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

CAS-2021-??

M.Sc. Thesis

Sound Zones with a Cost Function based on Human Hearing

Niels Evert Marinus de Koeijer B.Sc.

Abstract

TODO: Currently, this contains the draft of my thesis. Everything and anything can be changed as far as I'm concerned, I appreciate all feedback!

TU Delft

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

Sound Zones with a Cost Function based on Human Hearing

Subtitle Compulsory?

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

Advisors: ______________________

dr. M. Bo Møller

____

dr. P. Martinez Nuevo

Committee Members: ______________________

dr. M. Mastrangeli

____

dr. J. Martinez Castañeda

Abstract

TODO: Currently, this contains the draft of my thesis. Everything and anything can be changed as far as I'm concerned, I appreciate all feedback!

Acknowledgments

I would like to thank dr. M. Bo Møller, dr. P. Martinez Nuevo, dr.ir. R.C. Hendriks, and dr. J. Martinez Castañeda.

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

Contents

Introduction 1

1.1 Motivation

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 show, while an other 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 different sound content.

One class of algorithms that attempt to do so are known as 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 sound system to reproduce audio content in each zone. Using the principals of constructive and destructive interference, this is done in such a way that minimal content of one zone is audible in the others.

In the previously listed example, one zone would contain the audio of a show and another zone would contain music. The sound zone algorithm would then determine how to best use the sound system to reproduce the two zones. If the sound zone algorithm does a good job, both people can enjoy their audio content at its full potential without bothering one another. An image depicting the situation is given in Figure 1.1.

In practice however, sound zone algorithms do 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. In addition to this, reducing interference between zones can also come at the cost of reducing quality of the reproduced audio in the zones.

Improving sound zone algorithms is still an active topic of research. One recent approach is to include a model of the human auditory system to model how sound is perceived. Typically, sound zone algorithms model physical quantities such as

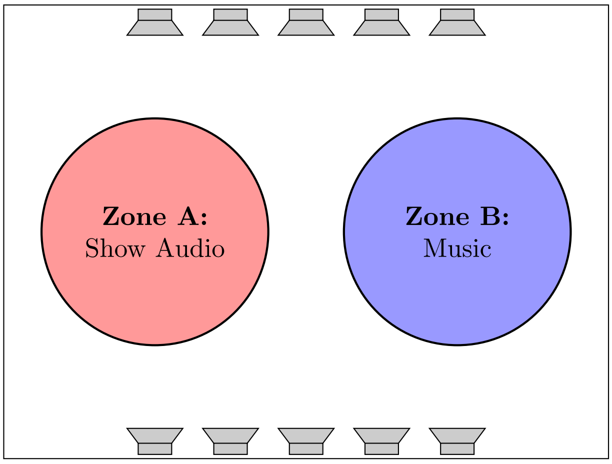

Figure 1.1: A room containing a sound system consisting of an array of loudspeakers and two zones. The goal of the sound zone algorithm is to control the sound system in such a way that the red zone contains the audio of a show, and the blue zone contains the music.

sound pressure, which doesn't always capture what is important for the perception of sound. 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. In this pursuit it was found that the quality of the reproduced audio in the zones [2]. Later, Lee et al. showed that including a perceptually-motivated weighting in the sound zone algorithm outperforms traditional algorithms [3, 4].

This work will seek to further explore this perceptual approach by proposing a novel perceptual sound zone algorithm. A perceptual sound zone algorithm will be constructed from scratch by considering what perceptual models and sound zone algorithms are best suited to integrate into one another.

1.2 Objectives and Organization

This section will state the goals of the thesis and organization of the rest of this document. Als stated in the motivation, the goal of the thesis is to construct a perceptual sound zone algorithm. The central question in this thesis is:

"How can perceptual models of the human auditory system be used to improve sound zone algorithms?"

This question is answered in three steps. First, a perceptual model is selected in chapter 2. Next, a sound zone algorithm is selected in chapter 3. Finally, the two are combined and the resulting algorithm is evaluated in chapter 4. The organization of the three chapters is given below.

1.2.1 Search and Implementation of a Perceptual Model

In order to construct a perceptual sound zone algorithm, a perceptual model is required. This begs the question: "What perceptual models exist?". To answer this question, this chapter starts with a summary of necessary psycho-acoustics background in section 2.2 followed by a literature review into state of the art perceptual models in section 2.3.

To select a perceptual model suitable for combination with a sound zone algorithm, one must consider what desired properties of such a model. These criteria and the resulting selection of one of the perceptual models from the literature review is given in section 2.4.

The selected perceptual model is then discussed in greater detail in section 2.5. Here, implementation details are given and analysis are performed to give the reader an intuition into the model.

1.2.2 Search and Implementation of a Sound Zone Approach

After selecting a perceptual model, a suitable sound zone approach must be selected. Before determining which sound zone approaches are suitable, a literature review is performed in section 3.2 to document what sound zone approaches exist.

The perceptual model will impose certain constraints on which sound zone algorithm is best suited for integration. As such, section 3.3 will reflect on these constraints to select one of the documented sound zone approaches as the most promising for integration.

The sections that follow will discuss implementation details of the selected sound zone approach. In section 3.4 a general sound zone data model will be given, formalizing the sound zone problem and laying the mathematical foundation. This mathematical foundation is then used in section 3.5 and section 3.6 to derive a sound zone algorithm based on the selected sound zone approach.

The derived sound zone algorithm will form the foundation on which the perceptual sound zone algorithm will be built. In addition to this, it will serve as a reference implementation with which the perceptual sound zone algorithm can be compared.

1.2.3 Implementation of a Perceptual Sound Zone Algorithm

Perceptual Model Review and Implementation 2

2.1 Introduction

The goal of this chapter is to find a suitable perceptual algorithm for integration with a sound zone algorithm.

To this end, this chapter is structured as follows. The chapter will begin in section 2.2 with some background information on the perceptual aspects of the human auditory system. This is done to ensure that the reader is well informed on the perceptual aspects discussed in later parts of the thesis.

Next, with the necessary psycho-acoustical background in place, a literature review into state of the art perceptual models is performed in section 2.3. The purpose of this review is to document candidates for perceptual models that are best suited for integration into sound zone algorithms. In addition to this, the reviewed models could also serve as potential candidates for use in the evaluation of the results of the perceptual sound zone algorithm, which will be done in chapter 4.

To aid in the selection of a perceptual model from the candidates found by the literature review, criteria are be defined in section 2.4. These criteria reflect desirable properties for a perceptual model to have for integration in a sound zone algorithm. The criteria are then in the same section to make a selection.

Next, in order to give the reader more a greater understanding, the selected perceptual model is discussed in more detail in section 2.5 by stating implementation details and describing its behavior. The chapter will then wrap up with some conclusions in section 2.6.

2.2 Perceptual Background

This section will provide some additional information on the human auditory system. This serves to give the reader the necessary background information in order to understand the rest of this chapter.

2.2.1 Workings of the Ear

Basilar Membrane, ERBS, ERB

2.2.2 Auditory Masking

Temporal, Spectral

2.2.3 Threshold of Quiet

2.3 Review of Perceptual Models

This section will document a literature review into perceptual models of the human auditory system. As stated in the introduction, this literature review is performed to determine which perceptual models are currently available in the state of the art. This review will then be later used in section 2.4 to select a perceptual model most suitable for integration into a sound zone algorithm.

For this literature review, the goal was to document the perceptual models that were either promising for the integration into algorithms or for evaluating the quality of the output of algorithms. These are models that attach some "score" or "rating" to the perceptual quality of input signals. These ratings can then be used in algorithms to obtain an optimal rating through optimization, or to determine the quality of results from later algorithms.

As such, the focus of the literature review is not on the latest findings in psychoacoustics, or models that accurately emulate the behavior of the human ear, such as the Dau model. Instead, two categories of perceptual models are considered.

First, "Objective Measures", which are discussed in subsection 2.3.1, which attempt to predict the perceptual quality ratings found in listening tests. And "Audio Coding" models, discussed in subsection 2.3.2, which are used to quantify how perceptually audible the artifacts of compression in audio are.

2.3.1 Objective Measures

In order to 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 these tests are performed is for the evaluation of listening aids, where they are used determine the speech intelligibility [5]. Other examples include determining which loudspeaker has the best 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 listening tests. This is very useful for algorithm developers for example, as they can get an indication of how well they are doing without having to perform a cumbersome test. Note however that a objective quality measure does not replace a listening test: it can only be used to give an indication.

The objective measures that will be considered take a reference and degraded audio stimuli as inputs. Most models then convert the input audio stimuli to a so-called internal representation, which models how the human auditory system perceives the stimuli. Various features are then derived from the internal representation of the stimuli. An estimator for teh quality of the stimuli is then created by fitting the features to the results of listening tests.

