investigated with the same measure. As the participants had to rate the disturbance of the glitches at different levels, this data can be evaluated (with the average ratings) for every sample over the playback level (RMS in this case). As expected, those curves are very different in their absolute placements, but all show a very similar shape, that is, up to a certain point, monotonously increasing.

To explain the drop that occurs for every sample at a certain point, we have to consider the playback levels of the individual participants. Before the actual trial began, each participant

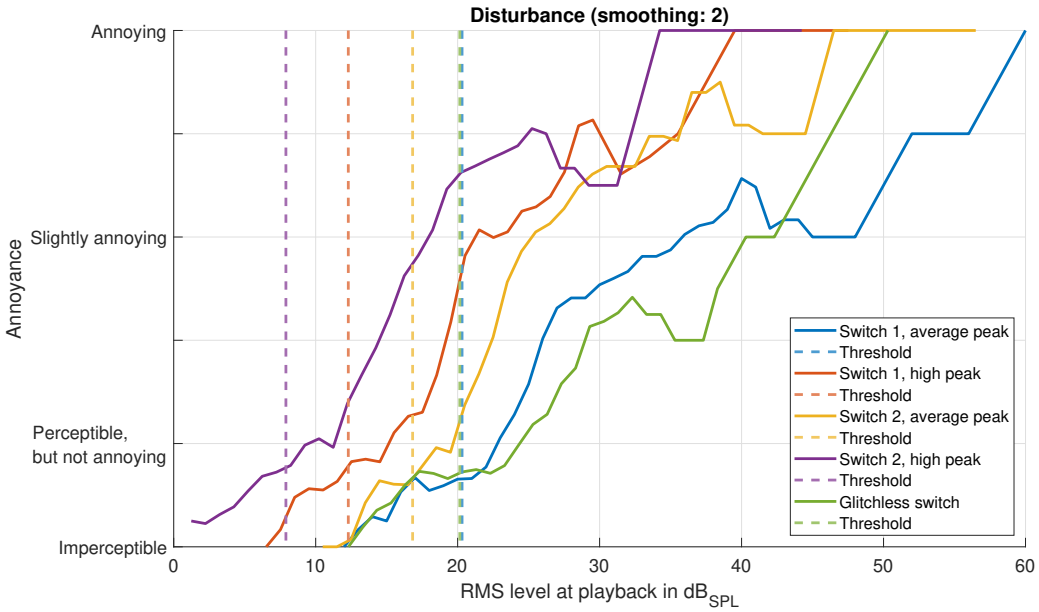


Figure 5.8: Average disturbance ratings per sample over their respective RMS playback level ranges. The spacing between the ratings represent the equidistant spacing used in the user interface for the listening test, including unlabeled intermediate values

had to set the level of two pure tones, which affected the initial playback level. In figure 5.1 (especially in the left plot) one can see that only a few participants started at higher levels than most of the other participants (trajectories are less dense). Consequently, only participants that started at higher levels could give ratings for such high levels, while the disturbance very soon decreases (for each participant individually) with decreasing level. Only after about 10 *dB* decrease, the other participants also were able to rate the disturbance and began with the highest disturbance rating, which increases the average again. This is supported by the variance in the rating behaviour over the playback levels for each sample (see figure 5.9) which increases towards the mid range of playback levels for each sample where participants with high initial playback levels are already at a lower disturbance rating, while participants that begin at this level start with high/the highest ratings. As expected, the variance decreases towards the lower and especially the higher end of the playback level axis.

This also highlights a different problem that occurs with this data, as the participants gave very different ratings compared to one another. Some participants mentioned they only rated between *Imperceptible* and *Perceptible, but not annoying* (the two lowest levels), as these were the descriptions that they felt were the most suitable descriptions for the heard artifacts. In contrast to this, there were also participants, who perceived even the slightest audible glitch as *Annoying*, so these participants will only use the ratings *Imperceptible* or to *Annoying*.

For further investigations it was decided to examine the signals at specific *disturbance thresholds* for their respective (peak) loudness. As the target is still to produce good quality microphones, ratings of *Slightly annoying* or worse were excluded for this analysis. As a still acceptable margin *Perceptible, but not annoying* is investigated as well as the next selectable value (just below *Slightly annoying*) as a still tolerable maximum.

