Circuits and Systems
Mekelweg 4,
2628 CD Delft
The Netherlands
http://ens.ewi.tudelft.nl/

CAS-2021-4480767

M.Sc. Thesis

Sound Zones with a Cost Function based on Human Hearing

Niels Evert Marinus de Koeijer B.Sc.

Abstract

Sound Zones with a Cost Function based on Human Hearing

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Electrical Engineering

by

Niels Evert Marinus de Koeijer B.Sc.
born in Delft, The Netherlands

This work was performed in:

Circuits and Systems Group
Department of Microelectronics & Computer Engineering
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Delft University of Technology

Delft University of Technology
Department of
Microelectronics & Computer Engineering

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"Sound Zones with a Cost Function based on Human Hearing"** by **Niels Evert Marinus de Koeijer B.Sc.** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: September 15, 2021

Chairman:

dr.ir. R.C. Hendriks

Daily Supervisor:

dr. J. Martínez-Castañeda

Committee Members:

dr. M. Mastrangeli

M. Bo Møller, PhD

dr. P. Martínez-Nuevo

Abstract

Acknowledgments

I would like to thank dr. J. Martínez-Castañeda , M. Bo Møller, PhD, dr.ir. R.C. Hendriks, and dr. P. Martínez-Nuevo for all their help and support during the project.

A special thanks to M. Bo Møller, PhD for often pushing me in the right direction through our numerous white board sessions. Secondly, I would also like to especially thank dr. J. Martínez-Castañeda for being a major help during, and outside of the project.

Niels Evert Marinus de Koeijer B.Sc.
Delft, The Netherlands
September 15, 2021

Contents

Abstract **v**

Acknowledgments **vii**

1 Introduction **1**
1.1 Preface: the Sound Zone Problem 1
1.2 Objectives and Organization 3
1.2.1 Creation of Perceptual Sound Zone Algorithms 3
1.2.2 Determining Benefits of Perceptual Sound Zone Algorithms . 3

2 Perceptual Model Review and Implementation **5**
2.1 Review of Perceptual Models from Literature 6
2.1.1 Review of Objective Measures 6
2.1.2 Review of Perceptual Models used in Audio Coding 10
2.2 Selection of Perceptual Model for Perceptual Sound Zone Algorithms 13
2.2.1 Desirable Properties of Perceptual Model for use in Perceptual Sound Zone Algorithms 13
2.2.2 Evaluating Reviewed Perceptual Models for use in Perceptual Sound Zone Algorithm 13
2.3 Discussion and Implementation of the Par Detectability Measure . . 15
2.3.1 High-Level Description of the Par Detectability Measure . . 15
2.3.2 Computation Details of the Par Detectability 16
2.3.3 Proposed Least-Squares Formulation of the Par Detectability 17

3 Sound Zone Approach Review and Implementation **21**
3.1 Sound Zone Problem Data Model 23
3.1.1 Room Topology 23
3.1.2 Defining Target Pressure 24
3.1.3 Realizing Sound Pressure through the Loudspeaker 25
3.1.4 Choice of Target Pressure 25
3.2 Review of Sound Zone Approaches 27
3.2.1 Pressure Matching 28
3.2.2 Acoustic Contrast Control 29
3.3 Proposed Perceptual Sound Zone Approach 31
3.3.1 Analysis of Mathematical Properties of Proposed Least Squares formulation of Par Detectability Measure 31
3.3.2 Introducing Proposed Perceptual Sound Zone Approach . . . 32

4 Implementation of Proposed Perceptual Sound Zone Algorithms **35**
4.1 Proposed Short-Time Frequency-Domain Reformulation of the Pressure Matching Approach 36
4.1.1 Proposed Block-Based Pressure-Matching Algorithm 36

4.1.2 Proposed Block-Based Frequency-Domain Pressure-Matching Algorithm . . . 39
4.2 Proposal of Proposed Perceptual Pressure Matching Algorithms . . 42
4.2.1 Proposed Unconstrained Perceptual Pressure Matching Algorithm . . . 42
4.2.2 Proposed Constrained Perceptual Pressure Matching Algorithm 43

5 Evaluation of Perceptual Sound Zone Algorithms 45
5.1 Evaluation Methodology . . . 46
5.1.1 Room Setup . . . 46
5.1.2 Content Selection . . . 46
5.1.3 Simulation Outputs . . . 47
5.1.4 Evaluation Criteria . . . 47
5.2 Evaluation of Proposed Algorithms . . . 48
5.2.1 Evaluating Unconstrained Perceptual Pressure Matching . . 49
5.2.2 Evaluating Constrained Perceptual Pressure Matching . . . 50

6 Conclusion 53

A Calibration of the the Par Detectability Measure 59
A.1 Relating Digital Representation and Sound Pressure Level . . . 59
A.2 Determing Calibration Constants . . . 60

B Extra Results 63

1 Introduction

1.1 Preface: the Sound Zone Problem

Sound systems are used world wide to fill rooms with enjoyable audio content. Problems arise however when multiple people in the same room want to enjoy different audio content at the same time.

For example, one person may want to enjoy a movie on the television, while another may want to listen to their music. If they are in the same room, their desires clash: neither person can fully enjoy their chosen activity without disturbing the other. In short, the interference of multiple source of audio leads to a situation where both individual experiences are diminished.

In recent years, attempts have been made to solve this problem by controlling the spatial reproduction of sound in such a way that different areas in a room have distinct content.

One class of algorithms that attempt to do so are sound zone algorithms [1]. Sound zone algorithms partition the space of the room into multiple so-called sound zones. Each sound zone is assigned different audio content. The sound zone algorithms decides how to use the loudspeakers of the sound system to reproduce said audio content in every zone. Using the principals of constructive and destructive interference, this is done in such that there is minimal interference between zones. That is to say: the audio content of each zone is not audible in the others.

In the previously listed example, one zone would contain the audio of a movie and another zone would contain music. An image depicting the situation is given in Figure 1.1. The sound zone algorithm determines how to best use the sound system to reproduce these two zones. In the ideal case, both people can now enjoy the full potential of their audio content, without bothering one another.

In practice however, the sound zone algorithm will not always do a perfect job. The performance of algorithms depends on the environment and the available sound system. Depending on the situation, the interference between zones can typically only be reduced by so much. As such, audio content of one zone is often still audible in other zones.

Improving sound zone algorithms is thus still an active topic of research. One recent approach is to include a model of the human auditory system, which models how sound is perceived by humans. Typically, sound zone algorithms use sound pressure, which is a physical quantity characterizing the sound. Sound pressure does not always accurately describe what is important for the perception of sound.

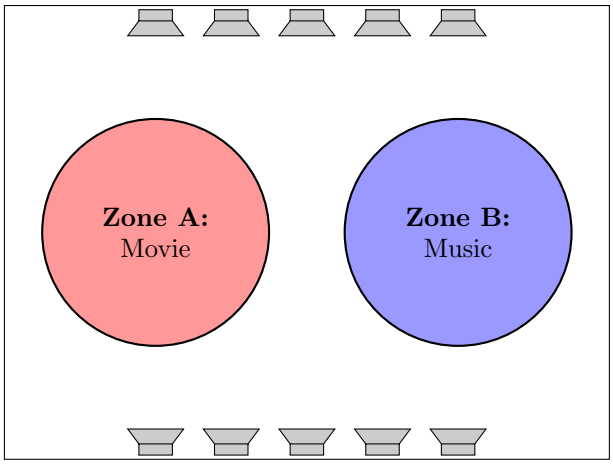


Figure 1.1: A room containing a sound system consisting of an array of loudspeakers and two zones. The goal of the sound zone algorithm is to control the sound system in such a way that the red zone contains the audio of a movie, and the blue zone contains the music.

As such, including a perceptual model may allow the algorithm to focus on the parts of the audio content that matter perceptually.

Early results show that the perceptual sound zone approach is promising. Recent work by Donley et al. explored including the absolute threshold of hearing, which models the lowest sound pressure humans can hear, into sound zone algorithms. This pursuit found an increased quality of the reproduced audio in the zones [2]. Other work by Lee et al. showed that including a perceptually-motivated weighting in the sound zone algorithm outperforms traditional algorithms [3, 4].

This work seeks to further explore this perceptual approach by proposing a novel perceptual sound zone algorithm and exploring the benefits of including perceptual information.

1.2 Objectives and Organization

This section states the goals of the thesis and organization of the rest of this document.

This thesis seeks to answer two research questions:

- **RQ1:** *"How can auditory perceptual models be included in sound zone algorithms?"*
- **RQ2:** *"What are the benefits of including auditory perceptual models in sound zone algorithms?"*

What follows is how these research questions will be answered, alongside the structure of the rest of this document.

1.2.1 Creation of Perceptual Sound Zone Algorithms

The first research question RQ1,

"How can auditory perceptual models be included in sound zone algorithms?"

is answered in chapter 2, chapter 3, and chapter 4. These chapters document the design of a perceptual sound zone algorithm. The chapters are structured as follows.

- First, in chapter 2 a literature review is performed to determine which perceptual models are suitable for use in a sound zone algorithm. In this pursuit, one perceptual model is found to be the most promising. This model is selected for use in the perceptual sound zone algorithm and is discussed in further detail in this chapter.
- Next, in chapter 3 motivates a perceptual sound zone approach. This is done by reviewing literature to determine which sound zone approaches exist and reflecting on the mathematical properties of the selected perceptual model.
- Finally, in chapter 4 implements two perceptual sound zone algorithms using proposed perceptual approach.

1.2.2 Determining Benefits of Perceptual Sound Zone Algorithms

The second research question RQ2,

"What are the benefits of including auditory perceptual models in sound zone algorithms?"

is answered in chapter 5. This is done by comparing the perceptual sound zone algorithms derived in answering RQ1 with a reference sound zone algorithm.

Perceptual Model Review and Implementation

2

TODO: Add required perceptual information.

Chapter Structure

The goal of this chapter is to find a suitable perceptual algorithm for the creation of a perceptual sound zone algorithm. This chapter is structured as follows.

- This chapter begins with a literature review into perceptual models is given in section 2.1. The purpose of this review is to document candidates for the perceptual model that will be used in the perceptual sound zone algorithm. In addition to this, the reviewed models could also serve as potential candidates for use in the evaluation of the perceptual sound zone algorithm that will be proposed in chapter 4.
- To perform the selection of a perceptual model from the candidates discussed in the literature review, criteria reflecting desirable properties for the model for use in sound zones are be defined in section 2.2. The criteria are then used to select a perceptual model.
- Afterwards, the selected perceptual model is discussed in more detail in section 2.3 by stating implementation details and describing its behavior.

2.1 Review of Perceptual Models from Literature

This section documents a literature review into perceptual models of the human auditory system. This review is used in section 2.2 to determine which perceptual model is most suitable for use in a perceptual sound zone algorithm.

Especially promising are models that attach some "score" or "rating" to the perceptual quality of input signals. These ratings can be used in algorithms to obtain an optimal rating through optimization. In addition to this, they can also be used to quantify the quality of results from later algorithms.

As such, the focus of the literature review is not on the latest findings in the field of psycho-acoustics, or models that most accurately emulate the behavior of the human ear, but rather on algorithms that quantify quality.

To this end, in this section, two categories of perceptual models are considered. First in subsection 2.1.1, "objective measures" are discussed. These are models which attempt to predict the perceptual quality ratings from listening tests. Next, perceptual models from "audio coding" are discussed in subsection 2.1.2. These models are used to quantify how audible the artifacts of audio compression are.

2.1.1 Review of Objective Measures

In order to objectively determine the perceived quality of audio one approach is to use listening tests. These are tests in which subjects are asked to rate a property (or properties) of a set of audio stimuli. One example where listening tests are used is for the evaluation of speech intelligibility of hearing aids [5]. Another example is determining which loudspeaker has higher perceived sound quality.

Performing listening tests is however often cumbersome due to the large amount of human labour involved. This motivates the use of objective quality measures, which attempt to predict the outcomes of these objective listening tests. This is very useful for algorithm developers, as they can get an indication of how well they are doing without having to perform a labour intensive listening test [5].

Note however that a objective quality measure does not replace a listening test: it can only be used to give an indication. Findings should always be confirmed with listening tests.

The objective measures that are considered in this review take a reference and degraded audio stimuli as inputs. Most of the discussed models take the following approach. First, input stimuli are converted to their so-called internal representations, which models how the human auditory system transforms the stimuli. Various features are then derived from this internal representation. The features are then mapped to a prediction of the results of a listening test.

These objective quality measures are promising for integration into sound zone algorithms as they summarize the quality of a signal into a single value, which can be potentially optimized for. It stands to reason that if an objective quality measure correlates with audio quality, optimizing over such a measure could improve

the sound quality of sound zone algorithms.

As such, this section explores various objective measures. This is done by considering three classes different objective measures, namely: measures that quantify the quality and intelligibility of speech audio and measures for the general quality of audio.

2.1.1.1 Review of Objective Speech Quality Measures

There have been a number of attempts to create objective measures to quantify the perceived quality of speech. In this section three objective speech quality measures are discussed. Namely the Perceptual Evaluation of Speech Quality (PESQ) [6] measure, Perceptual Objective Listening Quality Assessment (POLQA) [7]. measure, and Virtual Speech Quality Objective Listener (ViSQOL) [8, 9] measure.

- PESQ is a metric which attempts to determine the perceived quality of speech. It was standardized by the International Telecommunication Union (ITU-T) in 2001. PESQ is computed by first applying an auditory transform that maps the reference and degraded speech into a time-frequency representation that models the perceived loudness of the signals. From this internal representation, so-called symmetric and asymmetric disturbances are determined by computing differences between the time-frequency bins of the reference and degraded speech. A non-linear average is then taken to obtain the average disturbance per time bin. These averaged disturbances are then mapped to the outcomes of listening test outcomes through linear combination [6].

- POLQA is speech quality metric which was standardized by the International Telecommunication Union (ITU-T) in 2011. It was intended to be the successor of PESQ, with the improvement of having more accurate predictions on a wider range of distortions. POLQA works with a similar internal representation to PESQ, but computes distortion in a different way as to be capable of handing global temporal compression and expansions [7].

- ViSQOL is a metric developed in 2012 in a collaboration between Trinity College and Google. ViSQOL uses a different internal representation than PESQ and POLQA as it uses neurograms rather than loudness representations. Neurograms contain the neural firing activity of the auditory nerve in time-frequency bins, and NSIM determines how similar the firing patterns of two neurograms are. The neurograms are then compared by means of the Neurogram Similarly Index Measure (NSIM). This similarity is then related to the outcomes of listening tests through a laplacian fit [8], which is then used to make predictions.

In general, PESQ, POLQA and ViSQOL require many steps to compute and were found difficult to optimize for due to conditional branches within the algorithms and many non-differentiable steps such as clipping [6, 7, 8]. Some attempts have been made however to reformulate PESQ in order to make it more tractable for optimization by approximating the disturbances by other functions [10].

2.1.1.2 Review of Objective Speech Intelligibility Measures

Intelligibility of speech is defined as the percentage of words identified correctly given a degraded speech signal. Objective speech intelligibility metrics seek to predict this percentage. In this section, two of these metrics are discussed. Namely, the Short Time Objective Intelligibility (STOI) [5] measure and the Speech Intelligibility In Bits (SIIB) [11] measure.

- STOI was proposed by Taal et al. in 2011 as a speech intelligibility metric that could make accurate predictions for a speech signals that were degraded time-frequency weighted distortions.

  For its internal representation, it finds a time-frequency internal representation through filtering the input stimuli with a filter bank consisting of 1/3 octave bands, and then segmented the filter taps into short time frames. Silent bins that do not contain speech are removed, and clipping is applied to limit the effect of one severely degraded time-frequency bin. The average correlation coefficient between the time-frequency bins of the internal representation of the reference and degraded segments is then computed, and averaged over all bins to determine the intelligibility [5].