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 sound zone algorithms.

As such, this section will explore various objective measures. This will be done by considering various classes different objective measures, namely "Objective Speech Quality Measures", "Objective Speech Intelligibility Measures", and "Objective Audio Quality Measures".

2.3.1.1 Objective Speech Quality Measures

There have been a number of attempts to create objective measures to quantify the quality of speech. In this section three objective speech quality measures will be 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 of the perceived loudness. From this internal representation, so-called symmetric and asymmetric disturbances are determined between the time-frequency bins of the reference and degraded speech. A non-linear average over the frequency bins 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 is meant to be the successor of PESQ, with the intention 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 the Neurogram Similarly Index Measure (NSIM) to make its predictions. 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. This similarity is then related to the outcomes of listening tests through a laplacian fit [8].

In general, PESQ, POLQA and ViSQOL require many steps to compute and were found difficult to optimize for due to many non-differentiable such as clipping and conditional branches within the algorithms. 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].

2.3.1.2 Objective Speech Intelligibility

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 will be 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 distorted through a distortion that can be modeled as a time-frequency weighting.

  It computes the internal representation of the reference and degraded speech signals by converting them into 1/3 octave bands, and then segmented into short time frames. The average correlation coefficient between the segments of the internal representation fo the reference and degraded segments is then computed, and averaged over all time segments and frequency bands 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. As such, the intelligibility metric 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. As such, SIIB is computed through the mutual information rate between a clean speech signal and the speech signal received by a listener.

  In order to compute the mutual information rate, the paper models the transmission of an intended message from speaker to listener as a communication channel. Among other aspects, this transmission channel includes a model of the human auditory system. [11]

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.

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. However, if the communication channel is approximated as gaussian, the mutual information can be computed in closed form, and SIIB becomes a differentiable measure.

2.3.1.3 Objective Audio Quality Measures

The previous objective quality metrics are both intended for evaluating speech. In this section, a number of objective quality metrics will be discussed that are designed instead for perceived audio quality. 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.

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. As ViSQOLAudio is similar to ViSQOL with some small adjustments, it is similarly difficult to optimize for.

2.3.1.4 Distraction Model

One especially promising objective measure is the distraction proposed by Francombe et al. in 2015 [?]. 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 [?]. 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 was maximally distracted.

The target-interferer pairs and ratings resulted in a dataset. This dataset was then used to fit a model which predicted the distraction given novel a target-interferer pair. The model consisted of taking a linear combination of 5 features which could be computed through the audio files of the target and the interferer.

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 [?]. 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.

On face value, the distraction model is promising to optimize over. However, while easy to compute, the real-time distraction model by Rämö et al. 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.

2.3.2 Perceptual Models from Audio Coding

The second class of perceptual models that will be considered are the perceptual models used in audio coding. Audio coding algorithms attempt to find an low-bitrate representation of an audio input signal, as a form of compression. This process is usually lossy, as reducing the bitrate introduces errors. These errors can be a detriment to the listening experience.

As such, most audio coding algorithms use a perceptual model to quantify how disturbing the 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 [14]. The perceptual 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 is used to 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 could be to define sound zone algorithms that minimize a distortion function for example.

2.3.2.1 ISO MPEG Models

The ISO/IEC 11172-3 standard specifies a coded representation for audio files [15], and a decoder for said representation. 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 [16].

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 different frequency bands, modeling 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 [17], 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 [14], as such examples exist from which optimization schemes could be inspired.

2.3.2.2 Par Detectability

In 2005, van der Par et al. proposed a novel perceptual model designed for use in audio coding [17]. The model defines a distortion measure which determines the "detectability" of a distortion signal in presence of a 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.

The proposed method 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 [14].

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 an 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 [18, 19].

2.3.2.3 Taal Detectability

A paper from 2012 by Taal et al. proposed a novel perceptual model [14] which also introduce the detectability of a distortion signal in presence of a masking signal. In a way, paper proposes an 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 results 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 an L2-norm, which makes it a good candidate for optimization. The computational demand was however shown to be higher than the Par Detectability [14], especially for larger number of input samples.

2.4 Selection of Perceptual Model

In section 2.3 a literature review of various models of the human auditory system was provided. This section will determine criteria with which one of the discussed perceptual models can be selected for use in integration into a sound zone algorithm. The other perceptual models will still be of use however during the evaluation of the results of the perceptual sound zone algorithm that will be derived in chapter 4.

The structure of this section is as follows. First, subsection 2.4.1 will discuss the criteria that will inform the decision. Next, in subsection 2.4.2 said criteria will be used to select a suitable perceptual models and reflect on this choice.

2.4.1 Criteria for Selecting Perceptual Model

This section will define desirable criteria for the perceptual model for integration with a sound zone algorithm. Two criteria can be distinguished.

1. **Mathematical Tractability:**
   It is a desirable property for the perceptual model to be easy to include in optimization problems mathematically. As will be shown in chapter 3, many sound zone algorithms are posed as (convex) optimization problems.

   Therefore if a perceptual model is too complicated to be integrated mathematically into optimization problems are undesirable. For example, if the computation of a perceptual model involves conditional branching (i.e. if-statements), their integration into an existing optimization problem may be difficult.

   Furthermore, as most sound zone algorithms can be posed as a convex optimization problem, it is also desirable that the perceptual models preserve this convexity. Convexity is a desirable property for optimization, as it guarantees that the optimization problem has a single, global optimal value, rather than many locally optimal values [20].

2. **Computational Overhead:**
   It is desirable for the inclusion of the perceptual model to add minimal computational load to the sound zone algorithm. As such, the additional overhead of the perceptual model should not increased the run time of the sound zone algorithm by many orders of magnitude.

2.4.2 Selection of Perceptual Model

This section will use the criteria defined by subsection 2.4.1 in order to select a perceptual model. In the literature review given in section 2.3, two classes of perceptual model were considered: perceptual models from audio coding and objective audio measures.

All objective audio measures were found to be mathematically untractable, as all models are both non-differentiable and non-convex functions of their input signals. As such, they are difficult to integrate into existing sound zone algorithms. However, as the objective audio measures predict the outcomes of listening tests, they are uniquely suited for evaluation of perceptual sound zone algorithm that is to be proposed in chapter 4.

All models from audio coding are mathematically tractable and of low computational overhead. The ISO MPEG models were found to be less promising than the Par and Taal detectability, as it does not immediately define a cost function which can be optimized over: the models only determine how much noise can be added per band.

As such, the decision is between the Par and Taal detectability. The Taal detectability takes into account temporal properties of the input signal in it's perceptual model. 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.

As such, the Par detectability will be selected for use in the sound zone algorithm. Note however that surface through inspection Taal and Par detectability seem sufficiently similar that it is likely possible to use the Taal detectability in place of the Par detectability in the algorithms proposed in chapter 2. It is also interesting to note that the Taal detectability is computed in the time domain, whereas the Par detectability is computed in the frequency domain. Exploring the possibilities of the Taal detectability will however be left to future work and not further explored in this work.

2.5 Implementation of Selected Perceptual Model

In section 2.4, it was determined that the Par detectability was the most suited model for integration with sound zone algorithm of all perceptual models considered in the literature review given in section 2.3. In this section, the perceptual model is considered in greater detail in order to give the reader a greater understanding of the implementation and behavior of the model.

This section is organized as follows. In order to optimize over the model, a suitable expression for it must be found.

2.5.1 Introduction to Par Detectability

In this section, a high-level description of the Par detectability will be given. This is done to give the reader an understanding of the model before going into greater detail. The Par detectability defines a function $D(x[n], \varepsilon[n])$. Here, $x[n] \in \mathbb{R}^{N_x}$ is the the masking signal, and $\varepsilon[n] \in \mathbb{R}^{N_x}$ is the disturbance signal.

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

It is assumed that a human is listening to both the masking signal and the disturbance signal at the same time. The detectability can then be understood as the probability that a human listener can detect the disturbance signal $\varepsilon[n]$ in presence of the masking signal $x[n]$ [17]. The signal $x[n] \in \mathbb{R}^N$ is referred to as the masking signal because it masks the disturbance signal $\varepsilon[n]$ to a degree.

The metric 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 $x[n]$. 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 $\varepsilon[n]$.

2.5.2 Computation Details of the Par Detectability

This section will explore calculating the Par detectability. The first thing to note about the Par detectability is that it operates in the frequency domain. 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.

The Par detectability starts by computing an internal representation of the input signals $X[k]$ and $\mathcal{E}[k]$. This internal representation models how the input signals appear in the human auditory system.

