Design and Evaluation of Perceptually Pleasant Calibration Signals for Automated Loudspeaker Localisation

- Sound and Music Computing -

Master Thesis

Peter Ahrendt

Aalborg University
School of Information and Communication Technology

This document was created using LaTeX with Overleaf. All the figures were produced using Matlab.

School of Information and Communication Technology
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Design and Evaluation of Perceptually Pleasant Calibration Signals for Automated Loudspeaker Localisation

Theme:
Master Thesis

Project Period:
Spring Semester 2018

Project Group:
SMC181030

Participant(s):
Peter Ahrendt

Supervisor(s):
Jesper Kjær Nielsen

Copies: 0

Page Numbers: 102

Date of Completion:
May 28, 2018

Abstract:

It was attempted to find a pleasant calibration signal for automated loudspeaker localisation. A simple use case was defined where the time of arrival (TOA) should be estimated from one source to one receiver in corrupting noise and reverberation. Interviews with experts and research were used to determine necessary characteristics of a possible calibration signal. A self designed sound and other suitable sounds were rated in a listening experiment. The winner was tested in a virtual testing framework against a traditional signal. The traditional signal outperformed the pleasant one. Methods were applied to modify the pleasant signal in order to increase its performance. It was hypothesised that a wide spectral bandwidth or at least high frequencies are crucial for TOA estimation. High frequency pseudo-random noise was added to the pleasant signal according to and beyond its masking curves with an increase of TOA estimation performance but a decrease of perceptual quality.

Contents

Preface

This master thesis was made as a completion of the master's program in Sound and Music Computing at Aalborg University in close collaboration with my supervisor Jesper Kjær Nielsen. Furthermore, I got valuable input from Bang & Olufsen and Audiowise.

I would like to thank everybody who helped and supported me with this master thesis.

Aalborg University, May 28, 2018

Peter Ahrendt
<pahren16@student.aau.dk>

Chapter 1

Introduction

1.1 Problem Analysis

Users of audio systems with multiple loudspeakers typically place them according to the interior of their listening environment. In most cases those positions will not be optimal to achieve to best sound quality possible for a given listening position, the so called sweet spot. Furthermore, users will probably not consider how room specifics affect the sound perception.

Algorithms can compensate for imperfect loudspeaker positions and room acoustics. Information which has to be provided is distances from the loudspeakers to the listener or the room impulse response (RIR) at the listening position or both. To obtain this information, the audio system needs to be calibrated. State of the art systems typically use an external microphone which has to be placed at certain positions while a calibration sound is played back. The BeoLab 90 sound guide [3] from Bang & Olufsen for example requires that the external microphone has to be placed at three microphone positions per zone (the area where one is listening). RIRs are measured at each position. The Sonos Trueplay speaker comes with an app installed on an iPad which uses the build in microphone to record short sine sweep bursts emitted by the Sonos speaker. In both cases the calibration takes quite some time and seems to be tedious. The assumption therefore is that people will not do the calibration too often.

Ideally though, the calibration should be remade every time when something changes. That could be:

- changing the position of loudspeakers
- changing the listener position
- temperature changes

- furniture changes
- or open/close window/door.

One improvement compared to state of the art systems would be to ensure the best sound quality for every audio system setup in every environment any time. One idea to achieve this is to introduce automated audio system calibration. Automated audio system calibration can be triggered with turning on the system. The contribution then is that the user would not have to perform the calibration any more, but still enjoy the best sound quality possible for every listening situation. This would be especially useful for people who use portable loudspeakers and move them to different listening environments.

Automated audio system calibration could be performed using smart speakers. Smart speakers, like the Amazon's Alexa or Google Home, are loudspeakers with additional microphones which give a lot more opportunities than just playing back signals, hence the name smart speaker. Smart speakers usually have a compact size so that they can be moved around. The smart speaker could function on its own or be part of another audio system. In [17] a method is proposed to create a map of loudspeakers, the listener(s) and microphone subarrays. If a subarray would be integrated in a loudspeaker, one would call that a smart speaker. The resulting map can be used for moving the sweet spot to the listener's position, ensuring an optimal sound quality. Furthermore, it could be used for object based rendering specified in the MPEG-H standard [11] which enables immersive 3D-sound. Additionally, when creating sound zones, the impulse response from the loudspeaker to the zone is required. When the loudspeaker position, room characteristics and geometry and air characteristics are know, such an impulse response could be simulated.

Another way to improve the sound quality is to compute the transfer function (TF) from measuring the RIR and make adjustments based on that. For example, one could attenuate the low frequency response of a smart speaker when its near a wall or corner. Ideally, every technique which can improve the sound quality should be used.

For creating the map over loudspeakers and the listener in [17] every loudspeaker played back a pseudo-random noise. For measuring RIRs, well proven calibration signals are sine sweeps, maximum length sequences, pink or white noise. All of the above signals have a good performance in terms of loudspeaker localisation/RIR measurement but probably do not sound pleasant to most people. When an audio system calibration is performed every time the system is switched on and one of the above signals is used, though, it is heard frequently and therefore must sound

pleasant or at least not annoying. There is a trade-off between performance and pleasantness. "Nice" signals like speech or music are pleasant but perform insufficiently, traditional signals, e.g. the pseudo-random noise or a sine sweep, perform well but sound less pleasant. This was found in the work done previous to this thesis [1] were the time of arrival (TOA) estimation performance was tested for various signals.

1.2 Problem Formulation

The purpose of this master thesis is to design a calibration signal which fulfils the requirements on audio system calibration and at the same time sounds pleasant. With such a calibration signal automated audio system calibration could be done unobtrusively which would be different from state of the art systems.

1.2.1 Research Questions

The following use case is assumed: A map of N loudspeakers shall be created which all play back a desired calibration signal (instead of a traditional one). Based on that, the following research questions are asked:

1. What is a desired calibration signal for the aforementioned use case?
2. How well does the desired signal found in 1) and the traditional signals perform for source localisation and how robust are they to noise, reverberation, non-ideal loudspeakers and microphones, etc.?
3. How can one modify the desired calibration signals found in 1) so that they give
    a) the best possible localisation accuracy for a given amount of perceptual distortion, or
    b) the lowest perceptual distortion for a given localisation accuracy.

The problem is simplified to the case were the TOA is estimated from one source to one receiver in different conditions. With the TOA the distance between source and receiver can be calculated which is the first step towards creating a map. Furthermore, the TOA is an approximated version of the RIR which can be used to regulate the frequency response of loudspeakers. The above mentioned research questions can answered for the TOA estimation case.

The following chapter 2 starts by addressing the problem of finding desired calibration signals.

Chapter 2

Finding Desired Calibration Signals

This chapter deals with the problem of finding a desired calibration signal which can be used for automated audio system calibration, i.e. loudspeaker localisation. It regards the research question one in chapter 1. To simplify the case it is assumed that any sound would be possible to be used for calibration. Because the desired calibration signal is played back often, the following characteristics are deemed important:

- pleasantness
- annoyance
- simplicity
- duration

The calibration signal could also be designed in such a way that it is a signature and has a "wow-effect".

To receive input from experts on that field a meeting took place at Bang & Olufsen headquarters in Struer, Denmark and the CEO of the company Audiowise was interviewed.

2.1 Meeting at Bang & Olufsen

The meeting took place with two experts in the fields of user experience and user interactions from Bang & Olusfsen (B&O). B&O is a company which produces high end audio products, televisions and telephones. After explaining the idea of the project, the experts were asked: "What would be the perfect calibration signal for automated loudspeaker localisation in your opinion?" A first conclusion that was drawn quite quickly was that using speech is not a possible scenario. Possible

use of speech could be that the loudspeaker would count from one to the number of loudspeakers N or say "I am speaker N". But letting the speakers saying that would provide unnecessary information for the user. The user expects that the system has the information and uses it without the user to interact, so stating it would not have additional benefit. After ruling out speech, the choice was in favour of a musical signal. It became clear that there is no easy answer for how the musical signal should be characterised. The first question one has to answer is for which target group the calibration signal should be made. B&O for example tailors its products to a generally younger (B&O play) and generally older (B&O home) audiences. Products for the B&O play audience are cheaper, smaller and mobile, whereas products for the B&O home audience are high-end audio products which are bigger in general. For the B&O home audience a classical or a jazz musical sound would suit whereas for the B&O play audience it would need to be something more abstract or futuristic.

A main requirement according to the experts is that the sound needs to be short. It was considered that a user would switch on the audio system possibly five times a day, maybe even more. That means that the calibration needs to be done as fast as possible, which requires a short signal. Five seconds for example would already be too much. Furthermore, the sound cannot be annoying. The user should not be bothered by it after having it heard 5,000 times for example. The time when the audio system is switched on is a happy moment. The user wants to listen to music, so ideally the calibration sound needs to increase the desire for listening to it. Another thought from the B&O experts was that the calibration sound cannot compete with the music already playing. First play back the calibration sound then the music was the advise. Moreover, the sound should be something people should want to use to show off to their friends or family.

In addition to that, some attributes were mentioned which the sound should have. Note that B&O produces high-end audio products. Therefore, the attributes mentioned are intended to convey that feeling:

- well crafted
- associating beauty
- associating luxury
- rich
- beautiful
- everything is taken care of

2.2 Call Audiowise

Audiowise is a company which makes their living out of consulting companies regarding their audio representation. That means that the company helps increase sales and brand values. The CEO of Audiowise was asked what he would think would be a perfect sound for automated audio system calibration. As expected no definite answer can be given. There is no universal sound. Instead, it was pointed out that several considerations need to be made before starting to design or choose a sound.

As in the meeting with B&O the CEO of Audiowise underlined three steps to do before looking for a sound or designing one:

- Defining a target group. Which users do you want to address?
- What experience do you want users to have?
- And, what functionality should your sound have? What do you want to achieve with it?

Based on those criteria, the sound design process can begin. Additionally, it was pointed out that the market where the sound should be used has its own specifications. The Asian market for example has a different approach to sound than the European market. The interview ended with the advise that less is more. Rather take away something than adding more and more features to your sound.

2.3 Preceding Considerations for Sound Design

Based on the previous described interviews, it is indispensable for sound design to define a target group, know what experience a listener should have and clarify the functionality.

Functionality The sound is intended for the use of calibrating an audio system. Therefore, it needs to be simple and short. That is because the calibration cannot take long. Ideally, the audio system is started and after a couple of seconds ready for use. Designing the calibration signal is based on the case where the audio system consists only of a few or even just one loudspeaker(s). In this setup, it can be afforded when each loudspeaker plays back the calibration signal in turn or if the only loudspeaker plays back the signal. For now it is also assumed that all sounds would be suitable for audio system calibration. Modifying a designed sound that fulfils the calibration requirements is left as a future task.

Experience When hearing the calibration signal, the listener should associate high-quality with it. He or she needs to feel that all is taken care of. The user's wish to listen to music or other signals should be supported.

Target group The target group are customers who have a need for compact, portable audio equipment. They live in relatively small apartments. They move their audio equipment around. For instance would they bring it to a friend at his or her place and watch a movie together or listen to music. Depending on the listening environment, the smart speaker which is transported could be part of another audio system or just function alone. When the audio equipment is moved a lot, calibration makes most sense because the situation where the system is used, changes a lot. The target group does not buy the most expensive audio products but still is willing to pay above average to experience very good audio quality. The target group can be seen similar to the B&O play audience.

2.4 Summary

This chapter has started the process of sound design and being able to select appropriate sounds by obtaining knowledge from experts. Preceding considerations for designing a calibration signal were made. In the next chapter, literature regarding sound design is reviewed and used to design a calibration signal, which fulfils the expectations in terms of functionality, experience and target group. Subsequently, this signal together with other chosen sounds is used in a listening experiment to check if the expectations are indeed fulfilled.

Chapter 3

Sound Design

The goal of this chapter is to summarise the existing literature regarding sound design. The process of sound design shall be approached in a rather scientific way, identifying acoustic features which lead to desired attributes of the resulting sound. The most important desired attributes are that the sound is pleasant, not annoying and has musical aspects.

3.1 Literature review

3.1.1 Product-Sound Design

When looking into the field of sound design, much literature is found which deals which product sound design and how it can best represent the brand of the product. In [6] the distinction is made between consequential sound and intentional sounds. Consequential sounds are emitted when the product, one is interacting with, is used. For example, when one turns on a vacuum cleaner it produces a sound. Or if the button of a coffee machine is pressed, the machine emits a sound due to the milling of the coffee beans. Intentional sounds on the other hand are digitally added to the product to support its function. For instance, microwaves often beep when buttons are pressed.

The sound which shall be designed here would be an intentional sound, i.e. when switching on the audio system a sound is emitted which gives the user the information that the system is ready to go. The authors in [6] state that communication is a difficult task for both the sound designer and his or her clients. Those difficulties occur when sound shall be described verbally, as there is a lack of sound vocabulary. Therefore sounds are described using metaphors, emotions or onomatopoeia ("An onomatopoeia [...] is a word that phonetically imitates, resembles or suggests the sound that it describes"

(Wikipedia: https://en.wikipedia.org/wiki/Onomatopoeia)). Often the client does not know which sound he or she desires or cannot formulate his or her ideas. It is then difficult for the sound designer to create an appropriate sound. To ease this situation, a framework was proposed which supports communication between clients and sound designer using cards with descriptive attributes. The authors from [15] also endeavour to make the conceptualising phase for sound design easier. They, on the contrary, propose a sound sketching tool with physical objects which can be used to simulate every-day sounds. The tool is intended for inexperienced sound designers, supporting verbal communication with sound examples using a tangible user interface. Furthermore, the tool is rather concentrating on the design of consequential product sounds.

3.1.2 Semiotics in the Context of Product-Sound Design

In the scope of this thesis, though the idea and the requirements behind the desired sound are already clear. Deducting from the literature of product design it seems to be worthwhile to look into the semiotics in the context of product sound design. The book chapter [16] was studied to this end and the relevant information summarised as follows.

Sound design seems to be an artistic task where intuition plays a more important role than approaching the issue in a rational and systematic way. But sound designers make decisions based on the listener's perception. They want to communicate something. In that regard semiotics can help. It investigates cultural processes as communicative processes. In the context of sound, it investigates how the perception is affected by product sounds [16].

Semiotics is the theory of signs. Signs are physical objects which are given a specific meaning by convention. Signs could for example be traffic lights, number codes or speech. Signs have the purpose of communicating something. A traffic light for instance shows a certain colour, let us say red. This is perceived by the driver, interpreted and meaning is assigned to the object of perception. The driver understands that he or she has to stop the vehicle. The perception of a product sound is in principal similar, although more complex. A sound can also be seen as a sign carrier which carries information. Communication will fail if the code is not understood by the receiver. When one has never encountered traffic rules, a red light will not be understood. Consequently, there is the perception of an object of prior experience and meaning which evolves based on that [16].

Semiotics examine the processes of how objects are perceived and meaning is evolved from that. Humans generally have an expectation of objects they perceive. The more the perception is in line with the expectation the easier an object

of experience can be identified. The object of perception becomes a signifier for an object of experience, in other words, semiotics takes place. This happens highly dependent on individuals but there are also inter-individual relationships [16].

If the perception is not in line with what is expected, the data of perception is modified until its compatible with what is learned already, with the expectation. An example is synthetic speech: Ordinary speech is expected but synthetic speech is perceived. There is a new offer, a disequilibrium so that the data of perception has to be modified [16].

How can this be used for product-sound design? Ideally, the perception of the product-sound should lead to a positive meaning. The customer should think of high performance or good aesthetic characteristics. The user's expectations should be satisfied. The goal therefore is to meet the user's expectations with a product-sound which leads to the corresponding perception. The task of the sound designer is to create an image which supports product quality. The sound design is also a communication design. Specific information is conveyed for a specific purpose, using specific codes and media for a specific user group. Ultimately, the quality of the product results from judging the comparison of data of perception with data of expectation. The sound of a product is a sign carrier for that matter [16].

The challenge now is to design the sound so that the above is achieved. Listeners may concentrate on different aspects when they associate meaning with a specific product sound. Because this is different for every individual product-sound engineering is often seen as a black box. The author of the book section therefore tries to approach the topic in a scientific way. The sound designer must create an auditory event to associate meaning. He or she must create a mental image in the listener's mind. An example for this could be a hammer banging into wood. Due to the object of experience the listener will associate the sound with wooden material. A sound might also carry signs for cultural values, like tradition, luxury or originality. The meaning of all this is perceived in the so called Semiosphere which is a comprehensive context where product-sound is perceived and cognitively processed [16].

In the following, the takeaways for designing a sound which later can be used for calibration are summarised: The aim of sound design is that an acoustic event is created which is both novel and already known. Already known because it should enable the listener to quickly understand the meaning of the sound and its function. At the same time it should be novel because the sound designer wants to stimulate the listener's imagination and invite him or her to use the product, evoke curiosity. Presenting something novel also means that the sound cannot be

processed with available data of experience but needs to be modified first. This must be kept simple to a certain degree to ensure that basic relations and pattern of experience are not violated. Listeners should be able to treat the auditory event as a token of specific known type. This is specifically important when artificial signals are used. One way to do it would be to use artificial sounds in combination with natural sounds. This is then something novel and well known [16].

After talking about the semiotics of product sound design, the topic psychoacoustics in relation to sound quality is reviewed. The following section is based on [8].

3.1.3 Psychoacoustics and Sound Quality

Studies on psychoacoustics can help to find an optimum target sound for a specific product. With this goal in mind the following psychoacoustic quantities are worth mentioning:

Loudness The loudness of a product sound strongly affects the perceived sound quality of a product. Loudness is perceived based on spectral and temporal characteristics. A longer tone for example is perceived louder than a shorter tone, both having the same sound pressure level [8].

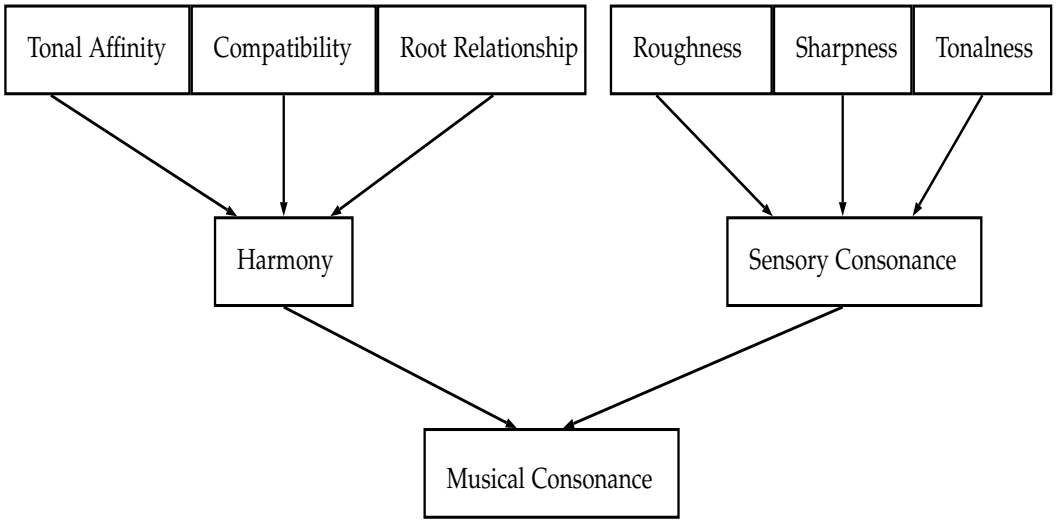
Sharpness Sharpness plays a dominant role in sound quality, too. It can be regarded as a measure of tone colour. Adding the right amount of sharpness to a sound can give it powerfulness, but adding too much yields the perception of aggressiveness. Sharpness can be reduced when low frequencies are added, but this must be done carefully because it causes the loudness to increase [8].

Roughness Roughness is caused by temporal variations of a sound. It reaches its maximum at frequencies around 70 Hz. Roughness can be a feature pointing to the sportiness of a car [8].

Fluctuation Strength Fluctuation strength is quite similar to roughness with the difference that it references the envelope fluctuation of fluent human speech, which fluctuates at 4 Hz. Consequently, this is where it has its maximum. 4 Hz corresponds to the number of syllables pronounced per second. The human hearing is very sensitive to that. The above mentioned quantities can be combined to calculate the overall annoyance of a sound [8].

Psychoacoustics can also be considered for sound design of musical sound effects. If roughness, sharpness and tonalness are chosen well, they lead to sensory consonance. Together with musical acoustics such as tonal affinity, compatibility

and root relationship which, put together in a right way to produce harmony, psychoacoustic quantities can produce musical consonance [9]. The following diagram taken from [9] visualises this relationship.



Tonal affinity is the relationship between tones. For example, the octave equivalence or fifth/fourth similarity. Compatibility describes a tonal piece of music or or tones which can be replaced by other compatible tones without seriously interfering with harmony. It sometimes is also called reversibility of chords, tolerability or interchangeability. Root relationship of tones are key notes which are assigned to musical chords. The perception of a root can be ambiguous, sometimes resulting in the perception of a virtual pitch. Tonalness is the amount of audible tonal components compared to noisy components [9].

In order to achieve sensory consonance roughness should be avoided and sharpness kept small. Tonalness should be high, meaning that noisy components are as small as possible. Contradicting to consequential product-sound design, loudness is not that important for sound design of musical effects. [9]

3.2 Sound Design Process

Based on the knowledge reviewed above the sound design process was started. The sound was supposed to be an abstract musical related piece with known and unknown elements. It should be an intentional sound. Care needs to be taken with adding unknown elements as the designed sound should be easy to understand and interpret. It shall be pleasant, not annoying and short. Sharpness and roughness shall be kept to a minimum. Low frequency content can be added to

reduce sharpness and because it is believed that this will increase sound quality in general. Additionally, noisy components should be minimised, too. Overall, the aim is to achieve sensory consonance. On the other hand all aspects which lead to harmony should be accomplished, too. In order to add associations to physical objects physically modelled sounds could be added.

The sound was designed using the digital audio workstation (DAW) Reason 9.5. With the DAW the built in polyphonic synthesiser Thor was used, specifically the Cosmo Pad 3.thor sound effect. The three notes B2, D#3 and F#3 were played in subsequent order and lasted exactly three seconds. Every note was played together with two of its same note two octaves higher so that three notes reaching over two octaves were played simultaneously three times.

Additionally to the notes, it was considered to add a logarithmic sine sweep or sawtooth sweep as a novel component. However the overall perceptually quality would have been reduced in the author's opinion so that this idea was dropped.

An excerpt was cut and added from a BeoLab 90 promotion video `https://www.youtube.com/watch?v=U5gufdPM8iY&feature=youtu.be`. The excerpt featured a low frequency rumble.

The notes were sounds which were well known. The unknown elements in the designed sound are rather subtle: The underlying low frequency rumble. Maybe the synthetic sound of the notes will be perceived as novel, too. But overall the known components are more dominant. The notes fitted together harmonically and it was taken care of that roughness and sharpness were kept low. Finally, the created sound had more tonal than noisy components. No physically modelled sounds were added, though.

3.3 Complementary Sounds

Besides the designed sound, other sounds were chosen which had characteristics crucial for a perceptually pleasant calibration sound. The sounds were chosen based on the reviewed literature and subjective opinion.

Apple HomePod start up sound The Apple HomePod pairing sound is played back when the system is switched on. It is a deep harmonic sound with some high frequency bell ringing sound which makes it interesting. The sound has a length of 4.6 seconds.

B&O bluetooth speaker start up sound This sound is used for B&O play's P2 bluetooth speaker when it is switched on. There are two subsequent tones, the latter higher than the former. The sound is very simple and familiar. It is also short: 0.6 seconds.

THX intro sound The well known THX intro sound was chosen for its massive sound which was thought to be associated with high quality. The author believes that people usually know this sound from cinemas which is in general an environment of good experience. The characteristic which makes it less suitable for calibration signal is its duration: 14.1 seconds.

3.4 Summary

This chapter has reviewed the literature on sound design. While the topic is rather an artistic and intuitive field, some rules from the field of semiotics and psychoacoustics could be found. The knowledge was applied in designing a sound.

In the next chapter the designed sound and other sounds will be compared against each other in a rating experiment.

Chapter 4

Rating Experiment

4.1 Experimental Design

In order to investigate the perception of possible calibration sounds and the perception of the related product (the audio system) by people a ratings experiment was designed. The designed test interface can be seen in figure 4.1. Attributes were chosen to describe both the sound perception and product perception, following the guidelines to construct a semantic differential scale [18, 12]. The rating scales itself were designed inspired by the MUSHRA standard to measure audio quality [14]. One scale was initialised for each presented sound ranging from 0 to 100. The attributes were positioned above and below the scales and are listed as follows:

- Unpleasant–Pleasant (Sound perception)
- Annoying–Not annoying (Sound perception)
- Long–Short (Sound perception)
- Boring–Exciting (Sound perception)
- Low quality–High quality (Product perception)
- Cheap–Expensive (Product perception)

The attributes were written in the following form: negative Attribute – positive Attribute. The underlying presumption was that when the sound perception would match all positive attributes, the user would have a good experience with the audio system. All attributes which were considered to be positive for the calibration sound were positioned above the scales, all negative attributes below in order to be consistent and avoid confusion. The positive attributes are equal to 100 on the scales, negative attributes equal to 0. The scales ranged on vertical lines and were

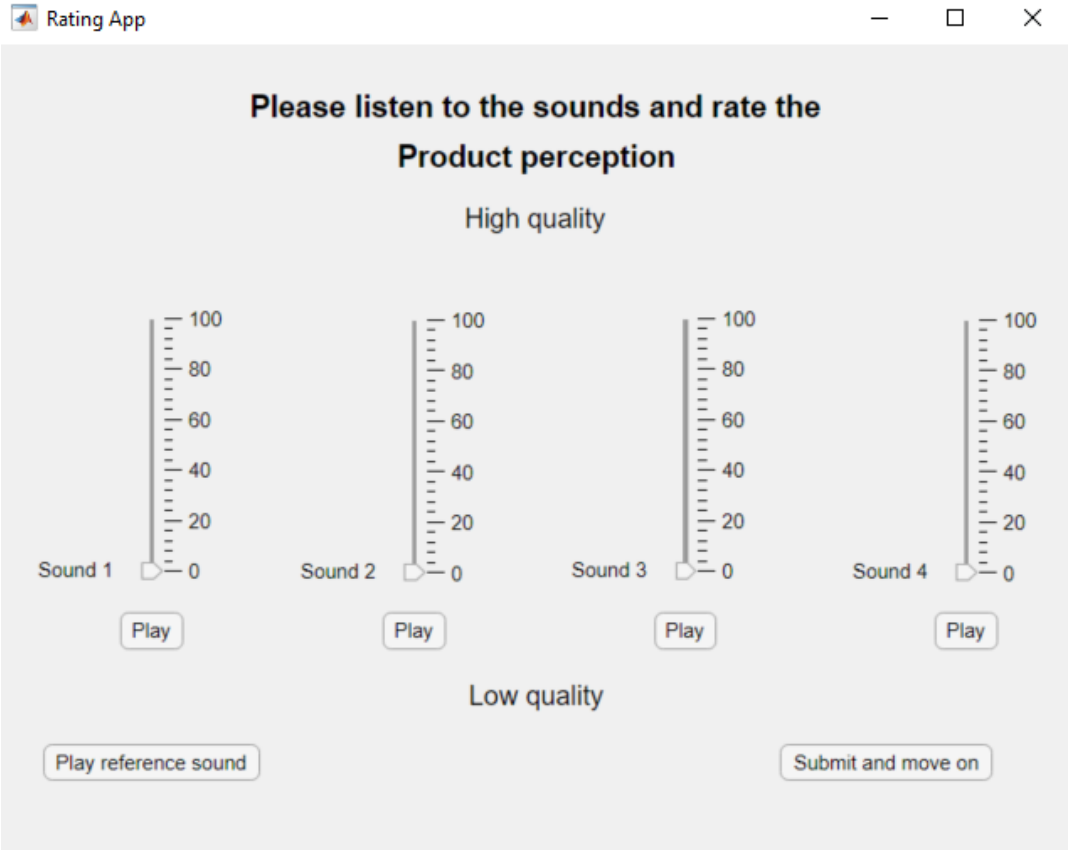


Figure 4.1: Testing interface rating experiment

positioned next to each other on a horizontal line. The scales were chosen according to the well proven MUSHRA standard [14] and to attempt the production of interval data to be able to use parametric statistical tests for post test analysis. The following rateable sounds were included:

- Apple HomePod pairing sound
- Bang & Olufsen bluetooth speaker start up sound
- THX intro sound
- and the designed sound described in chapter 3

Additionally, a reference sound was provided which referenced to the rating 0. All other ratings could only be equal or higher rated. The reference sound should be representative for the worst possible calibration sound in terms of a good sound and product perception, thus a bad user experience. Buttons enabled participants to play each sound. The test was done without a hidden reference and anchors.

The reasoning was that it would be too easy to detect those, thus not having any additional value.