- SIIB was introduced by Van Kuyk et al. in 2017 as a speech intelligibility metric that could be motivated through information theory through the mutual information rate. As such, SIIB is given in bits

  The idea behind SIIB is that the intelligibility of speech is related to the information shared between intended and degraded speech. SIIB models how transformation of the reference speech signal to the degraded speech signal as a transmission channel. Among other aspects, this transmission channel includes a model of the human auditory system [11]. This communication channel is then used to compute the mutual information rate.

Both STOI and SIIB are difficult to optimize for directly.

In STOI, the removal of silent regions and the clipping operator are non-differentiable operations. Furthermore, the computation of the correlation coefficient is a non-convex function of the degraded speech [5].

SIIB is in general non-convex and non-differentiable as it uses the Karhunen-Loève transform and a K-nearest neighbor estimator to compute the mutual information rate. However, if the communication channel is approximated as gaussian, the mutual information can be computed in closed form, and SIIB becomes a differentiable measure [11].

2.1.1.3 Review of Objective Audio Quality Measures

The previous objective quality metrics are both intended for evaluating speech. In this section, a number of objective quality metrics are discussed that are designed instead for evaluating the perceived quality of general audio . Namely, the Perceptual Evaluation of Audio Quality (PEAQ) [12] and ViSQOLAudio [13]. The latter is an adapted version of the ViSQOL speech quality measures.

- PEAQ is a audio quality metric standardized by the International Telecommunication Union (ITU-T). PEAQ estimates a quality grade by first computing an internal representation of the reference and degraded audio signals. This results in a time-frequency representation of the input stimuli from which a number of perceptually relevant feature, referred to by PEAQ as Model Output Variables (MOVs), are extracted. An example of these MOVs are the loudness of the noise or the bandwidth of the input stimuli. These MOVs are then mapped to the final audio quality grade through a neural network [12].
- In 2015, it was found that with some adjustments ViSQOL could be used to determine audio quality, which resulted in a new metric ViSQOLAudio. Among the adjustments were the removal of the voice activity detector included in ViSQOL and the use of a larger bandwidth to cover the entire spectrum of hearing from 50 Hz to 20000 Hz, rather than just the bandwidth of speech [13].

PEAQ, and ViSQOLAudio are both difficult to optimize for. A number of the MOVs computed in PEAQ, such as the partial noise loudness, are non-differentiable [12]. As ViSQOLAudio is similar to ViSQOL with some small adjustments, it is similarly difficult to optimize for [13].

2.1.1.4 Review of Distraction Model

One especially promising objective measure is the distraction proposed by Francombe et al. in 2015 [14]. This measure was designed with the application of sound zones in mind.

The distraction was determined to be the keyword that best describes the perceptual experience of interfering audio programs. This was determined through an elicitation study performed also performed by Francombe et al. in 2014 [15]. This prompted the creation of the model.

To create the model, a listening test was performed where the participants were subjected to audio-on-audio interference. The subjects were played a target audio stimuli they were instructed to focus listening to. At the same time, an interferer audio stimuli was played to distract the participant from the target. The participants were given a scale between 0 and 100 on which they were asked to rate how distracting the interference was when listening to the target program, where a 100 indicates that the interferer "overpowered" the target audio [14].

The target-interferer stimuli pairs and corresponding ratings resulted in a dataset. This dataset was then used to fit a model which predicted the distraction given novel a target-interferer stimuli pair. The model consisted of taking a linear combination of 5 features that were computed from the stimuli [14].

Computing said features could however not be performed in real time. The reason for this was that as the original distraction model is too computationally complex [16]. To this end, in 2017, Rämö et al. proposed a version of the distraction model that could be run in real-time. This was done by approximating the features of the original distraction model by computationally less complex alternatives. The

resulting real-time distraction model was found to be less precise, but could be run in 0.04% of the time of the original distraction model [16].

On face value, the real-time distraction model seems promising to optimize over. However, while easy to compute, the model is non-differentiable as the model uses piecewise functions and non-convex due to taking the logarithm of the square of the input signals. In addition to this, the model also performs operations that are difficult to express mathematically, such as counting the number of short-time blocks that exceed a certain threshold [16].

2.1.2 Review of Perceptual Models used in Audio Coding

The second class of perceptual models that are considered are the perceptual models used in audio coding. Audio coding algorithms attempt to find an low-bitrate representation of an audio input signal, which is a form of lossy compression. As such, audio coding algorithms typically introduces errors in doing so, which can be a detriment to the listening experience.

To minimize the impact of these errors, many audio coding algorithms use a perceptual model to quantify how disturbing the introduced distortions are. The perceptual model is used to introduce encoding errors in such a way that the audio output signal is perceptually indistinguishable from the audio input signal [17]. This model typically takes form of a distortion function which determines how audible the difference between a reference input audio signal and a distorted output audio signal is. This function can used to for example encode an input audio signal such that it has minimal distortion for a specified bitrate.

The perceptual models used in audio coding are promising for integration into a sound zone algorithm, as they are often mathematically tractable. As stated, these perceptual models typically take the form of some sort of distortion function that quantifies how perceptually disturbing the introduced artifacts are. One approach for example could be to define sound zone algorithms that minimize a distortion function.

As such this section explores three perceptual models from audio coding.

2.1.2.1 Review of Perceptual Models from ISO MPEG Standard

The ISO/IEC 11172-3 standard specifies a coded representation for audio files [18], and a corresponding decoder. An encoder said representation is not part of the standard. This is done deliberately, to allow for future improvements to the encoder, without having to change the standard [19].

The standard does however provide a number of examples of possible encoders, with increasing complexity. Alongside these example encoders, two psycho-acoustical models are included for use during the encoding process.

The psycho-acoustical models work by subdividing the input audio signal into frequency bands which correspond to the frequency bands in the human auditory system. The model then determines how much quantization noise can be added

separately per band without the noise becoming audible. As such, the model assumes that the distortion signal is noise-like [20], which is usually the case for quantization noise for audio coders.

The output of the psycho-acoustical model is thus the amount of noise that can be added per band. In the case of audio coding, this can then be used to control quantization noise. Note that this perceptual model does not come in the form of the earlier described distortion function. This technique has however been used for various signal processing purposes, such as audio watermarking [17]. As such, examples exist from which optimization schemes could be inspired.

2.1.2.2 Review of Par Detectability Measure

In 2005, van der Par et al. proposed a novel perceptual model designed for use in audio coding [20]. The model defines a distortion measure which determines the "detectability" of a distortion signal in presence of masking signal. That is to say, the function quantifies the degree to which a human is to detect a distortion signal. For audio coding purposes, this distortion signal is error introduced due to the audio compression.

Similarly to the ISO MPEG perceptual models, the Par detectability typically operates on short-time segments, typically in the order of 20 to 200 milliseconds. The proposed method however differentiates itself from the previously discussed ISO MPEG models in three ways.

Firstly, the paper uses newer findings from psycho-acoustic literature, namely spectral integration. In spectral integration, the masking effects from neighboring bands are taken into account when computing the masking effects. The psycho-acoustical models defined in the ISO MPEG standard does not do this as it effectively works independently per band [17].

Secondly, it assumes that the distortion signal is sinusoidal, rather than noise-like. As such, it is more effective in hiding sinusoidal distortion.

Thirdly and finally, the perceptual model is described as a distortion function which quantifies how detectable a disturbance stimuli is. The proposed distortion measure can be expressed as a squared L2-norm.

This mathematical tractability makes for easy integration into existing least-square problems. As such, the Par model has been used in many signal processing applications, examples ranging from speech enhancement to removing perceptually irrelevant sinusoidal components [21, 22].

2.1.2.3 Review of Taal Detectability Measure

A paper from 2012 by Taal et al. proposed a novel perceptual model [17] which introduces a alternative definition to the detectability defined in the Par model.

In contrast to the Par detectability, the Taal detectability measure takes temporal characteristics of a signal into account. The inclusion of temporal information allows for the suppression of "pre-echoes", which is an artifact that the Par model suffers

from. The "pre-echoes" artifacts arises from the assumption that the masking effects of the masking signal are stationary across time. As a result, audio coding algorithms may assume that audio content is masked while it is not, which could results in quantization noise not being masked.

In contrast to other temporal perceptual models, the Taal Detectability has a relatively low computational complexity. In addition to this, it can also be expressed as a squared L2-norm. The computational demand was however shown to be higher than the Par Detectability [17], especially for longer time segments.

2.2 Selection of Perceptual Model for Perceptual Sound Zone Algorithms

In section 2.1 a literature review discussing various perceptual models is given. This section determines which of these perceptual models from the review are suitable for use in a perceptual sound zone algorithm.

First, subsection 2.2.1 discusses desirable properties of the perceptual model for use in a perceptual sound zone algorithms. In subsection 2.2.2 these requirements are used to evaluate the models reviewed in section 2.1

2.2.1 Desirable Properties of Perceptual Model for use in Perceptual Sound Zone Algorithms

As is shown in chapter 3, many sound zone algorithms are posed as optimization problems. As such, it desirable for the perceptual model to be intergratable in optimization problems.

The goal optimization problems is typically to minimize or maximize a cost function, which is done by leveraging the (sub)differential of the function. Algorithms which contain conditional branching or complex operations cannot readily be integrated into cost functions, and are therefore less promising.

In addition to this, sound zone algorithms can often posed as convex optimization problems. This is a sub-class of optimization problems that guarantees that the optimizer is globally unique, rather than there being many local optima [23]. In addition to this, there are many efficient solvers available for convex optimization problems. Therefore, it is also preferable that the perceptual models can preserve convexity when integrated into cost functions.

2.2.2 Evaluating Reviewed Perceptual Models for use in Perceptual Sound Zone Algorithm

All objective audio measures discussed in subsection 2.1.1 were found to be difficult to optimize. As discussed, all models showed a degree of non-differentiability and non-convexity in their computation. As such, they are difficult to integrate into convex optimization problems and will not be used in the perceptual sound zone algorithm. However, as the objective audio measures predict the outcomes of listening tests, they will prove useful in the evaluation of the results.

From all three discussed perceptual models from audio coding, the perceptual models proposed by the ISO MPEG standard were found to be the least promising. As stated in subsection 2.1.2, this is due to the fact that these models do not define a cost function which can be optimized over: instead, only the noise that can be added per auditory band is determined.

As such, the decision is between the Par and Taal detectability. As stated in subsection 2.1.2, both models can be expressed through a squared L2-norm, which is

a convex function [23]. For this reason, it both the Par and Taal detectability are promising for the creation of a perceptual sound zone algorithm.

In contrast to the Par model, the Taal detectability takes into account temporal properties of the input signal. This is beneficial, as it will lead to a more accurate description of the masking properties of the input signals.

However, it has been shown to be at the cost of computational complexity. The Taal detectability has been shown to be take at least 2 times as much time to compute as the Par detectability, with this disparity seemingly growing as a function of input signal length [17].

In addition to this, the Taal model operates in the time-domain, whereas the Par model operates in the frequency-domain [20, 17]. Frequency-domain sound zone approaches have been shown to be less demanding computationally than time-domain approaches [24].

For the reasons given above, lower computational complexity, the Par detectability is used in the perceptual sound zone algorithm. Exploring the possibilities of the Taal detectability is left to future work and not further explored in this work.

2.3 Discussion and Implementation of the Par Detectability Measure

In section 2.2, it was determined that the Par detectability measure is the most suited model for the perceptual sound zone algorithm of all perceptual models considered in the literature review given in section 2.1. In this section, in order to give the reader a greater understanding of the model, the Par detectability measure is considered in greater detail.

This section is organized as follows. First, subsection 2.3.1 gives a high-level description of the Par detectability measure, providing an intuitive understanding and introducing some of the notation that is used. Next, the steps to computing the detectability are described in subsection 2.3.2. Finally, subsection 2.3.3 rewrites the detectability as a squared L2-norm.

2.3.1 High-Level Description of the Par Detectability Measure

In this section, a high-level description of the Par detectability measure is given. This is done to give the reader a basic understanding of the model before going into greater detail.

The Par detectability maps two input sequences to a positive real value, i.e. $D : (\mathbb{R}^{N_x}, \mathbb{R}^{N_x}) \mapsto \mathbb{R}^+$. The two input sequences are the masking signal $x[n] \in \mathbb{R}^{N_x}$ and the disturbance signal $\varepsilon[n] \in \mathbb{R}^{N_x}$. The detectability of these two sequences is denoted as $D(x[n], \varepsilon[n])$.

Imagine a human that is listening to both the masking signal $x[n]$ and the disturbance signal $\varepsilon[n]$ at the same time. The detectability $D(x[n], \varepsilon[n])$ can be understood as how easily a human listener can detect the disturbance signal $\varepsilon[n]$ in presence of the masking signal $x[n]$. The signal $x[n]$ is referred to as the masking signal because its masking properties are model to determine how well it masks the disturbance signal $\varepsilon[n]$.

For this interpretation to be accurate, the signals $x[n]$ and $\varepsilon[n]$ should be short-time signals. The paper uses a signal length of to 20 to 200 milliseconds. This is important, as the model assumes that the psycho-acoustical properties of $x[n]$ and $\varepsilon[n]$ are stationary.

The measure is normalized in such a way that the detectability $D(x[n], \varepsilon[n])$ is equal to 1 when the disturbance signal $\varepsilon[n]$ is "just noticeable" in presence of masking signal $x[n]$. That is to say: if the detectability is 1, the disturbance is on the verge of being noticeable and not noticeable.

The detectability $D(x[n], \varepsilon[n])$ can also attain a value larger than 1. The larger values of the detectability correspond with an increased perceived presence of the disturbance signal $\varepsilon[n]$.

2.3.2 Computation Details of the Par Detectability

This section explores calculating the Par detectability. The first thing to note about the Par detectability is that it is computed using the frequency domain representations of its inputs [20]. To this end, let $X[k]$ and $\mathcal{E}[k]$ denote the frequency domain representations of the masking signal $x[n]$ and the disturbance signal $\varepsilon[n]$ respectively.

After determining the frequency domain representations, the Par detectability computes an internal representation of the input signals $X[k]$ and $\mathcal{E}[k]$. This internal representation models how the input stimuli appear to the human auditory system. For the Par detectability measure, this is modeled by filtering the input stimuli.

Two subsequent filters are applied. The first filter models how parts of the ear filter the incoming sound with an outer- and middle-ear filter $H_{\text{om}}[k]$. Next, a 4^{th} order Gammatone filter bank is applied, modeling the frequency-place transform that occurs in basilar membrane inside of the ear [20].

The Gammatone filter bank consists of N_g filters. The frequency domain representation of each individual filter is denoted by $\Gamma_i[k]$, for $1 \leq i \leq N_g$. The filters in the filter bank $\Gamma_i[k]$ have a bandwidth given by the equivalent rectangular bandwidth (ERB) and center frequencies given by the corresponding equivalent rectangular bandwidth number scale (ERBS). Expressions for the gammatone filters $\Gamma_i[k]$ are provided by the original paper [20].

After filtering, the power per Gammatone filter tap is computed. Let M_i and S_i denote the output power of the i^{th} filter tap for the masking signal $X[k]$ and the disturbance signal $\mathcal{E}[k]$ respectively. This output power can be understood as the amount of power perceived per frequency band of the human ear. The relationship between the input quantities and the output power of the filter taps can be given as follows:

$$M_i = \frac{1}{N_x} \sum_{k=0}^{N_x-1} |H_{\text{om}}[k]|^2 \, |\Gamma_i[k]|^2 \, |X[k]|^2 \tag{2.1}$$

$$S_i = \frac{1}{N_x} \sum_{k=0}^{N_x-1} |H_{\text{om}}[k]|^2 \, |\Gamma_i[k]|^2 \, |\mathcal{E}[k]|^2 \tag{2.2}$$

The output powers can then be used to define the within-channel detectability D_i per filter tap i. This can be thought of the detectability per frequency band of the human ear, and is defined as follows:

$$D_i = \frac{N_x S_i}{N_x M_i + C_a} \tag{2.3}$$

Here, C_a is a calibration constant that ensures that the absolute threshold of hearing is predicted correctly. This can be understood by considering the case where no masking signal $x[n]$ is present, in which case $M_i = 0$ for all i. If not for the calibration constant C_a, the detectability of any non-zero disturbance $\varepsilon[n]$ would

be infinite. In order to take the frequency-dependence of the threshold of hearing into account, the previously described outer- and middle ear filters are defined as the inverse of the threshold of hearing [20].

