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

CAS-2021-4480767

M.Sc. Thesis

Sound Zones with a Cost Function based on Human Hearing

Niels Evert Marinus de Koeijer B.Sc.

Abstract

With the aid of an array of loudspeakers, sound zone algorithms seek to reproduce multiple distinct zones of audio inside an enclosure. Typical approaches determine the loudspeaker inputs by optimizing over a cost function that models the sound pressure inside the enclosure. However, recent methods propose cost functions that include a perceptual model of the human auditory system, which further models the perception of sound. This thesis investigates such an approach by proposing a framework within which sound zones are constructed through optimization over a perceptual model. The framework is used to propose two perceptual sound zone algorithms: unconstrained and constrained perceptual pressure matching. Simulations of the proposed algorithms and a reference algorithm are presented to determine the benefits of including auditory-perceptual information in sound zone algorithms. From this, it is found that the unconstrained perceptual approach outperforms the reference in terms of various perceptual measures. In addition, it is found that adding perceptual constraints to the optimization problem allows for control of sound zones which correlates well with other perceptual quality measures.

TU Delft

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

Sound Zones with a Cost Function based on Human Hearing

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Electrical Engineering

by

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

This work was performed in:

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

Delft University of Technology

Delft University of Technology
Department of
Microelectronics & Computer Engineering

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

Dated: September 16, 2021

Chairman: ______________________

dr.ir. R.C. Hendriks

Daily Supervisor: ______________________

dr. J. Martínez-Castañeda

Committee Members: ______________________

dr. M. Mastrangeli

____

M. Bo Møller, PhD

____

dr. P. Martínez-Nuevo

Abstract

With the aid of an array of loudspeakers, sound zone algorithms seek to reproduce multiple distinct zones of audio inside an enclosure. Typical approaches determine the loudspeaker inputs by optimizing over a cost function that models the sound pressure inside the enclosure. However, recent methods propose cost functions that include a perceptual model of the human auditory system, which further models the perception of sound. This thesis investigates such an approach by proposing a framework within which sound zones are constructed through optimization over a perceptual model. The framework is used to propose two perceptual sound zone algorithms: unconstrained and constrained perceptual pressure matching. Simulations of the proposed algorithms and a reference algorithm are presented to determine the benefits of including auditory-perceptual information in sound zone algorithms. From this, it is found that the unconstrained perceptual approach outperforms the reference in terms of various perceptual measures. In addition, it is found that adding perceptual constraints to the optimization problem allows for control of sound zones which correlates well with other perceptual quality measures.

Acknowledgments

I want to thank dr. J. Martínez-Castañeda, M. Bo Møller, PhD, dr.ir. R.C. Hendriks, and dr. P. Martínez-Nuevo for all their help and support during the project. A special thanks to M. Bo Møller, PhD for often pushing me in the right direction through our numerous whiteboard sessions. Secondly, I would like to give a special thanks to dr. J. Martínez-Castañeda for being a significant help during and outside the project.

In addition to this, I would like to thank Cassandra Grützner and especially Dorottya Hauk for their help proofreading the thesis and for many good times during the project.

Niels Evert Marinus de Koeijer B.Sc.
Delft, The Netherlands
September 16, 2021

Contents

Chapter 1

Introduction

1.1 Preface: the Sound Zone Problem

Sound systems are used worldwide to fill rooms with enjoyable audio content. Problems arise, however, when multiple people in the same room want to enjoy different audio content simultaneously.

For example, one person may want to enjoy a movie on the television, while others 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 audio sources can lead to a situation where both individual experiences are diminished.

In recent years, attempts have been made to solve this problem by controlling the spatial reproduction of sound in such a way that different areas in a room have distinct content without interfering with each other. This is typically done by controlling the sound pressure created by an array of loudspeakers.

One attempt at solving this problem is through the use of 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 decide how to use the sound system's loudspeakers to reproduce the specified audio content in each zone. Using the principles of constructive and destructive interference, this is done in such a way that there is minimal interference between zones [1]. That is to say: the audio content of each zone is not audible in the others.

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

In practice, however, the sound zone algorithm will not always do a perfect job [2].

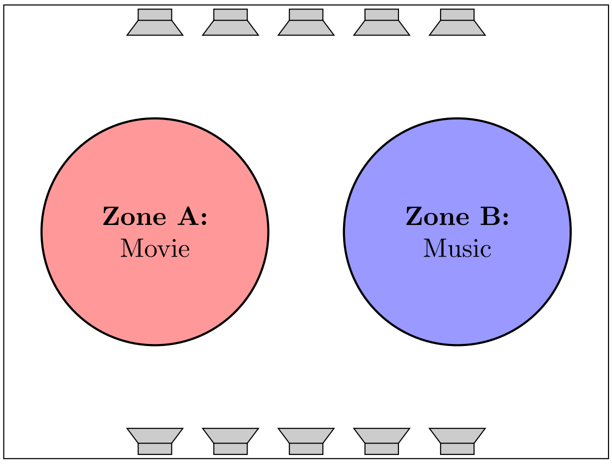


Figure 1.1: A room containing a sound system consisting of an array of loudspeakers and two zones. The goal of the sound zone algorithm is to control the sound system in such a way that the red zone contains the audio of a movie, and the blue zone contains the music with minimal interference.