The Par detectability models this filtering in two steps. The first step models how parts of the ear filter the incoming sound with an outer- and middle-ear filter $H_{\text{om}}[k]$. Next, a 4th order Gammatone filter bank is applied, modeling the filtering of the basilar membrane inside the ear [17]. Note that the exact same filtering is applied to both $X[k]$ and $\mathcal{E}[k]$.

The Gammatone filter bank consists of N_g filters, which will denoted by $\Gamma_i[k]$, for $1 \leq i \leq N_g$. To model the frequency-place transform that occurs in the the basilar membrane, 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). A possible expression for the gammatone filters $\Gamma_i[k]$ are given by the original paper [17].

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 of the for $X[k]$ and $\mathcal{E}[k]$ respectively. 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 for $1 \leq i \leq N_g$. This can be thought of the detectability per filter tap. The within-channel detectability 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.

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:

$$\begin{aligned} D(x[n], \varepsilon[n]) &= C_s L_{\text{eff}} \sum_{i=0}^{N_g} D_i \qquad (2.4) \\ &= C_s L_{\text{eff}} \sum_{i=0}^{N_g} \frac{\sum_{k=0}^{N_x-1} |H_{\text{om}}[k]|^2 \, |\Gamma_i[k]|^2 \, |\mathcal{E}[k]|^2}{\sum_{k=0}^{N_x-1} |H_{\text{om}}[k]|^2 \, |\Gamma_i[k]|^2 \, |X[k]|^2 + C_a} \qquad (2.5) \end{aligned}$$

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.

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 inaudible due to the threshold of hearing determined by 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 $S_i \gg C_a$ for all i, the detectability becomes large, as 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. What follows is the discussion of the calibration of the Par model in subsection 2.5.3, where calibration constants C_a and C_s are determined. Afterwards, the model above will be given as a least-squares formulation in order to ease the integration into optimization in subsection 2.5.4.

2.5.3 Calibration of the the Par Detectability

This section will describe the calibration of the Par detectability. A correct calibration of the Par detectability must satisfy the following:

1. The just noticeable disturbance signal must result in a detectability $D(x[n], \varepsilon[n])$ of 1.
2. The detectability must take the threshold of hearing into account correctly.

In order meet these requirements, calibration constants C_a and C_s are chosen in a certain way. Before a discussion on calibration can occur, it is important to first discuss the relationship between the input signals $x[n]$ and $\varepsilon[n]$ and reproduced sound pressure level.

2.5.3.1 Relating Digital Representation and Sound Pressure Level

TODO: Rewrite. 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 [21]:

$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} + 0.001\left(\frac{f}{1000}\right)^{4} - 6.5\exp\left[-0.6\left(\frac{f}{1000} - 3.3\right)^{2}\right] \quad (2.6)$$

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 -32786 and 32787. 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:

$$X_{\mathrm{dB}}(f) = 10\log_{10}(|X(f)|^2) + O_{\mathrm{dB}} \quad (2.7)$$

Here, $X_{\text{dB}}(f)$ 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.

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 32787 corresponds to e.g. a sound pressure level of 100 dB SPL.

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:

$$\mathcal{F}\left\{A\cos\left(2\pi f_0 t\right)\right\} = A\delta\left(f - f_0\right) \tag{2.8}$$

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 the previously defined relationship results in the following definition for O_{dB}.

$$O_{\text{dB}} = 10\log_{10}\left(|A|^2\right) - B_{\text{FS}} \tag{2.9}$$

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

2.5.3.2 Determing Calibration Constants

TODO: Rewrite. There are various ways of calibrating this model, but this section will discuss the method of calibrating that is given in the original paper [17]. 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 [17]. To model this, consider the following masking and disturbance signals.

$$x_{\text{JND}}[n] = A_{70}\cos\left(2\pi f_0 n/f_s\right) \tag{2.10}$$

$$\varepsilon_{\text{JND}}[n] = A_{52}\cos\left(2\pi f_0 n/f_s\right) \tag{2.11}$$

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 relationship defined in the previous section can be used to convert between the two. Thus, $\varepsilon_{\text{JND}}[n]$ must be just noticeable in presence of $x_{\text{JND}}[n]$, which corresponds to the following equation:

$$D(x_{\text{JND}}[n], \varepsilon_{\text{JND}}[n]) = 1 \tag{2.12}$$

The second requirement is that the threshold of hearing must be included correctly. The threshold of hearing is the verge between audible and inaudible sound. To this end, consider the following masking and disturbance signals:

$$x_{\text{THR}}[n] = 0 \tag{2.13}$$

$$\varepsilon_{\text{THR}}[n] = A_{\text{tq}} \cos\left(2\pi f_0 n / f_s\right) \tag{2.14}$$

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 its dB SPL representation is equal to 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. This corresponds to the second equation:

$$D(0, \varepsilon_{\text{THR}}[n]) = 1 \tag{2.15}$$

The system of equations defined by the previously derived equations can be solved through the bisection method. To see how this is done, the reader is referred to the original paper [17].

2.5.4 Least-Squares Formulation of the Par Detectability

TODO: Rewrite. This section will rewrite the previously introduced detectability into a least-squares representation. This representation will allow for easier integration into existing sound zone algorithms.

Introduce the following matrices and vectors:

$$\begin{array}{lclccccc} \mathbf{H}_{\text{om}} & = & \text{diag}([& H_{\text{om}}[0], & H_{\text{om}}[1], & \ldots, & H_{\text{om}}[N_x - 1] &]^T) \\ \mathbf{\Gamma}_i & = & \text{diag}([& \Gamma_i[0], & \Gamma_i[1], & \ldots, & \Gamma_i[N_x - 1] &]^T) \\ \mathbf{x} & = & [& X[0], & X[1], & \ldots, & X[N_x - 1] &]^T \\ \boldsymbol{\varepsilon} & = & [& \mathcal{E}[0], & \mathcal{E}[1], & \ldots, & \mathcal{E}[N_x - 1] &]^T \end{array} \tag{2.16}$$

The detectability can now be expressed as follows:

$$D(x[n], \varepsilon[n]) = C_s L_{\text{eff}} \sum_{i=0}^{N_g} \frac{\boldsymbol{\varepsilon}^{\text{H}} \mathbf{\Gamma}_i^{\text{H}} \mathbf{H}_{\text{om}}^{\text{H}} \mathbf{H}_{\text{om}} \mathbf{\Gamma}_i \boldsymbol{\varepsilon}}{\mathbf{x}^{\text{H}} \mathbf{\Gamma}_i^{\text{H}} \mathbf{H}_{\text{om}}^{\text{H}} \mathbf{H}_{\text{om}} \mathbf{\Gamma}_i \mathbf{x} + C_a} \tag{2.17}$$

$$= \boldsymbol{\varepsilon}^{\text{H}} \left(C_s L_{\text{eff}} \sum_{i=0}^{N_g} \frac{\mathbf{\Gamma}_i^{\text{H}} \mathbf{H}_{\text{om}}^{\text{H}} \mathbf{H}_{\text{om}} \mathbf{\Gamma}_i}{\mathbf{x}^{\text{H}} \mathbf{\Gamma}_i^{\text{H}} \mathbf{H}_{\text{om}}^{\text{H}} \mathbf{H}_{\text{om}} \mathbf{\Gamma}_i \mathbf{x} + C_a} \right) \boldsymbol{\varepsilon} \tag{2.18}$$

Now define diagonal matrix $\mathbf{W}$ as follows:

$$\mathbf{W} = \sqrt{C_s L_{\text{eff}} \sum_{i=0}^{N_g} \frac{\mathbf{\Gamma}_i^{\text{H}} \mathbf{H}_{\text{om}}^{\text{H}} \mathbf{H}_{\text{om}} \mathbf{\Gamma}_i}{\mathbf{x}^{\text{H}} \mathbf{\Gamma}_i^{\text{H}} \mathbf{H}_{\text{om}}^{\text{H}} \mathbf{H}_{\text{om}} \mathbf{\Gamma}_i \mathbf{x} + C_a}} \tag{2.19}$$

As such,

$$D(x[n], \varepsilon[n]) = \boldsymbol{\varepsilon}^{\text{H}} \mathbf{W}^{\text{H}} \mathbf{W} \boldsymbol{\varepsilon} \tag{2.20}$$

$$= ||\mathbf{W} \boldsymbol{\varepsilon}||_2^2 \tag{2.21}$$

As can be seen, the detectability can be expressed as a weighted squared L2-norm of the disturbance signal. The weighting is entirely determined by the masking signal $x[n]$.