The total detectability $D(x[n], \varepsilon[n])$ can then be computed as the scaled sum of all within channel detectabilities. It is defined as follows:

$$D(x[n], \varepsilon[n]) = C_s L_{\text{eff}} \sum_{i=0}^{N_g} D_i \tag{2.4}$$

$$= C_s L_{\text{eff}} \sum_{i=0}^{N_g} \frac{\sum_{k=0}^{N_x-1} |H_{\text{om}}[k]|^2 |\Gamma_i[k]|^2 |\mathcal{E}[k]|^2}{\sum_{k=0}^{N_x-1} |H_{\text{om}}[k]|^2 |\Gamma_i[k]|^2 |X[k]|^2 + C_a} \tag{2.5}$$

Here, C_s is a calibration constant chosen such that a just noticeable disturbance signal results in a detectability of $D(x[n], \varepsilon[n]) = 1$. The constant L_{eff} is the integration time of the human auditory system. It is chosen equal to the segment length of $x[n]$ and $\varepsilon[n]$ in milliseconds.

In order to further understand detectability, consider the behavior of the expression of the detectability $D(x[n], \varepsilon[n])$ above. Imagine that the spectrum of the masking signal is much larger than the disturbance signal, i.e. $X[k] \gg \mathcal{E}[k]$ for all frequency bins k. In this case, the detectability of $\varepsilon[n]$ will be small due to the masking of the masking signal $x[n]$ or due to the threshold of hearing (determined by the calibration constant C_a).

Conversely consider the case that the spectrum of the masking signal is much smaller than the disturbance signal, i.e. $X[k] \ll \mathcal{E}[k]$ for all frequency bins k. In this case, the resulting detectability is determined greatly by the calibration coefficient C_a:

- If the total energy of the filtered disturbance signal is much larger than the calibration constant $S_i \gg C_a$ for all i, the detectability becomes large. This models the case that the disturbance signal is large relative to the threshold of hearing.
- Alternatively, if $S_i \ll C_a$ for all i, the disturbance signal is inaudible due to the threshold of hearing and the detectability will be low accordingly.

This concludes the analysis of the Par model. The determination of the calibration constants C_a and C_s is discussed in Appendix A.

2.3.3 Proposed Least-Squares Formulation of the Par Detectability

This section will rewrite the previously introduced detectability into a least-squares representation. This representation is more mathematically tractable than Equation 2.5 and thus will allow for easier integration into existing sound zone algorithms.

To obtain this expression, the sum of squares will be expressed as a L2-norm.

Consider the following rewrite of the detectability given in Equation 2.5:

$$\begin{aligned}
D(x[n], \varepsilon[n]) &= C_s L_{\text{eff}} \sum_{i=0}^{N_g} \frac{\sum_{k=0}^{N_x-1} |H_{\text{om}}[k]|^2 |\Gamma_i[k]|^2 |\mathcal{E}[k]|^2}{\sum_{k=0}^{N_x-1} |H_{\text{om}}[k]|^2 |\Gamma_i[k]|^2 |X[k]|^2 + C_a} \\
&= \sum_{i=0}^{N_g} \left(\frac{C_s L_{\text{eff}}}{||H_{\text{om}}[k]\Gamma_i[k]X[k]||_2^2 + C_a} \right) \sum_{k=0}^{N_x-1} |H_{\text{om}}[k]|^2 |\Gamma_i[k]|^2 |\mathcal{E}[k]|^2 \\
&= \sum_{k=0}^{N_x-1} \left(\sum_{i=0}^{N_g} \frac{C_s L_{\text{eff}} |\Gamma_i[k]|^2}{||H_{\text{om}}[k]\Gamma_i[k]X[k]||_2^2 + C_a} \right) |H_{\text{om}}[k]|^2 |\mathcal{E}[k]|^2 \\
&= \sum_{k=0}^{N_x-1} |W_x[k]|^2 |\mathcal{E}[k]|^2 \\
&= ||W_x[k]\mathcal{E}[k]||_2^2
\end{aligned}$$

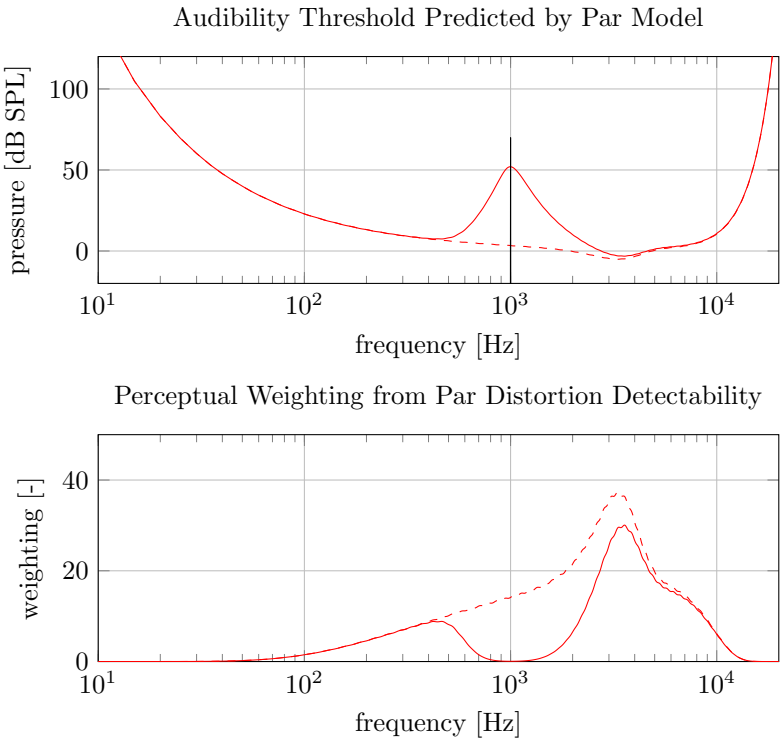
The rewrite above introduced perceptual weighting $W_x[k] \in \mathbb{R}^{N_x}$ describing the masking effects of the masking signal $x[n]$. The perceptual weighting $W_x[k]$ is defined as follows:

$$W_x[k] = \left(\sqrt{\sum_{i=0}^{N_g} \frac{C_s L_{\text{eff}} |\Gamma_i[k]|^2}{||H_{\text{om}}[k]\Gamma_i[k]X[k]||_2^2 + C_a}} \right) |H_{\text{om}}[k]| \tag{2.6}$$

Note from this formulation that the perceptual weighting is only a function of the masking signal $x[n]$.

Note also that the resulting detectability $D(x[n], \varepsilon[n])$ is a convex function of the disturbance signal $\varepsilon[n]$. The frequency-domain representation $\mathcal{E}[k]$ is related to the time-domain representation $\varepsilon[n]$ through the DFT, which is a linear operator. The perceptual weighting of $\mathcal{E}[k]$ performed by $W[k]$ is also a linear operation. As such, $W[k]\mathcal{E}[k]$ is an affine function of $\varepsilon[n]$.

TODO: It might be nice to have a plot showing the masking threshold and the corresponding perceptual weighting matrix. This could show the reader that the threshold of quite results in low weighting for the low and high frequencies, and effects that the masking signal has on the weights...


Audibility Threshold Predicted by Par Model
pressure [dB SPL]
100
50
0
10^1
10^2
10^3
10^4
frequency [Hz]
Perceptual Weighting from Par Distortion Detectability
weighting [-]
40
20
0
10^1
10^2
10^3
10^4
frequency [Hz]

Sound Zone Approach Review and Implementation

3

An initial description of the sound zone problem was given in the introduction of the thesis. This section seeks to build on this description in order to provide the understanding necessary for the rest of this work.

As mentioned in the introduction, the problem that sound zones seeks to solve is the reproduction of multiple types of audio content in the same room with minimal interference. This way, multiple people can enjoy different audio content without disturbing one another.

Controlling the spatial distribution of sound is done by controlling the audio that is produced by an array of loudspeakers. The space inside the enclosure is divided up into multiple zones. Each zone is assigned target sound pressure that we would like to have reproduced inside of it. This target sound pressure could be various audio content, for example music or the sound of a movie.

To understand this principal, consider the example given by Figure 3.1. The loudspeakers array that is present in the room is to be controlled by the sound zone algorithm in such a way that the desired content is reproduced in each zone. As mentioned, this is to be done in a way that results in minimal interference, e.g. it is undesirable to be able to hear content B when inside the red zone.

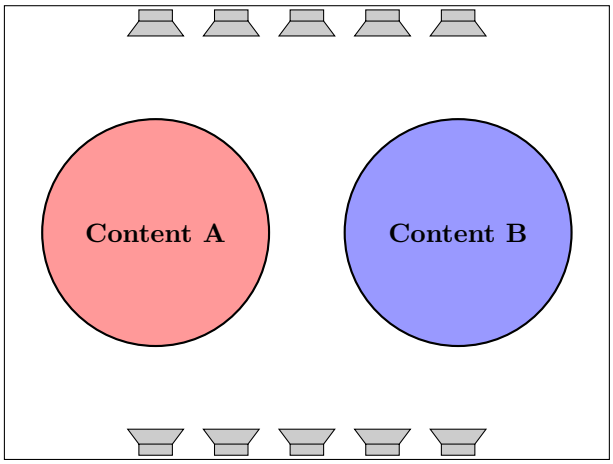


Figure 3.1: A birds-eye view of a room is depicted. It is divided into two zones: a red zone and a blue zone. Each zone is assigned different content: content A and content B. In the northern and southern parts of the room, a loudspeaker array is mounted on the walls.

There are various approaches to solving the sound zone problem. An important concept that is used frequently is that bright zones and dark zones.

Sound zone problems are typically decomposed into a separate subproblem for every

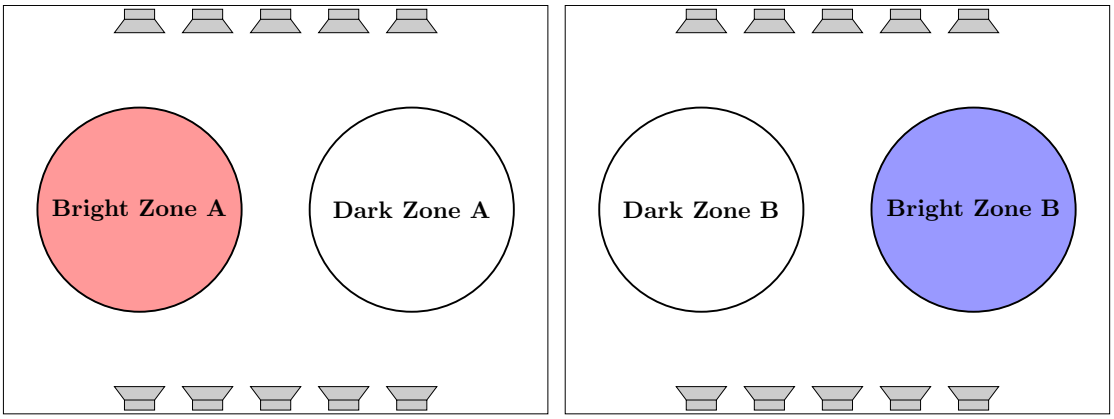


Figure 3.2: A birds-eye view of a room is depicted twice containing two different sound zone problems.

zone. Each one of these subproblems considers only two zones: one bright zone and one dark zone. The goal of each subproblems is to reproduce a specified target sound pressure in the bright zone while restricting the sound pressure in the dark zones.

The combination of all subproblems provides a solution to the sound zone problem. To ease the understanding of this concept, consider an example of this decomposition is given in Figure 3.2.

Here, a decomposition of the example given in Figure 3.1 into two bright-dark zone pairs. For the first problem, the goal is to reproduce content A in bright zone A while minimizing the amount of sound pressure in dark zone A. Similarly for the second problem: reproduce content B in bright zone B while minimizing the amount of sound pressure in dark zone B. Combining the two solutions results in a solution with content reproduced in both zones with minimal interference between zones.

This concludes the introduction to the sound zone problem.

Chapter Structure

The goal of the rest of this chapter is to motivate the proposal of a perceptual sound zone approach which uses the Par distortion detectability introduced in chapter 2. It is structured as follows:

- This chapter begins in section 3.1 with the presentation of a mathematical framework which can be used to describe the sound zone problem.
- This mathematical framework is then used to describe the two main sound zone approaches, “Pressure Matching” and “Acoustic Contrast Control”, in section 3.2.
- Finally, in section 3.3 will motivate the proposed perceptual sound zone approach. This is done by reflect on the mathematical properties of the Par detectability and the sound zone approaches review in section 3.2.

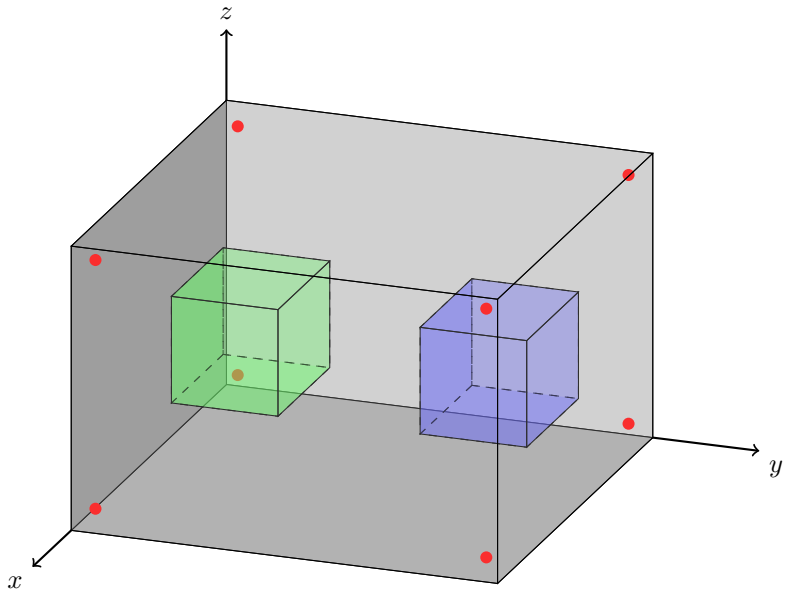


Figure 3.3: The room $\mathcal{R} \subset \mathbb{R}^3$ containing the zones $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{R}$ depicted in green and blue respectively. The room contains $N_L = 8$ loudspeakers, which are denoted by the red dots in the corners of the room.

3.1 Sound Zone Problem Data Model

In the previous section, the sound zone problem was introduced heuristically. In this section a mathematical framework for a room containing sound zones will be introduced. This framework will be used later in the derivation of the sound zone algorithms in section 3.2.

The contents of this section are as follows. First, subsection 3.1.1 develops a spatial description of a room containing two zones and a loudspeaker array. Then, subsection 3.1.2 defines the objective of the sound zone algorithm formally as realizing target sound pressure at discrete points in the room.

The relation between the sound pressure in the room and loudspeaker input signals will then be given in subsection 3.1.3, completing the mathematical framework. This is then used in subsection 3.1.4 to select a suitable target sound pressure which will be used in the remainder of this thesis.