The performance of algorithms depends on the environment and the available sound system. Depending on the chosen zones, number and position of the loudspeakers, and the room, the interference between zones can typically only be reduced by so much. As such, the audio content of one zone is often still audible in other zones [2].

Improving sound zone algorithms is thus an active topic of research. One recent approach is to include a model of the human auditory system, which models how humans perceive sound. Typically, sound zone algorithms use sound pressure [1], which is a physical quantity characterizing the sound. Sound pressure does not always accurately describe what is important for the perception of sound. Including a perceptual model may allow the algorithm to focus on reproducing and canceling the parts of the audio content that matter perceptually.

Early results show that a perceptual sound zone approach is promising. Recent work by Donley et al. explored including the absolute threshold of hearing, which models the lowest sound pressure humans can hear, into sound zone algorithms. This pursuit found an increased quality of the reproduced audio in the zones [3]. Other work by Lee et al. showed that including a perceptually motivated weighting in the sound zone algorithm outperforms traditional algorithms [4, 5].

This work seeks to explore this perceptual approach further. This is done by proposing novel perceptual sound zone algorithms based on the pressure matching approach. The proposed method is then used to determine the benefits of including perceptual information into sound zone algorithms.

1.2 Objectives and Organization

As stated in the preface, this thesis investigates the methodology and the benefits of including perceptual information in sound zone algorithms. To this end, the work in this thesis seeks to answer two research questions:

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

The answers to these questions are summarized in the conclusion given by Chapter 6. What follows is a description of the approach that is taken in answering these research questions, alongside the structure of the rest of this document.

1.2.1 Creation of Perceptual Sound Zone Algorithms

The first research question RQ1,

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

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

- First, in Chapter 2 a literature review is performed to determine which perceptual models are suitable for use in a perceptual sound zone algorithm. In this pursuit, one perceptual model is found to be the most promising and discussed in further detail.
- Next, in Chapter 3 a perceptual sound zone framework is proposed, which uses the selected perceptual model. This framework is motivated through a literature review of existing sound zone approaches and by reflecting on the mathematical properties of the perceptual model.
- Finally, in Chapter 4 the proposal of two perceptual sound zone algorithms that employ the proposed perceptual sound zone framework is discussed.

1.2.2 Determining Benefits of Perceptual Sound Zone Algorithms

The second research question RQ2,

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

is answered in Chapter 5, where the perceptual sound zone algorithms derived in answering RQ1 are analyzed and compared with a non-perceptual reference sound zone algorithm.

Chapter 2

Review and Implementation of Perceptual Models

In the field of psycho-acoustics significant research has been done in characterizing the auditory perception and time-frequency analysis capabilities of the human ear [6]. From this understanding, several perceptual models have been proposed which aim to model the perception of auditory stimuli by humans [7].

Perceptual models are employed for various purposes. Objective audio quality measures, for example, are perceptual models which aim to predict the perceived quality of audio [8]. In another example, perceptual audio coding uses models of auditory perception to minimize the perceived artifacts introduced when performing the compression of audio [9].

In general, many perceptual models operate on a time-frequency internal-ear representation of the input stimuli, obtained by applying an analysis filter bank. Among other effects, the filtering performed by the human ear is often taken into account at this stage [7, 10]. This representation is then used to determine its perceptually relevant aspects of the input stimuli [9].

One aspect often used in perceptual models are the various auditory masking properties of the input stimuli [9]. In general, auditory masking refers to the effects one sound has on the perception of other sounds [6]. In simultaneous masking, for example, one loud tone may overpower a tone of a similar frequency, rendering the latter tone inaudible [6].

Another aspect that is often used is the "threshold of hearing", which determines the minimum sound pressure level that can be perceived by a human [9]. Combining this principle with the masking properties, one can define the "masking threshold" of input stimuli. This threshold determines the sound pressure level required for other stimuli to be audible to a human observer in the presence of the input stimuli [6] and is often used in perceptual models [7, 10]. For example, the threshold of hearing is used in perceptual audio coding to make the coding artifacts inaudible [9].

This chapter will motivate the use of the "Par distortion detectability" as the perceptual model used in perceptual sound zone algorithm framework proposed in Chapter 3, which is subsequently used in Chapter 4 to propose two perceptual sound zone algorithms. In doing so, several other perceptual measures are discussed in detail, some of which are used in Chapter 5 to evaluate the performance of the proposed algorithms.

The structure of this chapter is given as follows.

- This chapter begins with Section 2.1 which documents a review of possible candidate perceptual models from literature for use in the perceptual sound zone framework.
- Next, Section 2.2 motivates the selection of one of the reviewed candidates, namely the “Par distortion detectability”, as the perceptual model for use in the proposed perceptual sound zone framework.
- Finally, the implementation and behavior of the “Par distortion detectability” is discussed in more detail in Section 2.3.

2.1 Review of Perceptual Models from Literature

This section documents a review of perceptual models that are promising for use in perceptual sound zone algorithms.