The testing interface was designed with the appdesigner in Matlab Version: 9.4.0.813654 (R2018a). The testing interface picked randomly attributes until all attributes were rated for the sounds, which were available for listening all the time. When ratings were made the participant could submit this and new attributes were presented. With each new attribute the order of the sounds was randomly changed so that participants would listen to each sound at least once per attribute pair. When the experiment was finished, the results were plotted which is exemplified in figure 4.2. The plotted results were intended to be used for a subsequent discussion

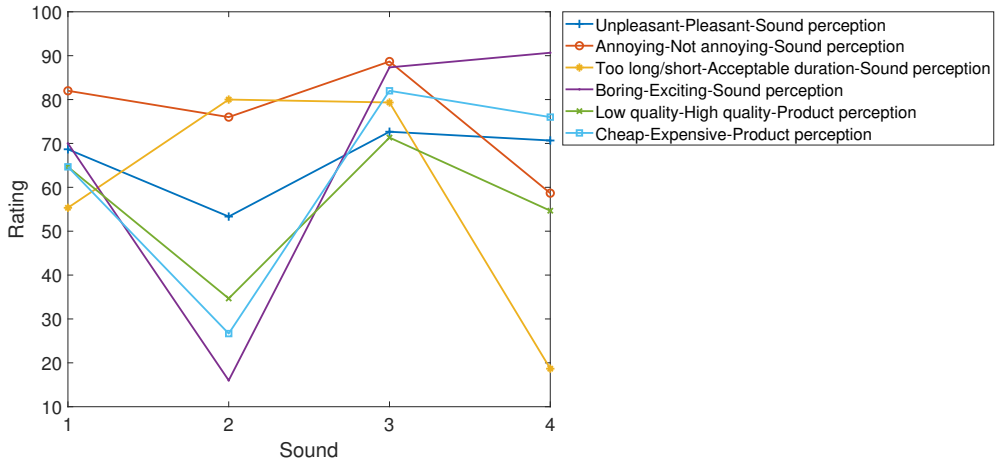


Figure 4.2: Plotted test results after finishing the rating experiment

with each participant of the experiment to understand their choices and give the opportunity for comments.

4.2 Experimental Procedure

The experiment was conducted in the sound lab in the CREATE building of Aalborg University. The test interface run on a laptop which was connected to two bioacoustics dynamics loudspeakers via a Roland Quad Capture sound card. Figure 4.3 shows a photograph of the testing environment. 14 participants took the experiment. After greeting, the experiment was explained with small varieties as follows:

Imagine the following situation: You are at home and switch on your audio system because you want to listen to music. Before you can do that, you will hear a start up sound.

Figure 4.3: Testing environment in the sound lab

After the start-up sound has ended, your audio system is ready to use.
Your are going to listen to four examples of such a start up sound and rate these based on different attributes making use of four different scales. The attribute which is above the scales is equal to the rating 100 on the scales, the attribute below equal to zero. You are supposed to rate the sounds in comparison to each other and based on a reference sound. The reference sound is scoring the minimum rating for each attribute, that is zero. In most cases the attributes are related to the sound perception of the start up sound itself. In two cases, though, they are related to the product perception. With 'product', I mean your audio system in your home. One thing which is very important to consider is that you would listen to the start up sound very often. Let us say that you want to listen to music five times a day, then you would also have to listen to the start up sound five times a day.
We start with a training interface. It looks exactly like the real one with the same attributes but has different sounds, which are simple sinusoids. You can use the training interface to explore the handling and get comfortable. When you are ready, we proceed with the real test. [Participants trains until he or she is ready.]
Here is the actual test interface. It looks the same as the training interface with the same attributes but now you are going to listen to different sounds. Please remember the situ-

ation you are in: you are at home. It is the best time of the day because you are about to listen to your favourite song, but before you can do that, you will hear the start up sound of your audio system.

After the participants finished the experiment, they were shown their ratings and asked if they had any comments about the sounds. This usually evolved in a little discussion.

4.3 Feedback from the Participants

For many participants a short duration of the start up sound was very important. Some even said that they would prefer no start up sound at all. Comments about each sound are described in detail as follows:

B&O bluetooth speaker sound Many participants said the sound is very simple and therefore not annoying. That also means, on the other hand, that it does not convey a high quality perception. It is not special. Some even said that it is boring. Participants liked the duration of the sound. Because it is so short, they would not mind hearing it often. Another comment was that it does not interrupt the routine of switching on the audio system. It does not grab attention.

Apple HomePod start up sound It was noticeable from the discussion after the experiment that most people preferred this sound. It was described as short but still with good quality. It has a good balance in terms of acoustic features. Something of everything was a comment. Also, it has a good bandwidth. Participants could imagine listening to it five times a day.

THX intro sound Almost every participant knew this sound. Either they really liked it or they did not. Many participants said that it was too long for the situation where it should be used in. But also many participants said that they associate it with high quality. Some said the sound is too dramatic for a living room environment. It triggers action which is counter productive when one wants to relax and listen to music. Although the sound was linked to the situation where you watch a movie in the cinema, which is generally a good experience, the THX sound was found inappropriate for casual occasions. Furthermore, almost everyone said that it is too long for the intended purpose.

Designed sound Despite applying knowledge from research on sound design, the created sound was not perceived as good as intended. The majority of participants associated it with low quality and not interesting to listen to. They would not want to listen to it often. Very few participants liked it in the given context.

4.4 Data Analysis and Results

Each of the 14 participants produced six ratings for six pairs of attributes for four sounds. Consequently, 84 ratings were available. The means for all attributes and sounds over the 14 observations are depicted in figure 4.4. It can be seen that

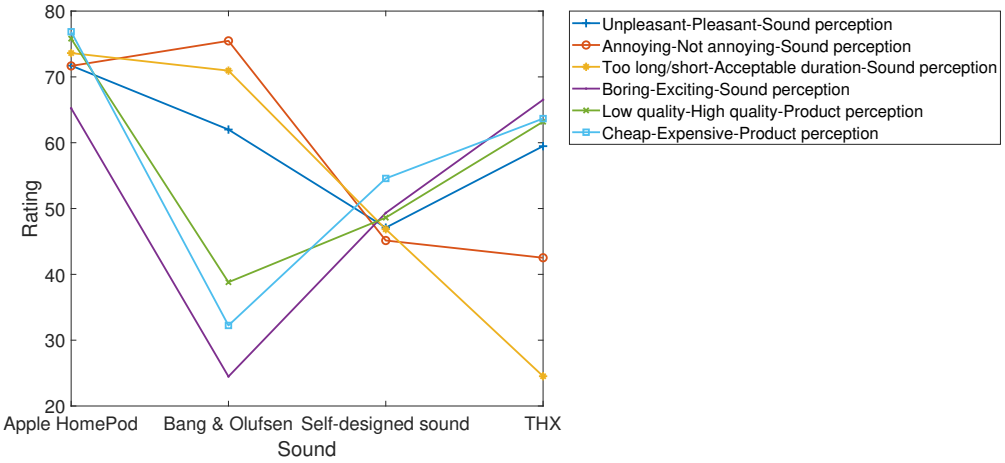


Figure 4.4: Averaged ratings for 14 observations

the ratings are scattered close together for the Apple HomePod and self-designed sound. For the Band & Olufsen sound, though, the attributes "annoying–not annoying", "too long/short–acceptable duration" and "unpleasant–pleasant" are located in the upper figure range, whereas "low quality–high quality", "cheap–expensive" and "boring–exciting" are located in the lower part of the figure. Same goes for the attributes "annoying–not annoying" and "too long/short–acceptable duration" for the THX sound which are positioned lower than the rest of the attributes. That means for example that an exciting sound not necessarily has to be pleasant.

A principal component (PC) analysis was executed to examine the underlying structure of the data. The result was that 58.95 % of the variance can be explained with one PC, 96.73 % with two PCs. From that, it can be concluded that participants used mostly two underlying factors to rate the sounds. Indeed, the attributes "too long/short–acceptable duration" and "annoying–not annoying" can be grouped together as well as "low quality–high quality", "cheap–expensive" and "boring–exciting" which can be seen in figure 4.5. The attribute "unpleasant–pleasant" stands alone. Based on the PC analysis one could argue that the data could be combined into two dependent variables with the assumption that they both influence the perception of the sounds. But because all attributes were deemed important for a possible calibration signal and described either a positive or negative user

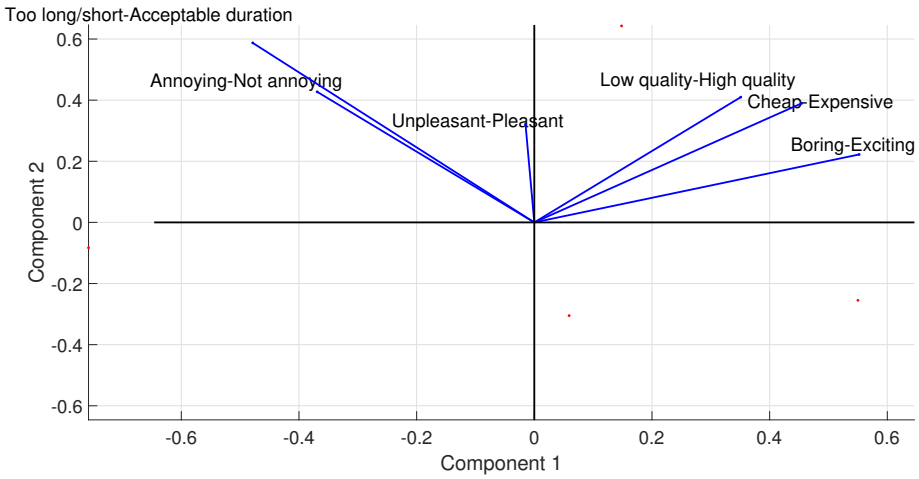


Figure 4.5: Bi-plot principal component analysis

experience, they were averaged for each participant and sound so that each participant contributed four ratings. Those ratings were concatenated for all participants.

It was checked if ratings for each sound were normally distributed (Chi-square goodness-of-fit test) and if they had equal variances (Bartlett test). This was the case. A box-plot of the data is depicted in figure 4.6. It can be noted that the

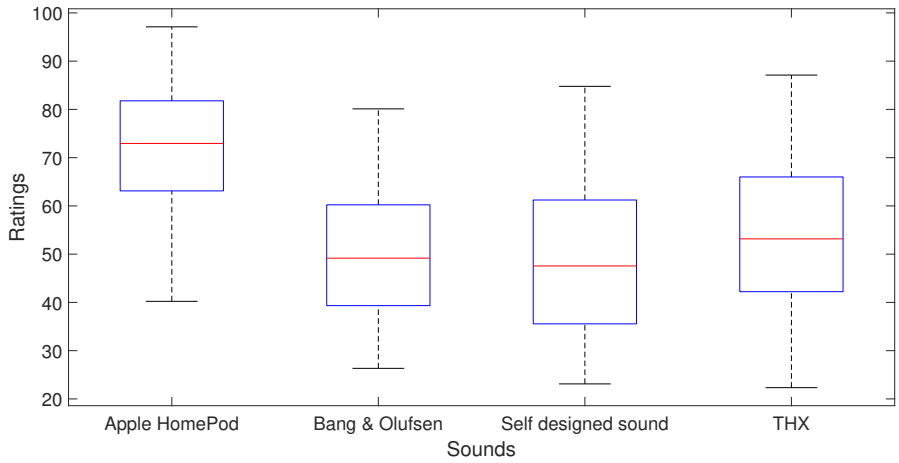


Figure 4.6: Box-plot for combined ratings

median of the Apple HomePod sound is the highest whereas all other medians are lower and not very different from each other.

A parametric one-way ANOVA was run to check if means from the different

sounds were different from each other. The results were df = 3, MS = 1688.78, F = 5.74 and p-value = 0.0018. That means that at least one mean of sound rating is different from other means. Consequently, a multi comparison test was conducted using the Bonferroni method to examine possible differences between single means. The conclusion is that the Apple HomePod sound is significantly higher rated than the rest of the sounds. No significant differences are between their means.

4.5 Summary

In this chapter, a test interface was designed to let 14 participants rate and compare four different kinds of calibration sounds using six pairs of attributes to describe the sound. It was elaborated how the experiment was conducted. Subsequently, the data was combined and analysed and results were presented. The Apple HomePod sound was significantly higher (better experience) rated than all other sounds.

The outcome of the conducted experiment was one possible calibration sound which was perceived as perceptually more pleasant than other possible calibration sounds. In the next chapter this sound, the Apple HomePod sound, will be tested against a traditional signal regarding its TOA estimation performance. With the end of this chapter, research question one in chapter 1 is answered.

Chapter 5

Signal Comparison

In the last chapter, the Apple HomePod start up sound was defined as a perceptually pleasant sound compared to others, which can be possibly used as a calibration signal for automated audio system calibration. Besides the pleasantness of a signal, it also needs to fulfil certain requirements in order to produce accurate enough results during the calibration process. Here, the requirements for audio system calibration are defined as TOA estimation from one source to one receiver in some kind of noise and possible reverberation. That means that the source (loudspeaker/smartspeaker) plays back the calibration signal which is recorded by the receiver. That could be the smart speaker itself, recording its own signal after it has been reflected off a wall or recording a signal from another loudspeaker. The TOA from source to receiver is then estimated. This chapter addresses research question two defined in chapter 1.

5.1 Testing Framework

TOA estimation in white Gaussian noise was already implemented in the testing framework in [1]. The maximum likelihood estimator was used. One could run a Monte-Carlo simulation where the noise was randomly added for every iteration. It was also possible to compare different signals for different SNRs.

The framework was extended in the following way: an interface was implemented which enabled the user to select the signals Apple HomePod start up sound, pseudo-random noise and sine sweep. The interface can be seen in figure 5.1. One, two or all of them could be selected so that they would be compared against each other.

Furthermore, one could add different noise types: traffic noise (recorded from the inside of an apartment), babble noise, impulse noise (from a restaurant) and

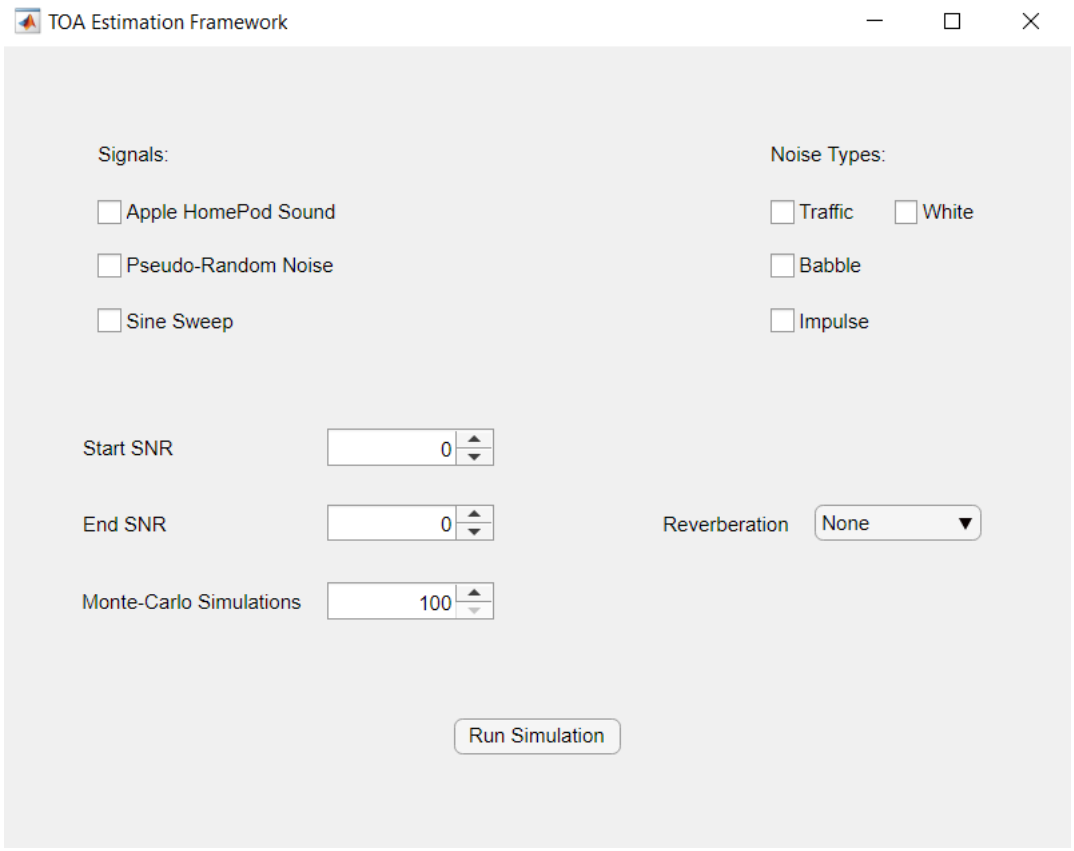


Figure 5.1: Testing framework interface TOA estimation performance

white noise. The noise samples were downloaded from `https://www.soundjay.com/index.html` and `https://freesound.org/`. Again, one could select one noise type, several or all of them which would result in a noise-mix.

It was also possible to define start and end SNRs and the number of Monte-Carlo simulations. The user could specify if reverberation should be applied to the tested signal-noise mix. This was possible in three different levels: low, moderate and high. Low reverberation corresponded to T60 times of 0.5 to 0.2 seconds in figure 5.2. It can be imagined as a damped living room with lots of carpet on the floor, curtains on the walls and furniture.

Moderate reverberation corresponded to T60 times of 1.1 to 0.7 seconds in figure 5.3. This can be imagined modern design living room with lots of glass all around and solid walls and furniture.

High reverberation corresponded to T60 times of 1.5 to 1.2 seconds in figure 5.4.

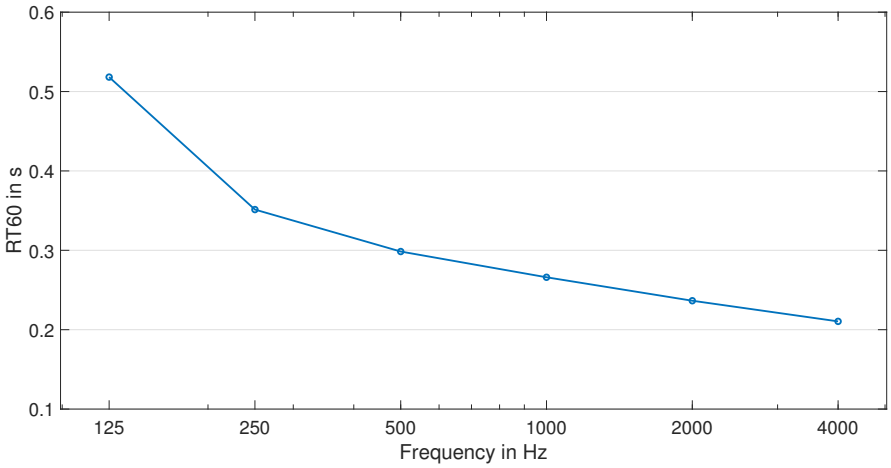


Figure 5.2: Reverberation times for a low-level reverberation room

This is a rather extreme case which could be an empty living room or a room with very few objects in it.

One could also select a custom reverberation of a specific room. When a reverberation level was selected RIRs were loaded and convolved with the signal and noise type. The RIRs were generated using MCRoomSim [22]. The room itself with source and receiver positions can be seen in appendix A.

After everything was selected, the user could start the calculations. The following pseudo code shows what was happening afterwards. All calculations were done in Matlab R2018a version: 9.4.0.813654.

1. Read noise type(s) and create combination of noise types, if more than one noise type
2. Normalise noise mix
3. Convolve with RIR if selected
4. For all signals to be compared
    a) Read signals and normalise
    b) Delay signals or convolve (convolution already delays the signal)
    c) Calculate signal power
    d) For all SNRs
        i. Calculate resulting noise power

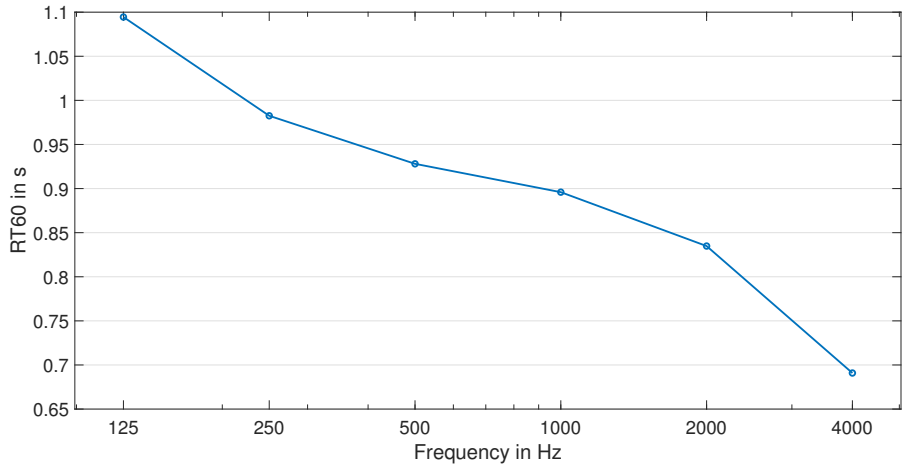


Figure 5.3: Reverberation times for a moderate-level reverberation room

ii. For number of Monte-Carlo simulations
   j. Randomly delay noise mix
   jj. Create signal/noise mix
   jjj. Estimate TOA

e) Calculate root mean square error (RMSE) between true delay and TOA estimations

5. Plot results

5.1.1 Testing Results

The perceptually most pleasant sound from chapter 4, the Apple HomePod start up sound, was tested against the pseudo-random noise as a representative for a traditional signal for RIR measurement and loudspeaker localisation. The tests were performed for SNRs -10 to 15 dB and 1000 Monte-Carlo simulations. The noise types traffic noise, babble noise, impulse noise and a combination of those noise types were tested. White noise was not tested because it was deemed as an unlikely scenario. For each variation of noise type none, low, moderate and high reverberation was added.

All test results can be seen in appendix B. In the following, the most important findings are discussed and corresponding figures are shown. The RMSE from the TOA estimation was transformed into meters using the sampling frequency of 44,100 Hz and a speed of sound of 343.21 m/s.

First of all, it became clear that the performance of the pseudo-random noise in

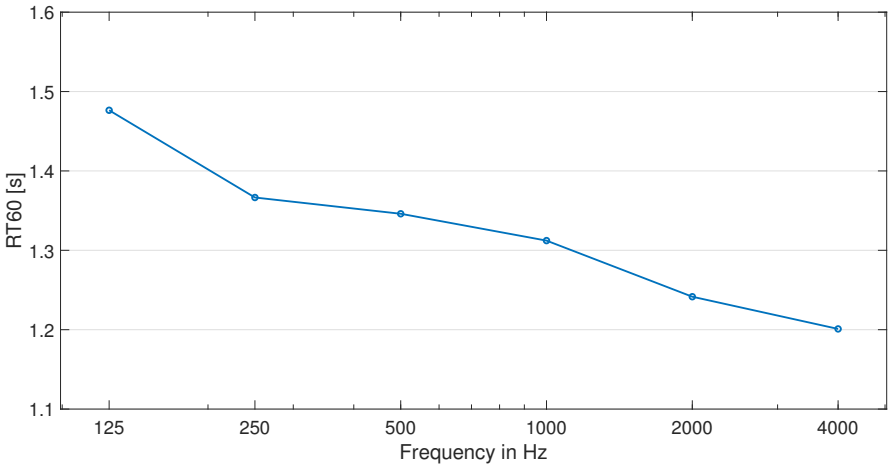


Figure 5.4: Reverberation times for a high-level reverberation room

terms of TOA estimation is better (lower RMSE) than for the Apple HomePod start up sound for all tested situations. This was to be expected. It can be seen in figure 5.5. For example, the circled curve for the Apple sound is above (higher RMSE)

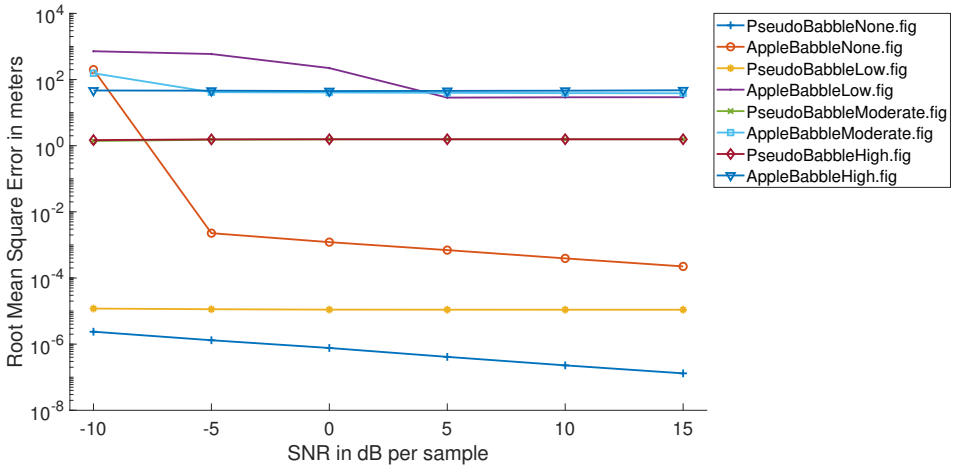


Figure 5.5: TOA estimation babble noise for no, low, moderate and high reverberation

the blue crossed curve for the pseudo-random noise (both no reverberation). Especially large is the difference between the pointed purple curve (Apple) and the star yellow curve (noise) for low reverberation. But also the curves for moderate and high reverberation for the pseudo-random noise are below the ones from the apple sound. Furthermore, it can be seen that only Apple no reverberation and pseudo-random noise no and low reverberation have an acceptable RMSE (below 1 cm for the Apple sound and below 0.1 mm for the pseudo-random noise). All

other curves lie above 1 meter which was defined as a failed TOA estimation.

It can also seen from figure 5.5 that the RMSE from both Apple sound and pseudo-random noise increases with reverberation. A large increase in RMSE can be noted for the pseudo-random noise from low reverberation to higher reverberation. There is not much difference any more from moderate to higher reverberation. Also the Apple sound is worsening from no reverberation at all to reverberation. What is weird here, is that the RMSE is only lower for the Apple sound with low reverberation from 5 dB on and higher. Below that, the RMSE is higher for low reverberation than for moderate and high reverberation. To investigate this further, another calculation was run were only the Apple sound was tested for 5 to 30 dB and again 1000 Monte-Carlo simulations. In figure 5.6 the situation is depicted were the noise was a combination of all noise types for low, moderate and high reverberation. It

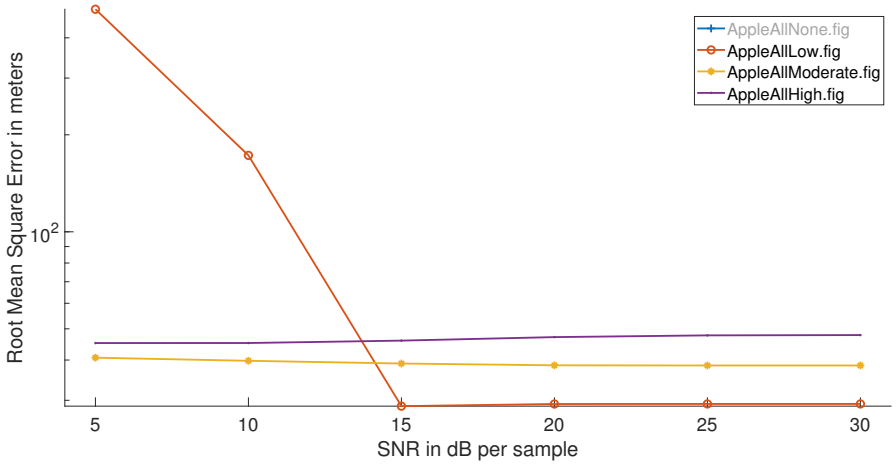


Figure 5.6: TOA estimation for a combination of all noise types and low, moderate and high reverberation (higher SNRs)

becomes clear that after a threshold SNR the RMSE is increasing in order with low, moderate and high reverberation. This is also true for every other noise type with the difference that the threshold SNR varies.

Another observation is that the curves converge when reverberation is present. With no reverberation the Apple sound and pseudo-random noise RMSEs decrease continuously.

In the next figure 5.7, no reverberation and RMSEs of Apple sound and pseudo-random noise for different noise types is depicted. It can be seen that the noise type has some influence on the RMSEs but does not change the overall structure. It changes the SNR of the Apple sound, though, from that on the RMSE is accept-

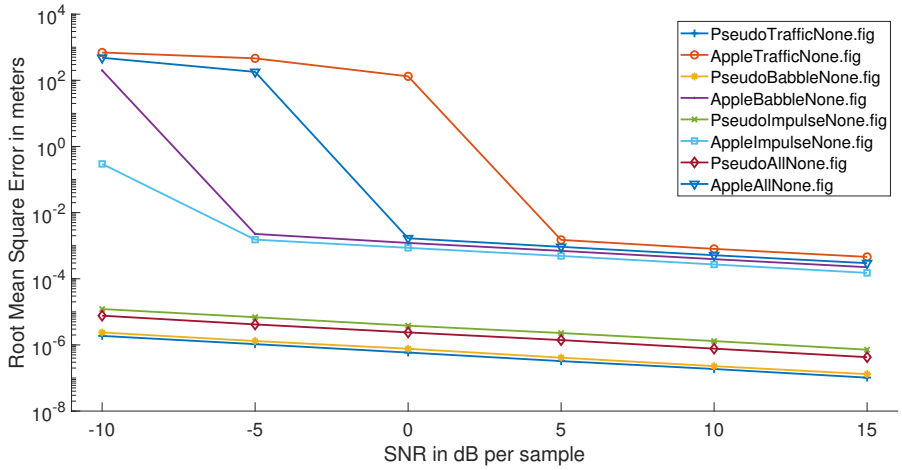


Figure 5.7: TOA estimation for all noise types and no reverberation

able. For example, with babble and impulse noise, the RMSE of the Apple sound is below 1 cm at -5 dB and higher and therefore acceptable whereas with with traffic noise, it is acceptable at 5 dB and higher. The RMSE of the pseudo-random noise is always acceptable.