3.1.1 Room Topology

A room $\mathcal{R}$ can be modeled as a closed subset of three dimensional space, $\mathcal{R} \subset \mathbb{R}^3$. The two non-overlapping zones $\mathcal{A}$ and $\mathcal{B}$ are contained within the room $\mathcal{R}$, i.e. $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{R}$ where $\mathcal{A} \cap \mathcal{B} = \emptyset$. In general, the room can contain any number of zones, but this thesis will focus on the two zone case. In addition to the zones, the room $\mathcal{R}$ also contains N_L loudspeakers, whose locations are modeled as discrete points. An example of a possible room, loudspeakers and pair of zones are visualized in Figure 3.3.

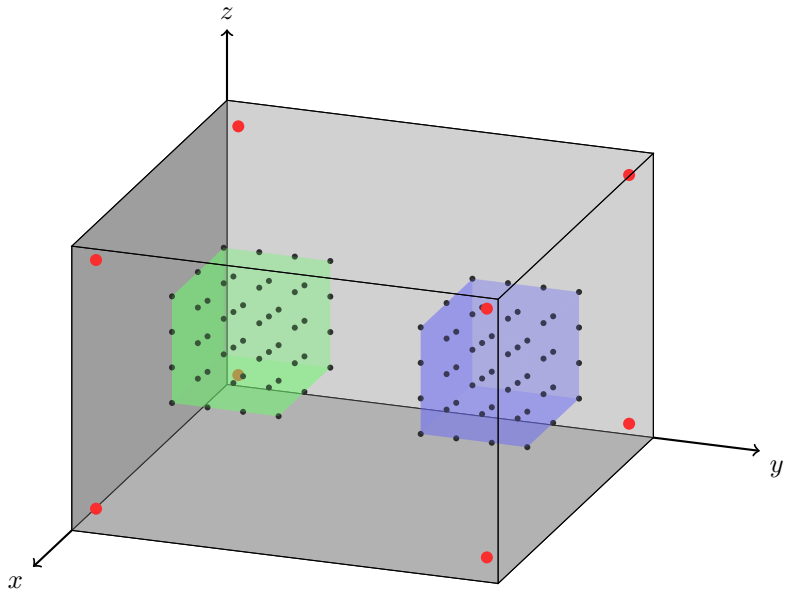


Figure 3.4: The previously introduced room $\mathcal{R}$ with zones $\mathcal{A}$ and $\mathcal{B}$ discretized.

The goal of the sound zone algorithm is to use the loudspeakers to realise a specified target sound pressure in the space described by zones $\mathcal{A}$ and $\mathcal{B}$. This is to be done in such a way that there is minimal interference between zones, meaning that target sound pressure intended for one zone should not be audible in the other zones.

The sound field generated by loudspeakers can be controlled by specifying their input signals. As such, the goal of the sound zone algorithm is finding loudspeaker input signals in such a way that specified target sound pressure is attained.

The rest of this section will focus on formalizing this notion mathematically.

3.1.2 Defining Target Pressure

As mentioned, the goal of the sound zone algorithm is to realize a specified target sound pressure in the different zones $\mathcal{A}$ and $\mathcal{B}$ in the room $\mathcal{R}$.

Currently, the zones are given as continuous regions in space. However, most sound zone approaches will instead discretize the zones by sampling the continuous zones $\mathcal{A}$ and $\mathcal{B}$ into so-called control points. The sound pressure is then controlled only in these control points.

Thus, we discretize zones $\mathcal{A}$ and $\mathcal{B}$ into a total of N_a and N_b control points respectively. Let A and B denote the sets of the resulting control points points contained within zones $\mathcal{A}$ and $\mathcal{B}$ respectively.

Now let $t^m[n]$ denote the target sound pressure at control point m in either A or B, i.e. $m \in A \cup B$. Our goal is thus to realize $t^m[n]$ in all control points $m \in A \cup B$ using the loudspeakers present in the room.

3.1.3 Realizing Sound Pressure through the Loudspeaker

The sound pressure produced by the loudspeakers can be controlled by specifying their input signals. Mathematically speaking, let $x^{(l)}[n] \in \mathbb{R}^{N_x}$ denote the loudspeaker input signal of length N_x for the l^{th} loudspeaker. For now, it is assumed that the loudspeaker input signals are of finite length. In a later part of the thesis, a short-time formulation will be given which supports infinite length sequences.

As such, the goal of the sound zone algorithm is to find loudspeaker inputs $x^{(l)}[n]$ such that the target sound pressure $t^m[n]$ is realized for all $m \in A \cup B$.

In order to do so, a relationship must be established between the loudspeaker inputs $x^{(l)}[n]$ and the resulting sound pressure at control points $m \in A \cup B$. This relationship can be modeled by room impulse responses (RIRs) $h^{(l,m)}[n] \in \mathbb{R}^{N_h}$.

The RIRs $h^{(l,m)}[n]$ determine the sound pressure at control point m due to playing loudspeaker signal $x^{(l)}[n]$ from loudspeaker l. Mathematically, let $p^{(l,m)}[n] \in \mathbb{R}^{N_x+N_h-1}$ represent said sound pressure. It can be defined as follows:

$$p^{(l,m)}[n] = \left(h^{(l,m)} * x^{(l)}\right)[n] \tag{3.1}$$

Here, the $*$ operator is used to denote linear convolution. The achieved sound pressure $p^{(l,m)}[n]$ only considers the contribution of loudspeaker l at reproduction point m. Let $p^{(m)}[n] \in \mathbb{R}^{N_x+N_h-1}$ denote the total sound pressure due to all N_L loudspeakers, which can be expressed as the sum over all contributions $p^{(l,m)}[n]$ as follows:

$$p^{(m)}[n] = \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * x^{(l)}\right)[n] \tag{3.2}$$

With the data model completed, the goal of the sound zone algorithm can be restated formally. Namely, the goal is to find $x^{(l)}[n]$ such that the achieved sound pressure $p^{(m)}[n]$ attains the target sound pressure $t^{(m)}[n]$ for all control points $m \in A \cup B$.

3.1.4 Choice of Target Pressure

The target sound pressure $t^{(m)}[n]$ describes the desired content for a specific control point m. So far, the choice of target sound pressure $t^{(m)}[n]$ has been kept general. In this section, a choice for the target pressure will be made and motivated.

Assume that the users of the sound zone system have selected desired playback audio signals $s_{\mathcal{A}}[n] \in \mathbb{R}^{N_x}$ and $s_{\mathcal{B}}[n] \in \mathbb{R}^{N_x}$ that they wish to hear in zone $\mathcal{A}$ and $\mathcal{B}$ respectively. In order to accommodate the wishes of the user, the target sound pressure is chosen as follows:

$$\begin{aligned} t^{(m)}[n] &= \sum_{l=0}^{N_L} \left(h^{(l,m)} * s_{\mathcal{A}}\right)[n] \qquad \forall\, m \in A \\ t^{(m)}[n] &= \sum_{l=0}^{N_L} \left(h^{(l,m)} * s_{\mathcal{B}}\right)[n] \qquad \forall\, m \in B \end{aligned} \tag{3.3}$$

This choice for the target pressure can be understood as the sound pressure that arises in a certain zone when playing only the desired playback audio for that zone from the loudspeaker array. For example, when in zone $m \in A$, the target sound pressure is set equal to the sound pressure corresponding to the sound pressure that arises when playing only $s_{\mathcal{A}}[n]$ from the loudspeaker array.

The motivation for choosing this target is that it physically attainable in each zone seperately with the given loudspeakers and room.

3.2 Review of Sound Zone Approaches

The two main approaches in sound zone literature are pressure matching (PM) and acoustic contrast control (ACC). One of these two approaches will be used in the perceptual sound zone algorithm to be introduced in chapter 4. This section will mathematically introduce both approaches with the previously derived data model with the goal of sketching their mathematical properties. This will then later be used in section 3.3 to determine which is suitable for integration with the Par detectability discussed in chapter 2.

In the typical sound zone approach the sound zone problem is divided up into subproblems as described in **??**. The resulting loudspeaker input signals $x^{(l)}[n]$ are determined for a single bright-dark zone pair: the loudspeaker input signals are found such that the a target audio is achieved in the bright zone, while leakage is minimized in the dark zone. If the solution for multiple zones is desired, then multiple problems must be solved and their resulting loudspeaker input signals combined.

There is another approach however. In a multi-zone approach, the loudspeaker input signals are instead determined for jointly for all zones, rather than decomposing into bright-dark zone pairs. This is the approach that will be taken in this thesis, as it was found to be more general and simple to present. For simplicity, this thesis will limit the number of zones to two. The approach is however generalizable to any multiplicity of zones.

In a two zone multi-zone approach, the loudspeaker input signals $x^{(l)}[n]$ will be decomposed into two parts as follows:

$$x^{(l)}[n] = x_{\mathcal{A}}^{(l)}[n] + x_{\mathcal{B}}^{(l)}[n] \tag{3.4}$$

Here, $x_{\mathcal{A}}^{(l)}[n]$ and $x_{\mathcal{B}}^{(l)}[n]$ are the parts of the loudspeaker input signal responsible for reproducing the target sound pressure in zone $\mathcal{A}$ and $\mathcal{B}$ respectively.

Through this decomposition, it is possible to consider the sound pressure that arises due to the separate loudspeaker input signals:

$$p_{\mathcal{Z}}^{(m)}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * x_{\mathcal{Z}}^{(l)} \right)[n] \tag{3.5}$$

Where $\mathcal{Z} \in (\mathcal{A}, \mathcal{B})$ represents either zones. Here, $p_{\mathcal{A}}^{(m)}[n]$ and $p_{\mathcal{B}}^{(m)}[n]$ can be understood to be the pressure that arises due to playing loudspeaker input signals $x_{\mathcal{A}}^{(l)}[n]$ and $x_{\mathcal{B}}^{(l)}[n]$ respectively at a specified control point. The total achieved sound pressure at control point m is then given by the addition of the two achieved sound pressures:

$$p^{(m)}[n] = p_{\mathcal{A}}^{(m)}[n] + p_{\mathcal{B}}^{(m)}[n] \tag{3.6}$$

What follows is using this decomposition to describe a multi-zone variant of both a pressure matching approach in subsection 3.2.1 and acoustic contrast control approach in subsection 3.2.2.

3.2.1 Pressure Matching

The "Pressure Matching" (PM) is widely used in the literature to solve the sound zone problem. In this section, a "Multi-Zone Pressure-Matching" (MZ-PM) algorithm will be derived using the previously derived data model.

In pressure matching approaches, one attempts to control the output of the loudspeaker array in such a way that the resulting sound pressure in the zone matches the specified target sound pressure for that zone, while simultaneously minimizing the sound pressure that results in other zones as to minimize the interference or crosstalk between zones [1, 25].

The idea in this approach is to chose $x_{\mathcal{A}}^{(l)}[n]$ and such that the resulting achieved pressure $p_{\mathcal{A}}^{(m)}[n]$ attains the target sound pressure $t^{(m)}[n]$ in all control points $m \in A$. At the same time however, $p_{\mathcal{A}}^{(m)}[n]$ should result in minimal achieved sound pressure in all control points $m \in B$. Any sound pressure resulting from $x_{\mathcal{A}}^{(l)}[n]$ in zone $\mathcal{B}$ is can be understood as leakage, or cross-talk between the zones.

Similar arguments can be given for $x_{\mathcal{B}}^{(l)}[n]$.

An optimization problem that achieves this goal can be formulated as follows:

$$\underset{x_{\mathcal{A}}^{(l)}[n],\, x_{\mathcal{B}}^{(l)}[n]\,\forall\, l}{\arg\min} \quad \sum_{m\in A} \left|\left| p_{\mathcal{A}}^{(m)}[n] - t^{(m)}[n] \right|\right|_2^2 + \sum_{m\in A} \left|\left| p_{\mathcal{B}}^{(m)}[n] \right|\right|_2^2 + \tag{3.7}$$

$$\sum_{m\in B} \left|\left| p_{\mathcal{B}}^{(m)}[n] - t^{(m)}[n] \right|\right|_2^2 + \sum_{m\in B} \left|\left| p_{\mathcal{A}}^{(m)}[n] \right|\right|_2^2 \tag{3.8}$$

Here, the $||\cdot||_2^2$ operates denotes the squared L2-norm, which corresponds to taking the sum of the squares the input sequence.

To further understand how optimization problem operates, consider the following definitions:

$$\mathrm{RE}_{\mathcal{Z}}^{(m)} = \left|\left| p_{\mathcal{Z}}^{(m)}[n] - t^{(m)}[n] \right|\right|_2^2 \qquad \forall\ m \in Z \tag{3.9}$$

$$\mathrm{LE}_{\mathcal{Z}}^{(m)} = \left|\left| p_{\mathcal{Z}}^{(m)}[n] \right|\right|_2^2 \qquad \forall\ m \notin Z \tag{3.10}$$

Here, $\mathrm{RE}_{\mathcal{Z}}^{(m)}$ is the reproduction error for zone $\mathcal{Z} \in (\mathcal{A}, \mathcal{B})$ for a control point $m \in Z$. This is error corresponds to how well the achieved sound pressure $p_{\mathcal{Z}}^{(m)}[n]$ matches the target sound pressure $t^{(m)}[n]$ for a control point in the bright zone Z.

$\mathrm{LE}_{\mathcal{Z}}^{(m)}$ is the leakage error in zone $\mathcal{Z} \in (\mathcal{A}, \mathcal{B})$ for a control point $m \notin Z$. This can be understood as the sound pressure that "leaks" into control point m in zones other than $\mathcal{Z}$ when attempting to reproduce the target sound pressure $t^{(m)}[n]$ in zone $\mathcal{Z}$. This can be otherwise be considered as the "interference" or "cross-talk" between zones.

Using these new definition allows for the following rewrite of the optimization prob-

lem:

$$\underset{x_{\mathcal{A}}^{(l)}[n], x_{\mathcal{B}}^{(l)}[n] \,\forall\, l}{\arg\min} \quad \sum_{m \in A} \mathrm{RE}_{\mathcal{A}}^{(m)} + \sum_{m \in B} \mathrm{LE}_{\mathcal{A}}^{(m)} + \sum_{m \in B} \mathrm{RE}_{\mathcal{B}}^{(m)} + \sum_{m \in A} \mathrm{LE}_{\mathcal{B}}^{(m)} \tag{3.11}$$

From this it becomes clear that this approach results in trade-off between minimizing the reproduction errors $\mathrm{RE}_{\mathcal{Z}}^{(m)}$ and leakage errors $\mathrm{LE}_{\mathcal{Z}}^{(m)}$.

Some pressure matching approaches attempt to control this trade-off by introducing weights for the different error terms, or by adding constraints. Choosing constraints can however be challenging as the squared L2 pressure error does not always correlate well with how the error is perceived.

3.2.2 Acoustic Contrast Control

"Acoustic Contrast Control" is another widely used sound zone approach from literature. The acoustic contrast control (ACC) approach to sound zones attempts to maximize the acoustic contrast between the bright zone and the dark zone. Acoustic contrast is defined as the ratio of the total sound energy of the bright zone and the dark zone. Essentially, the goal is to maximize the difference in sound pressure level between the bright and dark zones.

In this section, a "Multi-Zone Acoustic Contrast Control" (MZ-ACC) algorithm will be described. As the previously described data model is in the time domain, this approach will take inspiration from a time-domain approach found in literature known as the broadband acoustic contrast control (BACC) approach [26, 27, 28].

In contrast to the MZ-PM approach, the MZ-ACC approach does not optimize directly over the loudspeaker input signals $x_{\mathcal{A}}^{(l)}[n]$ and $x_{\mathcal{B}}^{(l)}[n]$. Instead, it indirectly controls the loudspeaker input signals by optimizing over FIR filter coefficients $w_{\mathcal{A}}^{(l)}[n] \in \mathbb{R}^{N_w}$ and $w_{\mathcal{B}}^{(l)}[n] \in \mathbb{R}^{N_w}$. These filters are applied to the desired playback signals $s_{\mathcal{A}}^{(l)}$ and $s_{\mathcal{B}}^{(l)}$ respectively to form the final loudspeaker input signals.