Especially promising are models that attach some “score” or “rating” to the perceptual quality of input signals. These ratings can be used in algorithms to obtain an optimal rating through optimization. In addition to this, they can also be used to quantify the quality of the algorithms in Chapter 5. As such, the focus of the literature review is not on the latest findings in the field of psycho-acoustics or models that most accurately emulate the behavior of the human ear but rather on models that quantify a perceptual quality.

In addition to this, the review also focuses on the optimization tractability of the models. This information is used in Section 2.2 to motivate the use of the “Par distortion detectability” perceptual model in the proposed perceptual sound zone algorithm.

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

2.1.1 Review of Objective Measures

In order to objectively determine the perceived quality of audio, one approach is to use listening tests. These are tests in which subjects are asked to rate a property (or properties) of a set of audio stimuli. One example where listening tests are used is for the evaluation of speech intelligibility of hearing aids [11]. Another example is determining which loudspeaker has higher perceived sound quality.

Performing listening tests is, however, often cumbersome due to the large amount of human labor involved. This motivates the use of objective quality measures, which attempt to predict the outcomes of these objective listening tests. This is very useful for algorithm developers, as they can get an indication of how well they are doing without having to perform a labor-intensive listening test [11].

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

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

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

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

2.1.1.1 Review of Objective Speech Quality Measures

There have been a number of attempts to create objective measures to quantify the perceived quality of speech. In this section, three objective speech quality measures are discussed. Of these three measures, the Perceptual Evaluation of Speech Quality (PESQ) is used in the evaluation of the proposed perceptual sound zone algorithm in Chapter 5.

- Perceptual Evaluation of Speech Quality (PESQ) [12] is a measure that attempts to determine the perceived quality of speech. It was standardized by the International Telecommunication Union (ITU-T) in 2001.

  PESQ is computed by first applying an auditory transform that maps the reference and degraded speech into a time-frequency representation that models the perceived loudness of the signals. From this internal representation, so-called symmetric and asymmetric disturbances are determined by computing differences between the time-frequency bins of the reference and degraded speech. A non-linear average is then taken to obtain the average disturbance per time bin. These averaged disturbances are then mapped to the outcomes of listening test outcomes through linear combination [12].

- The Perceptual Objective Listening Quality Assessment (POLQA) [13] is a speech quality measure that was standardized by the International Telecommunication Union (ITU-T) in 2011. It was intended to be the successor of PESQ, with the improvement of having more accurate predictions on a broader range of distortions.

  POLQA works with a similar internal representation to PESQ but computes distortion in a different way as to be capable of handling global temporal compression and expansions [13].

- Virtual Speech Quality Objective Listener (ViSQOL) [14, 15] is a measure developed in 2012 in a collaboration between Trinity College and Google.

  ViSQOL uses a different internal representation than PESQ and POLQA as it uses neurograms rather than loudness representations. The neurograms are then compared through the Neurogram Similarly Index Measure (NSIM). 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 [14], which is then used to make predictions.

In general, PESQ, POLQA, and ViSQOL require many steps to compute and are difficult to optimize for due to conditional branches within the algorithms and many non-differentiable steps such as clipping [12, 13, 14]. 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 [16].

2.1.1.2 Review of Objective Speech Intelligibility Measures

Intelligibility of speech is defined as the percentage of words identified correctly given a degraded speech signal. Objective speech intelligibility measures seek to predict this percentage. In this section, the Short-Time Objective Intelligibility (STOI) [11] measure and the Speech Intelligibility In Bits (SIIB) [17] measure are discussed. Both measures are used in the evaluation of the proposed perceptual sound zone algorithm in Chapter 5.

- The Short-Time Objective Intelligibility (STOI) [11] was proposed by Taal et al. in 2011 as a speech intelligibility measure that could make accurate predictions for speech signals degraded by time-frequency weighted distortions.

  For its internal representation, it finds a time-frequency internal representation through filtering the input stimuli with a filter bank consisting of 1/3 octave bands, and then segmented the filter taps into short time frames. Silent bins that do not contain speech are removed, and clipping is applied to limit the effect of one severely degraded time-frequency bin. The average correlation coefficient between the time-frequency bins of the internal representation of the reference and degraded segments is then computed and averaged over all bins to determine the intelligibility [11].

- The Speech Intelligibility In Bits (SIIB) [17] was introduced by Van Kuyk et al. in 2017 as a speech intelligibility measure that could be motivated through the mutual information rate from information theory. As such, SIIB is given in bits.

  The idea behind SIIB is that the intelligibility of speech is related to the information shared between intended and degraded speech. SIIB models how the reference speech signal transforms to the degraded speech signal as a transmission channel. Among other aspects, this transmission channel includes a model of the human auditory system [17]. This communication channel is then used to compute the mutual information rate.

Both STOI and SIIB are difficult to optimize for directly. In STOI, the removal of silent regions and the clipping operator are non-differentiable operations. Furthermore, the computation of the correlation coefficient is a non-convex function of the degraded speech [11]. SIIB is in general non-convex and non-differentiable as it uses the Karhunen-Loève transform and a K-nearest neighbor estimator to compute the mutual information rate [17]. However, if the communication channel is approximated as Gaussian, the mutual information can be computed in closed form, and SIIB becomes a differentiable measure [17].