Figure 5.10 shows the loudness for the test signals at the according disturbance values and the problem of the very different perception and understanding of disturbing in this very controlled

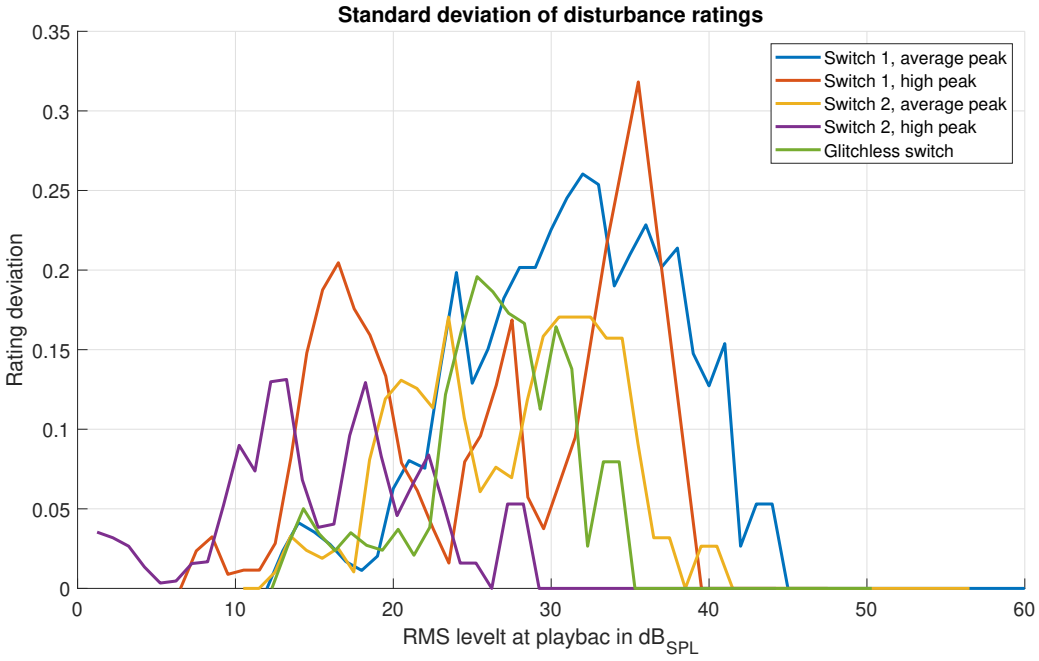


Figure 5.9: Standard deviation of the ratings over playback levels per audio file (Disturbance ratings are assigned equidistant values between zero and one). At the highest playback levels the ratings do not vary at all, which is due to single ratings in this range. Towards the lower levels (separate for every sample), the variance again declines, while in the mid range the ratings differ very strongly. This is a result off few participants with high initial playback levels, that rate the disturbance in the mid range less disturbing (due to the loud initial glitches), while other participants with lower initial playback levels, rate these disturbances the highest.

context becomes apparent. As the participants were only presented noise and glitches without any *carrier* signal (like speech or music) the participants could not produce a comprehensive understanding about the disturbance of signals at specific levels.