2.6 Conclusion

The goal of this chapter was to determine a suitable perceptual model for integration into a sound zone algorithm. This was done by first reviewing various objective quality measures and perceptual models from audio coding.

To select a suitable model, the mathematical tractability and the additional computational overhead of each model under review was considered. From this, it was found that the Par detectability was found to be the most promising perceptual model included in the review. Finally, the chapter concluded with an analysis and implementation of the Par detectability.

Sound Zone Approach Review and Implementation 3

3.1 Introduction

The goal of this chapter is find a suitable sound zone approach for integration with the perceptual model selected in chapter 2.

The chapter will start with section 3.2 which will provide a review of various sound zone approaches from literature to provide a selection to choose from.

After documenting the state of the art, one sound zone algorithm will be selected in section 3.3 as the most promising for combination with the perceptual model selected in chapter 2. This is done by reflecting on the mathematical properties of the selected perceptual model.

After selecting a sound zone approach, the rest of the chapter will focus on the derivation and the implementation of a sound zone algorithm that takes said approach. This algorithm will serve two purposes. Firstly, it will serve as the basis upon which the perceptual sound zone algorithm will be built. Secondly, it is well suited as a reference with which the perceptual sound zone algorithm can be compared.

To derive said sound zone algorithm, section 3.4 will start by introducing the data model. This is followed by the implementation of the selected algorithm in section 3.5. The implementation is then extended in section 3.6 to operate in a short-time block-based fashion, giving the algorithm the potential to run in real-time. The chapter ends with a summary and concluding remarks in section 3.8.

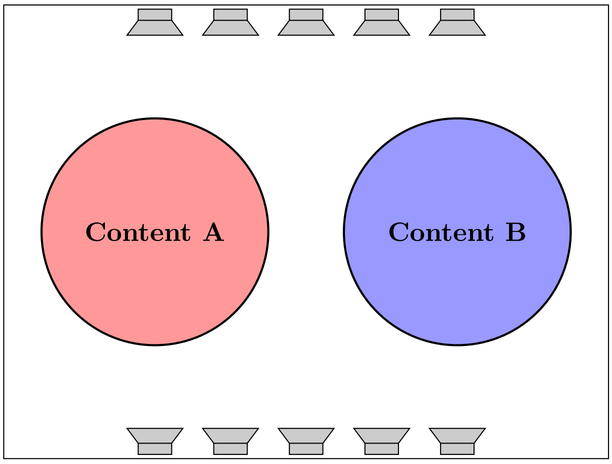

Figure 3.1: A room is divided into two zones: a red zone and a blue zone. Each zone is assigned different content: content A and content B. The loudspeakers array that is present in the room is to be controlled by the sound zone algorithm in such a way that the desired content is reproduced. As mentioned, 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.

3.2 Review of Sound Zone Approaches

This section will document a literature review done into various sound zone approaches from the state of the art. The goal here is obtain an overview of possible approaches to solving the sound zone problem to find an approach that is best suited for integration with the perceptual model selected in chapter 2.

3.2.1 The Sound Zone Problem

An initial description of the sound zone problem was given in the introduction. This section seeks to build on this description in order to provide the necessary background on sound zones to understand the rest of this work. Readers with prior familiarity with sound zones may wish to skip this chapter.

In sound zones, the goal is to control the spatial distribution of sound inside an enclosure. This 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. The reproduction is to be done in such a way that there is minimal interference between target sound pressures. To understand this, consider the example given by Figure 3.1.

An important concept for understanding sound zone literature is the concept of bright zones and dark zones. Sound zone problems are typically decomposed into multiple subproblems consisting of a bright zone and one or more dark zone(s). The goal is to reproduce a specified target sound pressure in the bright zone while restricting the sound pressure in the dark zones(s). In doing so, the target sound pressure is essentially reproduced locally in the bright zone. When the subproblems are recombined, the sound zone problem is solved as all target sound pressure is

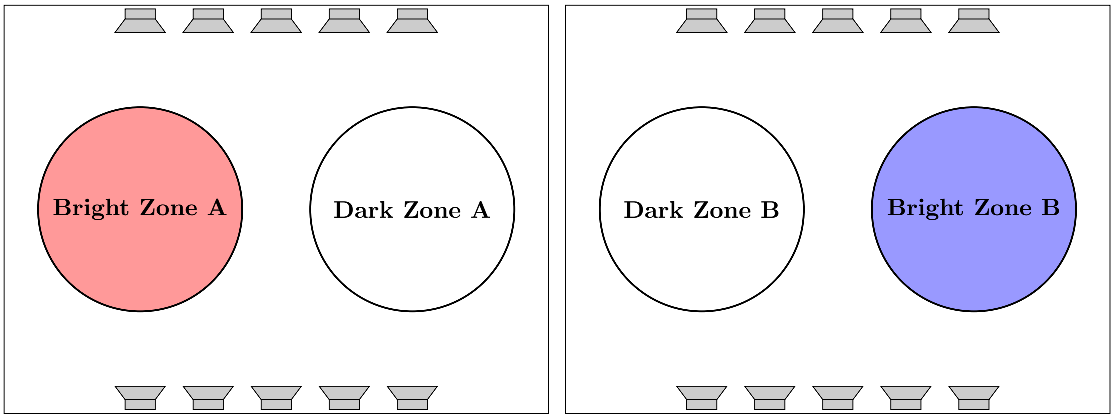

Figure 3.2: Decomposition of the example given in Figure 3.1 into two bright-dark zone pairs. For the first problem, the goal is to reproduce content A in bright zone A while minimizing the amount of sound pressure in dark zone A. Similarly for the second problem: reproduce content B in bright zone B 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.

reproduced locally in their respective zones. To ease the understanding of this concept, an example of this decomposition is given in Figure 3.2.

3.2.2 Sound Zone Approaches

With the sound zone problem sufficiently explained, this section will now cover various sound zone approaches.

3.2.2.1 Delay and Sum Beamforming

One traditional approach to creating sound zones is delay and sum (DS) beamforming [22]. In beamforming an array of loudspeakers is used to focus a “beam” of audio. This is done using the principals of constructive and destructive interference by playing with slight delays from each loudspeaker in the array (hence, the name “delay and sum”). The beam is constructed through constructive interference, whereas outside of the beam the audio content is partially canceled through destructive interference.

For sound zone applications, the angle of the beam is chosen such that the target audio is directed towards its respective bright zone. Beamforming is somewhat limited for sound zone applications as it limits the spatial control to an angled beam. The other sound zone algorithms that will be discussed in this section can be considered a generalization of beamforming, as their framework supports more spatial control.

3.2.2.2 Pressure Matching Approach

In pressure matching approaches (PM), one attempts to control the output of the loudspeaker array in such a way that resulting sound pressure in the zone matches the specified target sound pressure for that zone. While doing so, the PM approach attempts to minimize the sound pressure that results in other zones as to minimize the interference or crosstalk between zones [22, 1].

PM is usually implemented as an optimization problem which minimizes the least-squares error between the target and resulting sound pressures in the bright zone. The optimization problem often includes constraints which limit the resulting sound pressure in the dark zone(s). The minimization of the least-squares error has been done in both the frequency and the time domain.

3.2.2.3 Acoustic Contrast Control Approach

The acoustic contrast control (ACC) approach to sound zones attempts to maximize the acoustic contrast between the bright zone and the dark zone for each specified target sound pressure. The acoustic contrast is defined as the ratio of the acoustic potential energy of the bright zone and the dark zone. Just as with PM, ACC is defined as an optimization problem.

Typically, ACC is performed in the frequency domain independently per frequency band [22, 1]. However, Elliot et al. explored a broadband approach to ACC in 2011 which was further refined by Cai et al in 2014 called broadband acoustic contrast control (BACC) [23, 24]. This was done by solving in the time domain rather than the frequency domain.

ACC approaches typically have good acoustic contrast, but reproduce the target sound pressure less faithfully than PM approaches [4]. As such, hybrid approaches such as ACC-PM [25] and BACC-PM [26] have been proposed recently which allow for what is essentially a compromise between ACC and PM.

3.3 Selection of Sound Zone Approach

This section will focus on the selection of a sound zone approach suitable for integration with the perceptual model selected in chapter 2.

In chapter 2 the Par detectability was selected as the most promising perceptual model. In order to select a suitable sound zone approach from the approaches considered in the review given in section 3.2 the mathematical properties of this model will be considered.

3.3.1 Mathematical Properties of Detectability

Recall from section 2.5 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 $x[n] \in \mathbb{R}^{N_x}$. In section 2.5 it was also noted that the detectability is computed in the frequency domain. To this end, $X[k]$ and $\mathcal{E}[k]$ denote the frequency domain representations of $x[n]$ and $\varepsilon[n]$ respectively.