2.1.1.3 Review of Objective Audio Quality Measures

The previous objective quality measures are both intended for evaluating speech. In this section, two objective quality measures are discussed that are designed for evaluating the perceived quality of any audio stimuli.

- The Perceptual Evaluation of Audio Quality (PEAQ) [18] is an audio quality measure 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 several perceptually relevant features, referred to by PEAQ as Model Output Variables (MOVs), are extracted. An example of these MOVs is 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 [18].

- In 2015, it was found that, with some adjustments, the previously discussed ViSQOL measure could be used to determine audio quality. This resulted in a new measure, ViSQOLAudio [19].

  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 [19].

PEAQ and ViSQOLAudio are both difficult to optimize. A number of the MOVs computed in PEAQ, such as the partial noise loudness, are non-differentiable [18]. As ViSQOLAudio is similar to ViSQOL with some minor adjustments, it is similarly challenging to optimize.

2.1.1.4 Review of the Distraction Model

In an elicitation study performed by Francombe et al. in 2014 [20], “distraction” was determined to be the keyword that best describes the perceptual experience of interfering audio programs. Further research led to the proposal of a “distraction model”, which is capable of estimating how distracting an interferer stimulus is given a certain target stimuli [21]. This model was designed with the application of sound zones in mind and is as such an especially promising for use in the evaluation of the proposed algorithm in Chapter 5.

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 stimulus they were instructed to focus listening to. At the same time, an interferer audio stimulus was played to distract the participant from the target. The participants were given a scale between 0 and 100 on which they were asked to rate how distracting the interference was when listening to the target program, where a 100 indicates that the interferer “overpowered” the target audio [21].

The target-interferer stimuli pairs and corresponding ratings resulted in a dataset. This dataset was then used to fit a model which predicted the distraction given novel a target-interferer stimuli pair. The model consisted of taking a linear combination of 5 features that were computed from the stimuli [21].

Computing said features could, however, not be performed in real-time. The reason for this was that the original distraction model is too computationally complex [22]. 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 less computationally 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 [22].

At face value, the real-time distraction model seems promising to optimize. However, while easy to compute, the model is non-differentiable as the model uses piecewise functions and non-convex due to taking the logarithm of the square of the input signals. In addition to this, the model also performs operations that are difficult to express mathematically, such as counting the number of short-time blocks that exceed a certain threshold [22].

2.1.2 Review of Perceptual Models used in Audio Coding

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

To minimize the impact of these errors, many audio coding algorithms use a perceptual model to quantify how disturbing the introduced distortions are [9]. The perceptual model is used to introduce encoding errors in such a way that the audio output signal is minimally perceptually distinguishable from the audio input signal [10]. This model typically takes the form of a distortion function which determines how audible the difference between a reference input audio signal and a distorted output audio signal is. This function can be used to, for example, encode an input audio signal such that it has minimal distortion for a specified bitrate.

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

2.1.2.1 Review of Perceptual Models from ISO MPEG Standard

The ISO/IEC 11172-3 standard specifies a coded representation for audio files [23], and a corresponding decoder. An encoder for 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 [24].

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

The psycho-acoustical models work by subdividing the input audio signal into frequency bands that correspond to the frequency bands in the human auditory system. The model then determines how much quantization noise can be added separately per band without the noise becoming audible. As such, the model assumes that the distortion signal is noise-like [7], 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 [10]. As such, examples exist from which optimization schemes could be inspired.

2.1.2.2 Review of Par Distortion Detectability Measure

In 2005, van der Par et al. proposed a novel perceptual model designed for use in audio coding [7]. The model defines a distortion measure that determines the "distortion detectability" of a distortion signal in the presence of a masking signal. That is to say, the function quantifies the degree to which a human is to detect a distortion signal while also listening to the masking signal. For audio coding purposes, this distortion signal is the error introduced due to the audio compression.

Similarly to the ISO MPEG perceptual models, the Par detectability typically operates on short-time segments, typically in the order of 20 to 200 milliseconds [7]. The proposed method, however, differentiates itself from the previously discussed ISO MPEG models in three ways.

Firstly, the paper uses newer findings from psycho-acoustic literature, namely spectral integration. In spectral integration, the masking effects from neighboring bands are taken into account when computing the masking effects. The psycho-acoustical models defined in the ISO MPEG standard does not do this as it effectively works independently per band [10].

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 that quantifies how detectable a disturbance stimulus is.

The proposed distortion measure can be expressed as a squared L^2-norm, making it tractable for integration into existing least-square problems. As such, the Par distortion detectability has been used in many signal processing applications, examples ranging from speech enhancement to removing perceptually irrelevant sinusoidal components [25, 26].

2.1.2.3 Review of Taal Distortion Detectability Measure

A paper from 2012 by Taal et al. proposed a novel perceptual model [10] which introduces a alternative definition to the distortion detectability defined in the Par distortion detectability.