The overall structure also does not change with the noise type when reverberation is present seen in figure 5.8 for moderate reverberation. But again because

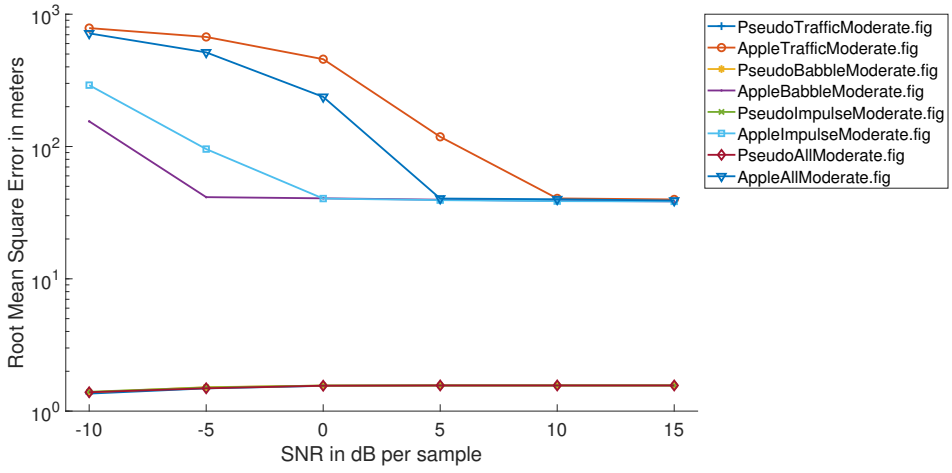


Figure 5.8: TOA estimation for all noise types and moderate reverberation

of the reverberation the performance of all signals breaks down and lies above 1 meter which is too much. This is actually explainable because the maximum likelihood TOA estimator assumes a free-field.

An observation which is against the expectation is that the RMSE increases with an increase of the SNR for pseudo-random noise with moderate and high reverberation regardless of the noise type. At some point it converges, though. This can be best seen in figure 5.9 for moderate reverberation. This behaviour of increasing

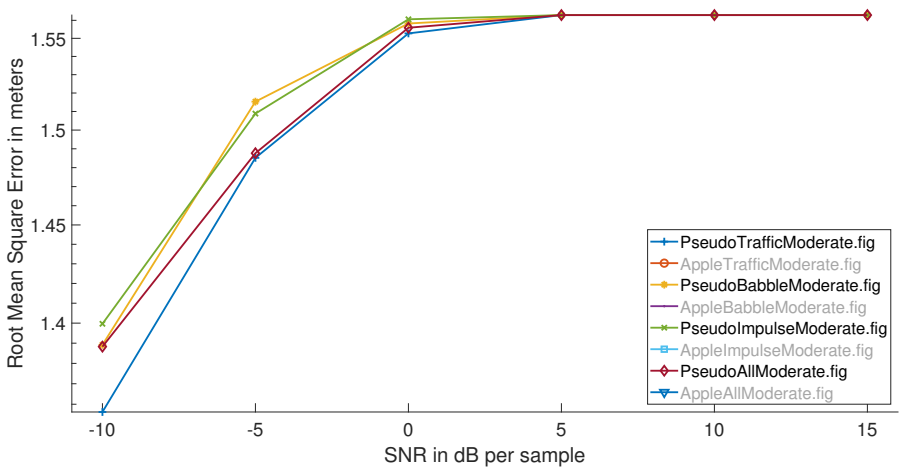


Figure 5.9: TOA estimation of pseudo-random noise for all noise types and moderate reverberation

RMSE can only slightly be observed for the Apple sound depicted in figure 5.10 for high low reverberation. However, if one looks at higher SNRs like in figure 5.11

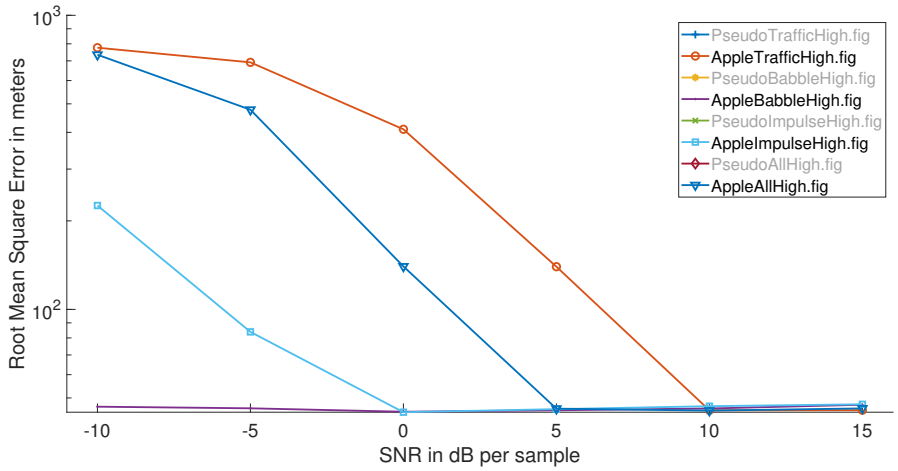


Figure 5.10: TOA estimation of Apple HomePod start-up sound for all noise types and high reverberation

the same behaviour as for the pseudo-random noise is visible. Also, the SNR of the signals converge does not change with the noise types after a high enough SNR,

probably because the noise loses its influence. Moreover, it became once more clear that an RMSE of around 48 meters can not be used as the simulated room is not even that big. Again, all results for the investigation for the Apple start-up sound

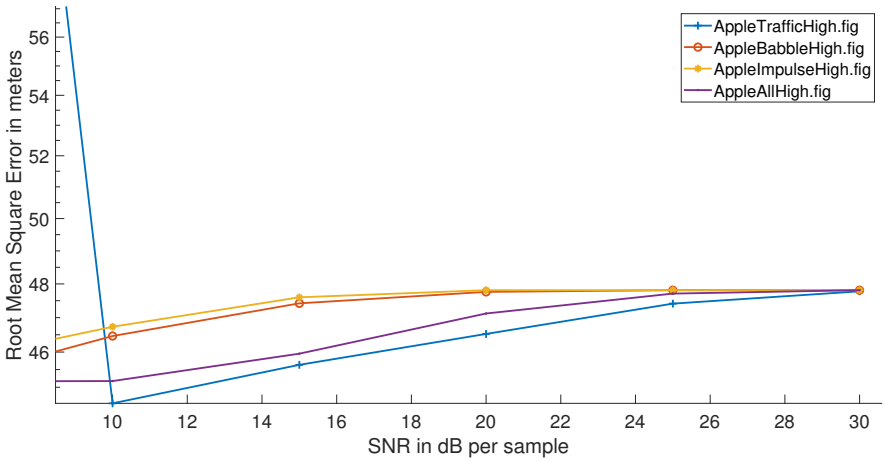


Figure 5.11: TOA estimation of Apple HomePod start-up sound for all noise types and high reverberation (higher SNRs)

for higher SNRs can be found in appendix B.

5.2 Conclusion

One take-away is that the performance of the Apple sound in terms of TOA estimation is worse than the one from the pseudo-random noise. (The performance of the sine sweep was not measured but will probably be similar.) This was expected. On the other hand, the Apple sound is perceptually pleasant whereas the pseudo-random noise is not. The difference in performance between the signals increases when reverberation is added but is not so much dependent on the corrupting background noise type.

Another take-away is that the TOA estimation is only useful in situations where no or low reverberation is present as those were the situations where the performance of the pseudo-random noise was acceptable. (For no reverberation the Apple sound also had an acceptable performance, depending on the SNR.) The noise type was not crucial. Assuming that the Apple sound can be modified so that its performance is similar to the one from the pseudo-random noise, those are the situations were it has to be tested. Another way to increase the performance of the Apple sound would be to adapt the TOA estimator but it was decided to concentrate on modifying the signal at this point. Doing so, the TOA estimator assumption can

be violated in two ways: a different noise than white noise is present or instead of a free-field, reverberation is added. In figures 5.12 and 5.13 the RMSEs of the Apple HomePod start up sound and pseudo-random noise are depicted for the two estimator violations: either having white noise and no, low reverberation (figure 5.12) or white noise and a combination of traffic, babble and impulse noise and no reverberation (figure 5.13). Basically, the conditions are as expected. Rever-

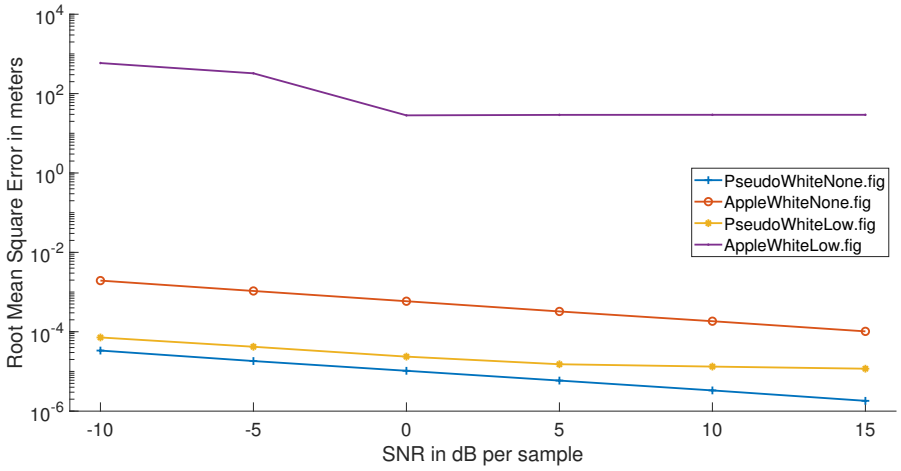


Figure 5.12: TOA estimation with white noise and no, low reverberation

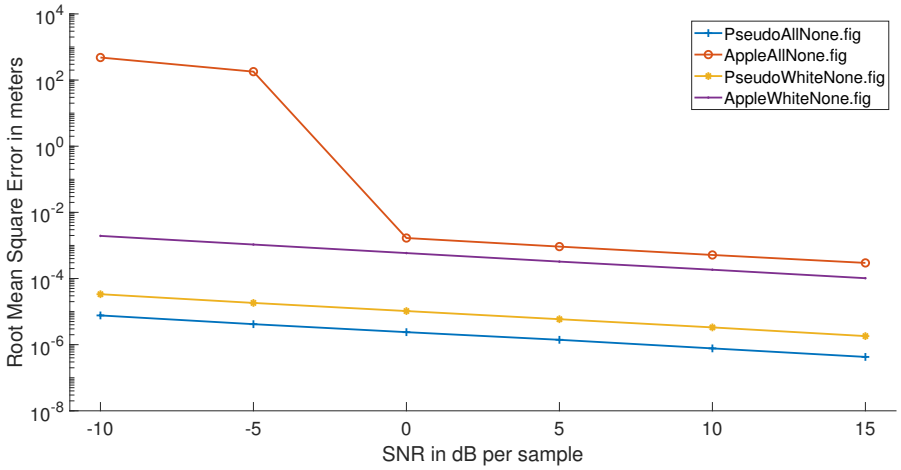


Figure 5.13: TOA estimation with white noise and a combination of noise types and no reverberation

beration increases the RMSE both for the pseudo-random noise and Apple sound, resulting in the Apple sound not being appropriate enough for TOA estimation. The combination of noise types increases the RMSE contradicting to white noise for the Apple sound but not for the pseudo-random noise. Here, the RMSE for the

noise type combination is lower. This was unexpected but in both cases the TOA estimation is accurate enough.

In figure 5.14 the estimator assumptions are violated in both ways: a combination of traffic, babble and impulse noise and low reverberation. Here, the pseudo-random noise is suitable for TOA estimation whereas the Apple HomePod start up sound is not. Those upper mentioned three situations will be the situations where

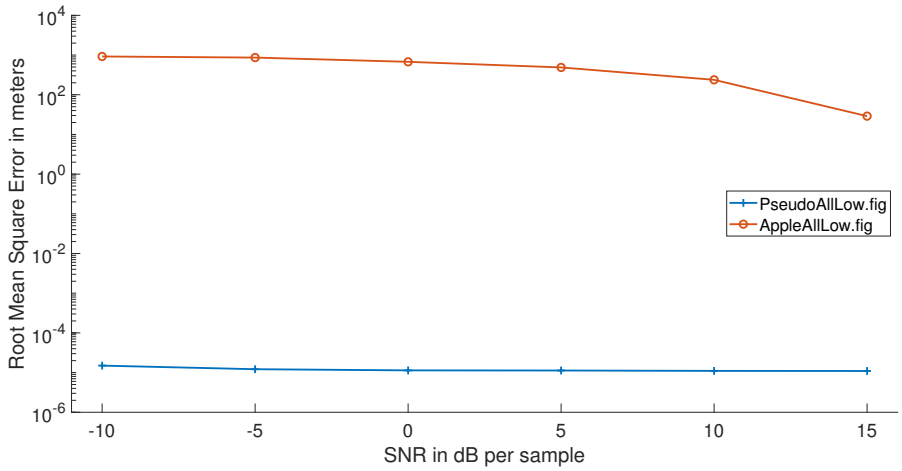


Figure 5.14: TOA estimation for a combination of traffic, babble and impulse noise and low reverberation

a modification of the Apple sound will be tested against the pseudo-random noise.

5.3 Summary

In this chapter, a testing framework was described and used for comparing Apple HomePod start up sound from the last chapter with the pseudo-random noise for different testing situations. The performance of the pseudo-random noise is always better than of the Apple sound. Three situations were defined where testing modifications of the Apple sound, to increase its performance, makes sense:

- white noise present and no, low reverberation
- no reverberation present and whit noise and a combination of traffic, babble and impulse noise
- a combination of traffic, babble and impulse noise and low reverberation present

The next task is now to modify the Apple sound so that its performance converges to the one from the pseudo-random noise in the situations explained above. Before

doing that, though, real-life measurements are made to check if their results will be similar to the ones from a simulation in a similar situation within the testing framework, thus determining if the testing framework delivers valid results. This will be the topic in the next chapter.

Chapter 6

Real-Life Measurements

In the last chapter, it was concluded through a virtual comparison that the Apple HomePod start up sound is performing worse than the pseudo-random noise. Furthermore, situations were defined where testing the Apple sound after modification, in order to increase its performance but also remain its pleasantness, against the pseudo-random noise makes sense. Before modifying the Apple sound, it has to be verified that the results from the theoretical testing framework used in chapter 5 can be reproduced in real-life measurements or the other way around. Therefore, real-life measurements were conducted and based on those the performance of the Apple sound and pseudo-random noise in terms of TOA estimation was tested. Consecutively, a similar environment like the one used for real-life measurements was simulated in the testing framework and the performance of the above mentioned signals was tested as well. The results were compared.

6.1 Real-Life Measurements Set Up

6.1.1 Recordings

The measurements were conducted in the sound lab in the CREATE building of Aalborg University. Sound was played back through in single bioacoustics dynamics loudspeaker and recorded through a Behringer ECM 8000 omnidirectional microphone. The distance between them amounted to 2.1 meters, the same as for the simulation in the testing framework. It has to be noted that the exact point of emission of the loudspeaker was not defined. Loudspeaker and microphone were connected to a PC through the PreSonus AudioBox 1818 VSL. A photo of the testing environment can be seen in figure 6.1. Previous to the test, audio files were created with signal noise mixes. The signals were the Apple HomePod start up sound and the pseudo-random noise. Based on the conclusion in chapter 5 that the noise type has not much influence on the performance of the signal, it was

Figure 6.1: Testing environment for real-life measurements

decided to use a combination of all noise types (traffic, babble and impulse noise) and white noise. The SNRs were chosen to be 10 and 15 dB to reach a state were the RMSE of the signals has already converged (also based on chapter 5). For each signal noise mix, ten audio files were produced with randomly delayed noise. In total 80 audio files were played back and recorded.

The measurements were conducted with the ITA-toolbox [7] in Matlab. It was played back and recorded simultaneously. Prior to the measurements, the delay of the PreSonus sound card was measured to see the exact delay according to the distance from the microphone to the loudspeaker in the recordings. The delay of the loudspeaker was not measured but an examination of the RIR matched the measured distance. Unfortunately, the created audio files were periodically delayed as well so that the delay of the recorded signal resulted in two times the distance, that is 4.2 meters. This was post-corrected.

6.1.2 Results

The recordings and original signals (Apple sound, pseudo-random noise) were used to estimate the TOA from loudspeaker to the microphone which can be used to calculate the distance. The was the defined measure of performance of the signal. Since the true distance and delay was known the RMSE was calculated for every of the ten different signal noise mixes for every different situation. The results are depicted in figure 6.2. The RMSE was transformed into distance using the sampling frequency of 44,100 Hz and a speed of sound of 343.21 m/s. It can be

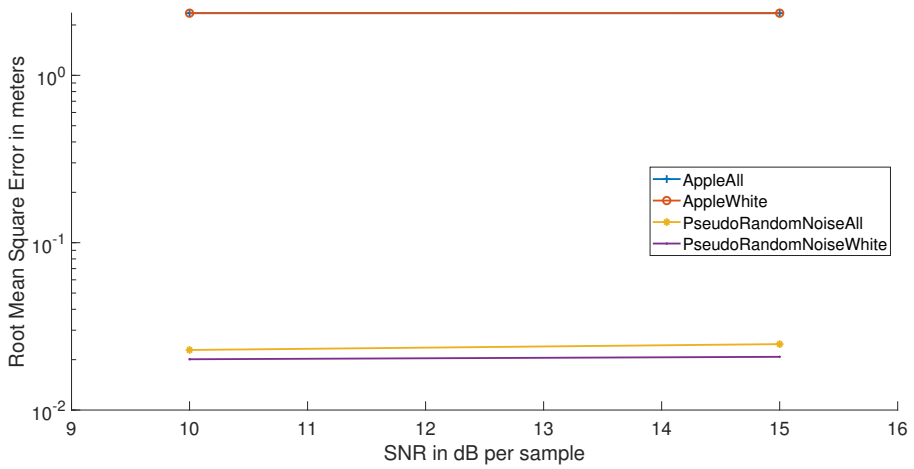


Figure 6.2: TOA estimation for real-life measurements

seen that the RMSE of the pseudo-random noise for white noise only and a combination of traffic, babble and impulse noise is lower than for the Apple HomePod start up sound for the same noise types. That is corresponding to previous results.

In figure 6.3 only the results of the pseudo-random noise is plotted. The RMSE ranges from around 2.1 cm to 2.5 cm. The RMSE for white noise is lower than for a combination of noise types. In both cases the RMSE is increasing with increasing SNR which was also observed in previous results. The RMSE was deemed still acceptable. The RMSE for the Apple sound lies around 2.3 meters. White noise and combination of noise types are barely different from each other. This can be seen in figure 6.4. This RMSE is not acceptable any more.

6.2 Testing Framework set up

In order to simulate an environment which is similar to the sound lab where the real-life measurements were conducted, the RIR of the sound lab had to be measured. The same set up was used as for the recordings, although with a different

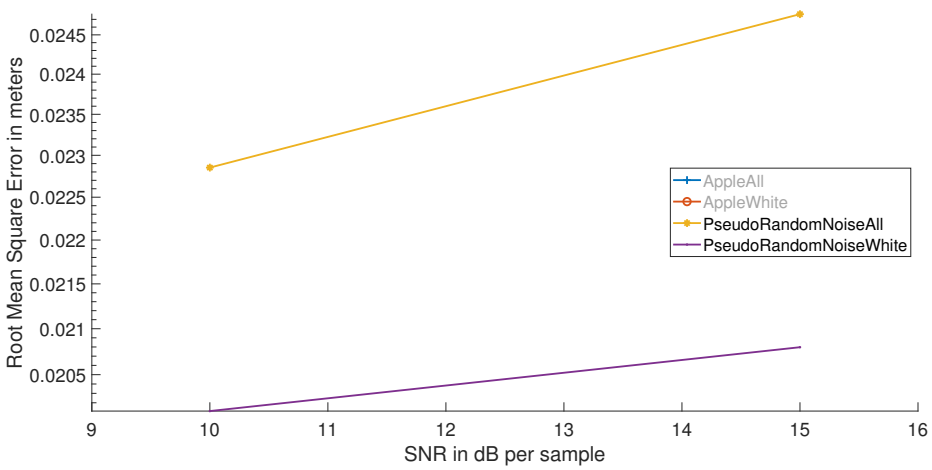


Figure 6.3: TOA estimation for real-life measurements, only pseudo-random noise

microphone: the G.R.A.S 40PP CCP Free-Field QC microphone. The RIR was also measured using the ITA-toolbox [7] using an exponential sine sweep ranging from 20 - 22,050 Hz. It was checked that the delay in the RIR matched the distance 2.1 meters from microphone to loudspeaker.

Later on, RIR measurements were again conducted with multiple loudspeaker and microphone positions following the standard [13]. The measurement process is described in detail in appendix C. One of the newly measured RIR was used to calculate the early decay times of the sound lab. Figure 6.5 depicts them. The early decay times from the sound lab are relatively low. At around 50 Hz is the highest time of around 0.7 seconds. with higher frequencies the times decrease with around 0.3 seconds from 1000 Hz on. The times are comparable with the low reverberation room simulation in figure 5.2. Here, the RT60 times ranged 0.5 seconds for lower frequencies (125 Hz) and 0.2 seconds for higher frequencies (4000 Hz). This is an argument for testing signal modifications in the low reverberation case in the testing framework because it the results could be repeated using real-life measurements in the sound lab. Furthermore, TOA estimation works in both situations with the pseudo-random noise.

After obtaining the RIR from the first measurement, it was used in the testing framework described in chapter 5. The signals Apple HomePod start up sound and pseudo-random noise were mixed with both a combination of traffic, babble and impulse noise and white noise from 10 to 15 dB and 1000 Monte-Carlo simulations. The measured RIR was loaded. Different signal-noise mixes were generated as for the real-life measurements.

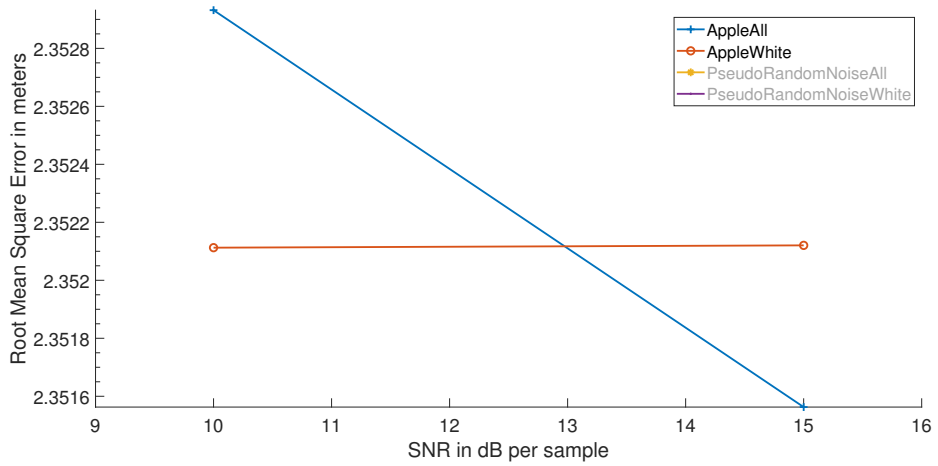


Figure 6.4: TOA estimation for real-life measurements, only Apple HomePod start up sound

6.2.1 Results

Figure 6.6 shows the results from the testing framework simulating the real-life measurements environment. Figure 6.6 looks similar to figure 6.2. The RMSE of the pseudo-random noise is lower than the RMSE of the Apple HomePod start up sound. The noise type did not change the RMSE a lot. Following the same procedure as before, only the RMSEs of the pseudo-random noise and Apple sound ins plotted, starting with the pseudo-random noise in figure 6.7. The RMSEs of the two noise types lie very close together, around 8.5 mm. The RMSE of the real-life measurements ranged from 2.1 cm to 2.5 cm. Contradicting to the combination of noise types for the real-life measurements, the RMSE for the pseudo-random noise for the testing framework decreased with increasing SNR even though with a very small amount.

The RMSEs for the testing framework and Apple sound lie around 2.4 meters, seen in figure 6.8. The noise types are marginal different from each other. The RMSEs from the real-life measurements lied around 2.3 meters. Again, the performance of the pseudo-random noise is acceptable whereas the performance of the Apple sound is not.

6.3 Comparison between real-life measurements and testing framework

The following table 6.1 lists the RMSE differences in centimetres between real-life measurements and the testing framework. It can be seen that the highest differ-

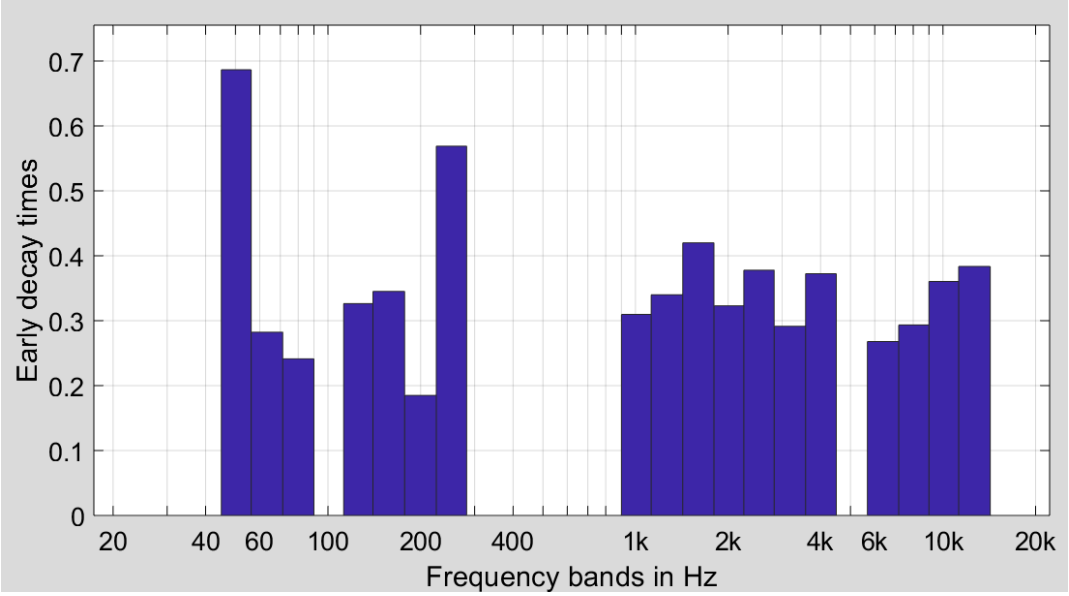


Figure 6.5: Early decay times sound lab Aalborg University

Table 6.1: RMSE differences in centimetres between real-life measurements and testing framework

	Combined Noise Types		White Noise	
	10 dB	15 dB	10 dB	15 dB
Apple HomePod start-up sound	2.9	3.0	3.0	3.0
Pseudo-random Noise	1.4	1.6	1.2	1.2

ence is 3.0 centimetres. That error is reasonably small and could be, for instance, caused by inaccuracies in measuring distances in the measurement environment, unknown location of the exact point of emission of the loudspeaker or unknown loudspeaker delay.

6.4 Summary

In this chapter, the performance of the Apple HomePod start up sound and the pseudo-random noise were compared both through real-life measurements and the testing framework from chapter 5, simulating the measurement environment. The results were similar with a maximal difference of 3.0 cm. The error was deemed small enough to conclude that the testing framework works accurate enough so that it can be used to test modified versions of the Apple sound with the goal to increase its performance. This is done in the next chapter. With the results

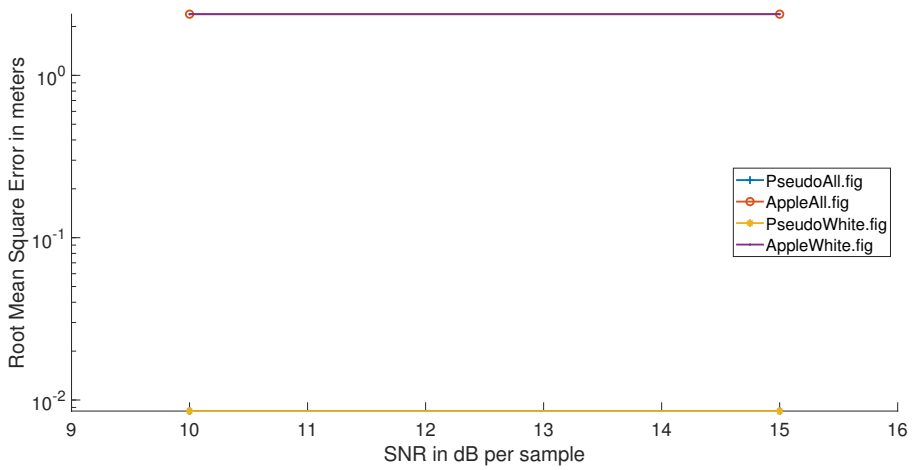


Figure 6.6: TOA estimation using the testing framework simulating the real-life measurements environment

in chapter 5 and the validation in this chapter, research question two form chapter 1 is answered, whereas it was not investigated how signals cope with non-ideal loudspeakers and microphones.