In subsection 2.5.4 it was found that the detectability could be expressed as in least-squares fashion as follows:

$$D(x[n], \varepsilon[n]) = ||\mathbf{W}\boldsymbol{\varepsilon}||_2^2 \tag{3.1}$$

Here, the matrix $\mathbf{W} \in \mathbb{R}^{N_x \times N_x}$ is a diagonal matrix that models the masking effects of $x[n]$. The vector $\boldsymbol{\varepsilon}$ contains the frequency domain representation $\mathcal{E}[k]$. As such, $\mathbf{W}$ weighs each frequency component contained $\boldsymbol{\varepsilon}$.

Note the following about the detectability:

1. It is computed in the frequency domain.
2. It operates on a short time scale (20 to 200 ms).
3. It is a convex function of the disturbance $\boldsymbol{\varepsilon}$.

3.3.2 Selecting an Appropriate Sound Zone Approach

TODO: Write this In section 3.2 three different sound zone approaches were discussed: delay and sum beamforming (DS), pressure matching (PM), and acoustic contrast control (ACC). It was found that the delay and sum beamformer framework was too limiting to include any perceptual information, thus PM and ACC remain.

As stated in section 3.2, the PM approach minimizes the squared error between the target sound pressure and the sound pressure reproduced by the loudspeaker array.

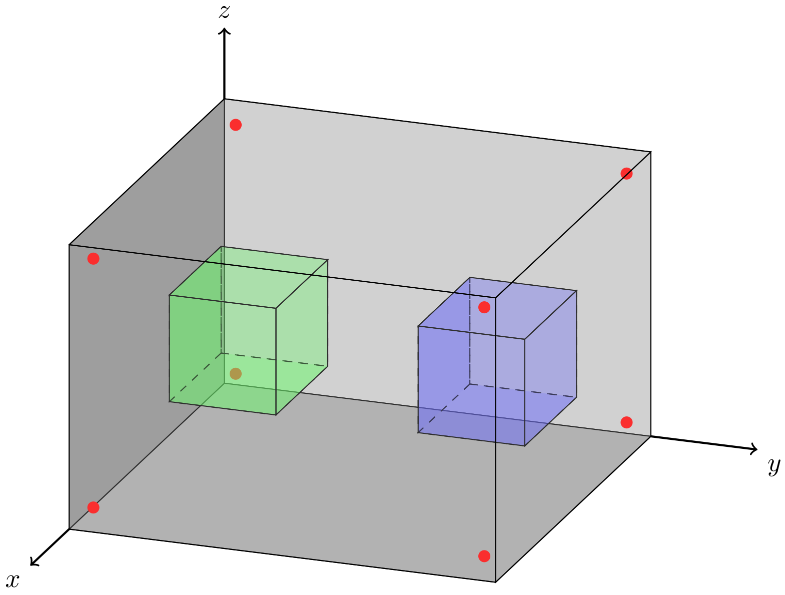

Figure 3.3: The room $\mathcal{R} \subset \mathbb{R}^3$ containing the zones $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{R}$ depicted in green and blue respectively. The room contains $N_L = 8$ loudspeakers, which are denoted by the red dots in the corners of the room.

3.4 Data Model

In this section a mathematical framework for a room containing sound zones will be introduced. This framework will be used in the derivation of the sound zone algorithms.

The contents of this section are as follows.

First, **??** introduces a spatial description of a room containing two zones and a loudspeaker array. Then, subsection 3.4.2 defines the objective of the sound zone algorithm 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 **??**, completing the mathematical framework. This is then used in **??** to select a suitable target sound pressure which will be used in the remainder of this thesis.

3.4.1 Room Topology

In this section, a description of the room in which sound zones are to be reproduced will be given. In general, the room can contain any number of zones, but this thesis will focus on the two zone case.

The room 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 R, i.e. $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{R}$ where $\mathcal{A} \cap \mathcal{B} = \emptyset$. In addition to the zones, the room $\mathcal{R}$ also contains N_L loudspeakers, which can be modeled as discrete points. The room, loudspeakers and zones are visualized in Figure 3.3.

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 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.

3.4.2 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 R.

Currently, the zones are given as continuous regions in space. Sound zone approaches will attempt to recreate a specified pressure in the entire region of space defined by $\mathcal{A}$ and $\mathcal{B}$. Other 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.

In this work, a pressure matching approach is used, and thus the latter approach will be taken. 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. The relationship between the loudspeaker input signals and the sound pressure is the topic of the next section.

3.4.3 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 for the l^{th} loudspeaker. 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:

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

The realized sound pressure $p^{(l,m)}[n]$ only considers the contribution of loudspeaker l at reproduction point m. Let $p^{(l)}[n] \in \mathbb{R}^{N_x+N_h-1}$ denote the total sound pressure due

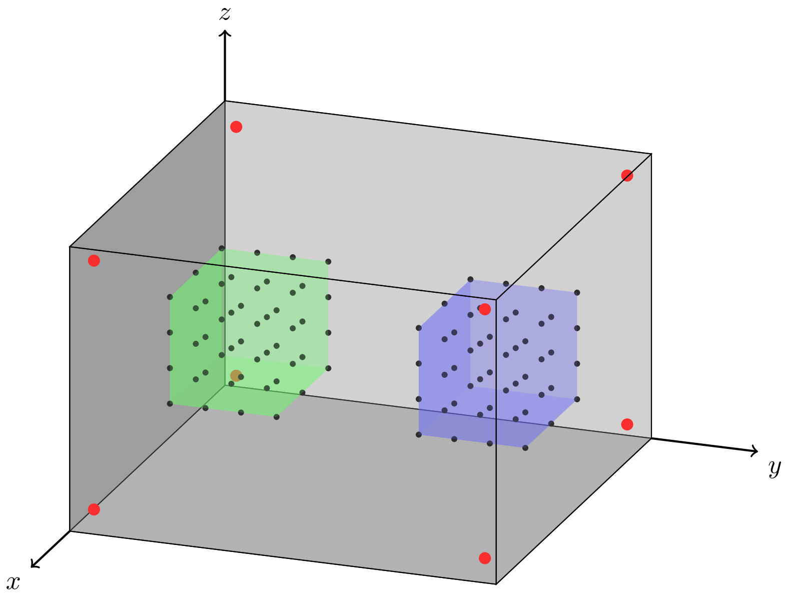

Figure 3.4: The previously introduced room $\mathcal{R}$ with zones $\mathcal{A}$ and $\mathcal{B}$ discretized.

to all N_L loudspeakers. It can now be expressed as the sum over all contributions as follows:

$$p^{(m)}[n] = \sum_{l=0}^{N_L} p^{(l,m)}[n] \tag{3.3}$$

$$= \sum_{l=0}^{N_L} \left(h^{(l,m)} * x^{(l)}\right)[n] \tag{3.4}$$

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

3.4.4 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 user of the sound zone system has 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:

$$t^{(m)}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * s_{\mathcal{A}}\right)[n] \qquad \forall\, m \in A \tag{3.5}$$

$$t^{(m)}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * s_{\mathcal{B}}\right)[n] \qquad \forall\, m \in B \tag{3.6}$$

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 with the given loudspeakers and room.

3.5 Implementation of Selected Sound Zone Algorithm

The “Pressure Matching” (PM) is widely used in literature to solve the sound zone problem. In this section, a “Multi-Zone Pressure Matching” (MZ-PM) algorithm will be derived. The motivation for introducing this algorithm is that it will be used as the foundation on which the perceptual sound zone algorithm will be built, as it was found that perceptual model was easily intergratable into the pressure matching framework.

In the typical PM approach, the resulting loudspeaker input signals $x^{(l)}[n]$ are determined for just a single zone. Here, the loudspeaker input signals are found such that the a target audio is achieved in one zone, while leakage is minimized to other zones. If the solution for multiple zones is desired, than multiple PM problems must be solved and their resulting loudspeaker input signals combined. In the MZ-PM approach, the loudspeaker input signals are instead determined for jointly for all zones.

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

$$x^{(l)}[n] = x^{(l)}_{\mathcal{A}}[n] + x^{(l)}_{\mathcal{B}}[n] \tag{3.7}$$

Here, $x^{(l)}_{\mathcal{A}}[n]$ and $x^{(l)}_{\mathcal{B}}[n]$ 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:

$$p^{(m)}_{\mathcal{Z}}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * x^{(l)}_{\mathcal{Z}} \right) [n] \tag{3.8}$$

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

$$p^{(m)}[n] = p^{(m)}_{\mathcal{A}}[n] + p^{(m)}_{\mathcal{B}}[n] \tag{3.9}$$

The idea in this approach is to chose $x^{(l)}_{\mathcal{A}}[n]$ and such that the resulting pressure $p^{(m)}_{\mathcal{A}}[n]$ attains the target sound pressure $t^{(m)}[n]$ in all $m \in A$.