In contrast to the approach proposed by van der Par et al. [7], the Taal distortion detectability measure takes temporal characteristics of the distortion and masking signals into account. The inclusion of temporal information allows for the suppression of “pre-echoes”, which is an artifact that the Par distortion detectability suffers from [10]. The “pre-echoes” artifacts arise 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, resulting in quantization noise not being masked.

In contrast to other temporal perceptual models, the Taal Detectability has a relatively low computational complexity.

The computational demand was, however, shown to be higher than the Par distortion detectability [10], especially for longer time segments.

2.2 Motivating Selection of Par Distortion Detectability

From the perceptual models discussed in the literature review given in Section 2.1, the Par distortion detectability is selected for use in the proposed perceptual sound zone framework, as it is found to be the most tractable for optimization. This section seeks to motivate this.

In Chapter 3 it is shown that sound zone algorithms are typically posed as optimization problems. The goal of optimization problems is typically to minimize or maximize a cost function, which is done by leveraging the (sub)differential of the function.

Furthermore, many approaches are posed as convex optimization problems. Convex optimization is a sub-class of optimization problems that guarantee that the optimizer is globally unique [27]. As such, one does not have to deal with many suboptimal local optima. In addition to this, there are many efficient solvers available for convex optimization problems.

As such, perceptual models which contain conditional branching or complex, non-convex operations which cannot readily be integrated into cost functions are less promising.

To this end, all the objective audio measures discussed in Section 2.1.1 are ruled out for use in the perceptual sound algorithm. As discussed, all models showed a degree of non-differentiability and non-convexity in their computation. They are challenging to integrate into convex optimization problems and are therefore not used in the proposed perceptual sound zone algorithm. They are, however, used in the evaluation of the proposed perceptual sound zone algorithm.

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

As such, the decision is between the Par and Taal distortion detectability, which are both expressed using a squared L2-norm, which is a convex function [27].

In contrast to the Par model, the Taal detectability takes into account the temporal properties of the input signal. This is beneficial, as it will lead to a more accurate description of the masking properties of the input signals. However, it has been shown to be at the cost of computational complexity. The Taal detectability has been shown to take at least two times as long to compute as the Par detectability, with this disparity seemingly growing as a function of input signal length [10].

In addition to this, the Taal model operates on time-domain versions of the input stimuli, whereas the Par model operates in the frequency-domain representations [7, 10]. Frequency-domain sound zone approaches are typically less demanding computationally than time-domain approaches [28].

As a lower computational complexity is desirable, the Par distortion detectability is used in the proposed perceptual sound zone algorithm. Exploring the possibilities of using the Taal detectability in a perceptual sound zone algorithm is found to be promising but is left to future work and not further explored.

2.3 Implementation and Analysis of the Par Distortion Detectability

The Par distortion detectability is the perceptual model used in the proposed perceptual sound zone framework. In this section, to give the reader a greater understanding of the model, the Par distortion detectability measure is considered in greater detail.

This section is organized as follows. First, Section 2.3.1 gives a high-level description of the Par distortion detectability, providing an intuitive understanding and introducing some of the notation that is used. Next, the steps for computing the distortion detectability are described in Section 2.3.2. Finally, Section 2.3.3 rewrites the distortion detectability into terms of a squared L^2-norm and provides some analysis of the behavior of the resulting representation.

2.3.1 High-Level Description of the Par Distortion Detectability

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

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

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

For this interpretation to be accurate, the signals $x[n]$ and $\varepsilon[n]$ should be short-time signals. The paper uses a signal length of 20 to 200 milliseconds [7]. This is important, as the model assumes that the psycho-acoustical properties of $x[n]$ and $\varepsilon[n]$ are stationary.

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

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

2.3.2 Computation Details of the Par Distortion Detectability

This section explores calculating the Par distortion detectability.k The first thing to note about the Par distortion detectability is that it is computed using the frequency domain representations of its inputs [7]. To this end, let $X[k]$ and $\mathcal{E}[k]$ denote the frequency domain representations of the masking signal $x[n]$ and the disturbance signal $\varepsilon[n]$ respectively.

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

Two subsequent filters are applied. The first filter models how parts of the ear filter the incoming sound with an outer- and middle-ear filter $H_{\text{om}}[k]$. Next, a 4^{th} order Gammatone filter bank is applied, modeling the frequency-place transform that occurs in basilar membrane inside of the ear [7].

The Gammatone filter bank consists of N_g filters. The frequency-domain representation of each individual filter is denoted by $\Gamma_i[k]$, for $1 \leq i \leq N_g$. The filters in the filter bank $\Gamma_i[k]$ have a bandwidth given by the equivalent rectangular bandwidth (ERB) and center frequencies are given by the corresponding equivalent rectangular bandwidth number scale (ERBS). Expressions for the gammatone filters $\Gamma_i[k]$ are provided by the original paper [7].

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

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

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

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

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

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

$\varepsilon[n]$ would be infinite. In order to take the frequency-dependence of the threshold of hearing into account, the previously described outer- and middle ear filters are defined as the inverse of the threshold of hearing [7].

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

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

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

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

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

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

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

The determination of the calibration constants C_a and C_s is discussed in Appendix A.