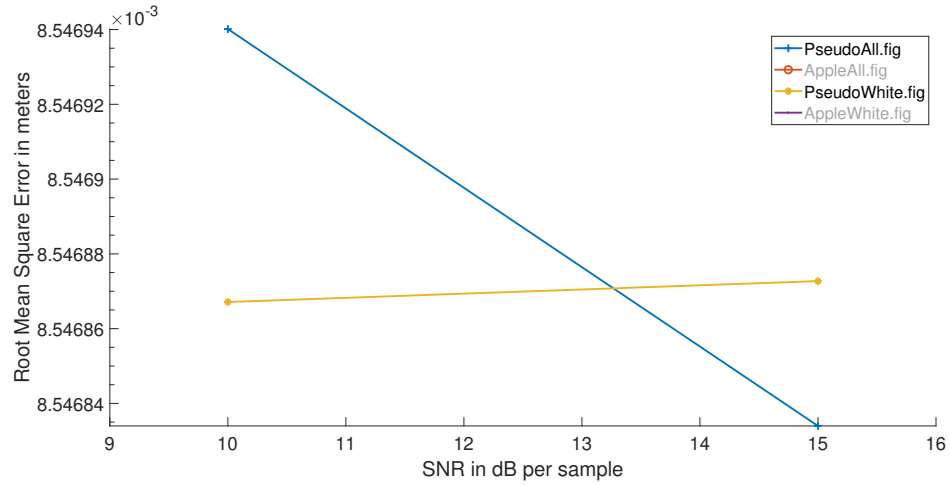


Figure 6.7: TOA estimation using the testing framework simulating the real-life measurements environment, only pseudo-random noise

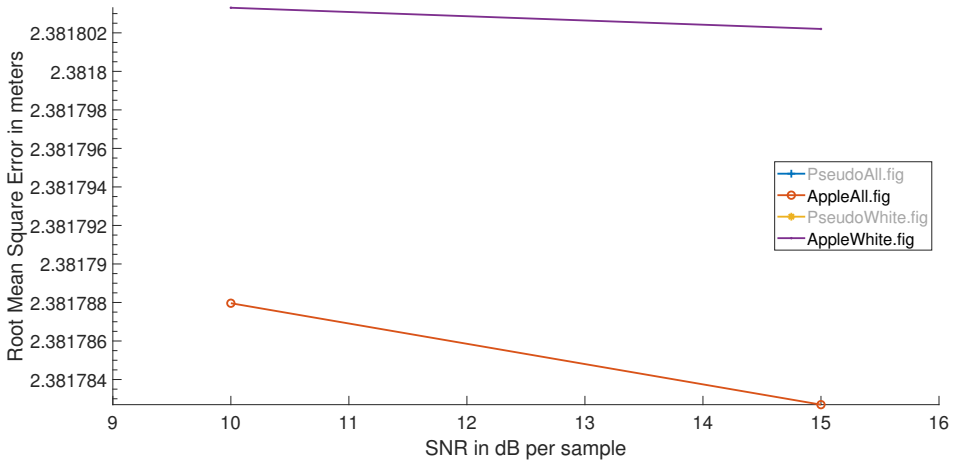


Figure 6.8: TOA estimation using the testing framework simulating the real-life measurements environment, only Apple HomePod start up sound

Chapter 7

Signal Modifications

After finding a pleasant signal, compared to others, which can possibly be used for automated audio system calibration, testing its performance in terms of TOA estimation against a traditional signal within a testing framework and conducting real-life measurements to ensure the validity of the testing framework results, this chapter addresses possible methods to modify the Apple HomePod start up sound. Doing this, research question three from chapter 1 will be in focus. The modification became necessary because the Apple sound did not deliver an acceptable accuracy in TOA estimation from one source to one receiver, which was defined to be the simplest case for automated audio system calibration. The performance of the traditional signal (pseudo-random noise) was always better, especially in the test scenario where reverberation is present. Only in the case where white noise with no reverberation was present, the performance of the Apple sound was acceptable throughout all tested SNRs. But also here, the pseudo-random noise outperformed the Apple sound by far.

In the following, different techniques are applied to the Apple sound. The modified Apple sound is then tested with the testing framework in the defined testing scenarios in chapter 5, to see how the modification affected its performance in comparison to the original Apple sound and pseudo-random noise. Best case scenario would be if the modification would cause a performance which is equal or close to the one from the pseudo-random noise with no or less loss in perceptual audio quality.

7.1 Spectral Envelope Estimation

The idea behind spectral envelope estimation is that the spectral envelope of a well performing signal in terms of TOA estimation is estimated and applied to the Apple sound. (It was also tried to estimate the spectral envelope of the Apple sound

and apply it to pseudo-random noise but that made the Apple sound unrecognisable) The assumption was that spectral characteristics of a well performing signal, e.g. frequency content, could be transferred to the Apple sound so that its performance would increase but still remain its pleasantness with a certain trade-off.

The following pseudo code explains how the spectral envelope estimation and application was implemented in Matlab. The estimation was based on the iterative cepstrum/true envelope, [2, 10, 20].

1. Load audio signal which shall be used for the spectral envelope (extract-signal)
2. Load audio signal where the spectral envelope should be applied to (apply-signal)
3. Calculate the short time Fourier transform (STFT) of the apply-signal
4. Estimate the spectral envelope of the whole extract-signal length using 50 iterations
5. Apply the estimated spectral envelope to each segment of the apply-signal
6. Reconstruct the apply-signal using the overlap add method

Firstly, the spectral envelope of the pseudo-random noise was estimated and applied to the Apple sound. A first test with just 100 Monte-Carlo iterations within the testing framework resulted in a bad performance, though. Consecutively, the spectral envelope of a recording of the author's unvoiced speech (ZHH-sound) was used because its performance was promising in the preceding semester project [1]. The results are shown in the next figures. Figure 7.1 depicts the results of the first test scenario (white noise and no, low reverberation present). It can be seen that the application of the spectral envelope almost did not change anything for the white noise, no reverberation case compared to the original sound. For the white noise, low reverberation case, the RMSE was decreased after 0 dB SNR. It has to be noted though, that there is still an error of approximately one meter for the modified Apple sound for 15 dB SNR, thus the modified signal is not suitable for TOA estimation and none of the signals certainly reach the performance of the pseudo-random noise.

Figure 7.2 depicts the second test scenario (no reverberation, white noise and a combination of noise types). Here, the application of the spectral envelope even worsened the performance of the modified Apple sound for the combination of noise types case as in that it reaches the performance of the original Apple sound at 10 dB SNR, contradicting to 0 dB SNR before. At those SNRs the performance is

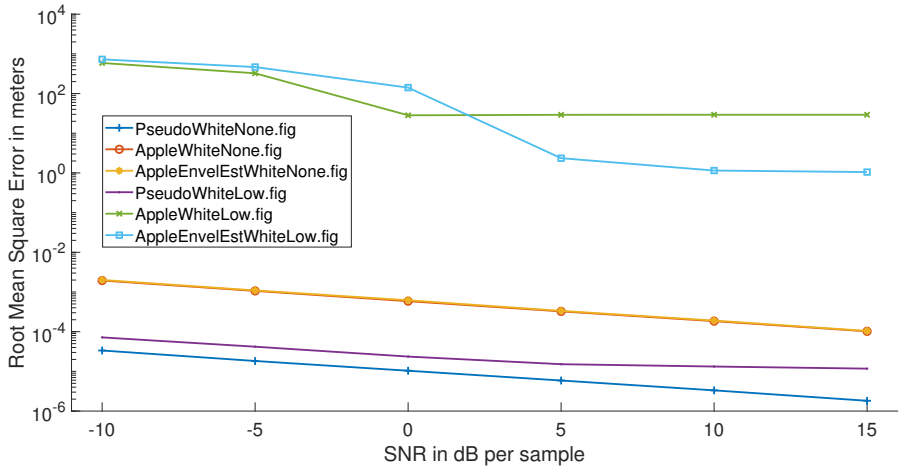


Figure 7.1: TOA estimation, first test scenario, spectral envelope estimation

acceptable with less than a centimetre, but as mentioned it was already acceptable with the original Apple sound.

The third test scenario was the case when a combination of noise types together with low reverberation is present. This is visualised in figure 7.3. The RMSE of the modified Apple sound is higher than the original Apple sound, thus no improvement.

The conclusion of the spectral envelope estimation method is that the modified Apple sound could not be improved to an acceptable performance.

7.2 Audio Coding

The principle of a lossy audio coding algorithm is that the size of the audio file is reduced with no perceptual audible decrease in audio quality. In order to achieve that, the masking curves of an audio file are computed segment wise. Then the bit rate of the file is reduced which causes increased quantisation noise. This quantisation noise is allocated to areas below the masking threshold so that it is not audible. The lossy audio coding principle was selected as a method to increase the TOA estimation performance of the Apple HomePod start up sound, because although the increased quantisation noise is not audible it will still be detected by the microphone and hopefully increase TOA estimation. This assumption is made because it was found in [1] that frequency content, especially high frequency content helps the maximum likelihood estimator when no violations are made. This was concluded based on the Cramer-Rao lower bound for that estimator. The MPEG 1

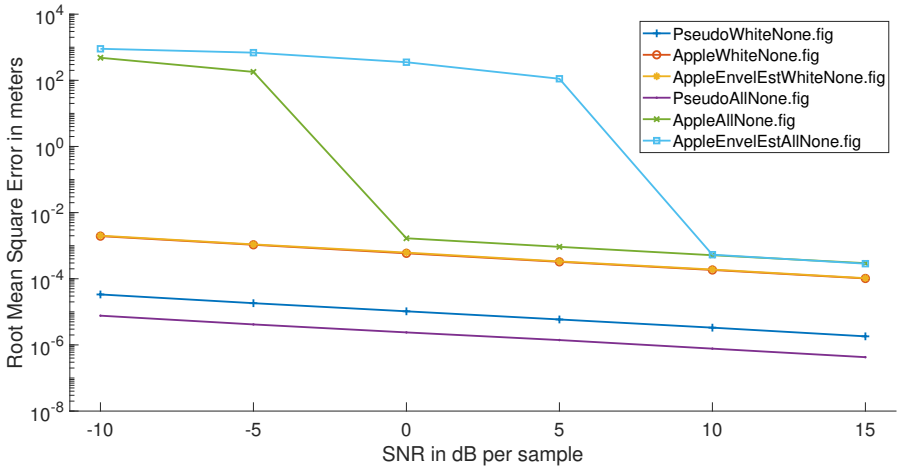


Figure 7.2: TOA estimation, second test scenario, spectral envelope estimation

layer 3 (mp3) [4] algorithm was used for audio coding. The bit rate was set to 32 kbits/second, fast encoding with 44,100 Hz sampling frequency and no lowpass filtering. A Matlab implementation from Dan Ellis (`https://se.mathworks.com/matlabcentral/fileexchange/13852-mp3read-and-mp3write`) was used to write the mp3-file from the original wav-file using the lame encoder.

The modified mp3 version of the original Apple sound was again tested in the three test scenarios. The results a depicted in the next figures. Figure 7.4 shows the test scenario one where white noise and no, low reverberation is present. It can be seen that the audio coding neither gives an improvement for the no reverberation case, nor for the low reverberation case. The RMSEs of the original and modified Apple sounds are similar. The RMSEs of the pseudo-random noise is much lower. Same goes for the second test scenario in figure 7.5. And again in the third test scenario (figure 7.6) no significant changes of the RMSEs could be found.

In can be concluded that the audio coding approach did not improve/change the performance of the Apple HomePod start up sound.

7.3 Masking Curves

Masking curves are computed during the mp3-algorithm in order to know at which frequencies and how much quantisation noise can be added to reduce the audio file. The only control one has over the amount of quantisation noise to be added is the resulting bit rate of the mp3-file. In order to have direct control of how much and what type of signal shall be added to the original signal, the masking curves

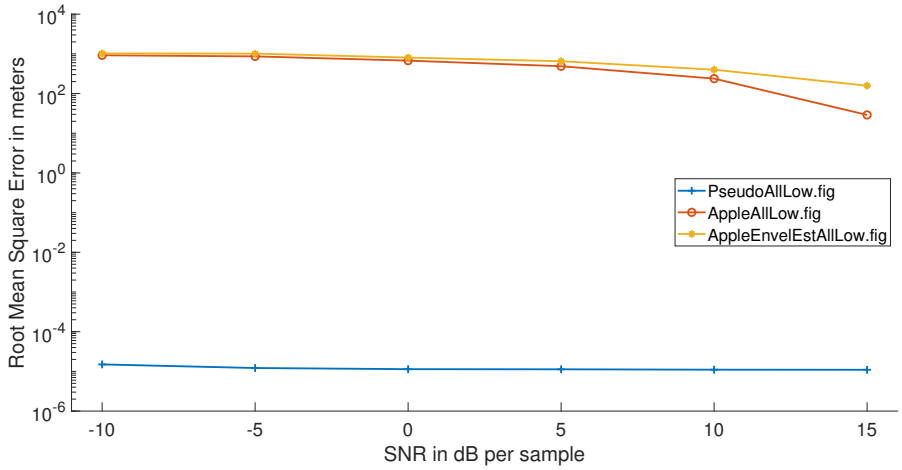


Figure 7.3: TOA estimation, third test scenario, spectral envelope estimation

were computed based on the psychoacoustic model of the MPEG 1 algorithm [4]. The goal was to add as much of the well performing pseudo-random noise to the Apple sound to increase its performance but still keep the noise inaudible or slightly audible.

7.3.1 Psychoacoustic Model

In the following, the MPEG 1 psychoacoustic model is described based on [21]. The algorithm is divided into five steps. Firstly, the signal is normalised and the STFT is performed. The normalisation is performed so that spectral components can be expressed in sound pressure level (SPL). The normalisation guarantees that a 4 kHz signal of +/- 1 bit amplitude is associated with an SPL near 0 dB which is close to the hearing threshold. If one has a full scale signal, for instance a sinusoid, it would result in an SPL near 90 dB.

The incoming signals are normalised according to the discrete Fourier transform (DFT) length N and the number of bits per sample b following equation:

$$x(n) = \frac{s(n)}{N(2^{b-1})} \tag{7.1}$$

Subsequently, the power spectral density (PSD) as part of the STFT is calculated with a certain hanning window size, DFT-length, and overlap with equation

$$P(k) = PN + 10\lg\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2\pi kn}{N}}\right|^2,\ 0 \leq k \leq \frac{N}{2} \tag{7.2}$$

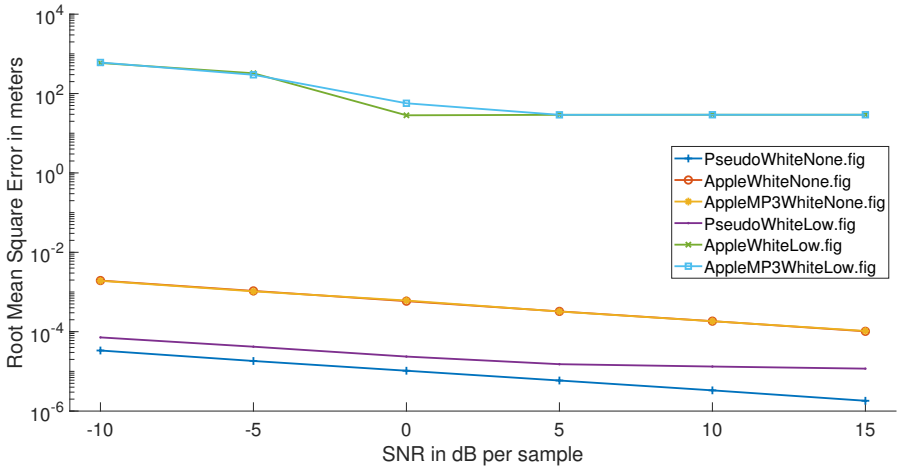


Figure 7.4: TOA estimation, first test scenario, audio coding

The power normalisation term PN is fixed at 90.302 dB.

The second step is the identification of tonal and noise maskers. Tonal maskers have a local maxima in the PSD that exceed neighbouring components within a certain Bark distance by at least 7 dB. The positions are defined as

$$S_T = \left\{ P(k) \left| \begin{array}{c} P(k) > P(k \pm 1), \\ P(k) > P(k \pm \Delta_k) + 7 \text{ dB} \end{array} \right. \right\} \tag{7.3}$$

where

$$\Delta_k \in \begin{cases} 2 & 2 < k < 63 \quad (0.17 - 5.5 \text{ kHz}) \\ [2,3] & 63 \leq k < 127 \quad (5.5 - 11 \text{ kHz}) \\ [2,6] & 127 \leq k \leq 256 \quad (11 - 20 \text{ kHz}) \end{cases} \tag{7.4}$$

Tonal maskers $P_{TM}(k)$ are then computed from the spectral peaks listed in S_T as follows

$$P_{\text{TM}}(k) = 10 \lg \sum_{j=-1}^{1} 10^{0.1P(k+j)} \text{ (dB)} \tag{7.5}$$

In words, for each neighbourhood maximum, energy from three adjacent spectral components centred at the peak are combined to form a single tonal masker. Single noise maskers are computed for each critical band and from the remaining spectral lines not used for tonal masker computation. They are defined as

$$P_{\text{NM}}(\bar{k}) = 10 \lg \sum_{j} 10^{0.1P(j)} \text{ (dB)}, \ \forall P(j) \notin \{P_{\text{TM}}(k,\ k \pm 1,\ k \pm \Delta_k\} \tag{7.6}$$

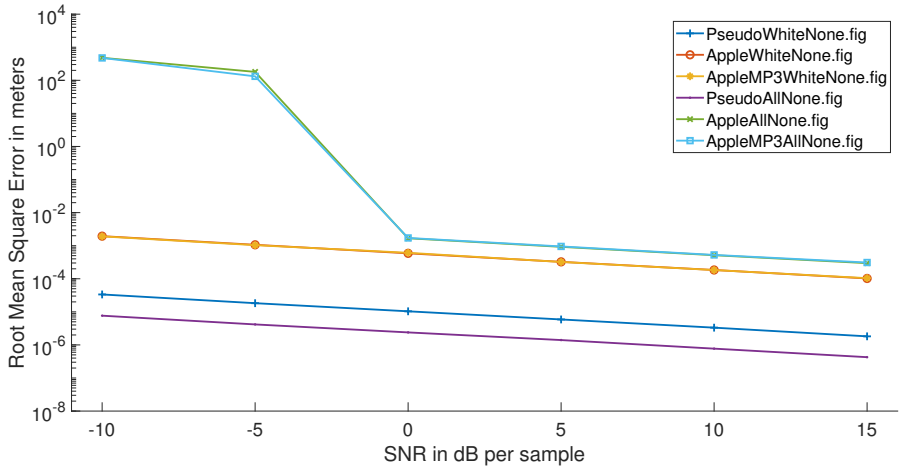


Figure 7.5: TOA estimation, second test scenario, audio coding

where

$$\overline{k} = \left(\prod_{j=l}^{u} j \right)^{1/(l-u+1)} \tag{7.7}$$

l and u are the critical band boundaries. As a consequence, residual energy within a critical band which is not associated with tonal maskers must be noisy. Or spectral components which have not contributed to tonal maskers, contribute to a single noisy masker for each band.

Step 3 is for decimation and reorganisation of previously computed maskers in order to reduce the number of maskers. First of all, all maskers which are below the absolute hearing threshold are discarded:

$$P_{\mathrm{TM,NM}}(k) \geq T_{\mathrm{q}}(k) \tag{7.8}$$

with $T_{\mathrm{q}}(k)$ being the absolute hearing threshold in quiet. Afterwards, a sliding 0.5 Bark-wide window is used to replace any pair of maskers within a 0.5 Bark distance by the stronger of the two. Then the frequency bins are reorganised according to the subsampling scheme:

$$\begin{gathered} P_{\mathrm{TM,NM}}(i) = P_{\mathrm{TM,NM}}(k) \\ P_{\mathrm{TM,NM}}(k) = 0 \end{gathered} \tag{7.9}$$

with

$$i = \begin{cases} k, & 1 \leq k \leq 48 \\ k + (k \bmod 2), & 49 \leq k \leq 96 \\ k + 3 - ((k-1) \bmod 4), & 97 \leq k \leq 232 \end{cases} \tag{7.10}$$

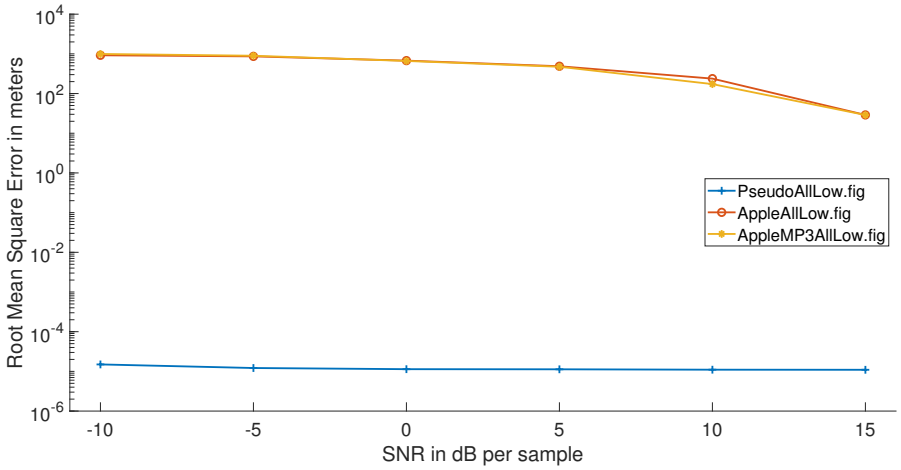


Figure 7.6: TOA estimation, third test scenario, audio coding

With this reorganisation, a reduction of masker bins without loss of masking components is achieved.

Step 4 is dedicated to calculate the individual masking thresholds. For every tonal or noise masker a masking threshold, representing the masking contribution at frequency bin i, is computed at the corresponding frequency bin j which was frequency bin i after reorganisation. The individual masking thresholds for tonal maskers are

$$T_{\mathrm{TM}}(i,j) = P_{\mathrm{TM}}(j) - 0.275\ \mathrm{Z_b}(j) + SF(i,j) - 6.025\ (\mathrm{dB\ SPL}) \tag{7.11}$$

$\mathrm{Z_b}(j)$ is the Bark frequency of bin j and $SF(i,j)$ is the spread of masking from masker bin j to masked bin i:

$$SF(i,j) = \begin{cases} 17\Delta_{\mathrm{Z_b}} - 0.4P_{\mathrm{TM}}(j) + 11, & -3 \le \Delta_{\mathrm{Z_b}} < -1 \\ (0.4P_{\mathrm{TM}}(j) + 6)\Delta_{\mathrm{Z_b}}, & -1 \le \Delta_{\mathrm{Z_b}} < 0 \\ -17\Delta_{\mathrm{Z_b}}, & 0 \le \Delta_{\mathrm{Z_b}} < 1 \\ (0.15P_{\mathrm{TM}}(j) - 17)\Delta_{\mathrm{Z_b}} - 0.15P_{\mathrm{TM}}(j), & 1 \le \Delta_{\mathrm{Z_b}} < 8 \end{cases} \quad (\mathrm{dB\ SPL}) \tag{7.12}$$

where

$$\Delta_{\mathrm{Z_b}} = \mathrm{Z_b}(i) - \mathrm{Z_b}(j) \tag{7.13}$$

The spread function $SF(i,j)$ approximates the spreading on the basilar membrane. The spread of masking is constrained to a 10-Bark neighbourhood for computational efficiency. The individual noise masker thresholds are computed as

$$T_{\mathrm{NM}}(i,j) = P_{\mathrm{NM}}(j) - 0.175\ \mathrm{Z_b}(j) + SF(i,j) - 2.025\ (\mathrm{dB\ SPL}) \tag{7.14}$$

The spread function $SF(i,j)$ is obtained in the same way.

Step 5 is the calculation of a global masking threshold for the current segment. For that, the individual masking thresholds for each frequency bin are combined to estimate a global masking threshold. This follows the assumption that masking effects are additive. The global masking threshold is defined as

$$T_{\mathrm{g}}(i) = 10\lg\left(10^{0.1T_{\mathrm{q}}(i)} + \sum_{l=1}^{L} 10^{0.1T_{\mathrm{TM}}(i,l)} + \sum_{m=1}^{M} 10^{0.1T_{\mathrm{NM}}(i,m)}\right)\ (\mathrm{dB\ SPL}) \qquad (7.15)$$

L and M is the total number of tonal and noise maskers respectively. The global masking threshold is a signal-dependent, power-additive modification of the absolute hearing threshold due to basilar spread of all tonal and noise maskers in the signal power spectrum for each frequency bin [21].

7.3.2 Code Implementation

An implementation of the psychoacoustic model was used from [19]. The following pseudo code explains how the masking threshold computation was used to add noise to the Apple sound:

1. Load the audio file which shall be modified, scale and normalise it
2. Compute the global masking thresholds segment wise according to the MPEG 1 psychoacoustic model
3. Load the audio file which shall be added to the signal, scale and normalise it
4. Compute amplitude values of SPL masking threshold values
5. Extract the start and end samples of the windows used for segment wise computation within the psychoacoustic model
6. Interpolate global masking thresholds for each segment for all frequency bins and multiply it with the noise to be added to the signal in the frequency domain
7. Transfer shaped noise back to time domain and add it to the signal within the start and end samples of the current segment window (computed earlier)

7.3.3 Results

The following figures depict again the performance of the modified Apple sound in terms of TOA estimation in comparison to the original Apple sound and pseudo-random noise for the three test scenarios. Figure 7.7 visualises the first test scenario where white noise and no, low reverberation is present. As can be seen, the

addition of pseudo-random noise to the Apple sound according to the masking thresholds almost did not change the TOA estimation RMSE at all. The same goes

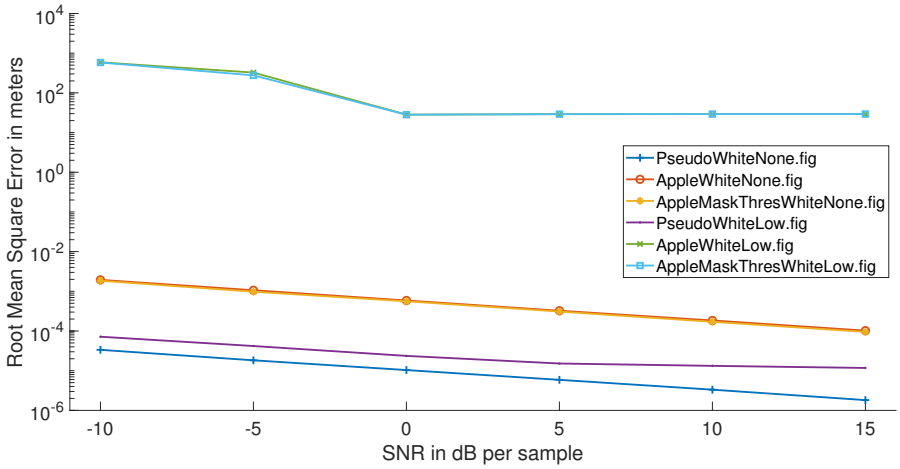


Figure 7.7: TOA estimation, first test scenario, masking thresholds

for the seconds test scenario with no reverberation and a combination noise types in figure 7.8. And finally, also for the third test scenario where a combination of

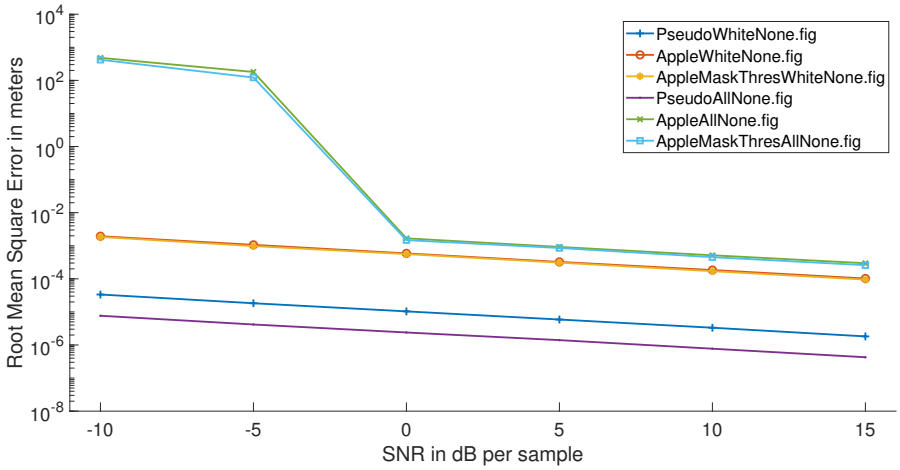


Figure 7.8: TOA estimation, second test scenario, masking thresholds

noise types and low reverberation is present, the addition of pseudo-random noise practically did not change the result (figure 7.9).