At the same time however, $p^{(m)}_{\mathcal{A}}[n]$ should not result in any sound pressure in all $m \in B$. Any sound pressure resulting from $x^{(l)}_{\mathcal{A}}[n]$ in zone $\mathcal{B}$ is essentially leakage, or cross-talk between zones. Similar arguments can be given for $x^{(l)}_{\mathcal{B}}[n]$.

In the MZ-PM approach, the loudspeaker input signals $x^{(l)}_{\mathcal{A}}[n]$ and $x^{(l)}_{\mathcal{B}}[n]$ that attain the target with minimal leakage can be found by minimizing the difference between the intended pressure and the realized pressure as follows:

$$\underset{x_{\mathcal{A}}^{(l)}[n],\, x_{\mathcal{B}}^{(l)}[n]\,\forall\, l}{\arg\min} \quad \sum_{m \in A} \left|\left| p_{\mathcal{A}}^{(m)}[n] - t^{(m)}[n] \right|\right|_2^2 + \sum_{m \in A} \left|\left| p_{\mathcal{B}}^{(m)}[n] \right|\right|_2^2 + \tag{3.10}$$

$$\sum_{m \in B} \left|\left| p_{\mathcal{B}}^{(m)}[n] - t^{(m)}[n] \right|\right|_2^2 + \sum_{m \in B} \left|\left| p_{\mathcal{A}}^{(m)}[n] \right|\right|_2^2 \tag{3.11}$$

Here, the first two terms can be understood as the reproduction error and the leakage for zone $\mathcal{A}$. Similarly, the last two terms are the reproduction error and leakage for zone $\mathcal{B}$. To make this more clear, the following definitions are introduced:

$$\mathrm{RE}_{\mathcal{Z}} = \sum_{m \in A} \left|\left| p_{\mathcal{A}}^{(m)}[n] - t^{(m)}[n] \right|\right|_2^2 \tag{3.12}$$

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

Here, $\mathrm{RE}_{\mathcal{Z}}$ is the reproduction error and $\mathrm{LE}_{\mathcal{Z}}$ is the leakage error in zone $\mathcal{Z} \in (\mathcal{A}, \mathcal{B})$. This allows for the following rewrite of the previously introduced optimization problem:

$$\underset{x_{\mathcal{A}}^{(l)}[n],\, x_{\mathcal{B}}^{(l)}[n]\,\forall\, l}{\arg\min} \quad \mathrm{RE}_{\mathcal{A}} + \mathrm{LE}_{\mathcal{A}} + \mathrm{RE}_{\mathcal{B}} + \mathrm{LE}_{\mathcal{B}} \tag{3.14}$$

From this it becomes clear that this approach results in trade-off between minimizing the reproduction errors $\mathrm{RE}_{\mathcal{Z}}$ and leakages $\mathrm{LE}_{\mathcal{Z}}$. 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 mean square pressure error is difficult to interpret.

The algorithm above will form the basis of the perceptual algorithms to be introduced in later chapters.

3.6 Block Based Approach

In the preceding section it is assumed that the desired playback signals $s_{\mathcal{A}}[n]$ and $s_{\mathcal{B}}[n]$ were known in their entirety. In practice however, this is not a valid assumption as a user can change the desired playback content in real-time. This is the case for example when a user changes the song they are playing on their system.

In reality, the sound zone system can only have knowledge of the most recent samples and all previous samples. In order to deal with this limitation, one option is to buffer a large number of incoming samples and apply the existing MZ-PM approach. However, this would introduce significant latency to the system.

Instead, a block-based approach can be used where the incoming samples of the desired playback signals are used in real-time as they become available. The system buffers a block of H incoming samples samples, and then solves the sound zone problem finding the newest block of loudspeaker input signals.

In addition to this, block-based approach is also practical for the integration of the perceptual model. The perceptual model is designed to operate on short time segments in the order of 20 to 200 milliseconds. Block-based approaches would allow the algorithm to operate on segments of this time scale.

For these reasons, this section will adapt existing Multi-Zone Pressure Matching approach introduced in **??** to accommodate for block based processing. This will be done by first modeling the knowledge limitation of the block-based processing on the desired playback signals.

3.6.1 Modeling Block Based Knowledge Limitations

For the block-processing based sound zone approach, the incoming samples of the desired playback signals $s_{\mathcal{Z}}[m]$ for $\mathcal{Z} \in \{\mathcal{A}, \mathcal{B}\}$ are assumed to be buffered into blocks of size H. This means that the system waits until it has a block of size H before revealing the new samples to the sound zone system.

This introduces a knowledge limitation in the desired playback signals, which requires the sound zone algorithm to be redesigned. In this section, this knowledge limitation will be modeled.

The sound zone system only has knowledge of all previous blocks and the most recent block, indexed by block index μ. The relation between the global time index n and block index μ is given as follows:

$$\mu(n) = \lfloor n/H \rfloor \tag{3.15}$$

Thus at a time n, up to and including the μ^{th} blocks of desired playback signals $s_{\mathcal{Z}}[n]$ are known.

Let the available knowledge of the desired playback signal with knowledge of blocks up to μ be denoted as $\tilde{s}^R_{\mathcal{Z}}[n,\mu]$. It can be defined as follows:

$$\tilde{s}^R_{\mathcal{Z}}[n,\mu] = \sum_{k=-\infty}^{\mu} \tilde{s}^R_{\mathcal{Z},k}[n] w^R_H[n-kH] \tag{3.16}$$

Here, $w^R_H[n] \in \mathbb{R}^H$ is a non-causal rectangular window with support $-H+1 \leq n \leq 0$. Furthermore, $\tilde{s}^R_{\mathcal{Z},k}[n]$ is the k^{th} block of the playback signal $s_{\mathcal{Z}}[n]$. As such, $\tilde{s}^R_{\mathcal{Z},k}[n] = s_{\mathcal{Z}}[n]$ for $-H+1+kH \leq n \leq kH$ (and zero elsewhere).

One way of interpreting the equation above is as a projection of $s_{\mathcal{Z}}[n]$ on a basis spanned by shifted non-overlapping rectangular windows $w^R_H[n-kH]$. Here, $\tilde{s}^R_{\mathcal{Z},k}[n]$ can be thought of as the coefficients for the basis functions resulting from the projection.

Expressing it this way reveals that a more general approach is possible: as shown, desired playback signals are projected onto non-causal non-overlapping rectangular windows of size H. This model can however be generalized to a larger class of windows.

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

$$\sum_{k=-\infty}^{\infty} w[n-kH] = 1 \quad \forall\, n \tag{3.17}$$

Projecting $s_{\mathcal{Z}}[n]$ onto a basis spanned by the shifted windows $w[n-kH]$ results in the following:

$$s_{\mathcal{Z}}[n] = \sum_{k=-\infty}^{\infty} s_{\mathcal{Z}}[n] w[n-kH] \tag{3.18}$$

$$= \sum_{k=-\infty}^{\infty} \tilde{s}_{\mathcal{Z}}[n] w[n-kH] \tag{3.19}$$

Here, $\tilde{s}_{\mathcal{Z},k}[n] = s_{\mathcal{Z}}[n]$ for $-N_w+1+kH \leq n \leq kH$ and zero elsewhere. The windows decimate the signal $s_{\mathcal{Z}}[n]$ into segments of size N_w. Due to the COLA condition, this the segments reconstruct the

This allows us to express the desired playback signal with knowledge up to block μ as follows:

$$\tilde{s}_{\mathcal{Z}}[n,\mu] = \sum_{k=-\infty}^{\mu} \tilde{s}_{\mathcal{Z},k}[n] w[n-kH] \tag{3.20}$$

This form will converge to the real desired playback signal as $\mu \to \infty$. Aside from being more general, this approach allows for use of overlapping windows. This was found to be beneficial, as overlapping windows were found to reduce edge effects in the resulting loudspeaker input signals.

3.6.2 Block Based Loudspeaker Input Signal Computation

As mentioned in the introduction, the block-based approach aims to compute the loudspeaker input signals at the same rate as the desired playback signals $s_{\mathcal{Z}}[n]$ are received. That is to say: when a new block μ of the desired playback signal is revealed, a new block μ of loudspeaker input signals need be computed.