2.3.3 Least-Squares Formulation of the Par Distortion Detectability

This section will rewrite the previously introduced detectability into a least-squares representation [10]. This representation is more mathematically tractable than Equation (2.5) and thus will allow for easier integration into existing sound zone algorithms.

To obtain this expression, the sum of squares will be expressed as a L^2 norm.

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

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

In this case, the norm is taken over the frequency-domain sequence indexed by k. The rewrite above introduced perceptual weighting $W_x[k] \in \mathbb{R}^{N_x}$ informed by the auditory masking effects of the masking signal $x[n]$. The entries of the perceptual weighting can be understood as the importance of those frequencies for the total detectability. The perceptual weighting $W_x[k]$ is defined as follows:

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

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

Note also that the resulting detectability $D(x[n], \varepsilon[n])$ is a convex function of the disturbance signal $\varepsilon[n]$. This can be seen as follows. The frequency-domain representation $\mathcal{E}[k]$ is related to the time-domain representation $\varepsilon[n]$ through the DFT, which is a linear operator. The perceptual weighting of $\mathcal{E}[k]$ performed by $W_x[k]$ is also a linear operation. As such, $W_x[k]\mathcal{E}[k]$ is an affine function of $\varepsilon[n]$. Finally, as the composition of an affine mapping and a convex function is convex [27], the Par detectability distortion is convex in $\varepsilon[n]$.

In order to gain a deeper understanding of the behavior of the perceptual weighting $W_x[k]$, consider Figure 2.1. The figure relates the auditory masking threshold and the corresponding perceptual weighting when a 1000 Hz tone at 70 dB SPL is used as masking signal $x[n]$. The top plot depicts the masking threshold, and the bottom plot depicts the corresponding perceptual weighting.

Recall that the masking threshold is the minimal sound pressure level that is required for an additional stimulus to be audible in the presence of the masking signal [6]. In addition to this, the threshold of hearing is also depicted to highlight the additional masking that occurs due to the masking signal.

As can be seen, the masking threshold peaks at 52 dB SPL at 1000 Hz. This implies that a different tone at the same frequency must be at least 52 dB SPL to be audible.

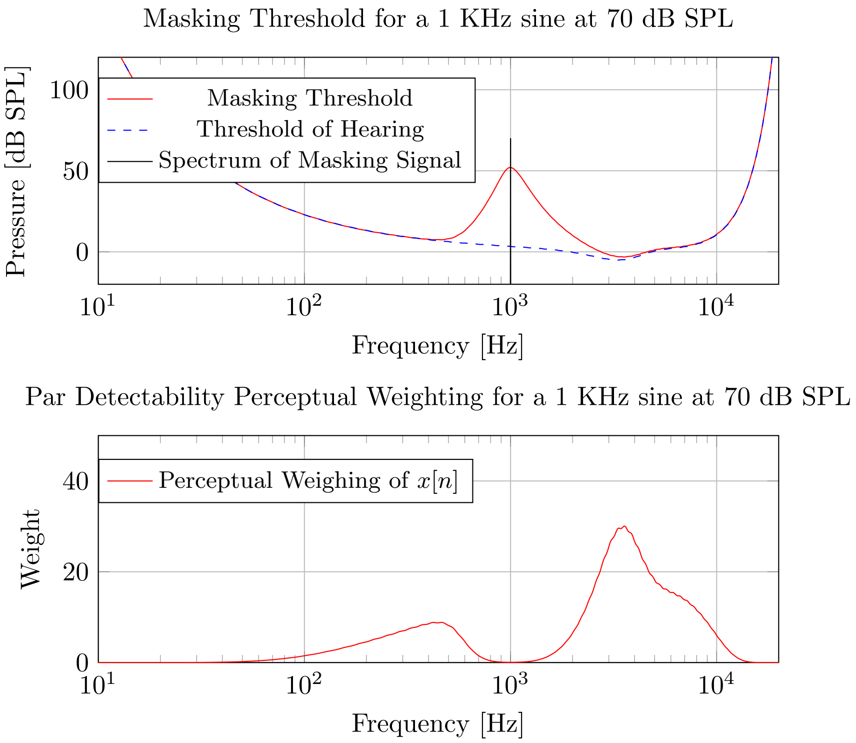


Figure 2.1: Depiction of the masking threshold and corresponding perceptual weighting function for a 1000 Hz tone with an amplitude of 70 dB SPL. The threshold of hearing is also depicted.

As depicted, this results in a low perceptual weighting at 1000 Hz, implying that a disturbance at this frequency is less detectable. Note also that the low and higher frequencies are also weighted lower due to the threshold of hearing. This implies that these frequencies are less important perceptually.