Concluding must be said that despite the effort of calculating the masking thresholds of the Apple HomePod start up sound and using them to add pseudo-random

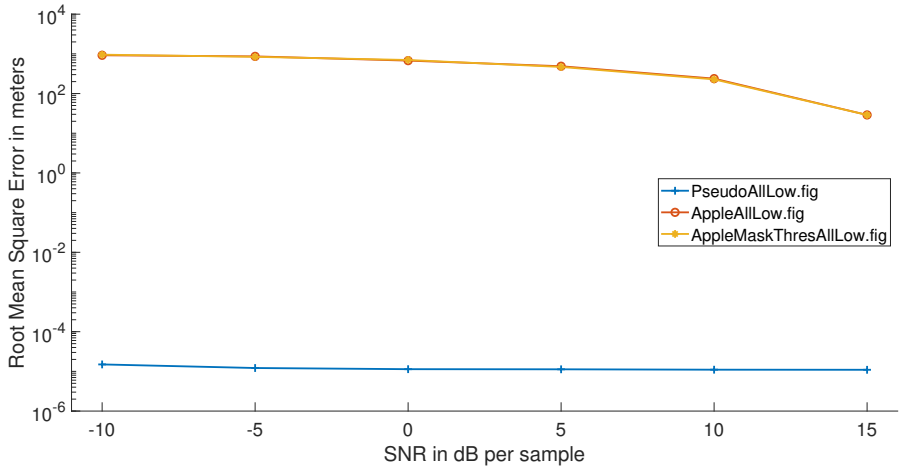


Figure 7.9: TOA estimation, third test scenario, masking threshold

noise below perception, no improvement in terms of TOA estimation could be achieved.

7.4 First Conclusion and Further Investigation

All of the three applied methods with the goal to increase the TOA estimation performance of the Apple sound failed. No acceptable RMSE could be achieved. Therefore, a new attempt with a different approach was made. This time, the idea was to estimate the spectral envelope from the Apple sound and apply it to the pseudo-random noise. This resulted in a signal which still sounded like the pseudo-random noise but more low-frequent according to the spectrum of the Apple sound which has most of its energy in low frequencies. The plan then was to add the signal to the Apple sound. It was hoped that although the pseudo-random noise is audible, it would not be noticed much because it has the same spectral envelope as the Apple sound. Before doing that, though, it was checked if the performance of the modified pseudo-random noise was good enough.

Figure 7.10 depicts the TOA estimation performance for the first test scenario with white noise and no, low reverberation. It is apparent that the RMSE from the modified pseudo-random noise did not change in comparison to the original Apple sound (asterisk-marker, yellow curve and circle-marker, red curve) for white noise and no reverberation. For low reverberation and white noise, it did change (square-marker, blue curve and cross-marker, green curve) but the RMSE of the modified pseudo-random noise is still around 5 meters, thus too high. No sufficient performance could be realised for the second and third test scenario either.

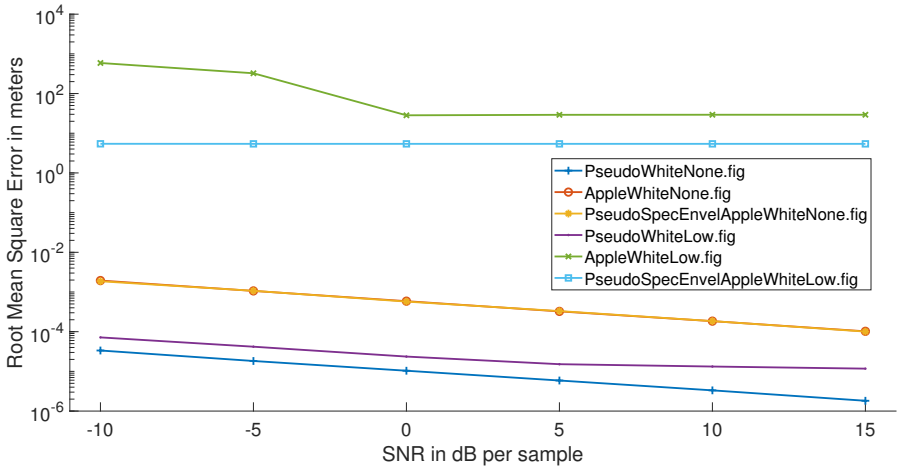


Figure 7.10: TOA estimation, first test scenario, pseudo-random noise with spectral envelope of the Apple HomePod start up sound

The figures can be found in appendix D.

It was hypothesised from this that high frequency content of the original pseudo-random noise is crucial for its performance. This observation is coherent with [5] where time differences of arrival of mobile station locations are estimated. The authors write that "spread spectrum signals, due to their wide bandwidth, are well suited for TOA estimation". That means that the Apple sound cannot reach the performance of the pseudo-random noise in terms of TOA estimation unless high frequency content is added.

Therefore, a new approach was tested to increase the TOA estimation performance of the Apple HomePod start up sound. Basis was again the calculation of the masking curves. But this time, only pseudo-random noise with high frequency content was added. The frequency range was chosen based on the exemplary masking curve of one segment visualised in figure 7.11. (The masking curves for all other segments were relatively similar to each other.) Here, one can see that the masking curve has its highest levels in lower frequencies until approx. 1000 Hz. After decreasing it increases again around 4000 Hz to 6000 Hz. The frequency content which was used began from approx. 8000 Hz until half the sampling frequency. In that range, the masking curve increases once more. Another reason for the choice that only pseudo-random noise of higher frequencies was added was that it was less audible and disturbing than low frequencies. This is important because in order to achieve an acceptable performance the pseudo-random noise has to be added beyond the masking curve (factor 33 (factor 1 would be according to the masking curve)).

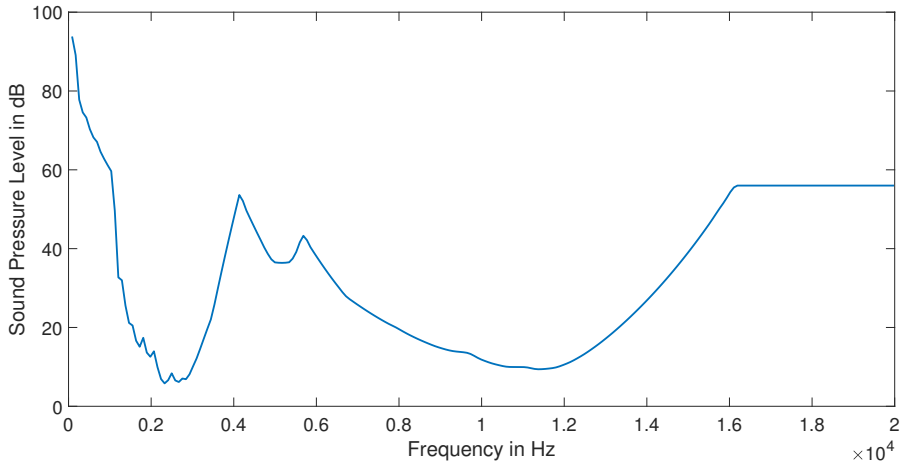


Figure 7.11: Masking curve based on the Apple sound of one segment

The TOA estimation performance of the newly modified Apple sound for the first test scenario with white noise and no, low reverberation is shown in figure 7.12. It can be seen that the RMSEs from the modified Apple sound decreased signif-

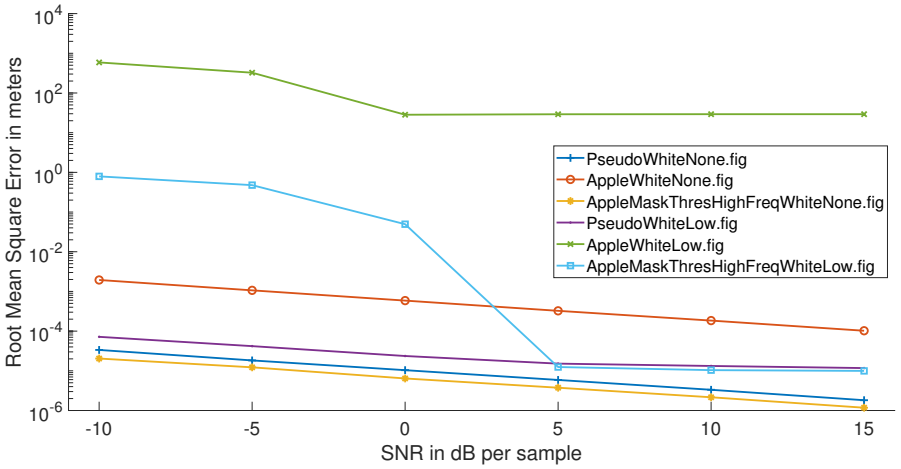


Figure 7.12: TOA estimation, first test scenario, adding only high frequency pseudo-random noise according to masking thresholds

icantly. In the case of white noise and no reverberation, the RMSE of the Apple sound is even lower than the one from the pseudo-random noise. When white noise and low reverberation is present, the RMSE of the modified Apple sound reaches the RMSE of the pseudo-random noise in that condition at 5 dB SNR.

Figure 7.13 visualises the second test scenario. Also in this case, the RMSE of

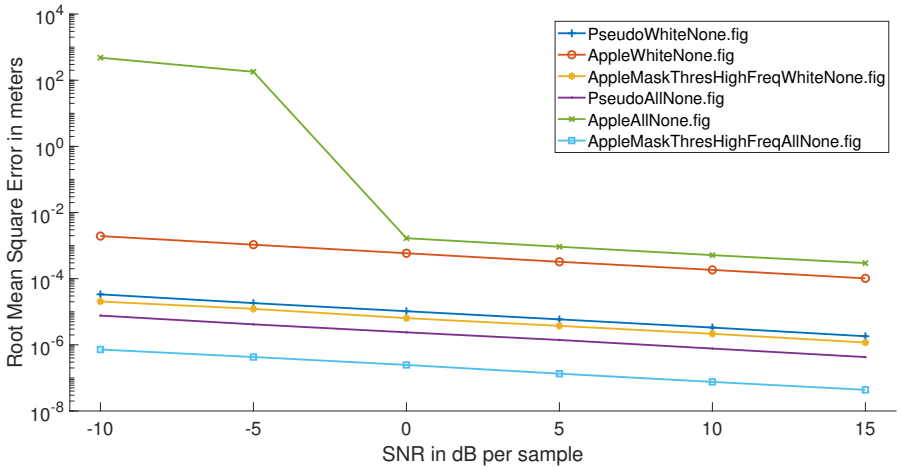


Figure 7.13: TOA estimation, second test scenario, adding only high frequency pseudo-random noise according to masking thresholds

the modified Apple sound decreased a lot, again below the pseudo-random noise RMSE.

And finally, the third scenario in figure 7.14. Again, the RMSE of the modified

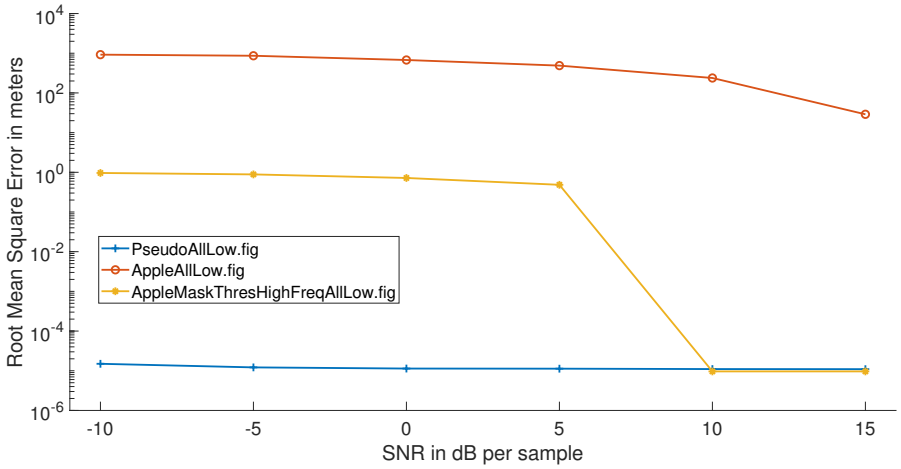


Figure 7.14: TOA estimation, third test scenario, adding only high frequency pseudo-random noise according to masking thresholds

Apple sound decreased significantly but reached the pseudo-random noise RMSE first at 10 dB SNR.

Another modified Apple HomePod start up sound was created where even more high frequency pseudo-random noise was added so that the performance of the modified Apple sound and original pseudo-random noise was relatively similar. The plots can be found in appendix D.

The new modification of the Apple sound came with a price. The added noise was audible. The Apple sound was still well audible too but one could clearly hear the noise in the background. It was not as annoying as low frequency noise but still decreased the perceived audio quality. With good will, one could say that the noise sounded a little bit like the crackling when small timbers are on fire. But overall the modified sound is not particularly pleasant and probably not sufficient for audio system manufactures, but it was the best trade-off which could be achieved so far. As for the case where even more noise was added (factor 58) the noise became more audible whereas the Apple sound was less audible. Note that the sound perception description was based on the author's subjective opinion only and was not based on a listening test.

One problem which could be occurring when only high frequency pseudo-random noise is added, is spatial aliasing. Spatial aliasing occurs when the receiver is a microphone array, which is mostly the case in smart speakers. When the recorded signal's frequency is too high and the microphones of the array are spaced too far from each other, ambiguities in the received signal can appear. For example if a signal of 10,000 Hz should be detected the spacing of the microphones must be less than 1.72 cm according to

$$d < \frac{c}{2 \cdot f} \tag{7.16}$$

with c being a speed of sound (e.g. 343.21 m/s) and f the frequency. In the case of TOA estimation with one source and receiver, this will not be problem and the modified Apple sound has not only high frequency content, which helps. But the problem has to be kept in mind when more complicated conditions shall be solved.

7.5 Final Conclusion

In this chapter, the Apple HomePod start up sound was modified in order to achieve a better TOA estimation performance. The methods spectral envelope estimation, audio coding and masking curve calculation with subsequent noise adding were applied to the Apple sound. The modified Apple sound was then tested with the validated testing framework in the defined testing scenarios described in chapter 5. The results was that none of the tested methods could increase the TOA estimation performance to an acceptable level. It was further hypothesised that with the spectral envelope of the original Apple sound, it is not possible to reach

an acceptable performance. Instead high frequency content is crucial. As a consequence, only high frequency pseudo-random noise was added to the original Apple sound beyond its masking thresholds. The result was a significant increase of TOA estimation performance to an acceptable level but with a decrease in perceptual quality (subjective opinion). With this result research question 3b) from chapter 1 was addressed.

Chapter 8

Conclusion

8.1 Project Summary

This master thesis has dealt with the problem of automated audio system calibration. A part of the calibration process is to estimate the locations of the loudspeakers and the listener. With that information, the so called sweet spot can be realised. Besides estimating the positions of the loudspeakers and the listener, another part of the calibration process is to compensate for room acoustics. A simplified version of that could be that one adjusts the low frequency response of a loudspeaker when it is close to the wall.

The idea was that the calibration process should automatically start when the audio system is switched on. The loudspeakers should play back a calibration signal which is recorded by microphones. Preferably, the microphones should be integrated in the loudspeakers which is the case in modern smart speakers. The calibration signal should be pleasant because it would be heard often. This was the focus of this work: finding a pleasant calibration signal which is also suitable for automated audio system calibration.

In order to solve this problem, it was simplified to the case where the time of arrival (TOA) should be estimated from one loudspeaker to one microphone. This information can be used to calculate the distance between them which is the first step to create a map over loudspeakers. The TOA estimation should work in corrupting noise, reverberation or both.

Firstly, effort has been made to define what characteristics a perceptually pleasant calibration signal should have to be suitable for automated audio system calibration. To that end, experts from Bang & Olufsen and Audiowise have been interviewed. A number of important characteristics were identified. The calibra-

tion sound should be pleasant, not annoying, not too long, exciting and being associated with high quality and expensiveness. Moreover, the functionality of the sound, the experience people should have with it and a target group was identified, too. The gained knowledge was used to further research the area of sound design. The goal was to be able to design and choose sounds which would fulfil the perceptual requirements on a calibration signal. Research was reviewed for product-sound design, semiotics in the context of product-sound design and psychoacoustics and sound quality. This knowledge was applied to design a sound. Furthermore, three other sounds were selected which were also possible candidates.

A listening experiment was conducted to find out how participants rated the chosen and designed sounds based on attributes important for a possible calibration sound. Participants rated the Apple HomePod start up sound significantly higher than the others. Consequently, this sound was chosen for further analysis.

After making sure that the perceptual qualities were the best of the chosen sounds, the TOA estimation performance was tested of the Apple sound against the traditional, well performing pseudo-random noise within a testing framework. In this framework, one could simulate different noise types and levels of reverberation and calculate an RMSE in meters for a number of Monte-Carlo simulations. The outcome was that the pseudo-random noise outperformed the Apple sound especially when reverberation was present. Additionally, three test scenarios were defined in which a modification of the Apple sound, with the intention to increase its TOA estimation performance, should be tested. The first test scenario consists of white noise and no, low reverberation, violating the TOA estimator in the free-field assumption. In the second test scenario, no reverberation was present and white noise and a combination of other noise types, violating the white noise assumption of the TOA estimator. In the third test scenario a combination of noise types and low reverberation were present. Both the free-field and white noise assumption of the TOA estimator were violated. In those test scenarios the pseudo-random noise delivered acceptable RMSEs so a modification of the Apple sound could theoretically increase its performance to an acceptable performance, too.

Subsequently, real-life measurements were completed to validate the results from the test framework. The compared results were, within a reasonable margin of error, similar. Thus, the testing framework delivered valid results and could be used for further testing.

Afterwards, different methods of modifying the Apple sound were applied, attempting to increase its TOA estimation performance. It was found that the spec-

tral content of the Apple sound makes a good TOA estimation impossible when reverberation is present. It was furthermore hypothesised that a wide spectral bandwidth is crucial for TOA estimation. So, high frequency pseudo-random noise was added to the original Apple sound according to its masking thresholds but well above them. The result was an acceptable TOA estimation performance but with a considerable loss of perceptual quality (the high frequency pseudo-random noise was clearly audible).

8.2 Future Work

The modification of the Apple sound has be be improved so that it retains its perceptual quality but still yields a better TOA estimation performance. To that end, the hypothesis that a wide spectral bandwidth is crucial for TOA estimation has to be investigated further. It would be interesting to find out whether one should attempt to obtain a flat spectral bandwidth or if only high frequencies are important. Furthermore, it could also be the case that just having low frequencies reduces the TOA estimation performance. Moreover, it has to be investigated how reverberation (especially in low frequencies) changes the situation. Maybe, low frequency content is only disadvantageous when reverberation is present, maybe it is disadvantageous is general.

Finally, when a method is found which increases the TOA estimation performance of the Apple sound to an acceptable level but does not or only slightly reduce the perceptual quality, the use case could be made more complicated, i.e. testing the performance when multiple loudspeakers are present. With this, other practical issues need to be addressed. How do one organise the playback over the loudspeakers? Shall they all play back the same signal subsequently? What happens if one has ten loudspeakers? Playing back the calibration signal ten times is clearly too often. Maybe a combination of signals which fit together could be played. But the duration of the whole process has to be short enough. If one finds a method to increase the TOA estimation of any signal then audio system calibration could be integrated in the normal use of the system modifying any signal which the user choose to play back. Maybe the modified calibration signal can even be used to obtain impulse responses of the listening environment. They could be used to adjust the frequency response of the loudspeakers depending on the room and their position. Another problem is to localise the listener being able to achieve the sweet spot. How does one do that without involving an action of the user other than operating the audio system in the usual way.

A different aspect one has not looked at in this work is how the TOA estimator influences the results. As mentioned above, it is only valid for the assumptions

that white Gaussian noise is present and the free-field assumption is valid. An adaptation of the estimator to a different situation might improve the results.

The problem of automated audio system calibration is far from being solved but the knowledge and results obtained through this work are useful for further research in that area.

Bibliography

[1] Peter Ahrendt. "Designing Calibration Signals for Loudspeaker Localisation". Aalborg, Denmark, 2017.

[2] D. Arfib et al. "Source-filter processing". In: *DAFX: Digital Audio Effects*. Wiley, 2011. Chap. 8.

[3] Bang & Olufsen A/S. *BeoLab 90 Technical Sound Guide*. 2017.

[4] Karlheinz Brandenburg and Gerhard Stoll. "ISO/MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio". In: *Journal of the Audio Engineering Society* 42.10 (Oct. 1994), pp. 780–792.

[5] Ali Broumandan et al. "Practical Results of Hybrid AOA/TDOA Geo-Location Estimation in CDMA Wireless Networks". In: *2008 IEEE 68th Vehicular Technology Conference*. IEEE, Sept. 2008, pp. 1–5.

[6] Maxime Carron et al. "Designing sound identity: providing new communication tools for building brands corporate sound". In: *Proceedings of the 9th Audio Mostly on A Conference on Interaction With Sound - AM '14*. Aalborg, Denmark: ACM Press, 2014, pp. 1–8.

[7] Pascal Dietrich et al. *MATLAB Toolbox for the Comprehension of Acoustic Measurement and Signal Processing*. Tech. rep. Institute of Technical Acoustics, RWTH Aachen University, 2013.

[8] Hugo Fastl. "Psycho-Acoustics and Sound Quality". In: *Communication Acoustics*. Berlin/Heidelberg: Springer-Verlag, 2005, pp. 139–162.

[9] Hugo Fastl and Eberhard Zwicker. "Examples of Application". In: *Psychoacoustics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 315–368.

[10] Robert M Gray and Lee. D. Davisson. *An introduction to statistical signal processing*. Vol. 1. 2004.

[11] Jurgen Herre et al. "MPEG-H 3D AudioThe New Standard for Coding of Immersive Spatial Audio". In: *IEEE Journal of Selected Topics in Signal Processing* 9.5 (Aug. 2015), pp. 770–779.

[12] Jayne Al-Hindawe. "Considerations when constructing a semantic differential scale." In: *La Trobe working papers in linguistics.* 9 (1996), pp. 41–58.

[13] *ISO 3382-2: Acoustics - Measurement of room acoustic parameters - Part 2: Reverberation time in ordinary rooms*. Berlin, 2008.

[14] *ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems*. 2015.

[15] Reinier J. Jansen, Elif Özcan, and René van Egmond. "PSST! Product Sound Sketching Tool". In: *Journal of the Audio Engineering Society* 59.6 (July 2011), pp. 396–403.

[16] Ute Jekosch. "Assigning Meaning to Sounds Semiotics in the Context of Product-Sound Design". In: *Communication Acoustics*. Berlin/Heidelberg: Springer-Verlag, 2005, pp. 193–221.

[17] Jesper Kjær Nielsen. "Loudspeaker and Listening Position Estimation using Smart Speakers". Aalborg, Denmark, 2017.

[18] Charles Egerton Osgood, George J. (George John) Suci, and Percy H. Tannenbaum. *The measurement of meaning*. Ninth. Urbana: University of Illinois Press, 1975, p. 342.

[19] Fabien A. P. Petitcolas. *MPEG psychoacoustic model I for MATLAB*. 2003. URL: `http://petitcolas.net/fabien/software/mpeg/`.

[20] Diemo Schwarz and Xavier Rodet. "Spectral Envelope Estimation and Representation for Sound Analysis Synthesis". In: *Proceedings of the ICMC*. Paris, 1999, pp. 351–354.

[21] Andreas Spanias, Ted Painter, and Venkatraman Atti. *Audio Signal Processing and Coding*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Jan. 2007.

[22] Andrew Wabnitz et al. "Room acoustics simulation for multichannel microphone arrays". In: *Proceedings of the International Symposium on Room Acoustics, ISRA*. Melbourne, Australia, 2010, p. 6.

Appendix A

Reverberation times MCRoomSim

For calculating the RIRs the room depicted in figure A.1 was simulated. It was

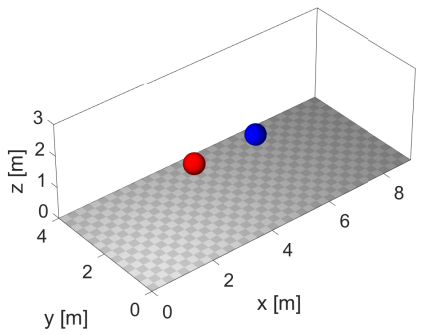


Figure A.1: Room simulated in MCRoomSim [22] with source and receiver locations for RIR calculation

9 meters long, 4 meters wide and 3 meters high. Source and receiver were 2.1 meters apart from each other and both omnidirectional. All walls had the same absorption coefficients and only specular reflections were enabled. Enabling diffuse reflections, too caused incomprehensible results. For example, performances were much better for higher reverberation than for lower.

Appendix B

Testing Results TOA Performance Estimation

B.1 Apple HomePod Start Up Sound versus Pseudo-Random Noise

Shown are results from the TOA estimation performance comparing the Apple HomePod start up sound and pseudo-random noise from -10 dB to 15 dB SNR for 1000 Monte-Carlo simulations for different noise types and reverberation.

B.1.1 Raw data

Figure B.1 depicts the comparison for traffic noise and no reverberation. Figure B.2

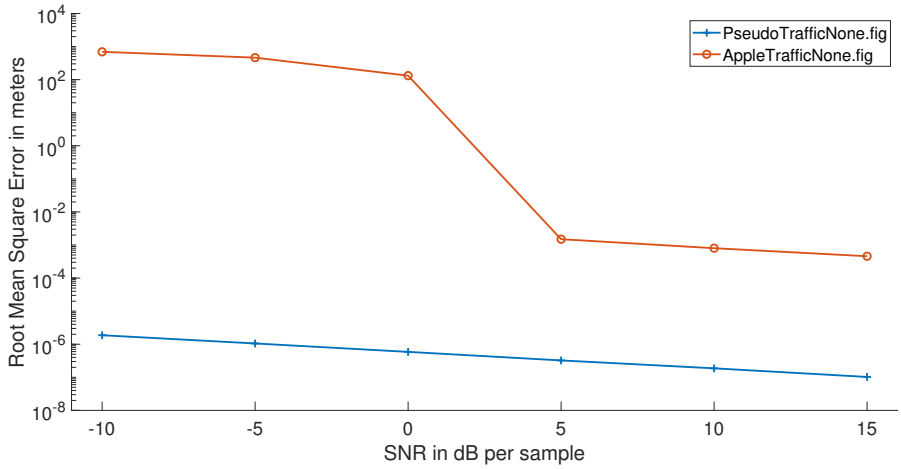


Figure B.1: TOA estimation for 1000 Monte-Carlo simulations, traffic noise, no reverberation

depicts the comparison for traffic noise and low reverberation. Figure B.3 depicts

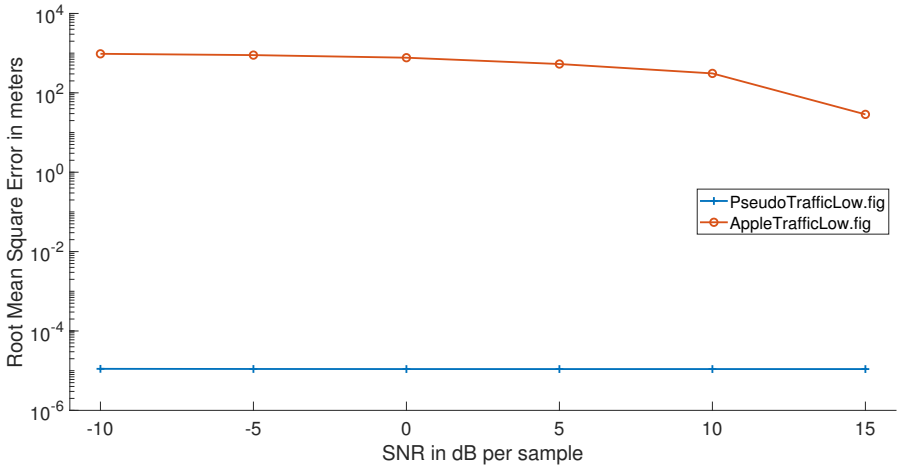


Figure B.2: TOA estimation for 1000 Monte-Carlo simulations, traffic noise, low reverberation

the comparison for traffic noise and moderate reverberation. Figure B.4 depicts the

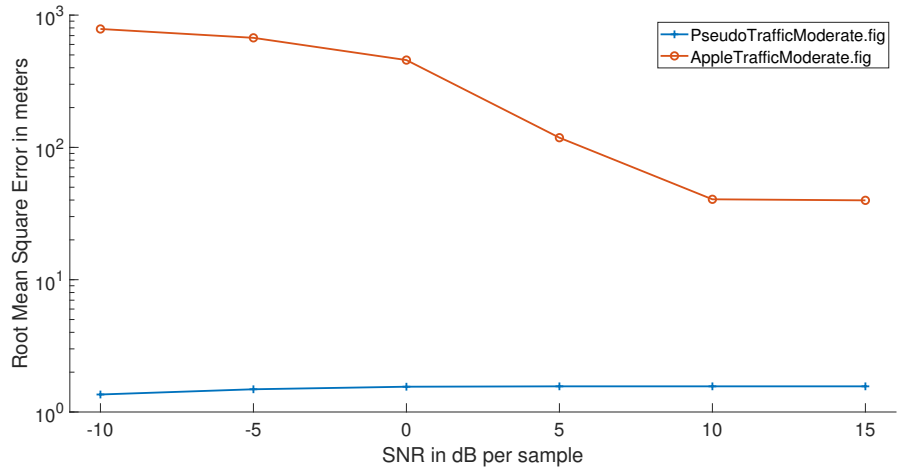


Figure B.3: TOA estimation for 1000 Monte-Carlo simulations, traffic noise, moderate reverberation

comparison for traffic noise and high reverberation. Figure B.5 depicts the comparison for babble noise and no reverberation. Figure B.6 depicts the comparison for babble noise and low reverberation. Figure B.7 depicts the comparison for babble noise and moderate reverberation. Figure B.8 depicts the comparison for babble noise and high reverberation. Figure B.9 depicts the comparison for impulse noise and no reverberation. Figure B.10 depicts the comparison for impulse noise and low reverberation. Figure B.11 depicts the comparison for impulse noise and moderate reverberation. Figure B.12 depicts the comparison for impulse noise and high

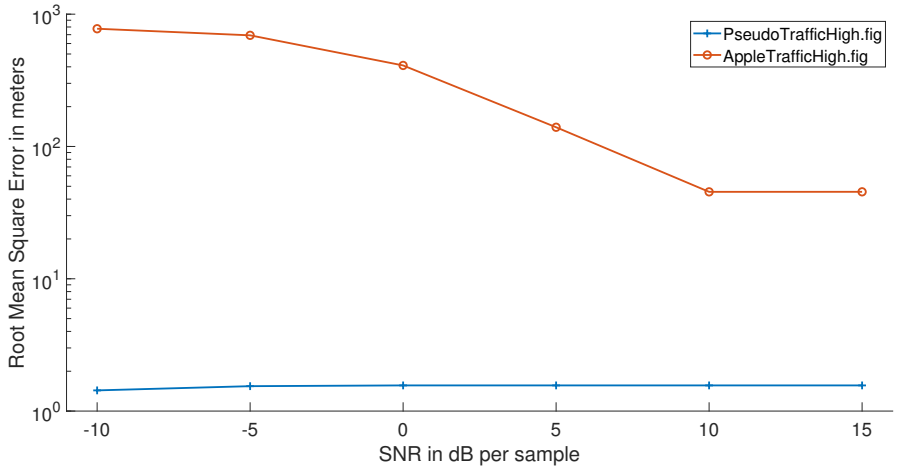


Figure B.4: TOA estimation for 1000 Monte-Carlo simulations, traffic noise, high reverberation

reverberation. Figure B.13 depicts the comparison for a combination of traffic, babble and impulse noise and no reverberation. Figure B.14 depicts the comparison for a combination of traffic, babble and impulse noise and low reverberation. Figure B.15 depicts the comparison for a combination of traffic, babble and impulse noise and moderate reverberation. Figure B.16 depicts the comparison for a combination of traffic, babble and impulse noise and high reverberation.