This relationship between the loudspeaker input signals and the filter coefficients is thus given as follows:

$$x_{\mathcal{Z}}^{(l)}[n] = \left(w_{\mathcal{Z}}^{(l)} * s_{\mathcal{Z}} \right)[n] \tag{3.12}$$

This definition also relates the filter coefficients to the resulting sound pressure through Equation 3.2.

As mentioned, the goal of the ACC approach is to maximize the acoustic contrast between bright and dark zones, which was defined as the ratio between the sound energy in the bright and dark zones. The total sound energy in a zone will be defined as the sum of squares of the sound pressure in a control point. As such, the acoustic contrast $\mathrm{AC}_{\mathcal{Z}}$ for a zone $\mathcal{Z}$ can be defined as follows:

$$\mathrm{AC}_{\mathcal{Z}} = \frac{\sum_{m \in Z} \left|\left| p_{\mathcal{Z}}^{(m)}[n] \right|\right|_2^2}{\sum_{m \notin Z} \left|\left| p_{\mathcal{Z}}^{(m)}[n] \right|\right|_2^2} \tag{3.13}$$

In an ACC approach the goal is to maximize the total acoustic contrast. Thus, consider the following optimization problem:

$$\underset{w_{\mathcal{A}}^{(l)}[n],\, w_{\mathcal{B}}^{(l)}[n]\,\forall\, l}{\arg\max} \quad \mathrm{AC}_{\mathcal{A}} + \mathrm{AC}_{\mathcal{B}} \tag{3.14}$$

As mentioned, the optimization is performed over the loudspeaker filter coefficients rather than over the loudspeaker input signals.

3.3 Proposed Perceptual Sound Zone Approach

In chapter 2 it was determined that the Par detectability is the perceptual model most suited for use in a perceptual sound zone algorithm. Previously, in section 3.2, two different sound zone techniques were discussed pressure matching (PM), and acoustic contrast control (ACC).

In this section proposes how the Par detectability measure can be combined with the previously discussed sound zone approaches. This is done by first reflecting on the mathematical properties of the Par detectability measure in subsection 3.3.1. Next, subsection 3.3.2 introduces the proposed perceptual sound zone approach.

3.3.1 Analysis of Mathematical Properties of Proposed Least Squares formulation of Par Detectability Measure

This section will reflect on the mathematical properties of the Par detectability.

Recall from section 2.3 that the detectability $D(x[n], \varepsilon[n])$ quantifies how detectable a disturbance $\varepsilon[n] \in \mathbb{R}^{N_x}$ is in presence of a masking signal $x[n] \in \mathbb{R}^{N_x}$. The Par detectability assumes that the time-scale of its inputs are short, in the order of 20 to 200 ms.

The detectability is computed using the frequency domain representation of its input signals. To this end, $X[k]$ and $\mathcal{E}[k]$ were introduced to denote the frequency domain representations of $x[n]$ and $\varepsilon[n]$ respectively.

In subsection 2.3.3 it was found that the detectability could be expressed in least-squares fashion as follows:

$$D(x[n], \varepsilon[n]) = ||W_x[k]\mathcal{E}[k]||_2^2 \tag{3.15}$$

Here, perceptual weighting $W_x[k]$ models the psycho-acoustical masking effects of the masking signal $x[n]$. The weights are applied to the frequency bins of the disturbance signal $\mathcal{E}[k]$, and the sum of squares is used determine the final detectability rating.

It was noted in subsection 2.3.3 that the Par detectability measure is a convex function of the disturbance signal $\mathcal{E}[k]$ when the masking signal is held constant. As such, one approach is to specify a sound zone algorithm which leverages the optimization over this disturbance signal in some way. This will be done by adopting a model for the disturbance $\mathcal{E}[k]$. For example, the disturbance could model the sound pressure error.

In summary, the detectability has the following properties that must be taken into account:

1. It is computed in the short-time frequency-domain.
2. It is convex when optimizing over the disturbance signal $\mathcal{E}[k]$.

3.3.2 Introducing Proposed Perceptual Sound Zone Approach

In section 3.2 two different sound zone approaches were discussed: pressure matching (PM) and acoustic contrast control (ACC). Currently, neither approach operates in the frequency domain on short-time scales. However, it has been shown in literature that formulations of both ACC and PM exist that satisfy this condition [29].

Either ACC or PM can be used. Through inspection, it was found that the Par detectability measure can be used to formulate a PM approach directly by modeling the errors minimized by PM as disturbance signals. As such, in this work the focus will be on the PM approach.

To see this, recall in the discussion of PM given in section 3.2 that PM was shown to minimize the sum of the reproduction error in the bright zone $\mathrm{RE}_{\mathcal{Z}}^{(m)}$ and the leakage to the dark zone $\mathrm{LE}_{\mathcal{Z}}^{(m)}$. The definitions of $\mathrm{RE}_{\mathcal{Z}}^{(m)}$ and $\mathrm{LE}_{\mathcal{Z}}^{(m)}$ are given by Equation 3.9 and Equation 3.10 respectively. Here, the reproduction error $\mathrm{RE}_{\mathcal{Z}}^{(m)}$ can be understood as the total error between the achieved and target sound pressure in the bright zone of zone $\mathcal{Z}$. The leakage $\mathrm{LE}_{\mathcal{Z}}^{(m)}$ can be understood as the sound pressure that arises in zones other than $\mathcal{Z}$, essentially "leaking" into other zones, when reproducing the target in the bright zone.

Both can be thus be considered as "errors" that we wish to minimize. Viewing it this way, one choice for model these errors as the disturbance signals $\varepsilon[n]$. To this end, consider the reproduction error detectability $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ and the leakage error detectability $\mathrm{LED}_{\mathcal{Z}}^{(m)}$. Fom now, ignore to the fact that the quantities for target and achieved sound pressure defined so far do not satisfy the short-time requirements. The perceptual errors are defined as:

$$\mathrm{RED}_{\mathcal{Z}}^{(m)} = D(t^{(m)}[n],\, p_{\mathcal{Z}}^{(m)}[n] - t^{(m)}[n]) \qquad \forall\ m \in Z \tag{3.16}$$

$$\mathrm{LED}_{\mathcal{Z}}^{(m)} = D(t^{(m)}[n],\, p_{\mathcal{Z}}^{(m)}[n]) \qquad \forall\ m \notin Z \tag{3.17}$$

The reproduction error detectability $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ is determined by the detectability of the sound pressure difference between the achieved sound pressure $p_{\mathcal{Z}}^{(m)}[n]$ and the target sound pressure $t^{(m)}[n]$ for the control point m. This sound pressure difference can be understood as the "error" in sound pressure.

The target sound pressure is used as the masking signal, in doing so the masking properties of the target sound pressure $t^{(m)}[n]$ is used. Note that this is an approximation: in reality, it cannot be assumed that the sound algorithm exactly attains the target sound pressure. In the ideal case, the masking properties of the total achieved sound pressure would be used instead.

However, this quantity depends on the loudspeaker input signals over which the optimization is performed, and is thus not available a priori. One approach would be to include the masking signal in the optimization, but this would violate convexity, as stated in subsection 2.3.3. As such, by the masking effects of the achieved pressure are modeled by those of the target sound pressure.

The $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ models how detectable the deviation from target sound pressure is One interpretation of minimizing $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ is finding the sound pressure difference which is least detectable.

Similarly, the leakage error detectability $\mathrm{LED}_{\mathcal{Z}}^{(m)}$ is determined by how detectable the sound pressure of zone $\mathcal{Z}$ is in other zones. Note as $m \notin \mathcal{Z}$, this definition considers all points outside of the bright zone for $\mathcal{Z}$. Here, the masking signal is again chosen as the target sound pressure of the control point in question. $\mathrm{LED}_{\mathcal{Z}}^{(m)}$ models how detectable the leakage of zone $\mathcal{Z}$ is in presence of the intended target sound pressure. Minimizing $\mathrm{LED}_{\mathcal{Z}}^{(m)}$ could thus be understood as finding the leakage that is minimally-detectable in presence of the target sound pressure.

These perceptually modified errors are found to be a promising way of combining pressure matching and the Par detectability, and are used to state perceptual sound zone algorithms in chapter 4. The short-time frequency-domain quantities required for the definitions of $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ and $\mathrm{LED}_{\mathcal{Z}}^{(m)}$ are introduced in this chapter.

Using the ACC approach to formulate perceptual sound zone algorithms is not explored further in this work, but is found to be promising future work.

Implementation of Proposed Perceptual Sound Zone Algorithms

4

In chapter 2 the Par detectability is selected as the most promising perceptual model for use in a perceptual sound zone algorithm. Next, chapter 3 various sound zone algorithms are discussed, ultimately leading to the proposal of a sound zone approach in section 3.3 that uses the Par detectability measure in order to create sound zones.

This chapter uses the proposed perceptual sound zone approach to propose and implement two perceptual sound zone algorithms.

Chapter Structure

This chapter is structured as follows.

- First, section 4.1 discusses the reformulation of the time-domain pressure matching approach given in section 3.2 to a short-time frequency-domain pressure-matching approach. This is necessary for the perceptual approach discussed in section 3.3 to be implementable.
- Next, section 4.2 discusses using the approach proposed in section 3.3 to formulate two perceptual sound zone algorithms.

4.1 Proposed Short-Time Frequency-Domain Reformulation of the Pressure Matching Approach

In section 3.3 a perceptual sound zone approach is proposed based on the pressure matching approach discussed in section 3.2. In this approach, the Par detectability measure is used to quantify the perceptual cost of sound pressure errors. In doing so, section 3.2 introduces the reproduction error detectability $\text{RED}_{\mathcal{Z}}^{(m)}$ and the leakage error detectability $\text{LED}_{\mathcal{Z}}^{(m)}$ per control point m introduced conceptually, but without defining them.

As noted, the original pressure matching approach from section 3.2 operates on full-length input sequences in the time-domain. The detectability however operates on short-time segments of 20 to 200 milliseconds in the frequency-domain. As such, in order to define the reproduction error detectability and the leakage error detectability, this section proposes a short-time frequency-domain pressure matching approach.

First in subsection 4.1.1 the existing pressure matching approach is reformulated to operate on short-time segments through a "block-based" approach. Next, subsection 4.1.2 adapts the short-time pressure matching algorithm to operate in the frequency domain.

4.1.1 Proposed Block-Based Pressure-Matching Algorithm

In order to operate on short-time segments, all quantities introduced in the data model from section 3.1 are converted to their short-time equivalent representations. This is done by expressing quantities using overlapping blocks of samples of the quantities.

Here, the blocks are each of size N_w and $N_w - H$ samples. The constant H denotes the hop size, which is the number of samples between each successive block.

First, the short-time equivalent representations of the desired playback signal $s_{\mathcal{Z}}[n]$ and the loudspeaker input signals $x_{\mathcal{Z}}^{(l)}[n]$ for zone $\mathcal{Z}$ and loudspeaker l are discussed.

In order to do so, $s_{\mathcal{Z}}[n]$ and $x_{\mathcal{Z}}^{(l)}[n]$ will be split up into multiple overlapping blocks by using shifted windows $w[n - kH]$.

The window $w[n] \in \mathbb{R}^H$ is a non-causal window with support $-N_w + 1 \leq n \leq 0$. Here, $w[n]$ is chosen such that it complies with the Constant Overlap Add (COLA) condition for a given hop size H. The COLA condition requires that the the sum of all H-shifted windows add to unity for all samples n. It is given as follows:

$$\sum_{k=-\infty}^{\infty} w[n - kH] = 1 \quad \forall\, n \tag{4.1}$$

Using the windows as defined above, consider the following representation of $s_{\mathcal{Z}}[n]$,

$$
\begin{aligned}
s_{\mathcal{Z}}[n] &= s_{\mathcal{Z}}[n] \sum_{k=-\infty}^{\infty} w[n-kH] \\
&= \sum_{k=-\infty}^{\infty} \tilde{s}_{\mathcal{Z},k}[n] w[n-kH]
\end{aligned}
\tag{4.2}
$$

and of $x_{\mathcal{Z}}^{(l)}[n]$,

$$
\begin{aligned}
x_{\mathcal{Z}}^{(l)}[n] &= x_{\mathcal{Z}}^{(l)}[n] \sum_{k=-\infty}^{\infty} w[n-kH] \\
&= \sum_{k=-\infty}^{\infty} \tilde{x}_{\mathcal{Z},k}^{(l)}[n] w[n-kH]
\end{aligned}
\tag{4.3}
$$

Here, $\tilde{s}_{\mathcal{Z},k}[n]$ and $\tilde{x}_{\mathcal{Z},k}^{(l)}[n]$ represent the content of the k^{th} blocks of the playback signal $s_{\mathcal{Z}}[n]$ and loudspeaker input signals $x_{\mathcal{Z}}^{(l)}[n]$.

As such, $\tilde{s}_{\mathcal{Z},k}[n] = s_{\mathcal{Z}}[n]$ and $\tilde{x}_{\mathcal{Z},k}^{(l)}[n] = s_{\mathcal{Z}}[n]$ for $-N_w + 1 + kH \leq n \leq kH$ and zero for all other samples n. One interpretation is that the windows decimate the signal, into segments of size N_w, which, due to the COLA condition, can be reconstructed perfectly.

One way of interpreting the equations above is as a projection of $s_{\mathcal{Z}}[n]$ and $x_{\mathcal{Z}}^{(l)}[n]$ on a basis of frames spanned by shifted overlapping windows $w[n-kH]$. Here, $\tilde{s}_{\mathcal{Z},k}[n]$ and $\tilde{x}_{\mathcal{Z},k}^{(l)}[n]$ can be thought of as the coefficients for the basis functions resulting from the projection.

Let $\tilde{s}_{\mathcal{Z}}[n,\mu]$ and $\tilde{x}_{\mathcal{Z}}^{(l)}[n,\mu]$ represent the desired playback signal and the loudspeaker input signals with the contributions up to and including the μ^{th} block. This can be expressed as follows:

$$
\tilde{s}_{\mathcal{Z}}[n,\mu] = \sum_{k=-\infty}^{\mu} \tilde{s}_{\mathcal{Z},k}[n] w[n-kH] \tag{4.4}
$$

$$
\tilde{x}_{\mathcal{Z}}^{(l)}[n,\mu] = \sum_{k=-\infty}^{\mu} \tilde{x}_{\mathcal{Z},k}^{(l)}[n] w[n-kH] \tag{4.5}
$$

This form will converge to the real desired playback signal as $\mu \to \infty$. As such, $\tilde{s}_{\mathcal{Z}}[n,\infty] = s_{\mathcal{Z}}[n]$ and $\tilde{x}_{\mathcal{Z}}^{(l)}[n,\infty] = x_{\mathcal{Z}}^{(l)}[n]$.

This representation is beneficial, as it can be used to show that the $\tilde{x}_{\mathcal{Z}}^{(l)}[n,\mu]$ can be computed recursively:

$$
\tilde{x}_{\mathcal{Z}}^{(l)}[n,\mu] = \tilde{x}_{\mathcal{Z},\mu}^{(l)}[n] w[n-\mu H] + \sum_{k=-\infty}^{\mu-1} \tilde{x}_{\mathcal{Z},k}^{(l)}[n] w[n-kH] \tag{4.6}
$$

$$
= \tilde{x}_{\mathcal{Z},\mu}^{(l)}[n] w[n-\mu H] + \tilde{x}_{\mathcal{Z}}^{(l)}[n,\mu-1] \tag{4.7}
$$

As the newest block depends on the previous blocks, this representation shows that $x_{\mathcal{Z}}^{(l)}[n]$ can be computed block-by-block.