Chapter 3

Perceptual Sound Zone Framework Proposal

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

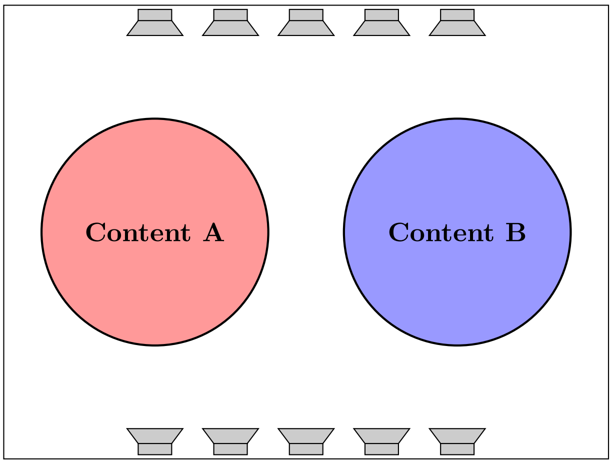


Figure 3.1: A birds-eye view of a room is depicted. The room is divided into two zones: a red zone and a blue zone. Each zone is assigned different content: content A and content B, respectively. In the northern and southern parts of the room, a loudspeaker array is mounted on the walls.

This section seeks to build on this description to provide the understanding necessary for the rest of this work.

Controlling the spatial distribution of sound is done by calculating the audio the loudspeakers must produce to approximate the desired sound field in the given space. The space inside the enclosure is divided up into multiple zones. Each zone is assigned target sound pressure that we would like to have reproduced inside of it. This target sound pressure could be any audio content, for example, music, the sound of a movie, or speech.

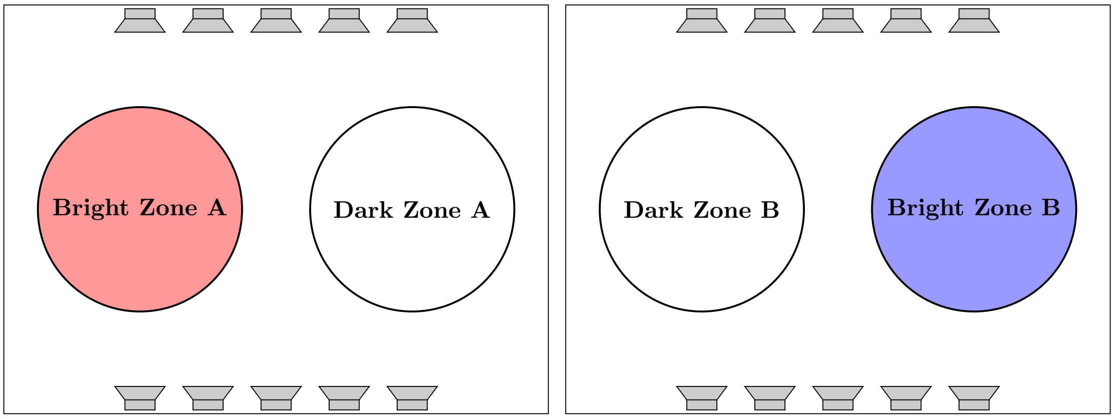


Figure 3.2: A birds-eye view of a room is given twice, each depicting two different sound zone problems. Combining the solutions to both subproblems results in a reproduction of sound with minimal interference between zones.

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

There are various approaches to solving the sound zone problem. Sound zone problems are typically decomposed into a separate subproblem for every zone. Each one of these subproblems considers only two zones: one bright zone and one dark zone. The goal of each subproblem is to reproduce a specified target sound pressure in the bright zone while restricting the sound pressure in the dark zones. The combination of both subproblems provides a solution to the sound zone problem.

To ease the understanding of this concept, consider an example of this decomposition is given in Figure 3.2. Here, a decomposition of the example given in Figure 3.1 into two bright-dark zone pairs.

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

The goal of the rest of this chapter is to motivate the proposal of a perceptual sound zone framework that makes use of the Par distortion detectability introduced in Chapter 2 in the construction of sound zones. This framework is then used to propose perceptual sound zone algorithms in Chapter 4.

- This chapter begins in Section 3.1 with the presentation of a mathematical model that can be used to describe the sound zone problem.
- This mathematical model is then used in Section 3.2 to describe the two main sound zone approaches, "pressure matching" and "acoustic contrast control".

- Finally, Section 3.3 motivates the proposed perceptual sound zone framework inspired by the pressure matching approach previously discussed.

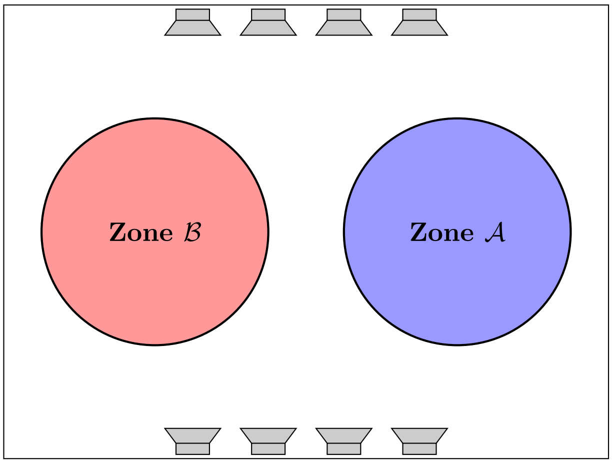


Figure 3.3: A birds-eye view of a room $\mathcal{R} \subset \mathbb{R}^3$ containing the zones $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{R}$ depicted in red and blue respectively. The room contains $N_L = 8$ loudspeakers, which are denoted by the red dots in the corners of the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

3.1 Sound Zone Problem Data Model

In the previous section, a high-level description of the sound zone problem is given. In this section, a mathematical framework for a room containing sound zones will be introduced. This framework will be used later in the description of the sound zone algorithms in Section 3.2.