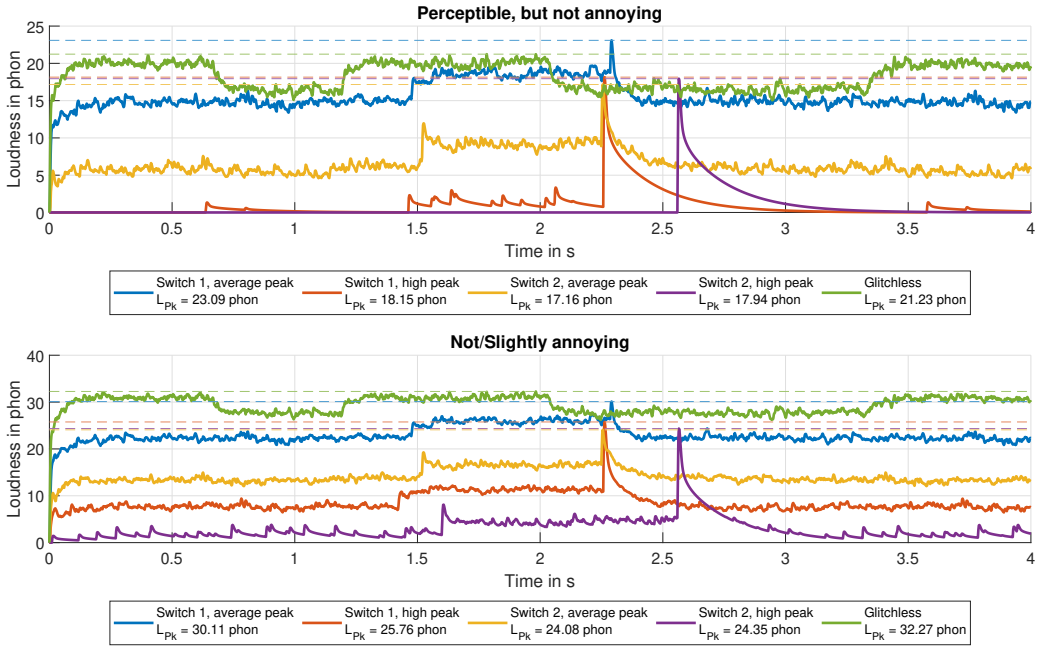


Figure 5.10: Loudness analysis for signals in audible range, where each signal is scaled to the levels corresponding to the disturbance values *Perceptible, but not annoying*, and the next intermediate value before *Slightly annoying*. The distinguishability starts to decrease, but there are discernable ranges to each disturbance value.

5.4 Derived measure for audibility and disturbance

Finally, to derive a set of applicable specification values, on can investigate the distribution and ranges of the peak loudness values of all tested signals. These thresholds are analysed and displayed in figure 5.11 for the audibility threshold as well as *Perceptible, but not annoying* and *Not annoying/Slightly annoying* (intermediate value).

Even though the results from figure 5.10 do not yield a simple threshold to discern between audibility and certain disturbance levels, we can see distinguishable ranges between those disturbance ratings. To use this as a design measure, the results in figure 5.11 should be interpreted as a range to classify for certain disturbance values, rather than thresholds.

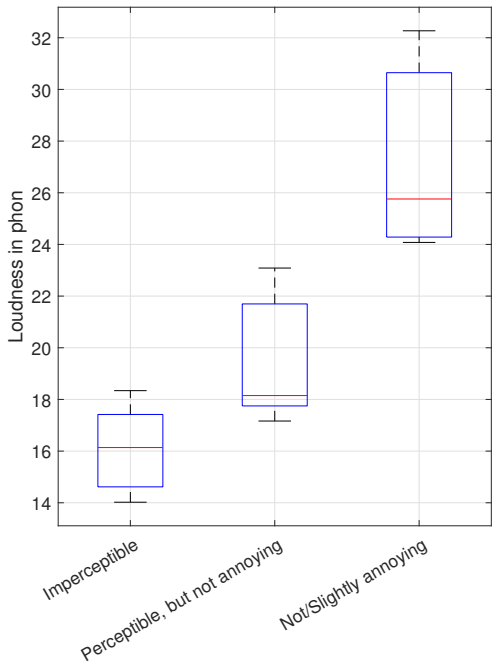


Figure 5.11: Loudness ranges for three disturbance thresholds, with discernable ranges from one another. For this representation, only the first three disturbance values are used (*Imperceptible*, *Perceptible, but not annoying* and *Not/Slightly annoying*, where the last is used to label the intermediate value between *Perceptible but not annoying* and *Slightly annoying)*. This was done, since disturbances above this level are already considered too much as well as the fact, that above this disturbance, the non-monotony starts for some samples.

6 Conclusion

With the results as discussed in section 5.2 and the requirements and targets for this research projects, the Loudness as defined by Zwicker [9] has proven to be a viable measure to predict the audibility of very short glitches. With regard to audibility the *Zwicker Loudness* also dominates in comparison with other perceptually motivated parameters such as the PEMO-Q and similar. Even though those parameters all apply similar psychoacoustic methods with similar computations, there are often many additional post processing steps involved, where the granular structure of how the stimulus is perceived is lost (i.e. the computation of the PSM in the PEMO-Q). Though in the case of very short bursts, such as the examined test signals are, exactly this micro structure is of importance.

While it seems to be a more sophisticated task to very exactly estimate the perceived overall *quality* of an audio signal, still the *Zwicker Loudness* is a parameter of great use, as already single glitches can have a disproportionally large influence on the overall perceived *quality* of a signal. Still, it is possible to derive certain loudness levels, that appear to be decisive in the context of disturbance and quality. Especially from figure 5.8 it becomes clear that the transition from *Imperceptible* to a disturbing audibility is a rather quick one leading to the conclusion that the margin for an acceptable disturbance is only a few phon above the audibility threshold (see figure 5.11).