B.1.2 Combined Data

In the next figures the data is combined as follows: first all TOA estimation RMSEs are depicted for each reverberation level and one noise type in the same figure. Then all TOA estimation RMSEs for each noise types are depicted together in the same plot for one reverberation level.

Figure B.17 depicts the comparison combined for traffic noise and none, low, moderate and high reverberation. Figure B.18 depicts the comparison combined for babble noise and none, low, moderate and high reverberation. Figure B.19 depicts the comparison combined for impulse noise and none, low, moderate and high reverberation. Figure B.20 depicts the comparison combined for a combination of traffic, babble and impulse noise and none, low, moderate and high reverberation. Figure B.21 depicts the comparison for traffic, babble and impulse noise and no reverberation. Figure B.22 depicts the comparison for traffic, babble and impulse noise and low reverberation. Figure B.23 depicts the comparison for traffic, babble and impulse noise and moderate reverberation. Figure B.24 depicts the comparison for traffic, babble and impulse noise and high reverberation.

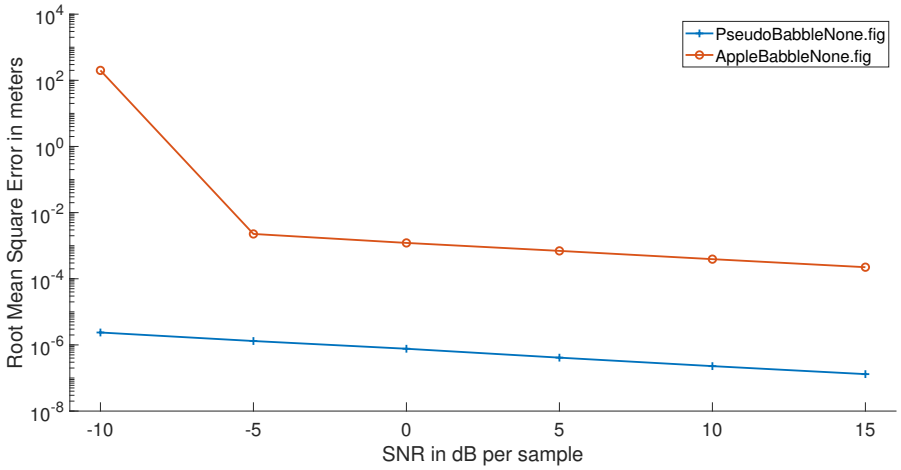


Figure B.5: TOA estimation for 1000 Monte-Carlo simulations, babble noise, no reverberation

B.2 Only Apple StartUp Sound

In the following a further investigation was conducted for TOA estimation performance for the the Apple HomePod start up sound only and SNRs from 5 to 30 dB for 1000 Monte-Carlo simulations for different noise types and reverberation.

B.2.1 Raw Data

Figure B.25 depicts the RMSE for traffic noise and no reverberation. Figure B.26 depicts the RMSE for traffic noise and low reverberation. Figure B.27 depicts the RMSE for traffic noise and moderate reverberation. Figure B.28 depicts the RMSE for traffic noise and high reverberation. Figure B.29 depicts the RMSE for babble noise and no reverberation. Figure B.30 depicts the RMSE for babble noise and low reverberation. Figure B.31 depicts the RMSE for babble noise and moderate reverberation. Figure B.32 depicts the RMSE for babble noise and high reverberation. Figure B.33 depicts the RMSE for impulse noise and no reverberation. Figure B.34 depicts the RMSE for impulse noise and low reverberation. Figure B.35 depicts the RMSE for impulse noise and moderate reverberation. Figure B.36 depicts the RMSE for impulse noise and high reverberation. Figure B.37 depicts the RMSE for a combination of traffic, babble and impulse noise and no reverberation. Figure B.38 depicts the RMSE for a combination of traffic, babble and impulse noise and low reverberation. Figure B.39 depicts the RMSE for a combination of traffic, babble and impulse noise and moderate reverberation. Figure B.40 depicts the RMSE for a combination of traffic, babble and impulse noise and high reverberation.

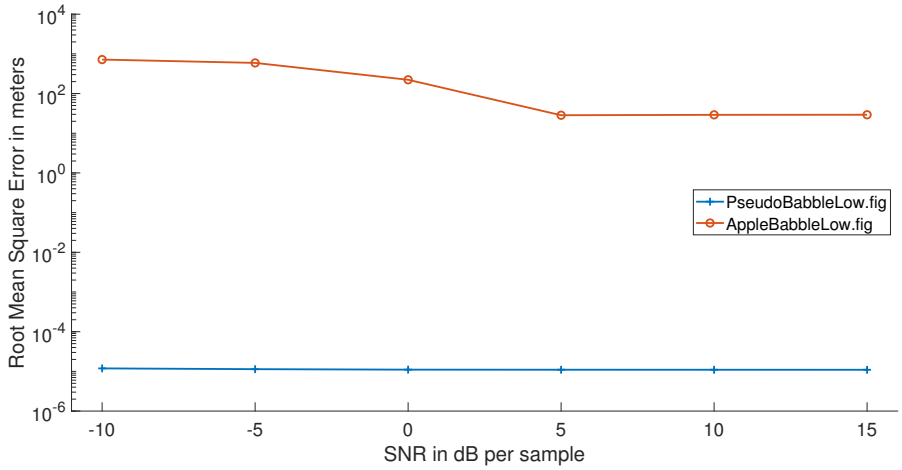


Figure B.6: TOA estimation for 1000 Monte-Carlo simulations, babble noise, low reverberation

B.2.2 Combined Data

In the next figures the data is combined as follows: first all TOA estimation RMSEs are depicted for each reverberation level and one noise type in the same figure. Then all TOA estimation RMSEs for each noise types are depicted together in the same plot for one reverberation level.

Figure B.41 depicts the RMSE for traffic noise and no, low, moderate and high reverberation. Figure B.42 depicts the RMSE for babble noise and no, low, moderate and high reverberation. Figure B.43 depicts the RMSE for impulse noise and no, low, moderate and high reverberation. Figure B.44 depicts the RMSE for a combination of traffic, babble and impulse noise and no, low, moderate and high reverberation. Figure B.45 depicts the comparison for traffic, babble and impulse noise and no reverberation. Figure B.46 depicts the comparison for traffic, babble and impulse noise and low reverberation. Figure B.47 depicts the comparison for traffic, babble and impulse noise and moderate reverberation. Figure B.48 depicts the comparison for traffic, babble and impulse noise and high reverberation.

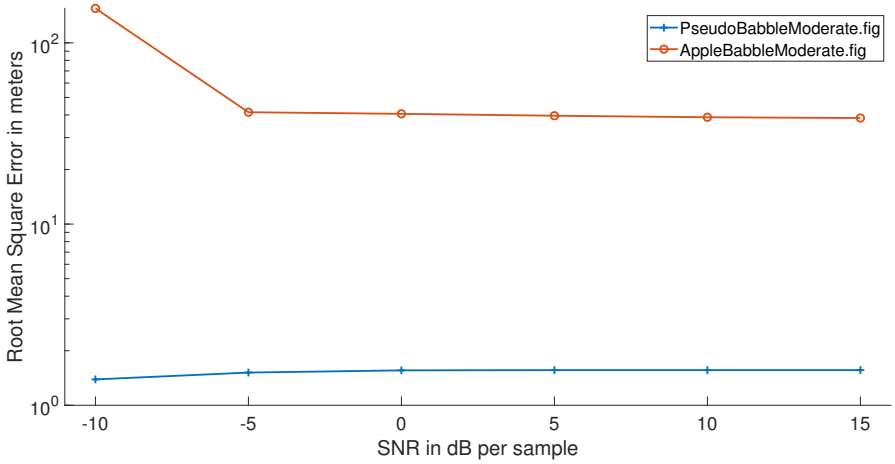


Figure B.7: TOA estimation for 1000 Monte-Carlo simulations, babble noise, moderate reverberation

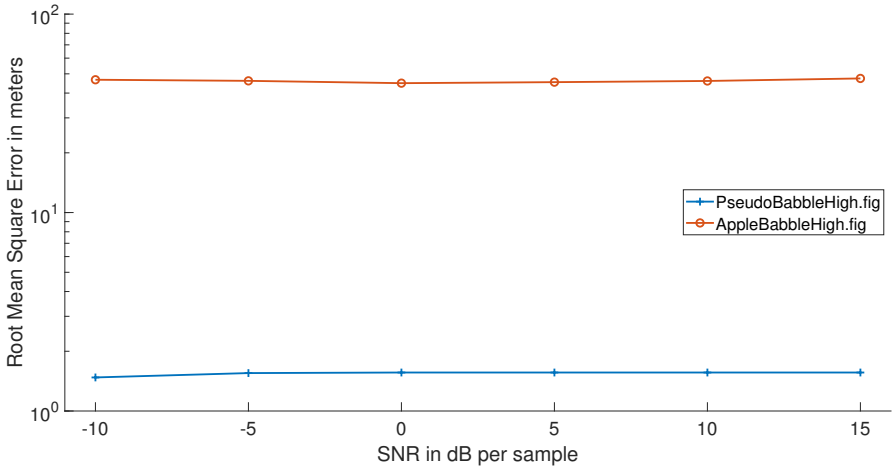


Figure B.8: TOA estimation for 1000 Monte-Carlo simulations, babble noise, high reverberation

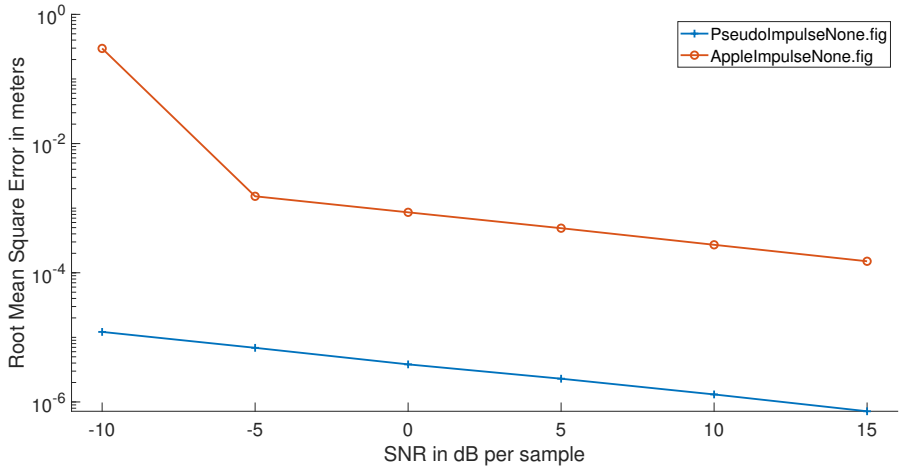


Figure B.9: TOA estimation for 1000 Monte-Carlo simulations, impulse noise, no reverberation

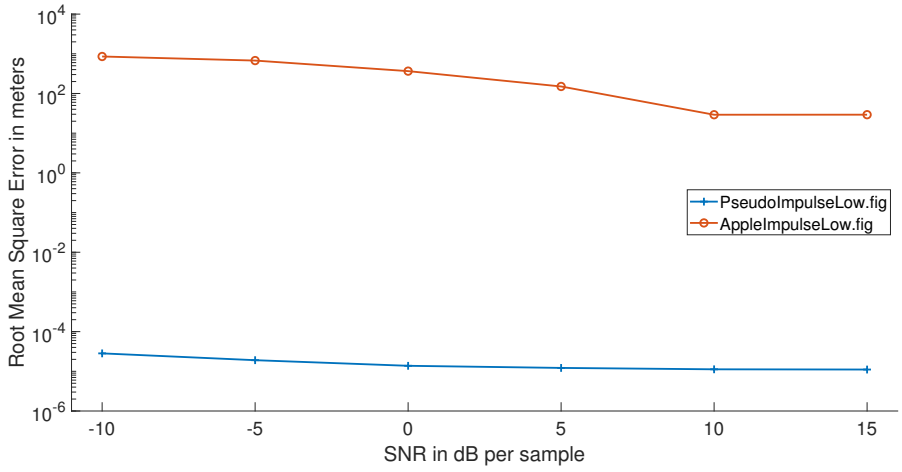


Figure B.10: TOA estimation for 1000 Monte-Carlo simulations, impulse noise, low reverberation

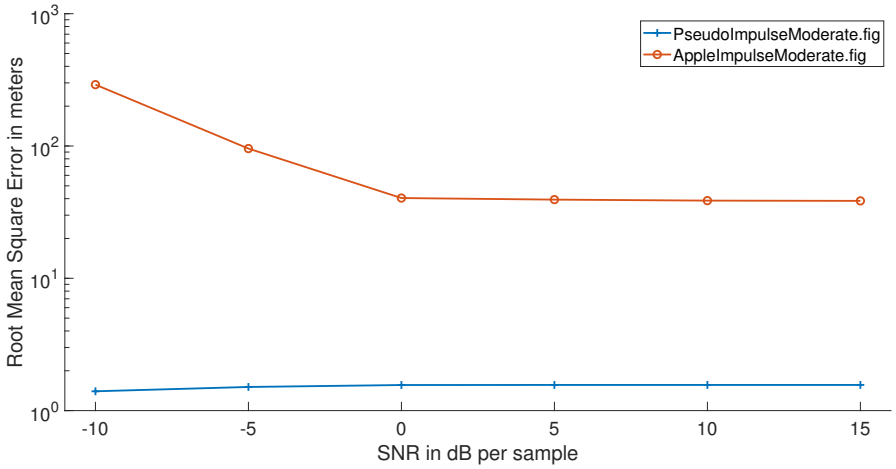


Figure B.11: TOA estimation for 1000 Monte-Carlo simulations, impulse noise, moderate reverberation

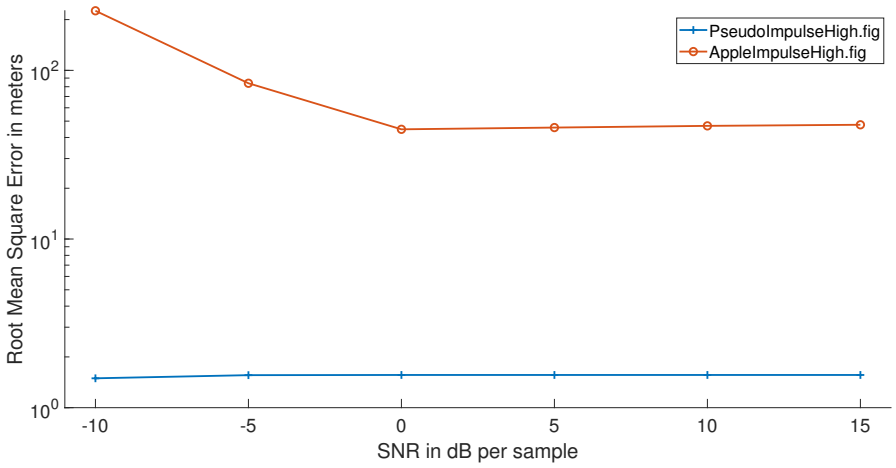


Figure B.12: TOA estimation for 1000 Monte-Carlo simulations, impulse noise, high reverberation

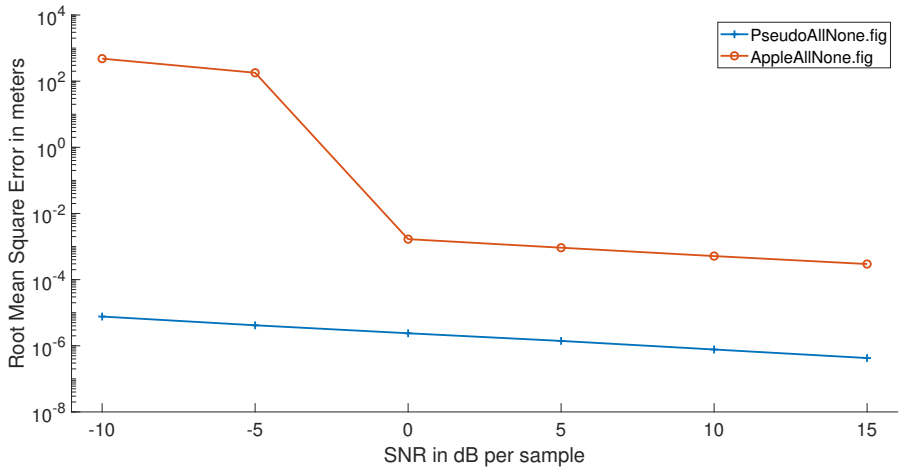


Figure B.13: TOA estimation for 1000 Monte-Carlo simulations, combination of traffic, babble and impulse noise, no reverberation

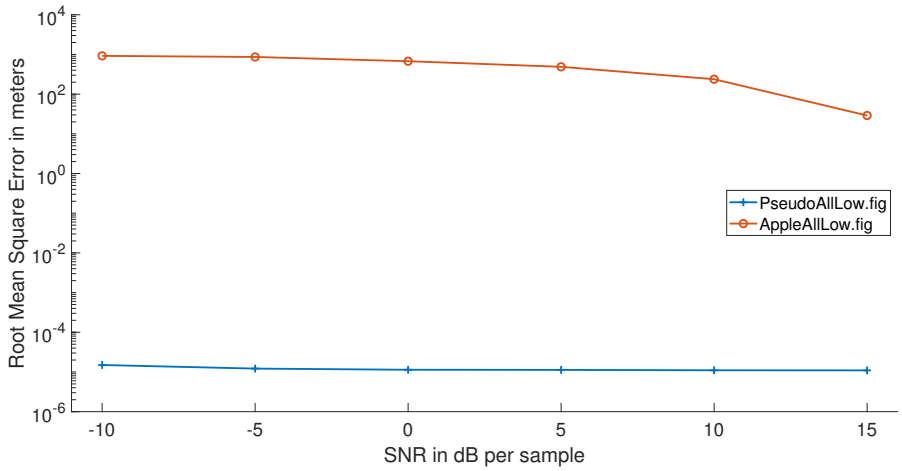


Figure B.14: TOA estimation for 1000 Monte-Carlo simulations, combination of traffic, babble and impulse noise, low reverberation

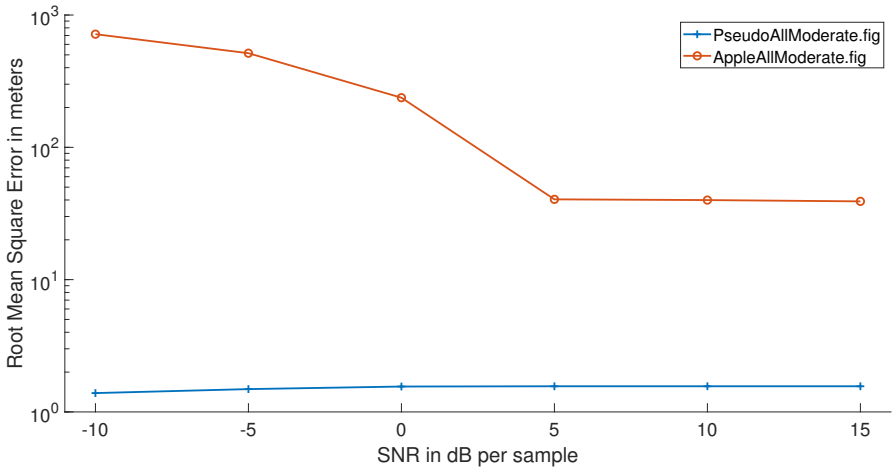


Figure B.15: TOA estimation for 1000 Monte-Carlo simulations, combination of traffic, babble and impulse noise, moderate reverberation

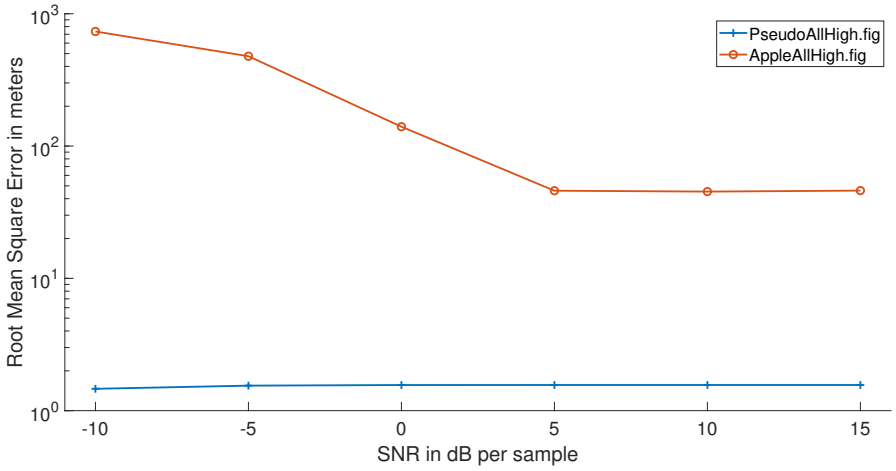


Figure B.16: TOA estimation for 1000 Monte-Carlo simulations, combination of traffic, babble and impulse noise, high reverberation

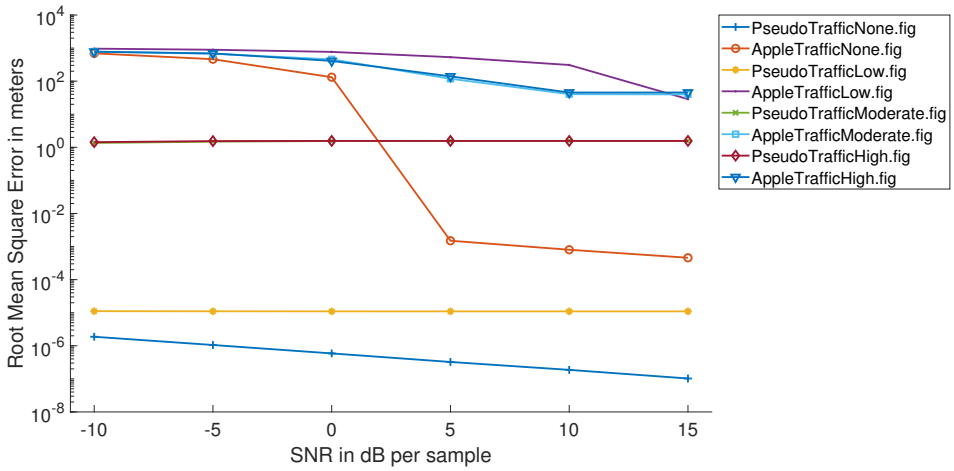


Figure B.17: TOA estimation for 1000 Monte-Carlo simulations, traffic noise, none, low, moderate and high reverberation

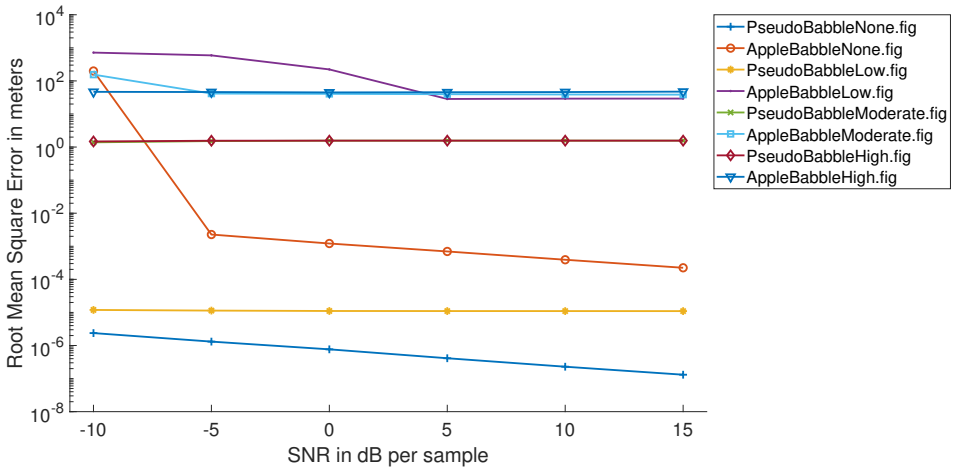


Figure B.18: TOA estimation for 1000 Monte-Carlo simulations, babble noise, none, low, moderate and high reverberation

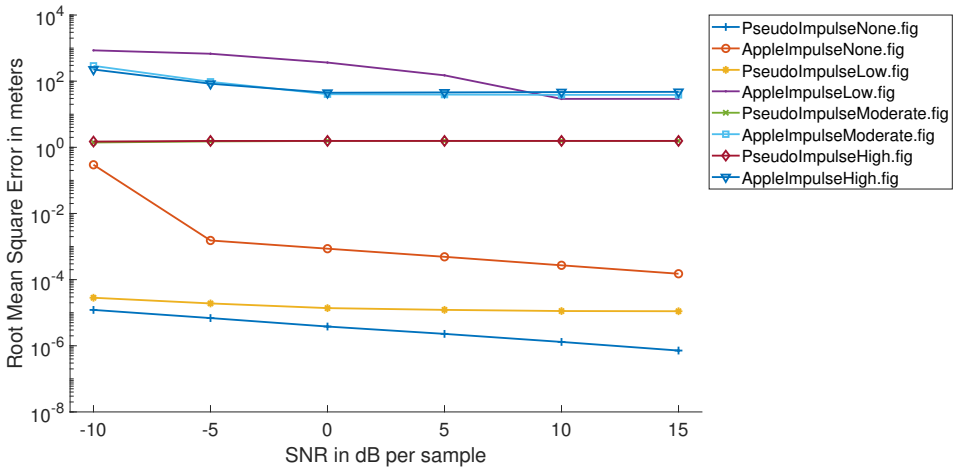


Figure B.19: TOA estimation for 1000 Monte-Carlo simulations, impulse noise, none, low, moderate and high reverberation

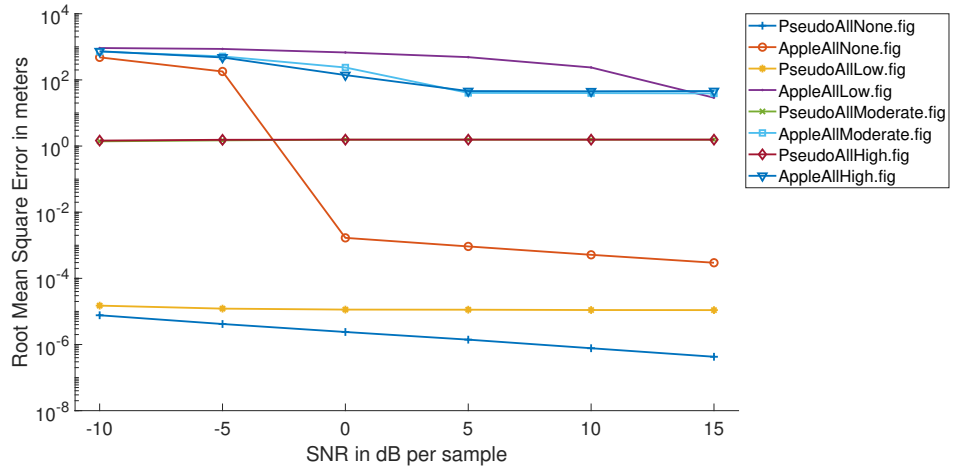


Figure B.20: TOA estimation for 1000 Monte-Carlo simulations, combination of traffic, babble and impulse noise, none, low, moderate and high reverberation