In order to compute the loudspeaker input signals in this block-wise fashion, a similar decomposition as for the desired playback signals is performed. Consider decomposing the loudspeaker input signals in blocks through the windows as follows:

$$\begin{aligned}\tilde{x}_{\mathcal{Z}}^{(l)}[n,\mu] &= \sum_{k=-\infty}^{\mu} \tilde{x}_{\mathcal{Z},k}^{(l)}[n]w[n-kH] && (3.21)\\ &= \sum_{k=-\infty}^{\mu-1} \tilde{x}_{\mathcal{Z},k}^{(l)}[n]w[n-kH] + \tilde{x}_{\mathcal{Z},\mu}^{(l)}[n]w[n-\mu H] && (3.22)\\ &= \tilde{x}_{\mathcal{Z}}^{(l)}[n,\mu-1] + \tilde{x}_{\mathcal{Z},\mu}^{(l)}[n]w[n-\mu H] && (3.23)\end{aligned}$$

The equation above shows that $\tilde{x}_{\mathcal{Z}}^{(l)}[n]$ can be computed by recursively computing its coefficients $\tilde{x}_{\mathcal{Z},k}^{(l)}$ and adding them. Note that this recursive property also holds for the desired playback signal $\tilde{s}_{\mathcal{Z}}[n]$.

The approach to find $\tilde{x}_{\mathcal{Z}}^{(l)}[n]$ is now as follows. When the μ^{th} block of the desired playback signals is revealed, this will allow for the calculation of the corresponding target sound pressure. Next, the μ^{th} coefficient $\tilde{x}_{\mathcal{Z},\mu}^{(l)}$ is computed such that $\tilde{x}_{\mathcal{Z}}^{(l)}[n]$ attains this target sound pressure.

How $\tilde{x}_{\mathcal{Z}}^{(l)}[n]$ and $\tilde{s}_{\mathcal{Z}}[n]$ relate to the resulting and target sound pressure respectively is the topic of the next section.

3.6.3 Block Based Sound Pressure

As discussed previously, the Multi-Zone Pressure-Matching (MZ-PM) algorithm attempts to control the loudspeaker input signals $x_{\mathcal{Z}}^{(l)}[n]$ for $\mathcal{Z} \in \{\mathcal{A}, \mathcal{B}\}$ such that the resulting sound pressure matches a specified target sound pressure $t^{(m)}[n]$ at all control points m.

Previous sections gave block-based versions of the loudspeaker input signals and the desired playback signals. In this section, block-based versions of the resulting and target sound pressure will be given.

The block-based target sound pressure will be denoted by $\tilde{t}^{(m)}[n,\mu]$, and 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 **??**:

$$\tilde{t}^{(m)}[n,\mu] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * \tilde{s}_{\mathcal{Z}}\right)[n] \tag{3.24}$$

$$= \sum_{l=0}^{N_L} \sum_{k=-\infty}^{\mu} \sum_{m=0}^{N_h-1} h^{(l,m)}[m]\tilde{s}_{\mathcal{Z},k}[n-m]w[n-kH-m] \tag{3.25}$$

$$= \tilde{t}^{(m)}[n,\mu-1] + \sum_{l=0}^{N_L} \sum_{m=0}^{N_h-1} h^{(l,m)}[m]\tilde{s}_{\mathcal{Z},\mu}[n-m]w[n-\mu H-m] \tag{3.26}$$

The definition above holds for all points $m \in Z$, 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 will be denoted by $\tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu]$, and 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 **??**. This results in the following:

$$\tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * \tilde{x}_{\mathcal{Z}}^{(l)}\right)[n] \tag{3.27}$$

$$= \sum_{l=0}^{N_L} \sum_{k=-\infty}^{\mu} \sum_{m=0}^{N_h-1} h^{(l,m)}[m]\tilde{x}_{\mathcal{Z},k}^{(l)}[n-m]w[n-kH-m] \tag{3.28}$$

$$= \tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu-1] + \sum_{l=0}^{N_L} \sum_{m=0}^{N_h-1} h^{(l,m)}[m]\tilde{x}_{\mathcal{Z},\mu}^{(l)}[n-m]w[n-\mu H-m] \tag{3.29}$$

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 Multi-Zone Pressure-Matching (MZ-PM) approach are known. The next section will use these quantities to state the block-based MZ-PM algorithm.

3.6.4 Derivation of Block-Based Multi-Zone Pressure-Matching

After translating the loudspeaker input signals and the target sound pressure into their block-wise counterparts, the Block-Based Multi-Zone Pressure-Matching (BB-MZ-PM) algorithm can be stated.

As mentioned before, the approach that will be taken is to compute the μ^{th} coefficient of the loudspeaker input signal $\tilde{x}_{\mathcal{Z},\mu}^{(l)}$ such that the 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 the most recent loudspeaker coefficients $\tilde{x}^{(l)}_{\mathcal{Z},\mu}$ are being controlled. The previous coefficients are assumed to have already been played. As such, they 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:

$$\underset{\tilde{x}^{(l)}_{\mathcal{A},\mu}[n],\,\tilde{x}^{(l)}_{\mathcal{B},\mu}[n]\,\forall\, l}{\arg\min} \quad \sum_{m\in A}\left|\left|\tilde{p}^{(m)}_{\mathcal{A}}[n,\mu]-\tilde{t}^{(m)}_{\mu}[n,\mu]\right|\right|_2^2+\sum_{m\in A}\left|\left|\tilde{p}^{(m)}_{\mathcal{B}}[n,\mu]\right|\right|_2^2+ \tag{3.30}$$

$$\sum_{m\in B}\left|\left|\tilde{p}^{(m)}_{\mathcal{B}}[n,\mu]-\tilde{t}^{(m)}_{\mu}[n,\mu]\right|\right|_2^2+\sum_{m\in B}\left|\left|\tilde{p}^{(m)}_{\mathcal{A}}[n,\mu]\right|\right|_2^2 \tag{3.31}$$

Note that this problem implicitly contains the target sound pressure and resulting sound pressure of the previous blocks $\mu - 1$ 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 $\tilde{x}^{(l)}_{\mathcal{A},\mu}[n]$ and $\tilde{x}^{(l)}_{\mathcal{B},\mu}[n]$ as new blocks $\tilde{s}_{\mathcal{A},\mu}[n]$ and $\tilde{s}_{\mathcal{B},\mu}[n]$ are revealed. The final loudspeaker input signals $\tilde{x}^{(l)}_{\mathcal{Z}}[n,\infty]$ can then be found by means of **??**.

3.7 Frequency Domain Conversion

In the previous section, the Block-Based Multi-Zone Pressure-Matching (BB-MZ-PM) algorithm was derived. When deriving this algorithm it was stated that it's advantages are twofold.

Firstly, one advantage of using this algorithm over its non block-based counterpart is that it can work in real-time. Secondly, the block-based approach can operate on short time-scales. This is useful, as the perceptual model that we wish to integrate operates on time-scales of the order of 20 to 200 milliseconds.

There is however an additional adjustment that needs to be made before the perceptual model can be integrated. Currently, the BB-MZ-PM algorithm minimizes L2-error in the time-domain, whereas the detectability is computed in the frequency domain. In order to integrate the perceptual model and the sound zone algorithm, they must operate in the same domain.

In this case, it was chosen to use convert the existing sound zone algorithm to the frequency domain. This was chosen as to remain close to the definition of the detectability. Approximating detectability in the time domain was not explored further and could potentially be of interest for future work.

To this end, this section will adjust the existing time domain BB-MZ-PM algorithm to an equivalent frequency domain formulation. This will be done by first introducing a suitable transformation relating the time and frequency domain quantities. The transformed quantities can than be used define the frequency domain version of the BB-MZ-PM algorithm.

3.7.1 Frequency Domain Transformation

A suitable transform to obtain the frequency domain representation of a signal is the DFT. However, it is important to take a number of precautions before applying the DFT directly, as the computation of the sound pressures used in the optimization problem introduced previously involves taking the linear convolution of the loudspeaker input signals.

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 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.

A suitable transform is thus given by a $N_w + N_h - 1$ point DFT.

$$X[k] = \sum_{n=0}^{N_w+N_h-2} x[n] \exp\left(\frac{-j2\pi kn}{N_w + N_h - 1}\right) \tag{3.32}$$

Here, $x[n]$ and $X[k]$ are the time- and frequency-domain representations of an arbitrary sequence.

3.7.2 Quantities the Frequency Domain

In this section, the quantities used in the Block-Based Multi-Zone Pressure-Matching approach will be converted to their frequency domain counterparts.

Essentially, this involves converting the sound pressures $\tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu]$ and $\tilde{t}^{(m)}[n,\mu]$ to their frequency domain versions given by $\tilde{P}_{\mathcal{Z},\mu}^{(m)}[k]$ and $\tilde{T}_{\mu}^{(m)}[k]$ respectively.