With the block-based equivalents of the desired playback signal $\tilde{s}_{\mathcal{Z}}[n,\mu]$ and the loudspeaker input signals $\tilde{s}_{\mathcal{Z}}[n,\mu]$ defined, the block-based equivalents of the target and achieved sound pressure $\tilde{t}_{\mathcal{Z}}[n,\mu]$ and $\tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu]$ can be computed:

- The block-based target sound pressure $\tilde{t}^{(m)}[n,\mu]$ can be defined by simply substituting the definition for the block-based desired playback signal $\tilde{s}_{\mathcal{Z}}[n,\mu]$ into the definition of the target pressure given by Equation 3.3:

$$\begin{aligned}\tilde{t}^{(m)}[n,\mu] &= \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * \tilde{s}_{\mathcal{Z}}[\mu]\right)[n] \\ &= \sum_{l=0}^{N_L-1} \sum_{k=-\infty}^{\mu} \left(h^{(l,m)} * \tilde{s}_{\mathcal{Z},k} w_k\right)[n] \\ &= \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * \tilde{s}_{\mathcal{Z},\mu} w_\mu\right)[n] + \tilde{t}^{(m)}[n,\mu-1]\end{aligned} \tag{4.8}$$

  Here, $w_k[n]$ is defined to be equal to $w[n-kH]$ and is introduced for notational convenience. The definition above holds for all points $m \in Z$, i.e. the points contained in zone $\mathcal{Z}$.

  As can be seen, the block based target sound pressure for the block μ can be computed recursively by adding the contribution of the newest block $\tilde{s}_{\mathcal{Z},\mu}[n]$ the target sound pressure of the previous block.

- The block-based resulting sound pressure $\tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu]$ can be defined by simply substituting the definition for the block-based loudspeaker input signals $\tilde{x}_{\mathcal{Z}}^{(l)}[n,\mu]$ into the definition of the resulting pressure given by Equation 3.2. This results in the following:

$$\begin{aligned}\tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu] &= \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * \tilde{x}_{\mathcal{Z}}^{(l)}[\mu]\right)[n] \\ &= \sum_{l=0}^{N_L-1} \sum_{k=-\infty}^{\mu} \left(h^{(l,m)} * \tilde{x}_{\mathcal{Z},k}^{(l)} w_k\right)[n] \\ &= \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * \tilde{x}_{\mathcal{Z},\mu}^{(l)} w_\mu\right)[n] + \tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu-1]\end{aligned} \tag{4.9}$$

  The definition above again holds for all points $m \in Z$.

  As can be seen, the block based resulting sound pressure for the block μ can also be computed recursively.

With this, all quantities required for the block-based formulation of the pressure matching approach are defined.

It is shown that all quantities can be computed recursively. This is used in the block-based pressure matching approach by computing the blocks of the loudspeaker input signal $x_{\mathcal{Z}}^{(l)}[n]$ one by one.

As such, the k^{th} loudspeaker input signal coefficient $\tilde{x}_{\mathcal{Z},\mu}^{(l)}[n]$ is computed such that the resulting resulting sound pressure $\tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu]$ best matches the target sound pressure $\tilde{t}^{(m)}[n,\mu]$.

Note that in this approach only newest loudspeaker coefficients $\tilde{x}_{\mathcal{Z},\mu}^{(l)}$ are being controlled. Thus, the previous coefficients $\tilde{x}_{\mathcal{Z},k}^{(l)}[n]$ for $-\infty \leq k \leq \mu - 1$ are held fixed.

The block-based optimization problem can be found by simply replacing all quantities in the previously derived optimization problem with their block-based counterparts. The problem is given as follows:

$$\begin{aligned}\underset{\tilde{x}_{\mathcal{A},\mu}^{(l)}[n],\,\tilde{x}_{\mathcal{B},\mu}^{(l)}[n]\,\forall\, l}{\arg\min} \quad & \sum_{m\in A}\left|\left|\tilde{p}_{\mathcal{A}}^{(m)}[n,\mu]-\tilde{t}_{\mu}^{(m)}[n,\mu]\right|\right|_2^2+\sum_{m\in A}\left|\left|\tilde{p}_{\mathcal{B}}^{(m)}[n,\mu]\right|\right|_2^2+ \\ & \sum_{m\in B}\left|\left|\tilde{p}_{\mathcal{B}}^{(m)}[n,\mu]-\tilde{t}_{\mu}^{(m)}[n,\mu]\right|\right|_2^2+\sum_{m\in B}\left|\left|\tilde{p}_{\mathcal{A}}^{(m)}[n,\mu]\right|\right|_2^2\end{aligned} \tag{4.10}$$

Note that this problem implicitly contains the target sound pressure and resulting sound pressure of the previous blocks $-\infty \leq k \leq \mu - 1$ due to the aforementioned recursive definitions. As a result, the history of what has been transmitted by the loudspeaker previously is included in the optimization.

The problem above is solved recursively for all loudspeaker input signal coefficients $\tilde{x}_{\mathcal{A},\mu}^{(l)}[n]$ and $\tilde{x}_{\mathcal{B},\mu}^{(l)}[n]$. The final loudspeaker input signals $x_{\mathcal{Z}}^{(l)}[n]$ can then be found by means of Equation 4.3.

4.1.2 Proposed Block-Based Frequency-Domain Pressure-Matching Algorithm

This section will adjust the block-based data model equivalent frequency domain formulation in order to propose a block-based frequency-domain pressure-matching algorithm. This is done by first introducing a transformation relating the time and frequency domain quantities.

A suitable transform is the discrete Fourier transform (DFT). However, it is important to take a number of precautions before applying the DFT directly. As shown in Equation 3.2 the computation of the sound pressures used in the optimization problem introduced previously involves taking the linear convolution of the loudspeaker input signals with the room impulse responses.

Time domain circular convolution can be computed in the frequency domain through the Hadamard product. Time domain circular convolution coincides with time domain linear convolution only if the two operands are zero-padded sufficiently. To be specific, both operands need be zero-padded to the length of the resulting linear convolution.

As such, the frequency domain transform requires this zero padding to be built in. The convolutions described in the previous chapter are between the window coefficients of size N_w and the room impulse responses of size N_h. Thus, the both must be zero padded to convolution length N_w+N_h-1 before going to the frequency domain.

Let $x[n]$ and $X[k]$ denote the time- and frequency-domain representations of an arbitrary sequence. A suitable transform is given by the following $N_w + N_h - 1$ point DFT:

$$X[k] = \sum_{n=0}^{N_w+N_h-2} x[n] \exp\left(\frac{-j2\pi kn}{N_w + N_h - 1}\right) \tag{4.11}$$

Converting the previously introduced block-based pressure matching to a frequency domain equivalent version essentially involves converting the sound pressures $\tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu]$ and $\tilde{t}^{(m)}[n,\mu]$ to their frequency domain counterparts, which are denoted by $\tilde{P}_{\mathcal{Z},\mu}^{(m)}[k]$ and $\tilde{T}_{\mu}^{(m)}[k]$ respectively.

This results in the following expressions.

$$\tilde{T}^{(m)}[k,\mu] = \tilde{T}^{(m)}[k,\mu-1] + \sum_{l=0}^{N_L} H^{(l,m)}[k]\tilde{S}_{\mathcal{Z},\mu}[k] \tag{4.12}$$

$$\tilde{P}_{\mathcal{Z}}^{(m)}[k,\mu] = \tilde{P}_{\mathcal{Z}}^{(m)}[k,\mu-1] + \sum_{l=0}^{N_L} H^{(l,m)}[k]\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k] \tag{4.13}$$

Here, $H^{(l,m)}[k] \in \mathbb{C}^{N_w+N_h-1}$ is the transformed version of the room impulse responses.

Furthermore, $\tilde{S}_{\mathcal{Z},\mu}[k] \in \mathbb{C}^{N_w+N_h-1}$ and $\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k] \in \mathbb{C}^{N_w+N_h-1}$ are the frequency domain versions of the desired playback signal and the loudspeaker input signal, which are defined as follows:

$$\tilde{S}_{\mathcal{Z},\mu}[k] = \sum_{n=0}^{N_w+N_h-2} \tilde{s}_{\mathcal{Z},\mu}[n]w[n-\mu H] \exp\left(\frac{-j2\pi kn}{N_w + N_h - 1}\right) \tag{4.14}$$

$$\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k] = \sum_{n=0}^{N_w+N_h-2} \tilde{x}_{\mathcal{Z},\mu}^{(l)}[n]w[n-\mu H] \exp\left(\frac{-j2\pi kn}{N_w + N_h - 1}\right) \tag{4.15}$$

Note that the window is implicitly included in the transformed quantities. This is done for ease of notation.

Using the previously derived quantities, it is possible express the frequency domain version of the block-based pressure matching approach as follows:

$$\underset{\tilde{x}_{\mathcal{A},\mu}^{(l)}[n],\, \tilde{x}_{\mathcal{B},\mu}^{(l)}[n]\,\forall\, l}{\arg\min} \quad \sum_{m\in A} \left|\left|\tilde{P}_{\mathcal{A}}^{(m)}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right|\right|_2^2 + \sum_{m\in A} \left|\left|\tilde{P}_{\mathcal{B}}^{(m)}[k,\mu]\right|\right|_2^2 + \tag{4.16}$$

$$\sum_{m\in B} \left|\left|\tilde{P}_{\mathcal{B}}^{(m)}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right|\right|_2^2 + \sum_{m\in B} \left|\left|\tilde{P}_{\mathcal{A}}^{(m)}[k,\mu]\right|\right|_2^2 \tag{4.17}$$

Note how the optimization is still performed over the time domain signal. This was done to constrain the loudspeaker input signal coefficient to size N_w, as that is an assumption made by the frame-based processing.

In principal, this introduces more complexity than solving directly over the frequency domain loudspeaker input coefficient $\tilde{X}^{(l)}_{\mathcal{Z},\mu}[k]$. This however introduces issues as it requires the truncation of the time-domain version to the first N_w samples.

Naively truncating N_w this way introduced artifacts. In experiments in which this approach is attempted the time-domain representation of the resulting frequency-domain loudspeaker input signals results in significant energy contained in the last $N_h - 1$ samples. If truncated, a significant portion of the signal energy would be disregarded, which serves as a possible explanation for the artifacts.

However, due to the computational benefits, formulating a frequency domain approach that minimizes the impact of or prevents these artifacts is found to be promising future work.

4.2 Proposal of Proposed Perceptual Pressure Matching Algorithms

Previously, section 3.3 noted that pressure matching approach can be formulated using the detectability. In this section, rather than optimizing the sum reproduction errors $\mathrm{RE}_{\mathcal{Z}}^{(m)}$ and leakage errors $\mathrm{LE}_{\mathcal{Z}}^{(m)}$, it is proposed to instead use the reproduction error detectability $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ and the leakage error detectability $\mathrm{LED}_{\mathcal{Z}}^{(m)}$ respectively. These can be understood to be perceptual alternatives to $\mathrm{RE}_{\mathcal{Z}}^{(m)}$ and $\mathrm{LE}_{\mathcal{Z}}^{(m)}$.

In section 3.3 the definition of these quantities is given using full-length input sequences. As noted, this is inaccurate as the detectability operate on short-time scale, and is only done this way in order to clearly convey the concept.

Henceforth, using the concepts introduced in section 4.1, let $\mathrm{RED}_{\mathcal{Z}}^{(m)}[\mu]$ and $\mathrm{LED}_{\mathcal{Z}}^{(m)}[\mu]$ denote the reproduction error detectability and leakage error detectability for block μ in control point m. These be understood as the detectability of the error introduced in due to the current block μ.

In order to define these error detectability quantities, recall that detectability is defined as follows:

$$D(x[n],\, \varepsilon[n]) = ||W_x[k]\mathcal{E}[k]||_2^2 \tag{4.18}$$

Using this definition alongside the short-time frequency domain definitions given in section 4.1, the definition of the reproduction error detectability $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ and the leakage error detectability $\mathrm{LED}_{\mathcal{Z}}^{(m)}$ can be given as follows:

$$\begin{aligned} \mathrm{RED}_{\mathcal{Z}}^{(m)}[\mu] &= D(\tilde{t}^{(m)}[n,\mu],\, \tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu] - \tilde{t}^{(m)}[n,\mu]) \\ &= \left|\left| W_{\tilde{t}^{(m)}[\mu]}[k] \left(\tilde{P}_{\mathcal{Z}}^{(m)}[k,\mu] - \tilde{T}^{(m)}[k,\mu] \right) \right|\right|_2^2 \end{aligned} \tag{4.19}$$

$$\begin{aligned} \mathrm{LED}_{\mathcal{Z}}^{(m)}[\mu] &= D(\tilde{t}^{(m)}[n,\mu],\, \tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu]) \\ &= \left|\left| W_{\tilde{t}^{(m)}[\mu]}[k] \left(\tilde{P}_{\mathcal{Z}}^{(m)}[k,\mu] \right) \right|\right|_2^2 \end{aligned} \tag{4.20}$$

Here, $W_{\tilde{t}^{(m)}[\mu]}[k]$ can be understood as the perceptual weighting informed by the masking properties of the frequency domain target $\tilde{t}^{(m)}[n,\mu]$.

What follows is the proposal of two perceptual sound zone algorithms using the proposed error detectabilities.

4.2.1 Proposed Unconstrained Perceptual Pressure Matching Algorithm

This section proposes an algorithm which minimizes the detectability of the total error. This is similar to the pressure matching approach introduced in section 3.2, in which the total error is minimized.

Consider the following optimization problem:

$$\begin{aligned}\underset{\tilde{x}_{\mathcal{A}}^{(l)}[n,\mu],\,\tilde{x}_{\mathcal{B}}^{(l)}[n,\mu]\,\forall\, l}{\arg\min} \quad & \sum_{m\in A} \mathrm{RED}_{\mathcal{A}}^{(m)}[\mu] + \sum_{m\in B} \mathrm{LED}_{\mathcal{A}}^{(m)}[\mu] + \\ & \sum_{m\in B} \mathrm{RED}_{\mathcal{B}}^{(m)}[\mu] + \sum_{m\in A} \mathrm{LED}_{\mathcal{B}}^{(m)}[\mu]\end{aligned} \tag{4.21}$$

The total detectability of the reproduction errors and the leakage errors is minimized by optimizing over the block-based representations of the loudspeaker input signals $\tilde{x}_{\mathcal{A}}^{(l)}[n,\mu]$ and $\tilde{x}_{\mathcal{B}}^{(l)}[n,\mu]$.

4.2.2 Proposed Constrained Perceptual Pressure Matching Algorithm

The previously discussed approach minimizes over the total detectability. In this section, a perceptual sound zone algorithm is proposed that introduces constraints the problem.

The motivation for this approach is that was found that the Par distortion detectability has a consistent perceptual interpretation. For example, as mentioned in section 2.3, a detectability of 1 will consistently imply "just noticeable" [20].

This makes choosing constraints for detectability easier than typical non-perceptual pressure matching approaches. These approaches typically directly constrain the sound pressure energy, for which it is difficult to determine constraints as the sound pressure energy does not have a consistent perceptual interpretation.

This motivates the proposal of a perceptually constrained sound pressure approach. In this approach, the reproduction error detectability will be constrained, while the leakage error detectability will be minimized in the cost function. To this end, the following optimization problem is defined:

$$\begin{aligned}\underset{\tilde{x}_{\mathcal{A}}^{(l)}[n,\mu],\,\tilde{x}_{\mathcal{B}}^{(l)}[n,\mu]\,\forall\, l}{\arg\min} \quad & \sum_{m\in B} \mathrm{LED}_{\mathcal{A}}^{(m)}[\mu] + \sum_{m\in A} \mathrm{LED}_{\mathcal{B}}^{(m)}[\mu] \\ \text{subject to} \quad & \mathrm{RED}_{\mathcal{A}}^{(m)}[\mu] \leq D_0 \quad \forall\, m \in A \\ & \mathrm{RED}_{\mathcal{B}}^{(m)}[\mu] \leq D_0 \quad \forall\, m \in B\end{aligned} \tag{4.22}$$

Here, D_0 is the maximum allowed detectability of the reproduction error.