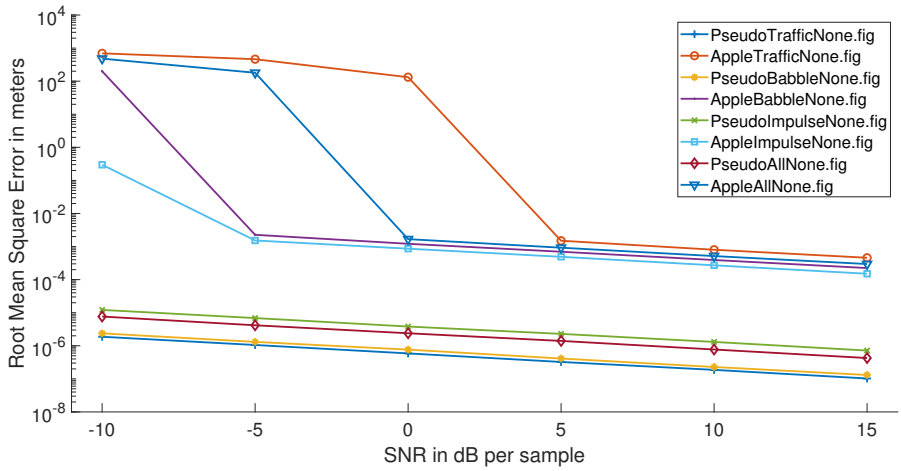


Figure B.21: TOA estimation for 1000 Monte-Carlo simulations, traffic, babble and impulse noise, no reverberation

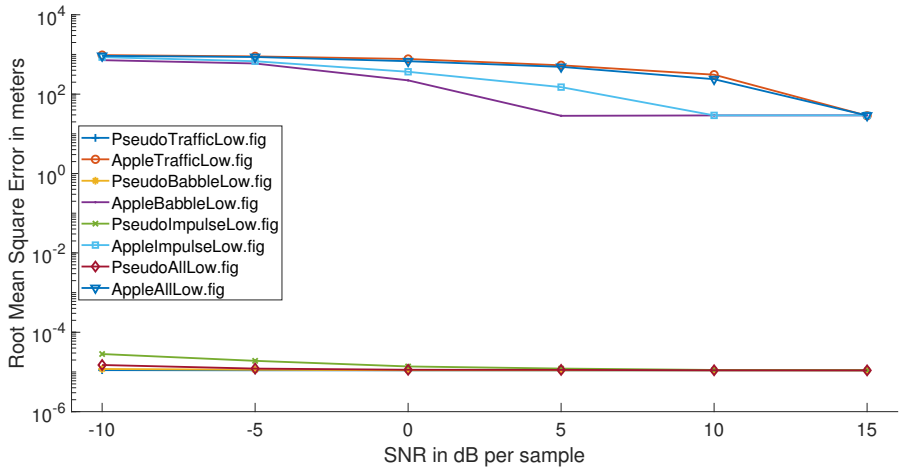


Figure B.22: TOA estimation for 1000 Monte-Carlo simulations, traffic, babble and impulse noise, low reverberation

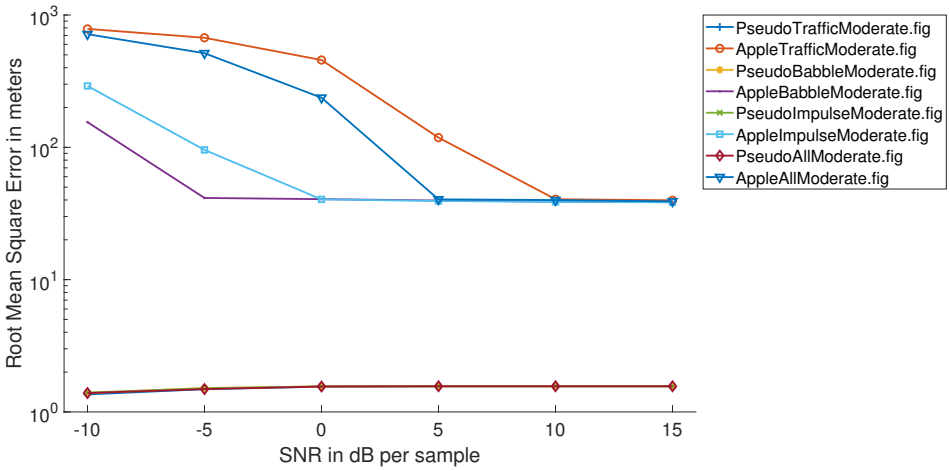


Figure B.23: TOA estimation for 1000 Monte-Carlo simulations, traffic, babble and impulse noise, moderate reverberation

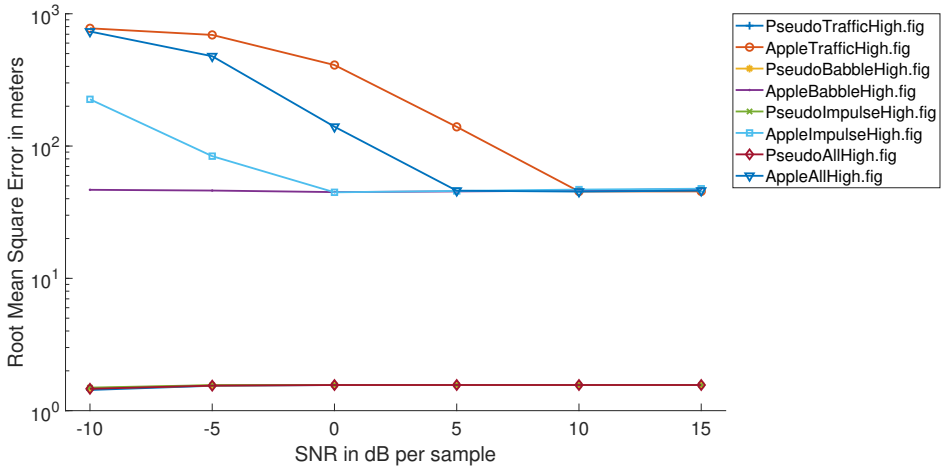


Figure B.24: TOA estimation for 1000 Monte-Carlo simulations, traffic, babble and impulse noise, high reverberation

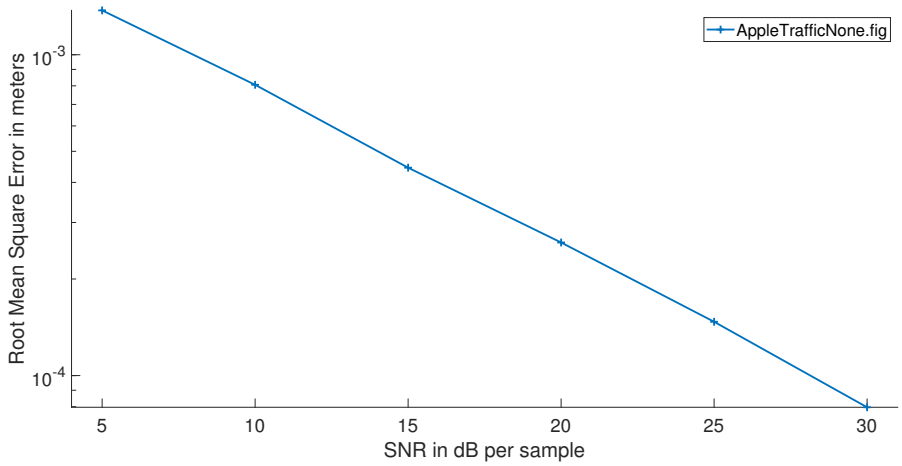


Figure B.25: TOA estimation for 1000 Monte-Carlo simulations, traffic noise, no reverberation

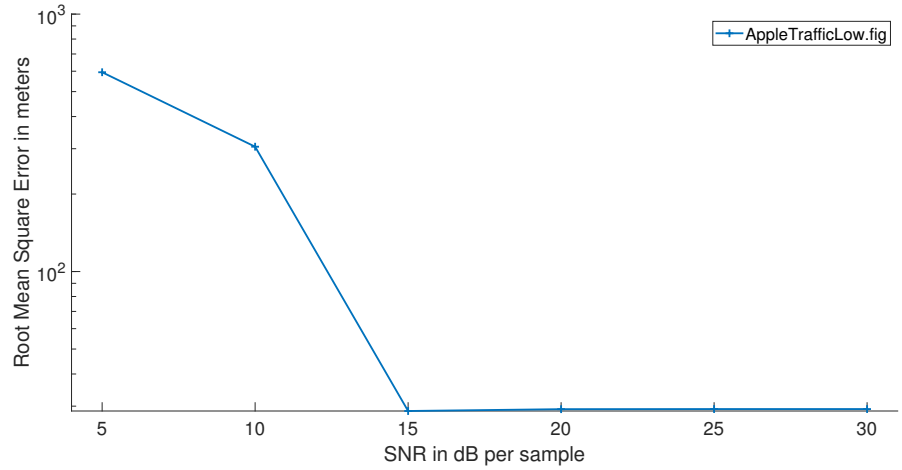


Figure B.26: TOA estimation for 1000 Monte-Carlo simulations, traffic noise, low reverberation

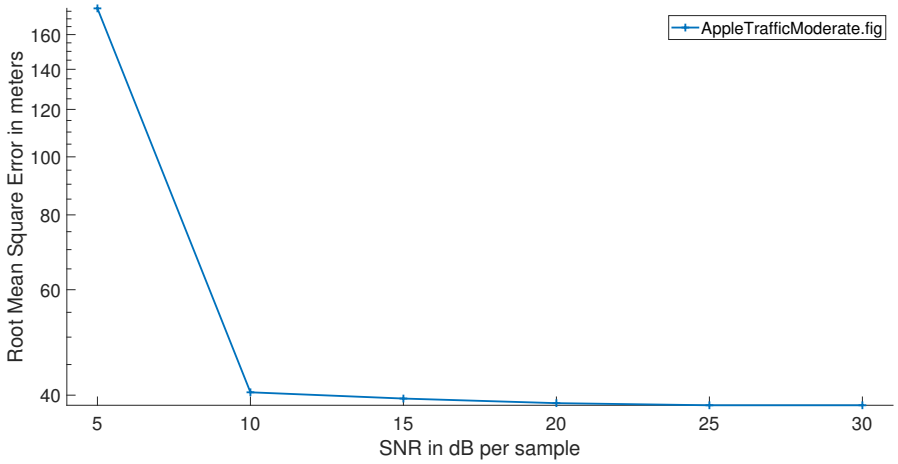


Figure B.27: TOA estimation for 1000 Monte-Carlo simulations, traffic noise, moderate reverberation

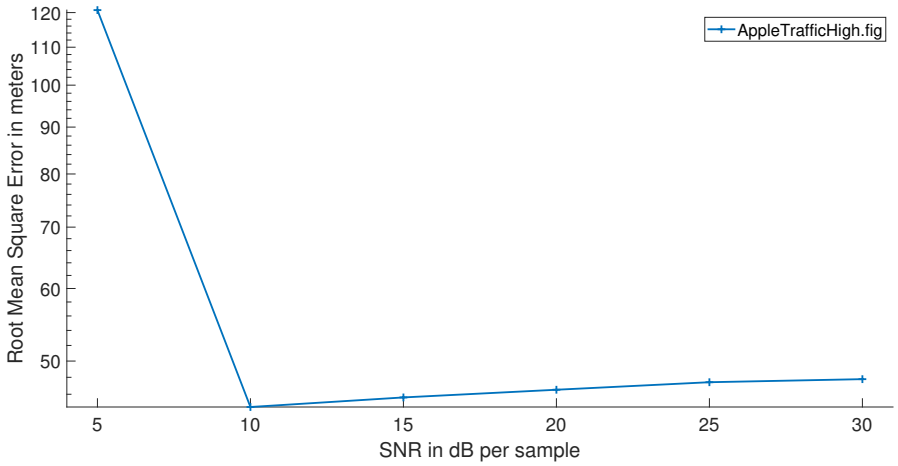


Figure B.28: TOA estimation for 1000 Monte-Carlo simulations, traffic noise, high reverberation

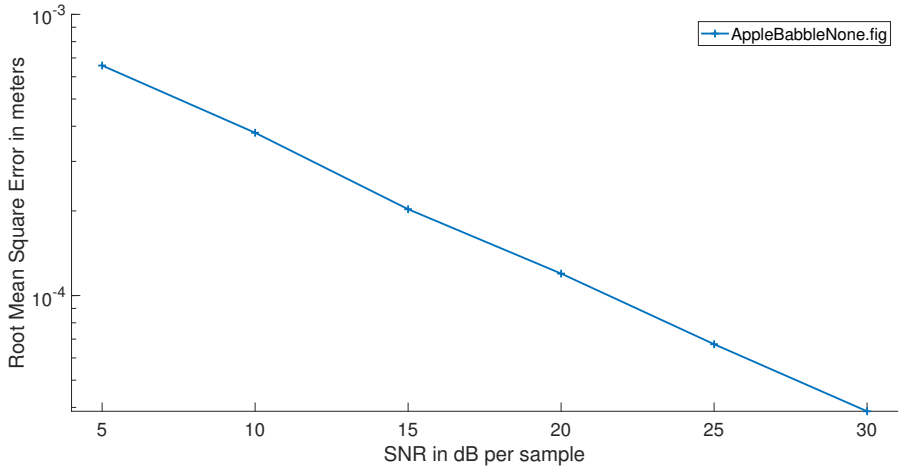


Figure B.29: TOA estimation for 1000 Monte-Carlo simulations, babble noise, no reverberation

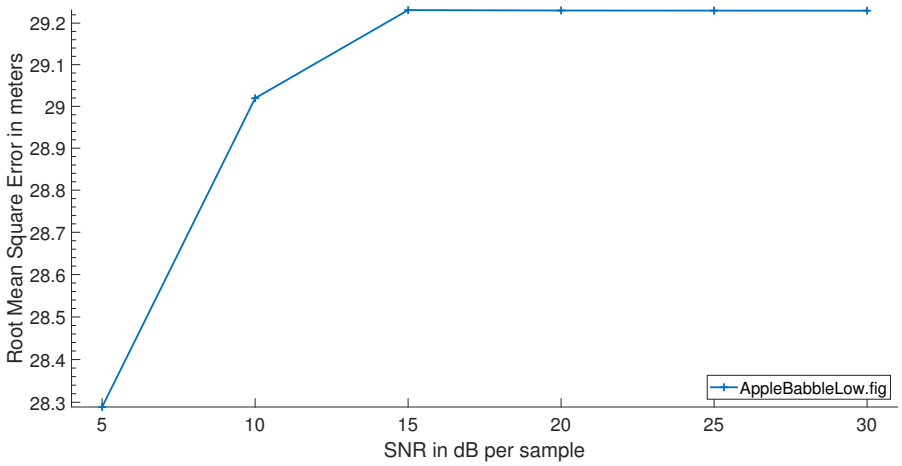


Figure B.30: TOA estimation for 1000 Monte-Carlo simulations, babble noise, low reverberation

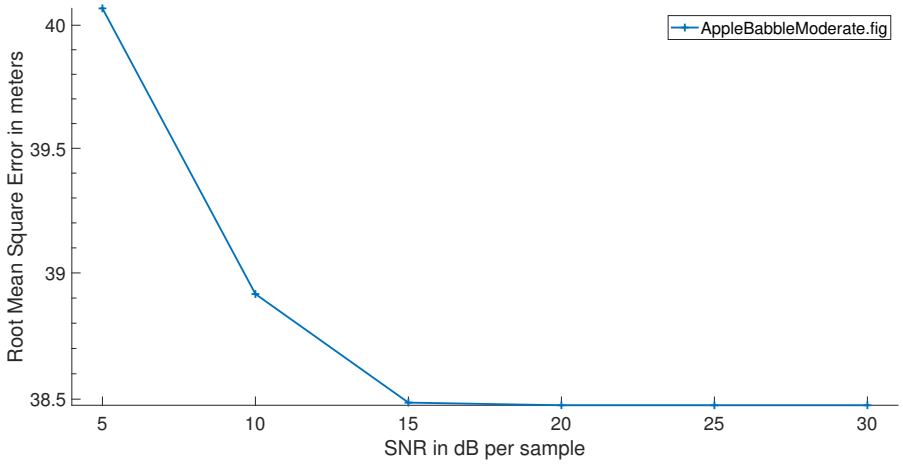


Figure B.31: TOA estimation for 1000 Monte-Carlo simulations, babble noise, moderate reverberation

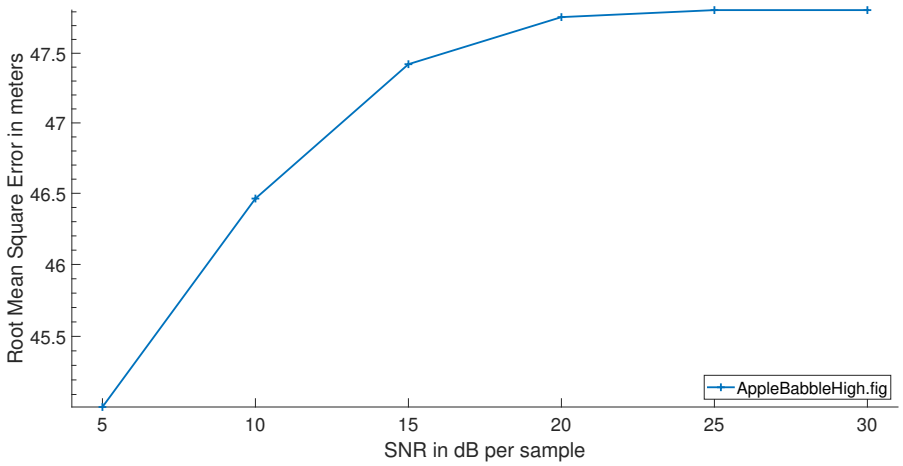


Figure B.32: TOA estimation for 1000 Monte-Carlo simulations, babble noise, high reverberation

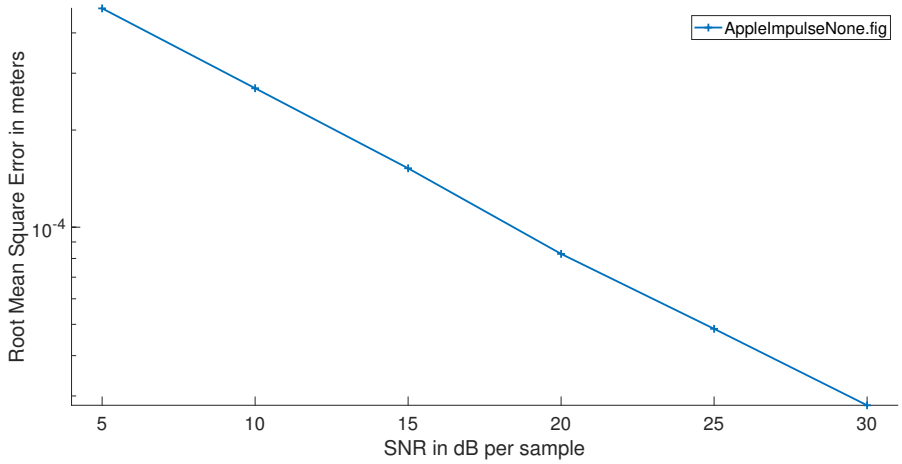


Figure B.33: TOA estimation for 1000 Monte-Carlo simulations, impulse noise, no reverberation

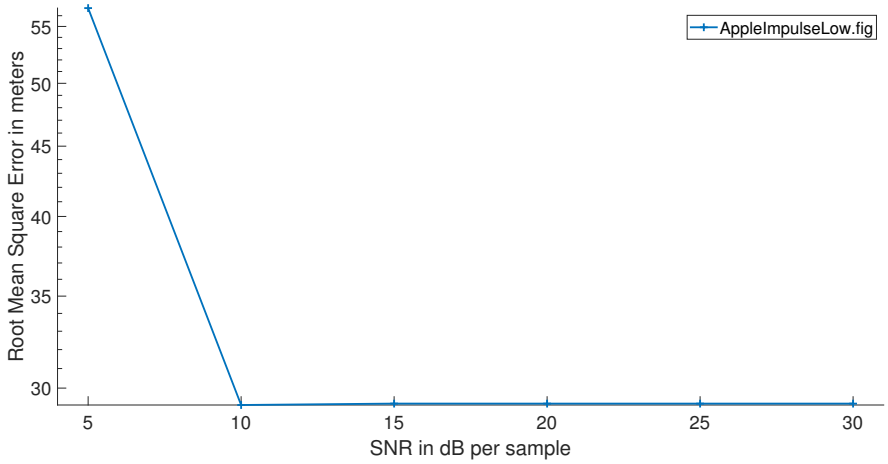


Figure B.34: TOA estimation for 1000 Monte-Carlo simulations, impulse noise, low reverberation

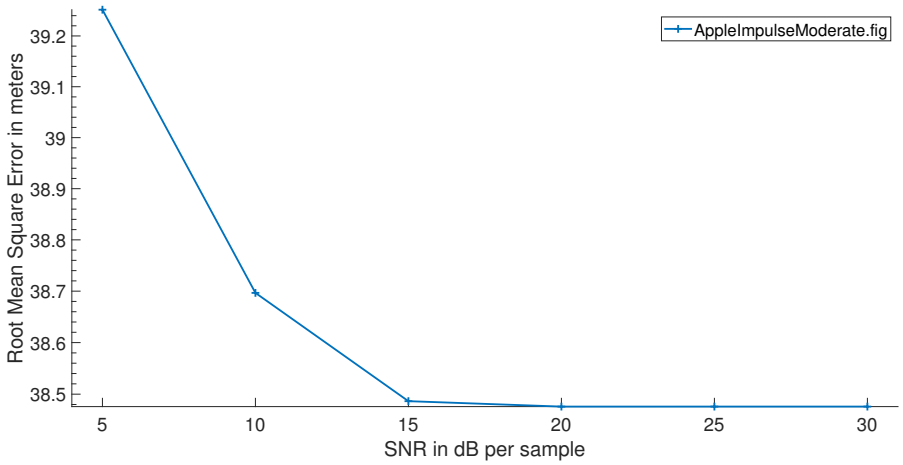


Figure B.35: TOA estimation for 1000 Monte-Carlo simulations, impulse noise, moderate reverberation

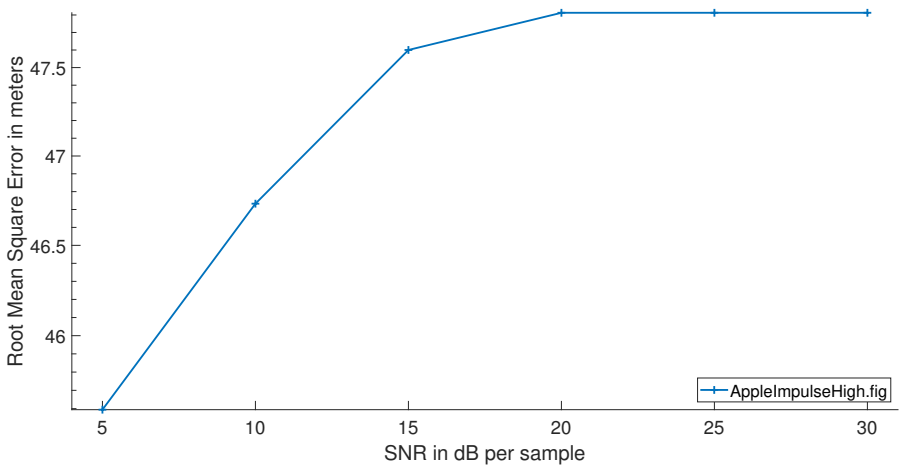


Figure B.36: TOA estimation for 1000 Monte-Carlo simulations, impulse noise, high reverberation

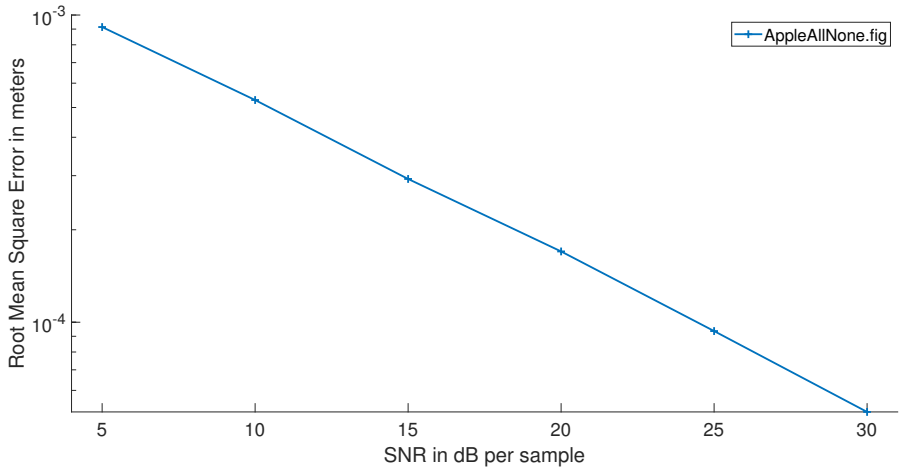


Figure B.37: TOA estimation for 1000 Monte-Carlo simulations, combination of traffic, babble and impulse noise, no reverberation

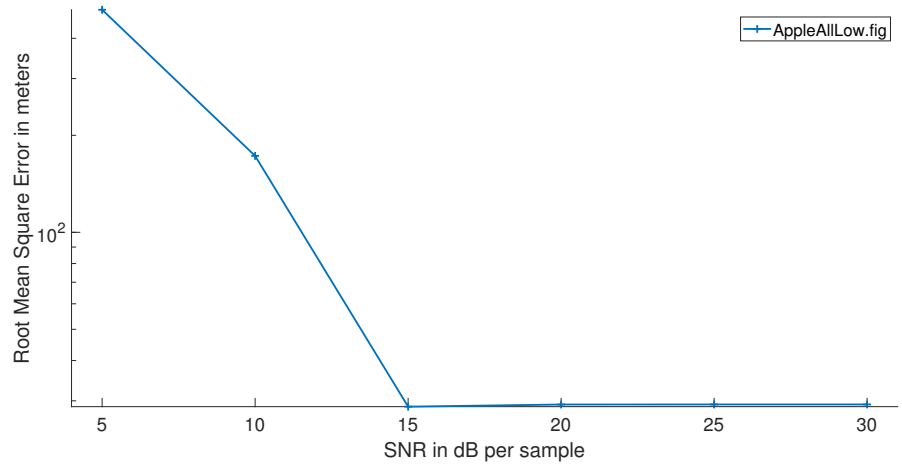


Figure B.38: TOA estimation for 1000 Monte-Carlo simulations, combination of traffic, babble and impulse noise, low reverberation

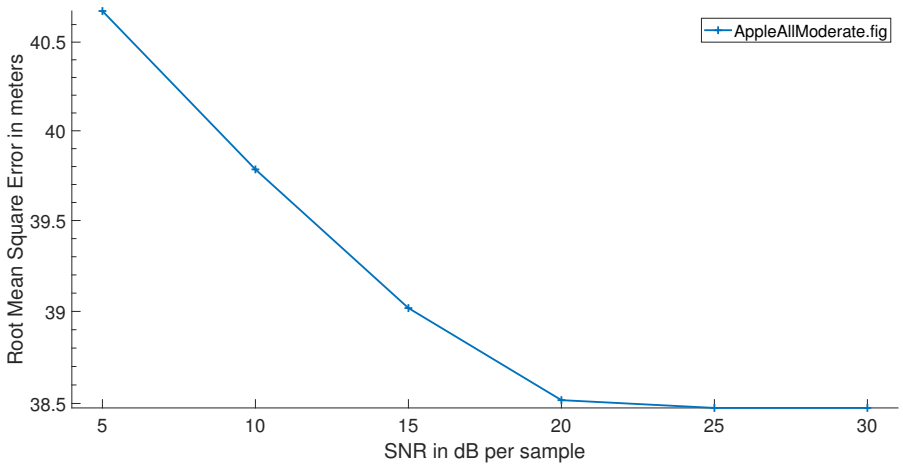


Figure B.39: TOA estimation for 1000 Monte-Carlo simulations, combination of traffic, babble and impulse noise, moderate reverberation