This can be done by applying the previously derived transform directly to these quantities. This results in the following expressions.

$$\tilde{T}^{(m)}[k,\mu] = \tilde{T}^{(m)}[k,\mu-1] + \sum_{l=0}^{N_L} H^{(l,m)}[k]\tilde{S}_{\mathcal{Z},\mu}[k] \tag{3.33}$$

$$\tilde{P}_{\mathcal{Z}}^{(m)}[k,\mu] = \tilde{P}_{\mathcal{Z}}^{(m)}[k,\mu-1] + \sum_{l=0}^{N_L} H^{(l,m)}[k]\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k] \tag{3.34}$$

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:

$$\tilde{S}_{\mathcal{Z},\mu}[k] = \sum_{n=0}^{N_w+N_h-2} \tilde{s}_{\mathcal{Z},\mu}[n]w[n-\mu H] \exp\left(\frac{-j2\pi kn}{N_w + N_h - 1}\right) \tag{3.35}$$

$$\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k] = \sum_{n=0}^{N_w+N_h-2} \tilde{x}_{\mathcal{Z},\mu}^{(l)}[n]w[n-\mu H] \exp\left(\frac{-j2\pi kn}{N_w + N_h - 1}\right) \tag{3.36}$$

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

3.7.3 Proposed Frequency Domain Approach

Using the previously derived quantities, it is possible express the frequency domain version of the Block-Based Multi-Zone Pressure-Matching (BB-MZ-PM) approach as follows:

$$\underset{\tilde{x}_{\mathcal{A},\mu}^{(l)}[n],\,\tilde{x}_{\mathcal{B},\mu}^{(l)}[n]\,\forall\, l}{\arg\min} \quad \sum_{m\in A} \left|\left|\tilde{P}_{\mathcal{A}}^{(m)}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right|\right|_2^2 + \sum_{m\in A} \left|\left|\tilde{P}_{\mathcal{B}}^{(m)}[k,\mu]\right|\right|_2^2 + \tag{3.37}$$

$$\sum_{m\in B} \left|\left|\tilde{P}_{\mathcal{B}}^{(m)}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right|\right|_2^2 + \sum_{m\in B} \left|\left|\tilde{P}_{\mathcal{A}}^{(m)}[k,\mu]\right|\right|_2^2 \tag{3.38}$$

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 $\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k]$. This however introduces issues as it requires the truncation of the time-domain version to N_w samples, which was found to introduce artifacts.

3.8 Conclusion

Implementation and Evaluation of a Perceptual Sound Zone Algorithm

4

4.1 Introduction

This chapter seeks to combine the selected perceptual model and the selected sound zone algorithm. In chapter 2, it was found that the Par detectability was best suited.

Based on the constraints posed by this perceptual model, chapter 3 found that a pressure matching approach was best suited. The chapter concluded with the implementation of a multi-zone pressure matching algorithm that can be used as a foundation to build the perceptual algorithm on.

This chapter is structured as follows.

4.2 Unconstrained Perceptual Pressure Matching

4.3 Perceptual Minimization Approach

Detectability is defined as:

$$D(x[n], e[n]) = ||\mathbf{w}(X[k]) \circ \mathbf{E}[k]||_2^2 \tag{4.1}$$

The detectability quantifies how detectable masked signal $e[n]$ is in presence of masker signal $x[n]$. It is calibrated such that $D(x[n], e[n]) = 1$ for a masker-masked signal pair Here, $\mathbf{E}[k]$ is a vector containing the DFT of the $e[n]$. The weighting vector $\mathbf{w}(X[k])$ is computed based on the psycho-acoustical masking properties of $x[n]$, which are calculated based on its frequency domain counterpart $X[k]$.

The idea now is to integrate this into the previously introduced cost function:

$$\underset{\tilde{x}_{\mathcal{A},\mu}^{(l)}[n], \tilde{x}_{\mathcal{B},\mu}^{(l)}[n] \,\forall\, l}{\arg\min} \quad \sum_{m \in A} \left|\left|\tilde{P}_{\mathcal{A}}^{(m)}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right|\right|_2^2 + \sum_{m \in A} \left|\left|\tilde{P}_{\mathcal{B}}^{(m)}[k,\mu]\right|\right|_2^2 + \tag{4.2}$$

$$\sum_{m \in B} \left|\left|\tilde{p}_{\mathcal{B}}^{(m)}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right|\right|_2^2 + \sum_{m \in B} \left|\left|\tilde{P}_{\mathcal{A}}^{(m)}[k,\mu]\right|\right|_2^2 \tag{4.3}$$

Essentially, the approach is to replace the norms in the equation above by the perceptually weighted equivalent. For each norm, we consider the masking effects determined by the target sound pressure for the respective point m. This results in the following cost function:

$$\underset{\tilde{x}_{\mathcal{A},\mu}^{(l)}[n], \tilde{x}_{\mathcal{B},\mu}^{(l)}[n] \,\forall\, l}{\arg\min} \quad \sum_{m \in A} \left|\left|\mathbf{w}\left(\tilde{T}^{(m)}[k,\mu]\right) \left[\tilde{P}_{\mathcal{A}}^{(m)}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right]\right|\right|_2^2 + \tag{4.4}$$

$$\sum_{m \in A} \left|\left|\mathbf{w}\left(\tilde{T}^{(m)}[k,\mu]\right) \left[\tilde{P}_{\mathcal{B}}^{(m)}[k,\mu]\right]\right|\right|_2^2 + \tag{4.5}$$

$$\sum_{m \in B} \left|\left|\mathbf{w}\left(\tilde{T}^{(m)}[k,\mu]\right) \left[\tilde{P}_{\mathcal{B}}^{(m)}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right]\right|\right|_2^2 + \tag{4.6}$$

$$\sum_{m \in B} \left|\left|\mathbf{w}\left(\tilde{T}^{(m)}[k,\mu]\right) \left[\tilde{P}_{\mathcal{A}}^{(m)}[k,\mu]\right]\right|\right|_2^2 \tag{4.7}$$

The cost function behaves the same way, except all terms are now psycho-acoustically weighted. The weighting vectors are chosen based on the target sound pressure in the relevant control point m. This essentially exploits the psycho-acoustical masking properties of the target signal.

4.4 Constrained Perceptual Pressure Matching

4.5 Perceptual Constraining Approach

It turns out that the perceptual minimization approach doesn't always trade-off nicely between interference suppression and reproduction error minimization. In order to do so, we can move certain terms to the constraints. That is done below:

$$\begin{align}
\underset{\tilde{x}^{(l)}_{\mathcal{A},\mu}[n],\, \tilde{x}^{(l)}_{\mathcal{B},\mu}[n]\, \forall\, l}{\arg\min} \quad & \sum_{m \in B} \left|\left| \mathbf{w}\left(\tilde{T}^{(m)}[k,\mu]\right) \left[\tilde{P}^{(m)}_{\mathcal{A}}[k,\mu]\right] \right|\right|_2^2 + \tag{4.8} \\
& \sum_{m \in A} \left|\left| \mathbf{w}\left(\tilde{T}^{(m)}[k,\mu]\right) \left[\tilde{P}^{(m)}_{\mathcal{B}}[k,\mu]\right] \right|\right|_2^2 \tag{4.9} \\
\text{subject to} \quad & \left|\left| \mathbf{w}\left(\tilde{T}^{(m)}[k,\mu]\right) \left[\tilde{P}^{(m)}_{\mathcal{A}}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right] \right|\right|_2^2 \leq Q \quad \forall m \in \mathcal{B} \tag{4.10} \\
& \left|\left| \mathbf{w}\left(\tilde{T}^{(m)}[k,\mu]\right) \left[\tilde{P}^{(m)}_{\mathcal{B}}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right] \right|\right|_2^2 \leq Q \quad \forall m \in \mathcal{A} \tag{4.11}
\end{align}$$

Here, we are minimizing the perceptually-weighted leakage, subject to constraints limiting the detectability of the reproduction error per point m.

4.6 Results

4.7 Conclusion

5
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)*, pp. 136–140, IEEE, 2019.

[4] T. Lee, J. K. Nielsen, and M. G. Christensen, "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, pp. 749–752, IEEE, 2001.

[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*, pp. 1–4, VDE, 2012.

[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] 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.

[15] 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," techreport 3, International Organization for Standardization, Geneva, CH, Oct. 1993.

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

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

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

[24] 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.

[25] M. F. S. Gálvez, S. J. Elliott, and J. Cheer, "Time domain optimization of filters used in a loudspeaker array for personal audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1869–1878, 2015.

[26] 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)*, pp. 491–495, IEEE, 2018.