Finally it should also be noted, that this listening test was performed in very *clinical* environments, where every external as well as expected influence (such as background noise or music) was controlled for by the setup of the listening test. So the results from these tests show the absolute lowest expected thresholds rather than what general consumer will perceive in everyday situations.

From a development point of view this may represent the *worst case scenario* as this testing procedure produces the strictest results (i.e. lowest thresholds), but it is nonetheless important to consider these situations, as one can not rely on the presence of background noise in every situation in order for a product to function as expected.

6.1 Outlook

For future undertakings in this general area of audio quality and audibility the results from this research should prove useful, as the understanding for perception could be strengthened. It may also be, that the potential usability to use the Loudness parameter as a predictor may be investigated in future research and could prove as a solid foundation for more perceptual measures.

Bibliography

[1] Georg v Békésy. "A new audiometer". In: *Acta oto-laryngologica* 35.5-6 (1947), pp. 411–422.

[2] Pablo M Delgado and Jürgen Herre. "Can We Still Use PEAQ? A Performance Analysis of the ITU Standard for the Objective Assessment of Perceived Audio Quality". In: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2020, pp. 1–6.

[3] *Berechnung des Lautstärkepegels und der Lautheit aus dem Geräuschspektrum - Verfahren nach E. Zwicker - Änderung 1: Berechnung der Lautheit zeitvarianter Geräusche*. Standard. Berlin, DE: Deutsches Institut für Nromung e.V., Mar. 2010.

[4] Gustav Theodor Fechner. *Elemente der psychophysik*. Vol. 2. Breitkopf u. Härtel, 1860.

[5] Zhipeng Feng, Ming Liang, and Fulei Chu. "Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples". In: *Mechanical Systems and Signal Processing* 38.1 (2013), pp. 165–205.

[6] Rickye S Heffner. "Primate hearing from a mammalian perspective". In: *The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology: An Official Publication of the American Association of Anatomists* 281.1 (2004), pp. 1111–1122.

[7] R. Huber and B. Kollmeier. "PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.6 (2006), pp. 1902–1911. DOI: 10.1109/TASL.2006.883259.

[8] *Acoustics — Normal equal-loudness-level contours*. Standard. Geneva, CH: International Organization for Standardization, Aug. 2003.

[9] *Acoustics - Methods for calculating loudness - Part 1: Zwicker method*. Standard. Geneva, CH: International Organization for Standardization, Nov. 2017.

[10] *Sensory analysis - General guidance for the selection, training and monitoring of assessors - Part 2: Expert sensory assessors*. Standard. Geneva, CH: International Organization for Standardization, June 2008.

[11] *Methods for the subjective assessment of small impairments in audio systems*. Standard. Geneva, CH: International Telecommunications Union, Radio Communication Sector of ITU, Feb. 2015.

[12] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. Standard. Geneva, CH: International Telecommunications Union, Feb. 2001.

[13] *Perceptual Objective Listening Quality Assessment*. Standard. Geneva, CH: International Telecommunications Union, Jan. 2011.

[14] J Suresh Kumar. "The psychology of colour influences consumers' buying behaviour–a diagnostic study". In: *Ushus Journal of Business Management* 16.4 (2017), pp. 1–13.

[15] Dominik Laurin Pürner. "Änderungen der peripheren und zentralen Schallverarbeitungsleistungen mit dem Alter". PhD thesis. Technische Universität München, 2019.

[16] Hugh Robjohns. "The end of the loudness war?" In: *Sound on Sound* (2014).

[17] Stuart Rosen and Peter Howell. *Signals and systems for speech and hearing*. Vol. 29. Brill, 2011.

[18] Norhashimah Mohd Saad, Abdul Rahim Abdullah, and Yin Fen Low. "Detection of heart blocks in ECG signals by spectrum and time-frequency analysis". In: *2006 4th Student Conference on Research and Development*. IEEE. 2006, pp. 61–65.

[19] Stanley Smith Stevens. "A scale for the measurement of a psychological magnitude: loudness." In: *Psychological Review* 43.5 (1936), p. 405.

[20] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and models*. Vol. 22. Springer Science & Business Media, 2013.