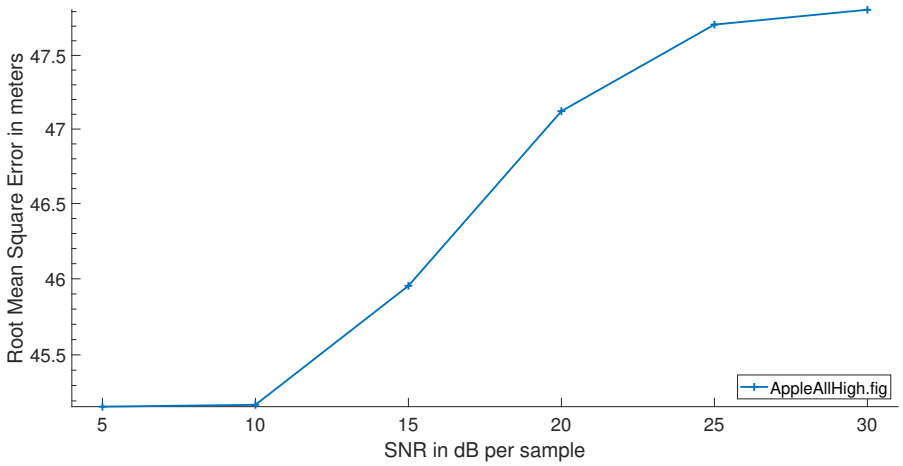


Figure B.40: TOA estimation for 1000 Monte-Carlo simulations, combination of traffic, babble and impulse noise, high reverberation

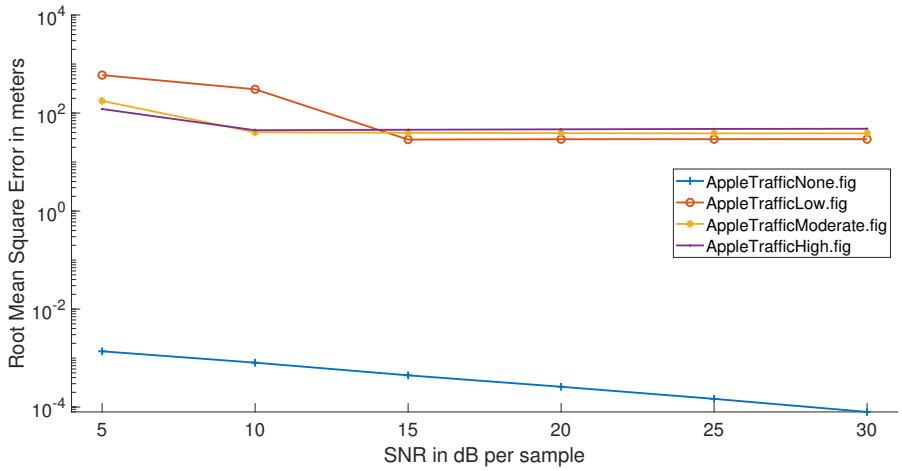


Figure B.41: TOA estimation for 1000 Monte-Carlo simulations, traffic noise, no, low, moderate and high reverberation

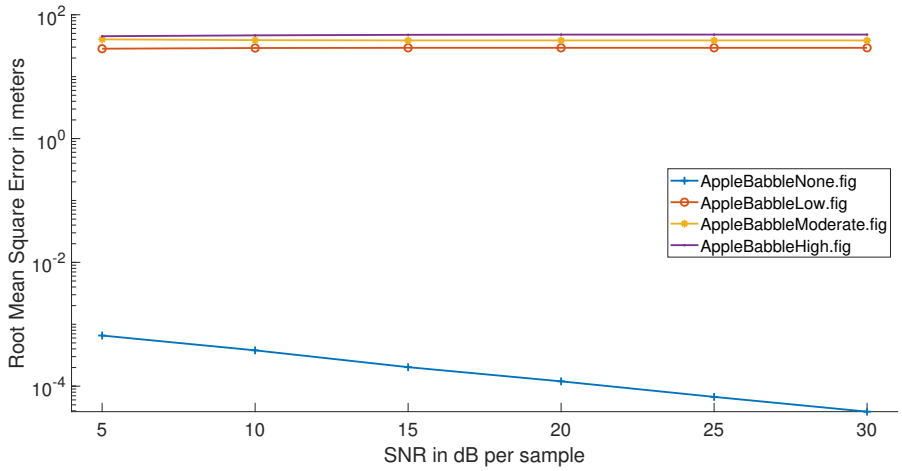


Figure B.42: TOA estimation for 1000 Monte-Carlo simulations, babble noise, no, low, moderate and high reverberation

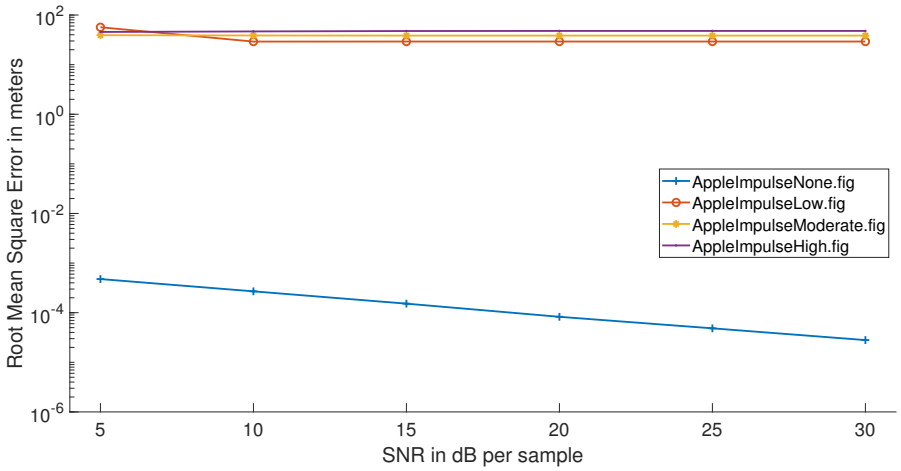


Figure B.43: TOA estimation for 1000 Monte-Carlo simulations, impulse noise, no, low, moderate and high reverberation

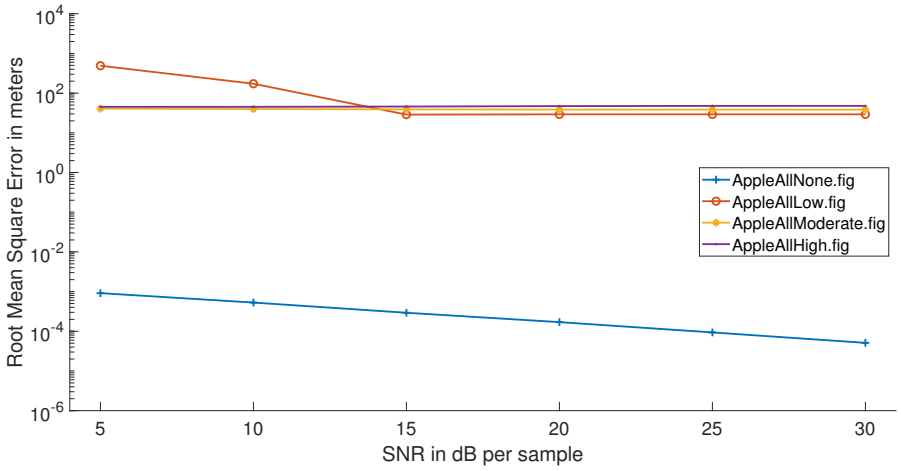


Figure B.44: TOA estimation for 1000 Monte-Carlo simulations, combination of traffic, babble and impulse noise, no, low, moderate and high reverberation

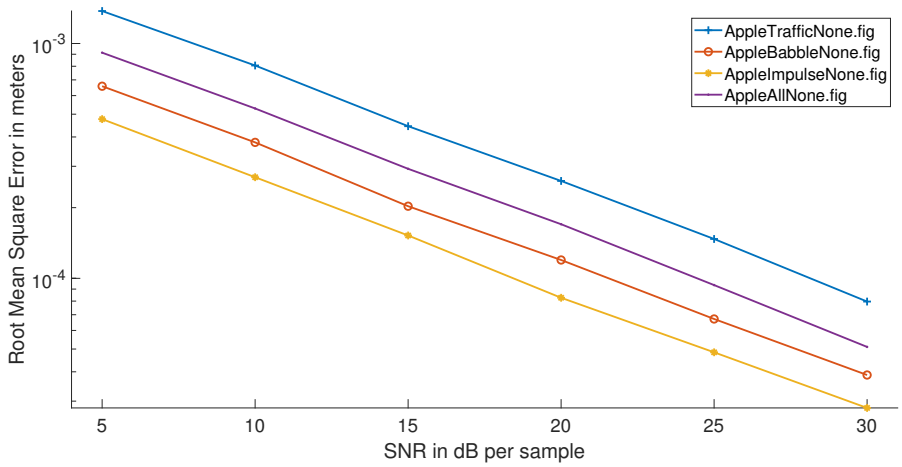


Figure B.45: TOA estimation for 1000 Monte-Carlo simulations, traffic, babble and impulse noise, no reverberation

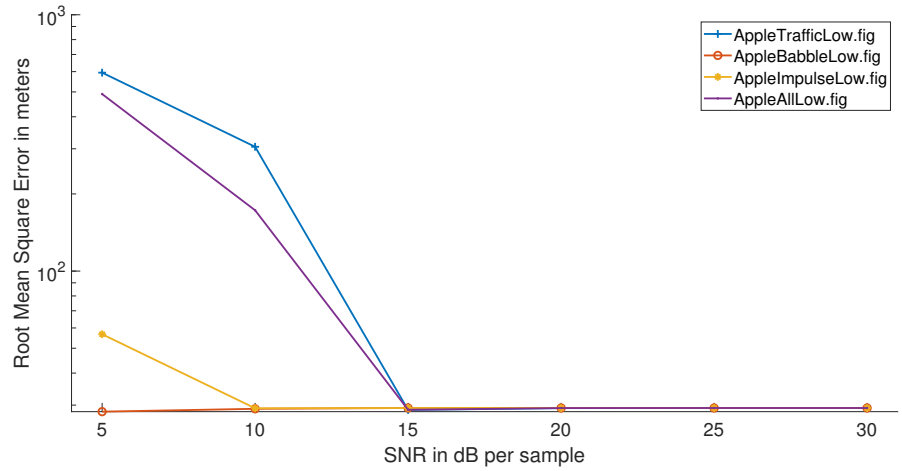


Figure B.46: TOA estimation for 1000 Monte-Carlo simulations, traffic, babble and impulse noise, low reverberation

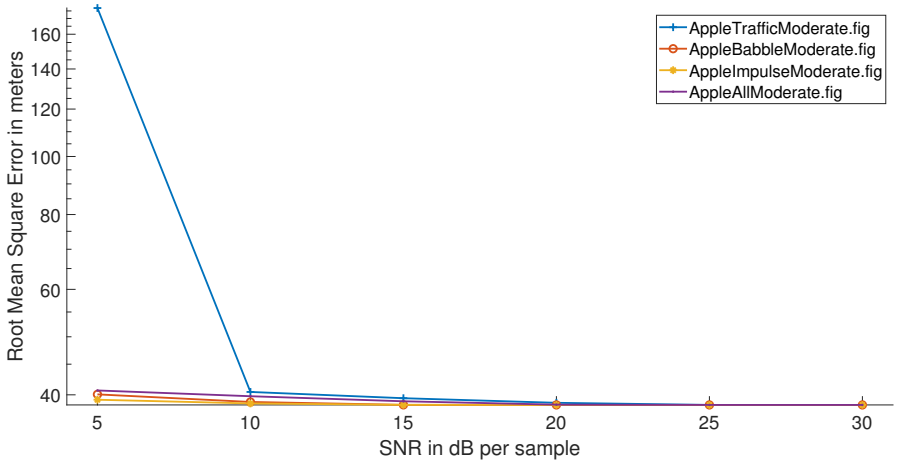


Figure B.47: TOA estimation for 1000 Monte-Carlo simulations, traffic, babble and impulse noise, moderate reverberation

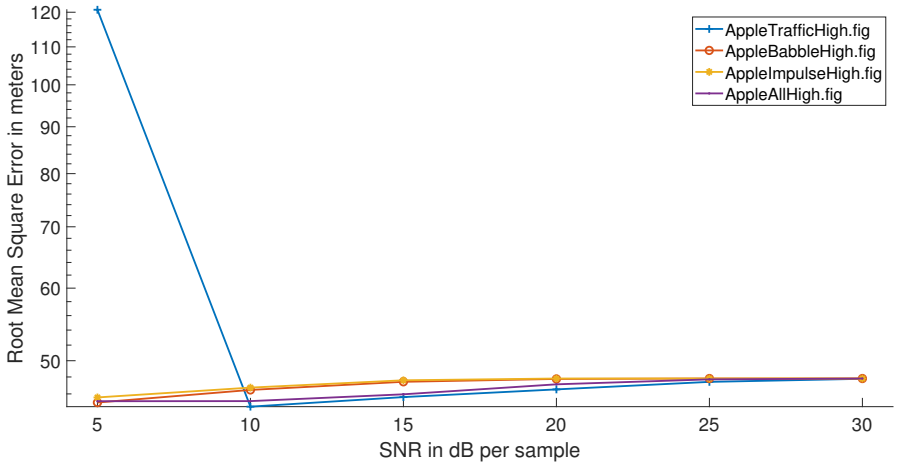


Figure B.48: TOA estimation for 1000 Monte-Carlo simulations, traffic, babble and impulse noise, high reverberation

Appendix C

RIR Measurement Sound Lab

C.1 Measurement Set Up

In the following, the measurement procedure of measuring the RIRs the sound lab in the CREATE building at Aalborg University in Aalborg is described. The loudspeaker used for RIR measurement was the Brüel and Kjær Omnisource type 4295. It was connected to an Yamaha natural sound stereo amplifier AX-492. The amplifier was connected to a PreSonus AudioBox 1818VSL which was finally connected to a laptop. The recording microphone was the G.R.A.S 40PP CCP Free-Field QC microphone. An exponential sine sweep was chosen for the excitation signal. The measurement set up was calibrated so that the distance from loudspeaker and microphone and also the delay caused by the measurement equipment was corrected, resulting in the RIR starting directly without any delay. The measurement procedure was conducted based on [13]. Two loudspeaker positions and were chosen. For every loudspeaker position three microphone positions were measured and every RIR at a specific set up was an average of three measurements. All measurement were done using the ITA-toolbox for Matlab [7] Microphone and loudspeaker had variable heights. Neither the loudspeaker, nor the microphone was less than one meter away from any objects, walls, floor or ceiling. The objects remaining in the room during the measurement were eight loudspeakers on a stand, a multichannel sound card and the measurement equipment on a movable table. Figure C.1 shows a photograph of the measured sound lab.

C.2 Room Impulse Responses

In total six RIRs were measured. Figure C.2 depicts one of them and figure C.3 depicts the corresponding transfer function. Figure C.4 depicts the calculated early decay time from one of the RIRs.

Figure C.1: RIR measurement set up sound lab

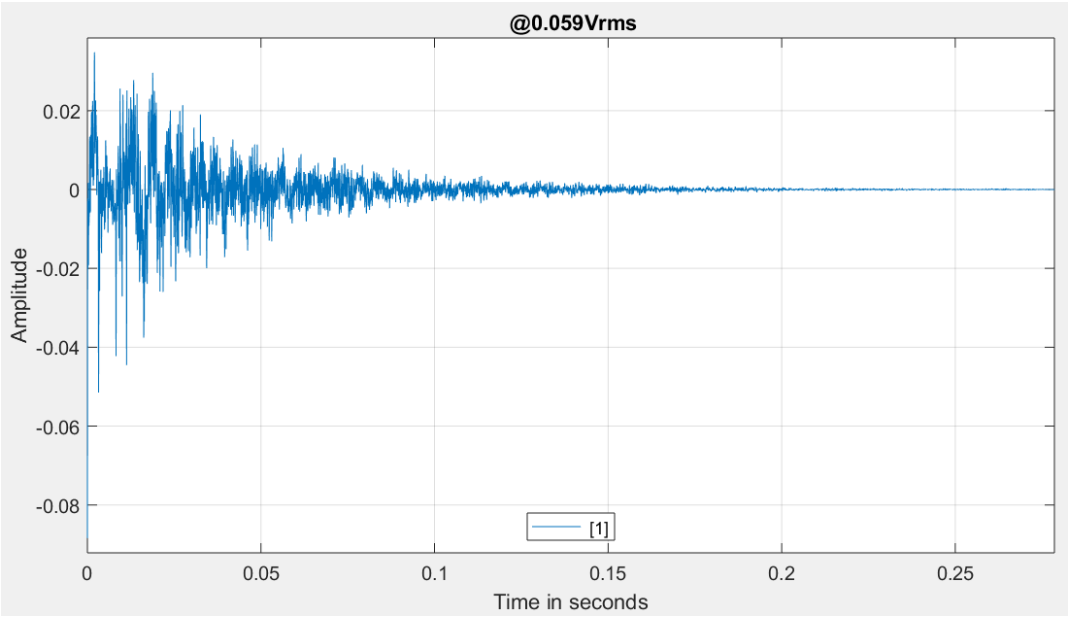


Figure C.2: RIR at one measured position

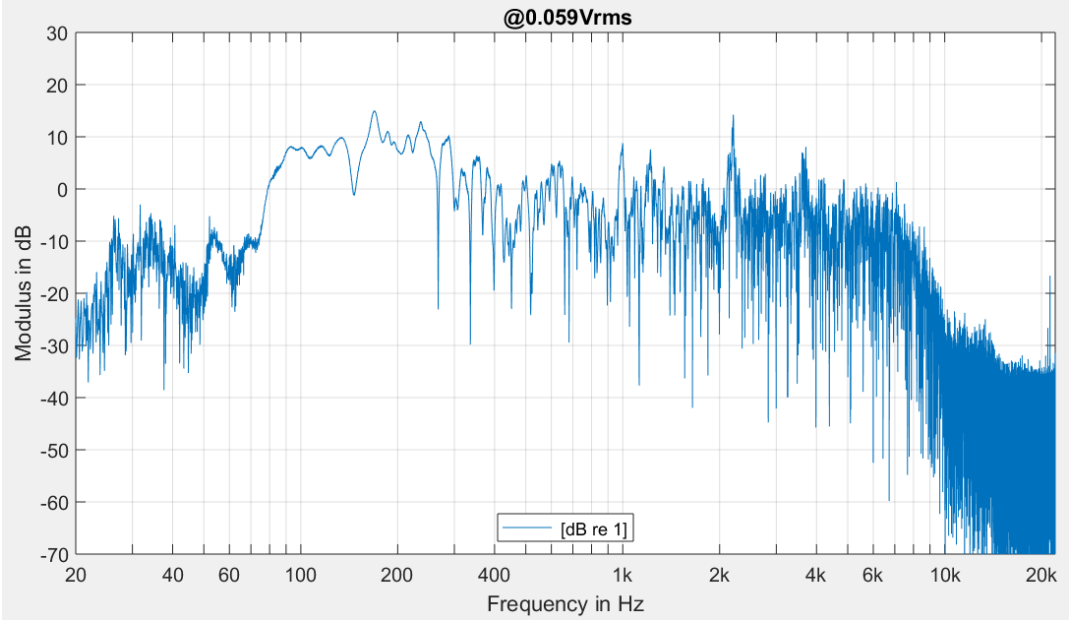


Figure C.3: Transfer function at one measured position

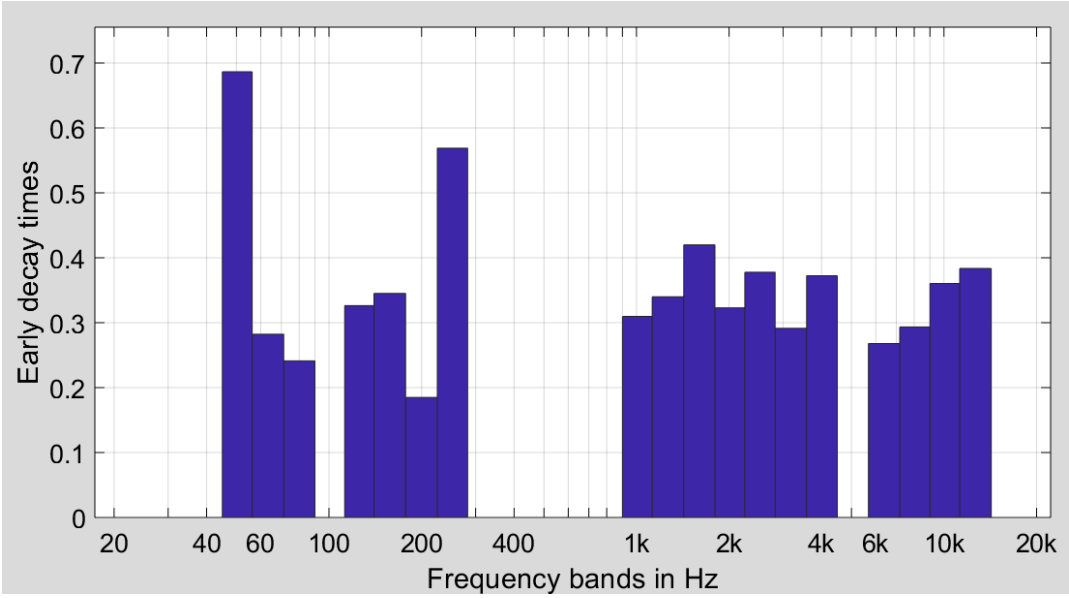


Figure C.4: Early decay time calculated from one RIR

Appendix D

Signal Modification TOA Estimation Performance Results

D.1 Pseudo-Random Noise with Spectral Envelope of Apple Sound

Figure D.1 depicts the TOA estimation performance for the second test scenario with no reverberation present and white noise and a combination of noise types. The only improvement is that the modified pseudo-random noise has already acceptable RMSEs for -10 and -5 dB whereas the original Apple sound has not. Fig-

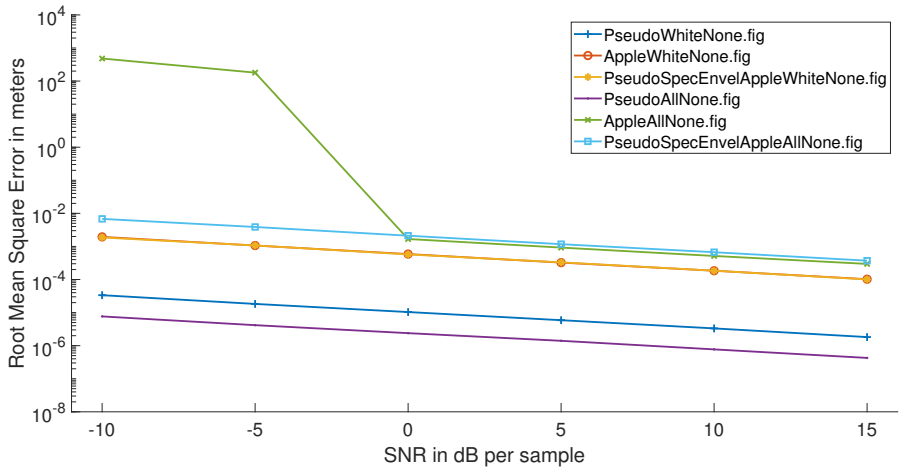


Figure D.1: TOA estimation, second test scenario, pseudo-random noise with spectral envelope of the Apple HomePod start up sound

ure D.2 depicts the third test scenario where a combination of noise types and low reverberation is present. The RMSE of the modified pseudo-random noise did

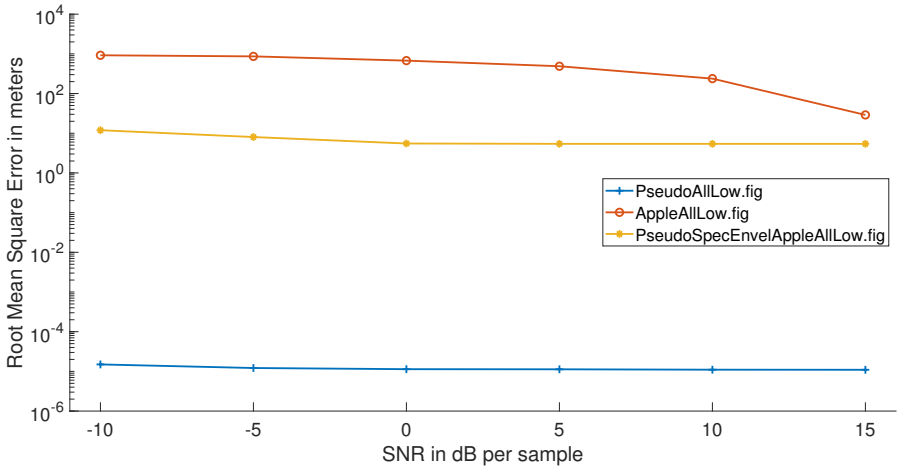


Figure D.2: TOA estimation, third test scenario, pseudo-random noise with spectral envelope of the Apple HomePod start up sound

decrease but not to an acceptable level.

D.2 Apple HomePod start up sound with high frequency pseudo-random noise

In the following figures, the TOA estimation is documented of the pseudo-random noise, original Apple sound and Apple sound with high frequency pseudo-random noise added according but above the masking threshold (factor 58 (factor 1 would be exactly according to the masking threshold)) within the three test scenarios. Figure D.3 depicts the first test scenario with white noise and no, low reverberation. Figure D.4 depicts the second test scenario with no reverberation and white noise and a combination of other noise types. Figure D.5 depicts the third test scenario with low reverberation and a combination of noise types.

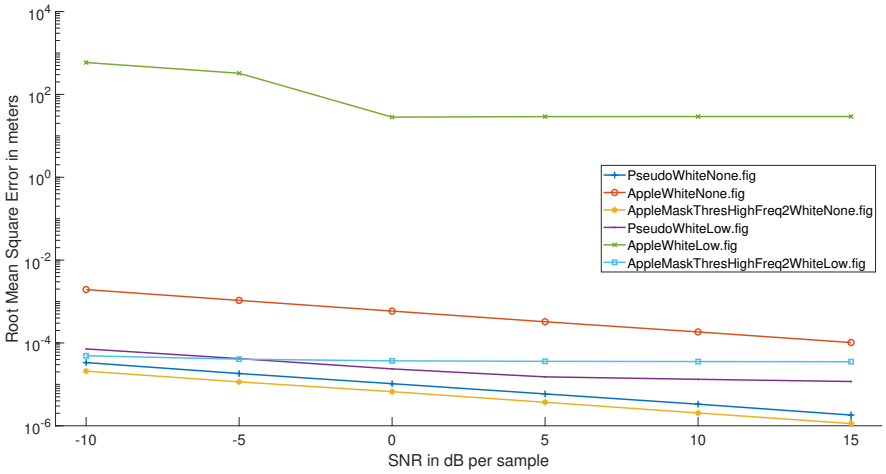


Figure D.3: TOA estimation, first test scenario, Apple sound with high frequency pseudo-random noise (factor 58)

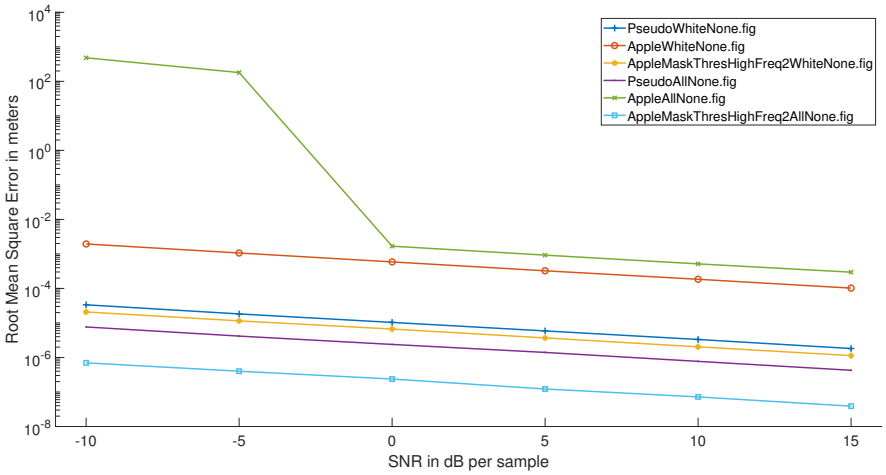


Figure D.4: TOA estimation, second test scenario, Apple sound with high frequency pseudo-random noise (factor 58)

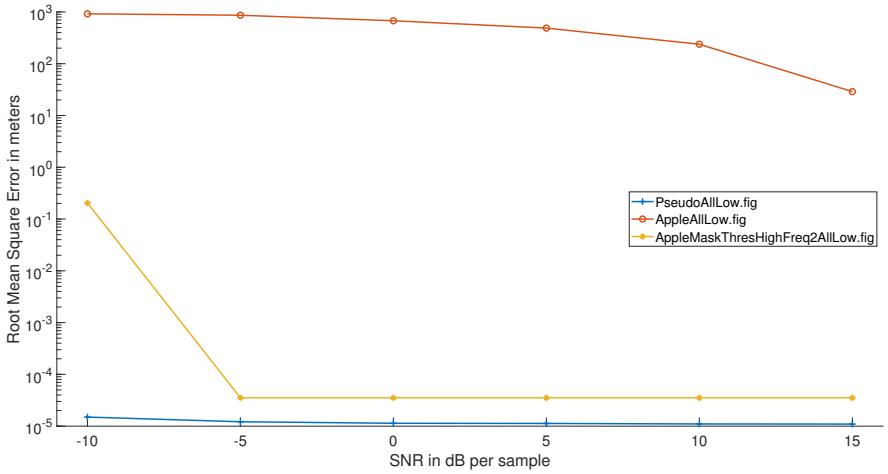


Figure D.5: TOA estimation, third test scenario, Apple sound with high frequency pseudo-random noise (factor 58)