The optimization problem trades-off between reproduction error and leakage: relaxing the constraint on the detectability of the reproduction error should result in more mistakes in the reproduced sound pressure., this relaxation should however also allow for a greater minimization of the leakage detectability.

It should be noted that constraining the leakage error detectability is also a valid choice. To the knowledge of the authors there is no precedent in literature in which a reproduction error is constrained. As such, constraining the reproduction error detectability is found to be a more interesting exploration.

Constraining the detectability of the leakage error is found to be promising future work.

Evaluation of Perceptual Sound Zone Algorithms

5

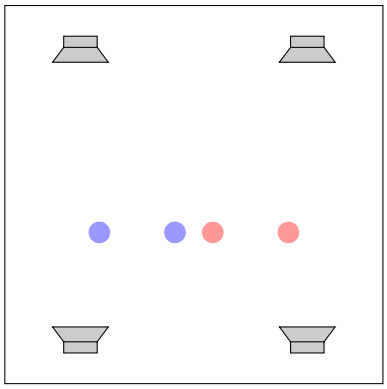


Figure 5.1: The room setup used in the simulations for the evaluation of the algorithms.

5.1 Evaluation Methodology

In chapter 3 and chapter 4 various sound zone algorithms were defined. First, in chapter 3 a non-perceptual pressure matching approach was introduced. In the following chapter two perceptual sound zone algorithms were introduced, namely unconstrained perceptual pressure matching and perceptually constrained pressure matching.

This section will discuss the general approach for evaluation of the previously derived algorithms. All algorithms will be evaluated by means of simulation. As all derived algorithms were found to be computationally intensive, many of the design consideration for the simulations were chosen such to keep the computational complexity low.

5.1.1 Room Setup

All sound zone algorithms were evaluated in a simulated square room of 5 by 5 meters, with a ceiling height of 3.4 meters. The room contained 4 loudspeakers placed in the corners of the room at a height of 1.2 meters.

Two zones each consisting of two control points are in the middle of the room. In order to obtain a challenging problem, the zones are placed in close proximity, which the two closest points 0.5 meters apart.

An image depicting the entire setup is given in Figure 5.1.

5.1.2 Content Selection

For the evaluation speech audio content was used. The motivation for this is that it is computationally intensive to run the algorithms at higher sampling rate such as 48 kHz, and speech signals can be represented well at lower sampling rates. In addition to this, the objective speech quality measures described in chapter 2 were found to be more robust. Many general audio quality measures were found to have little to no official openly available implementations.

A set of 4 speech signals is used for evaluation. For each experiment, a speech

signal was assigned to each zone. All possible combinations of the speech signals were made, resulting in 12 total experiments.

5.1.3 Simulation Outputs

After running the simulations, all output will be written to file. For ease of evaluation, the following sound pressures is available:

- **Target Sound Pressure per Control Point m:**
  This is the desired sound pressure per control point m. This can be understood as the sound pressure that the algorithm attempts to attain. In the derivations of chapter 3, this was denoted by $t^{(m)}[n]$.
- **Achieved Target Sound Pressure per Control Point m:**
  This is the target sound pressure achieved by the algorithm per control point m, excluding the interference coming from other zones. This quantity allows for the evaluation of the quality of intended sound pressure per control point, isolated from all interference. In the two zone case for any $m \in A$, this corresponds to $p_{\mathcal{A}}^{(m)}[n]$ in previous derivations.
- **Achieved Interfering Sound Pressure per Control Point m:**
  This is the interference sound pressure achieved by the algorithm per control point m, isolated the target sound pressure that was intended for that control point. This allows for analysis of the interference in isolation from the intended sound pressure. Again for the two zone case for any $m \in A$, this corresponds to $p_{\mathcal{B}}^{(m)}[n]$.
- **Total Achieved Sound Pressure per Control Point m:**
  This is the total sound pressure per control point m achieved by the algorithm. This can be understood as the sound pressure that would be experienced by a listener standing in the position of point m, and is equal to the sum over the achieved target sound pressure and the achieved interfering sound pressure previously discussed. In the derivations for the two zone case, this corresponds to to $p_{\mathcal{A}}^{(m)}[n] + p_{\mathcal{B}}^{(m)}[n]$ for any m.

The achieved target sound pressure and the achieved interfering sound pressure can thought of as decompositions of the total achieved sound pressure. This decomposition was found to be useful during evaluation.

5.1.4 Evaluation Criteria

In section 2.1 a literature review into various perceptual grading models from literature was discussed. As mentioned before, speech audio is used as the content for the sound zones. As such, it makes sense to use metrics that specialize in speech for evaluation.

For this reason, one of the metrics that will be used is the Perceptual Evaluation of Speech Quality (PESQ). As described in subsection 2.1.1, PESQ is a metric which grades the quality of a degraded speech signal with respect to a reference speech

signal. The resulting quality grade will be between 0 and 5, where 5 is the highest obtainable grade.

PESQ will be used to evaluate two aspects of the result.

- **PESQ of the Achieved Target Sound Pressure with respect to the Target Sound Pressure per Control Point m:**
  As described, the achieved target sound pressure is the sound pressure per control point m, without the interference due to the other zones.

  This will henceforth referred to as the **"Target PESQ"**. This quantity thus describes the quality of the intended sound pressure in isolation from interference.

- **PESQ of the Total Achieved Sound Pressure with respect to the Target Sound Pressure per Control Point m:**
  As previously described, the total achieved sound pressure is the sound pressure that would be audible to an individual present at control point m. This includes the intended sound pressure, but also the interference coming from other zones.

  This quantity will be referred to as the **"Overall PESQ"** and represents the quality of the total experience per control point m.

Another metric that will be used for evaluation is the Distraction model also introduced in subsection 2.1.1. This model grades how distracting an interferer is in presence of target audio. The grade uses a scale from 0 to 100, where 100 is considered maximally distracting.

The distraction will be used as follows:

- **Distraction of the Achieved Interfering Sound Pressure with respect to the Achieved Target Sound Pressure per Control Point m:**
  This quantifies how distracting the interfering sound pressure is when listening to the indented sound pressure per control point m. This is a way of quantifying how disturbing the interference of the resulting algorithms is.

  This will simply be referred to as the **"Distraction"** per control point from now on.

5.2 Evaluation of Proposed Algorithms

In section 4.2 two perceptual sound zone algorithms were introduced. First, an unconstrained perceptual pressure matching approach where the detectability of the sound pressure errors is minimized. Secondly, a constrained perceptual pressure matching approach which leverages the fact that the detectability has a consistent perceptual interpretation to constrain the detectability of the reproduction error.

This section will explore the results of both approaches.

Statistic		**Reference**	**Perceptual**
Overall PESQ	Mean	2.61	3.15
	Standard Deviation	0.247	0.240
Target PESQ	Mean	4.11	3.35
	Standard Deviation	0.065	0.254
Distraction	Mean	12.7	7.8
	Standard Deviation	7.46	5.89

Table 5.1: Summary of results for the evaluation of the unconstrained perceptual pressure matching approach, using the evaluation metrics defined in section 5.1.

5.2.1 Evaluating Unconstrained Perceptual Pressure Matching

In this section, the unconstrained perceptual pressure matching algorithm will be evaluated. This algorithm minimized the total detectability of the reproduction and the leakage error. As mentioned in section 5.1, the perceptual pressure matching approaches will be evaluated with respect to the reference pressure matching algorithm introduced in section 3.2. The unconstrained reference algorithm instead minimizes the total reproduction and leakage error.

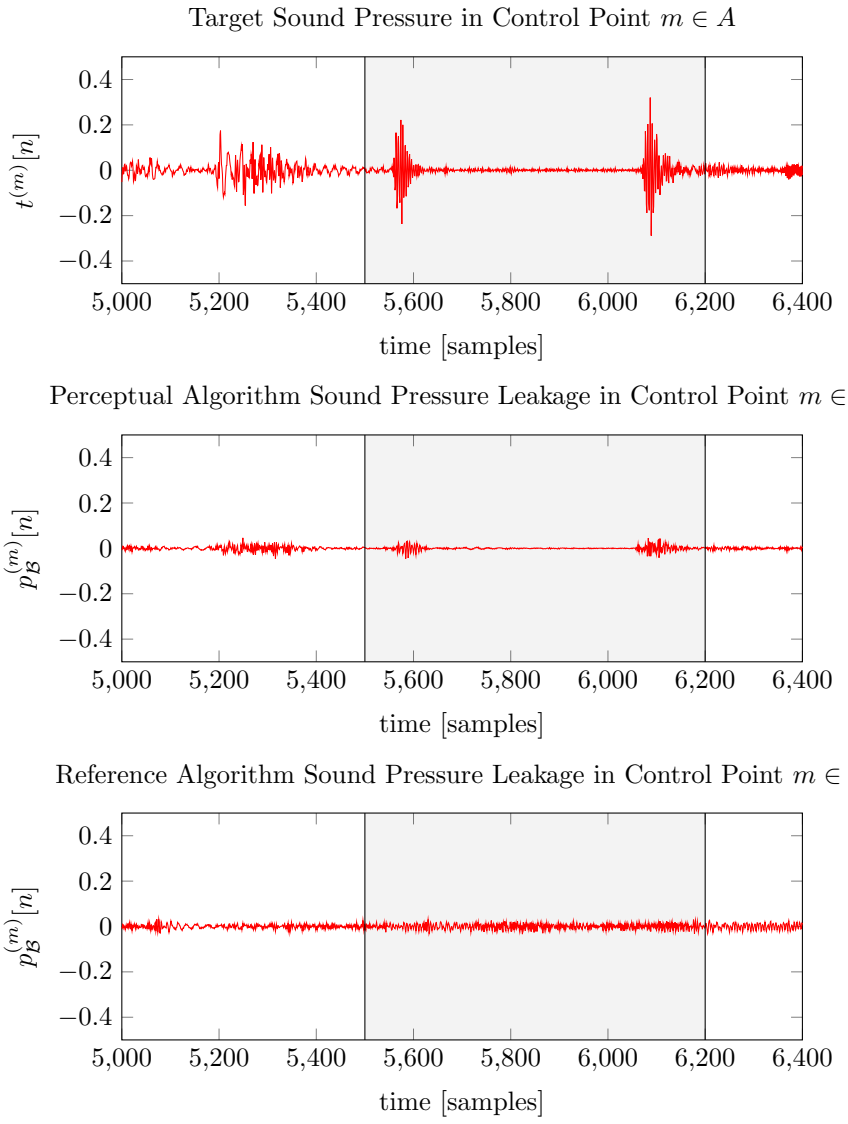
The experiment consists of 12 simulations which use speech signals for both zones. For evaluation, 3 metrics are used as introduced by section 5.1. First, Overall PESQ, which quantifies the quality of the total experience: achieved intended audio and interference. Secondly, Target PESQ, which quantifies the quality of just the achieved intended audio (excluding interference). Finally, Distraction, which determines how distracting the interference is. These three quantities were found to be representative for the sound zone experience.

The results of the experiment are summarized in Table 5.1. This table depicts the mean and standard deviation of the metrics taken over all 4 control points and all 12 experiments. From this, the following conclusions are drawn:

- The perceptual sound zone algorithm outperforms the reference sound zone algorithm in Overall PESQ. Thus, this implies that perceptual sound zone algorithm delivers greater quality of speech.
- The reference sound zone algorithm outperforms the perceptual sound zone algorithm in Target PESQ. This implies that the reference algorithm, when disregarding interference, delivers greater quality for the intended speech signal.
- The perceptual sound zone algorithm outperforms the reference with regards to Distraction. This can be understood as the reference results in more distracting sound zones than the perceptual sound zone algorithm.

From the results above, the Overall PESQ and the Distraction indicate that the perceptual sound zone algorithm seems to achieve an better overall experience. The reference sound zone algorithm achieves a greater Target PESQ, however after adding the interference, the Overall PESQ is lower.

This implies that the the perceptual sound zone algorithm makes a better over-



all perceptual trade-off between producing the target content and minimizing the leakage.

5.2.2 Evaluating Constrained Perceptual Pressure Matching

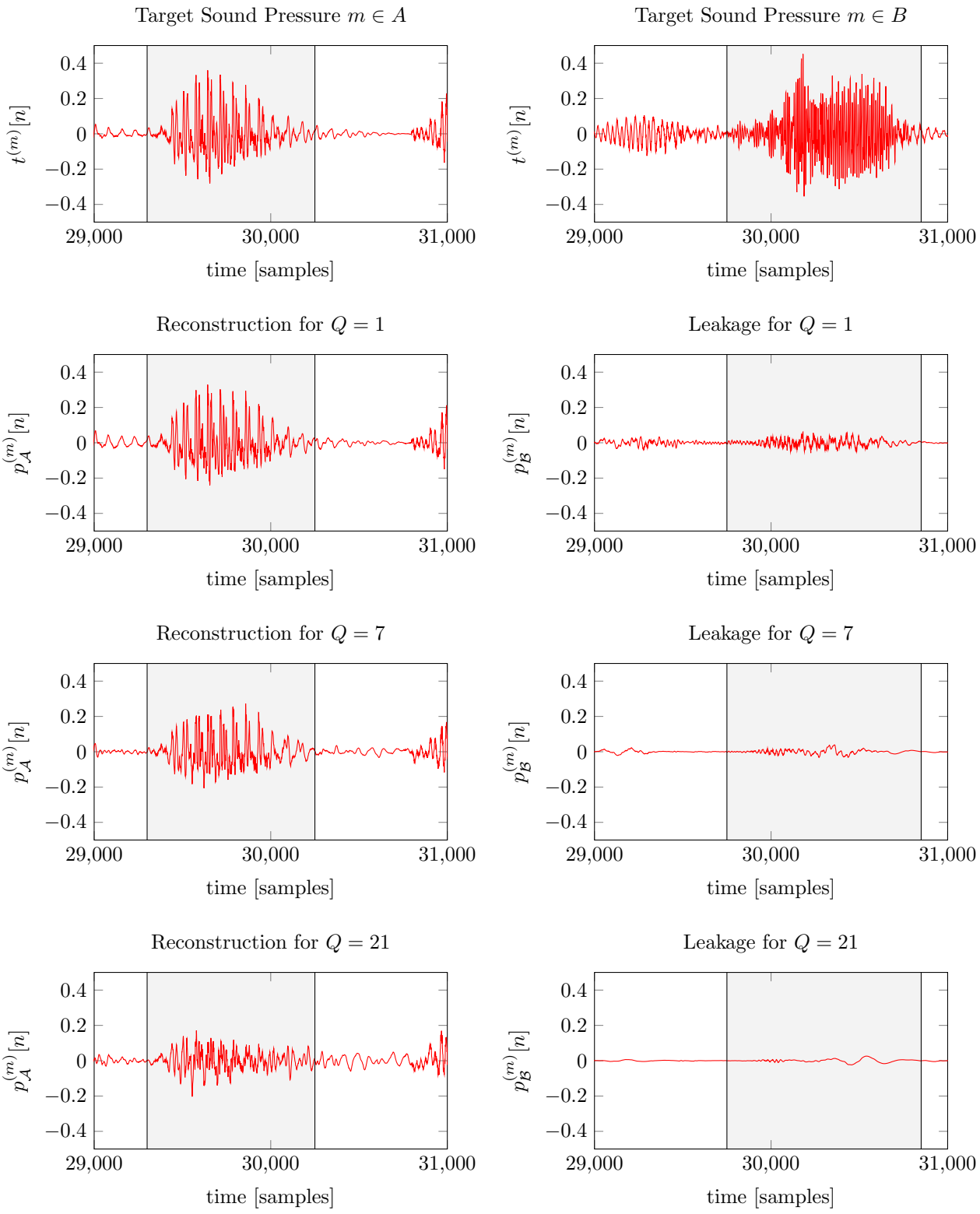


Figure 5.2: Plots depicting stuff.

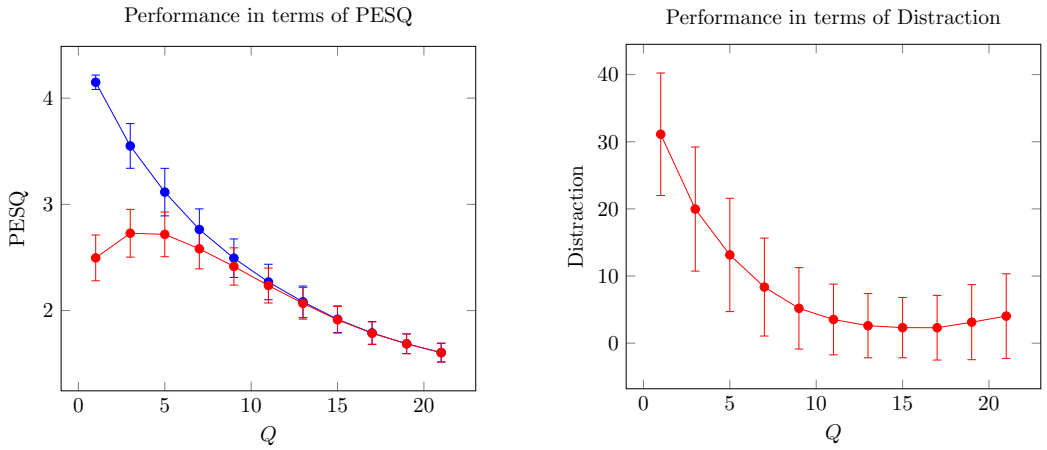

Performance in terms of PESQ
PESQ
4
3
2
0
5
10
15
20
Q
Performance in terms of Distraction
Distraction
40
30
20
10
0
0
5
10
15
20
Q

Figure 5.3

6 Conclusion

Bibliography

[1] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 81–91, 2015.

[2] J. Donley and C. H. Ritz, "Multizone reproduction of speech soundfields: A perceptually weighted approach," 2015.

[3] T. Lee, J. K. Nielsen, and M. G. Christensen, "Towards perceptually optimized sound zones: A proof-of-concept study," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 136–140.

[4] ——, "Signal-adaptive and perceptually optimized sound zones with variable span trade-off filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2412–2426, 2020.

[5] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[6] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[7] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[8] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "Visqol: The virtual speech quality objective listener," in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*. VDE, 2012, pp. 1–4.

[9] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," pp. 1–6, 2020.

[10] J. Kim, M. El-Kharmy, and J. Lee, "End-to-end multi-task denoising for joint sdr and pesq optimization," *arXiv preprint arXiv:1901.09146*, 2019.

[11] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2017.

[12] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "Peaq-the itu standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[13] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "Visqolaudio: An objective audio quality metric for low bitrate codecs," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL449–EL455, 2015.

[14] J. Francombe, R. Mason, M. Dewhirst, and S. Bech, "A model of distraction in an audio-on-audio interference situation with music program material," *Journal of the Audio Engineering Society*, vol. 63, no. 1/2, pp. 63–77, 2015.

[15] ——, "Elicitation of attributes for the evaluation of audio-on-audio interference," *The Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2630–2641, 2014.

[16] J. Rämö, S. Bech, and S. H. Jensen, "Real-time perceptual model for distraction in interfering audio-on-audio scenarios," *IEEE Signal Processing Letters*, vol. 24, no. 10, pp. 1448–1452, 2017.

[17] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1553–1564, 2012.

[18] I. J. S. 29, "Information technology — coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s — part 3: Audio," International Organization for Standardization, Geneva, CH, techreport 3, Oct. 1993.

[19] D. Pan, "A tutorial on mpeg/audio compression," *IEEE multimedia*, vol. 2, no. 2, pp. 60–74, 1995.

[20] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1–13, 2005.

[21] P. Balazs, B. Laback, G. Eckel, and W. A. Deutsch, "Time–frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 1, pp. 34–49, 2009.

[22] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, 2013.

[23] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[24] L. Vindrola, M. Melon, J.-C. Chamard, B. Gazengel, and G. Plantier, "Personal sound zones: A comparison between frequency and time domain formulations

in a transportation context," in *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019.

[25] M. Olik, J. Francombe, P. Coleman, P. J. Jackson, M. Olsen, M. Møller, R. Mason, and S. Bech, "A comparative performance study of sound zoning methods in a reflective environment," in *Audio Engineering Society Conference: 52nd International Conference: Sound Field Control-Engineering and Perception*. Audio Engineering Society, 2013.

[26] S. J. Elliott and J. Cheer, "Regularisation and robustness of personal audio systems," 2011.

[27] Y. Cai, M. Wu, L. Liu, and J. Yang, "Time-domain acoustic contrast control design with response differential constraint in personal audio systems," *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. EL252–EL257, 2014.

[28] M. B. Møller and M. Olsen, "Sound zones: On performance prediction of contrast control methods," in *Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control*. Audio Engineering Society, 2016.

[29] T. Lee, J. K. Nielsen, J. R. Jensen, and M. G. Christensen, "A unified approach to generating sound zones using variable span linear filters," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 491–495.

[30] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

Calibration of the the Par Detectability Measure

A

A calibration is necessary for the Par detectability to provide the correct output. In the previous section, it was shown that the constants C_a and C_s are used to this end. A correct calibration of the Par detectability must satisfy the following:

1. The just noticeable disturbance signal must result in a detectability of 1.
2. The threshold of hearing takes effect appropriately.

Both the concepts of just noticeable distortion and the threshold of hearing require knowledge of the sound pressure level of the stimuli. Thus, before determining the calibration coefficients, it is important to first discuss the relationship between the input signals $x[n]$ and $\varepsilon[n]$ and reproduced sound pressure level. This is the topic of section A.1. Afterwards, section A.2 discusses the determination of the coefficients C_a and C_s.

A.1 Relating Digital Representation and Sound Pressure Level

One difficulty of taking the threshold of hearing into account is that it is typically given in terms of sound pressure level (SPL), measured in dB. The one-sided spectrum of the threshold of hearing in dB SPL can be approximated by the following function [30]:

$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} + 0.001\left(\frac{f}{1000}\right)^{4} - 6.5\exp\left[-0.6\left(\frac{f}{1000} - 3.3\right)^{2}\right] \quad \text{(A.1)}$$

The signals $x[n]$ and $\varepsilon[n]$ are however given digital representation of audio.

For example, they might be given in a pulse code modulated (PCM) format within which they attain integer values between -32768 and 32767.

As such, to meaningfully integrate the threshold of quiet, the digital representation and the sound pressure levels must be related. This relationship can be modeled as follows:

$$X_{\mathrm{dB}}(f) = 10\log_{10}(|X(f)|^2) + O_{\mathrm{dB}} \quad \text{(A.2)}$$

Here, $X_{\mathrm{dB}}(f)$ is the dB SPL representation of a given spectrum $X(f)$. Furthermore, O_{dB} is an offset to ensure the digital representation corresponds to the correct sound pressure level. In order to use this relationship to determine the appropriate digital equivalent of the threshold in quiet, a definition of the offset O_{dB} must be determined.

One way of determining the offset O_{dB} is by relating the sound pressure level and the digital representation of a full-scale sinusoid. A full-scale sinusoid is a sinusoid that has an amplitude of the maximum value that can be attained in the digital representation.

In our previous example, one way of doing so would be to state that a full-scale sinusoid with amplitude 32767 corresponds to e.g. a sound pressure level of 100 dB SPL. The interpretation of this is that playing a full-scale sinusoid will result in a sound pressure of 100 dB SPL when played from the sound system.

To do so, let the digital representation of the full-scale sinusoid be modeled by a sinusoid with amplitude A and frequency f_0. Consider the one-sided fourier representation of the digital representation of this full-scale sinusoid:

$$\mathcal{F}\left\{A\cos\left(2\pi f_0 t\right)\right\} = A\delta\left(f - f_0\right) \tag{A.3}$$

It is assumed that playing the digital representation of this sinusoid results in a sound pressure level of A_{dB} dB SPL. Substituting these definitions into Equation A.2 results in the following definition for O_{dB}:

$$O_{\text{dB}} = 10\log_{10}\left(|A|^2\right) - A_{\text{dB}} \tag{A.4}$$

The offset fully defines the relationship between digital representation and sound pressure level, and allows for the conversion of the threshold of hearing to digital representation.

A.2 Determing Calibration Constants

There are various ways of calibrating this model, but this section will discuss the method of calibrating that is given in the original paper [20]. The given approach is to find the two unknowns C_a and C_s by solving a system of two equations that model the previously stated calibration requirements.

The first requirement is that a just noticeable disturbance signal must result in a detectability of 1. From perceptual literature it is known that a sinusoidal disturbance signal at a given frequency f_0 is just noticeable in presence of an in-phase sinusoidal masking signal that is 18 dB SPL louder [20]. To model this, consider the following masking and disturbance signals.

$$x_{\text{JND}}[n] = A_{70}\cos\left(2\pi f_0 n/f_s\right) \tag{A.5}$$

$$\varepsilon_{\text{JND}}[n] = A_{52}\cos\left(2\pi f_0 n/f_s\right) \tag{A.6}$$

Here, $x_{\text{JND}}[n]$ is a sinusoid with an amplitude A_{70}, which corresponds to 70 dB SPL. Furthermore, $\varepsilon_{\text{JND}}[n]$ is a sinusoid with an amplitude A_{52}, which is 18 dB SPL less. Note that the amplitudes are both given in digital representation, not sound pressure level representation. The digital representation amplitudes are found through Equation A.2.

Thus, $\varepsilon_{\text{JND}}[n]$ must be just noticeable in presence of $x_{\text{JND}}[n]$. This can be expressed as follows:

$$D(x_{\text{JND}}[n], \varepsilon_{\text{JND}}[n]) = 1 \tag{A.7}$$

This expression forms the first equation in the system of equations that can be solved to calibrate the Par detectability.

The second requirement is that the threshold of hearing must be included correctly. The threshold of hearing defines the sound pressure levels that are the verge between audible and inaudible sound as a function of frequency. To this end, consider the following masking and disturbance signals:

$$x_{\text{THR}}[n] = 0 \tag{A.8}$$

$$\varepsilon_{\text{THR}}[n] = A_{\text{tq}} \cos\left(2\pi f_0 n / f_s\right) \tag{A.9}$$

Here the masking signal $x_{\text{THR}}[n]$ is zero. The disturbance signal is a sinusoid of frequency f_0 with amplitude A_{tq}, which is chosen such that it attains the threshold of quiet at f_0, i.e. $T_q(f_0)$.

As the threshold of quiet is the verge between audible and inaudible sound, it is assumed that a disturbance signal in presence of no masking signal that has an amplitude equal to the threshold of quiet is just noticeable. Recall that for just noticeable stimuli the detectability must be equal to 1. This allows us to specify the second equation in the system of equations:

$$D(0, \varepsilon_{\text{THR}}[n]) = 1 \tag{A.10}$$

The system of equations defined by Equation A.7 and Equation A.10 can be solved through the bisection method. To see how this is done, the reader is referred to the original paper [20].

Extra Results

B

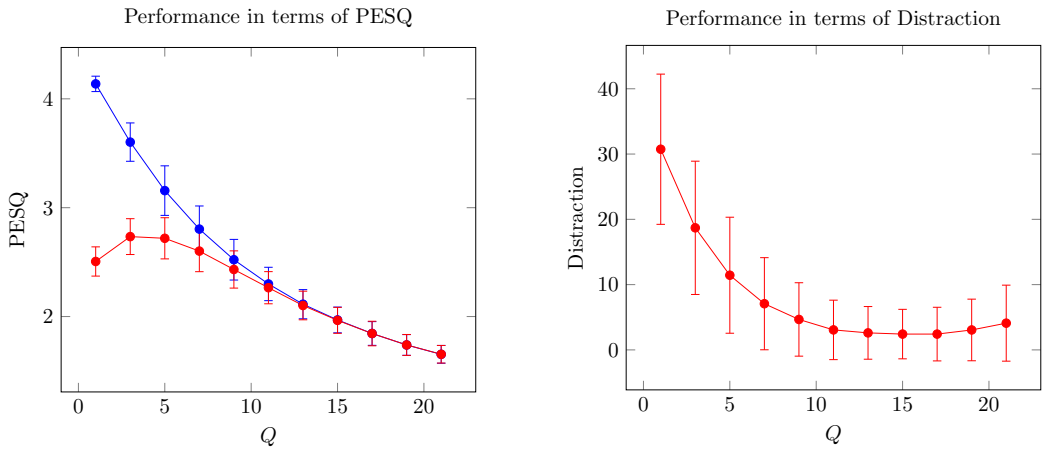


Figure B.1

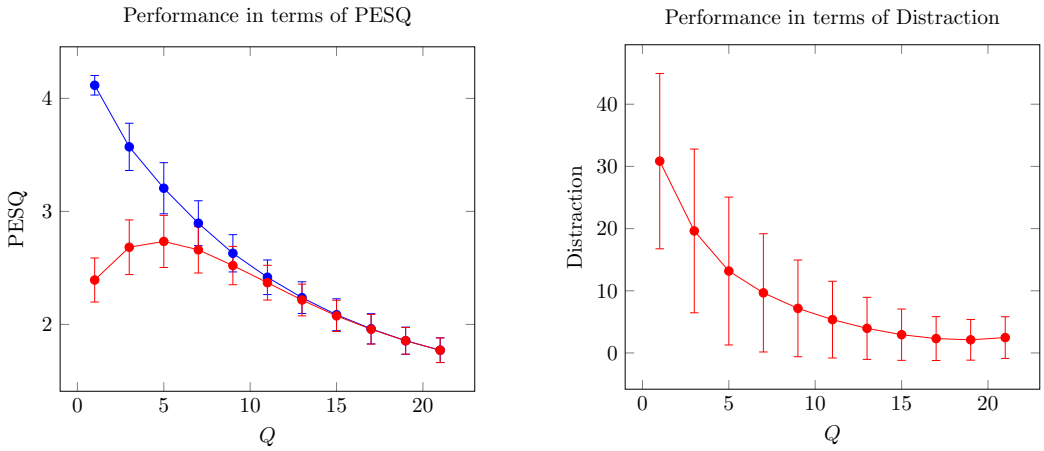


Figure B.2

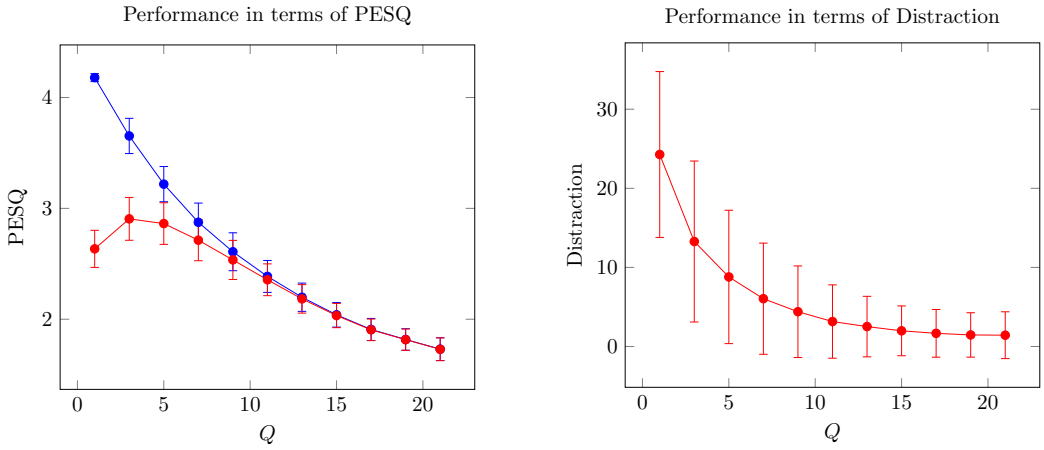


Figure B.3