The contents of this section are as follows. First, Section 3.1.1 develops a spatial description of a room containing two zones and a loudspeaker array. Then, Section 3.1.2 defines the objective of the sound zone algorithm formally as realizing a desired target sound pressure at discrete points in the room. Finally, Section 3.1.3 discusses a suitable target sound pressure, which is used in the remainder of this thesis.

3.1.1 Room Topology

A room $\mathcal{R}$ can be modeled as a closed subset of three dimensional space, $\mathcal{R} \subset \mathbb{R}^3$. The two non-overlapping zones $\mathcal{A}$ and $\mathcal{B}$ are contained within the room $\mathcal{R}$, i.e. $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{R}$ where $\mathcal{A} \cap \mathcal{B} = \emptyset$. That is, there is no intersection between zones.

In general, the room can contain any number of zones; however, this thesis focuses on the two-zone case without loss of generality. In addition to the zones, the room $\mathcal{R}$ also contains N_L loudspeakers, which are modeled as point sources. An example of a possible room, loudspeakers, and pair of zones are visualized in Figure 3.3.

The sound zone algorithm aims to use the sound pressure generated by the loudspeakers to realize 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. Thus, allowing for multiple distinct audio experiences

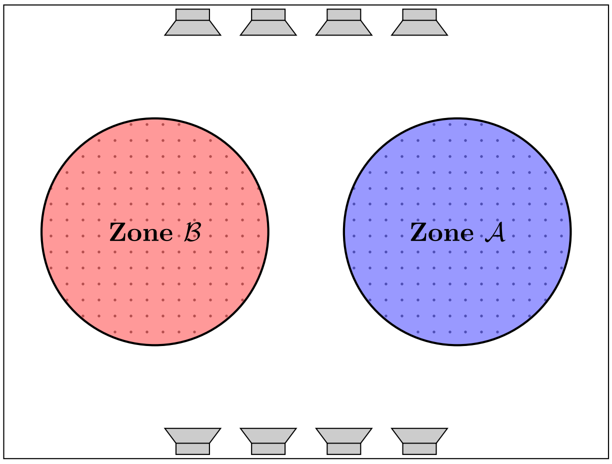


Figure 3.4: The previously introduced room $\mathcal{R}$ with zones $\mathcal{A}$ and $\mathcal{B}$ discretized.

in the room.

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

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

3.1.2 Defining Target and Achieved Pressure

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

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

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

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

To do so, a relationship must be established between the loudspeaker inputs $x^{(l)}[n]$ and the achieved sound pressure at control points $m \in A \cup B$. This relationship can be established by using a linear model based on room impulse responses (RIRs) $h^{(l,m)}[n] \in \mathbb{R}^{N_h}$ [29].

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 [1]:

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

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

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

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

3.1.3 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 to properly define the target pressure is given and motivated.

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

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

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

The motivation for choosing this target is that it is physically attainable in each zone separately with the given loudspeakers, their positions, and the room acoustics.

3.2 Review of Sound Zone Approaches

The two main approaches in sound zone literature are “pressure matching” (PM) and “acoustic contrast control” (ACC). Pressure matching is used as the main inspiration for the perceptual sound zone framework proposed in Section 3.3, which is in turn used to propose perceptual sound zone algorithms in Chapter 4. A description of acoustic contrast control is included for completeness. This section introduces and describes both approaches using the previously derived data model to sketch their mathematical properties.

Classically, the sound zone problem is divided up into subproblems as described in the introduction of this chapter. The resulting loudspeaker input signals $x^{(l)}[n]$ are determined for a single bright-dark zone pair: the loudspeaker input signals are found such that the target audio is achieved in the bright zone, while leakage is minimized in the dark zone. If a solution for multiple zones is desired, multiple problems must be solved independently and their resulting loudspeaker input signals combined [1].

There is another approach, however. In a multi-zone approach, the loudspeaker input signals are instead determined jointly for all zones, rather than decomposing into bright-dark zone pairs.

A multi-zone approach is taken in this thesis, as it is found to be more general. For simplicity, but without loss of generality, this thesis limits the number of zones to two. The approach is, however, generalizable to any multiplicity of zones.

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

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

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

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

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

where $\mathcal{Z} \in (\mathcal{A}, \mathcal{B})$ represents either zones. Here, $p_{\mathcal{A}}^{(m)}[n]$ and $p_{\mathcal{B}}^{(m)}[n]$ can be understood to be the achieved sound pressure that arises due to playing loudspeaker input signals $x_{\mathcal{A}}^{(l)}[n]$ and $x_{\mathcal{B}}^{(l)}[n]$ respectively.

The total achieved sound pressure at control point m is then given by the addition of the two achieved sound pressures:

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

This decomposition is used to describe a multi-zone variant of both a pressure matching approach in Section 3.2.1, and acoustic contrast control approach in Section 3.2.2.

3.2.1 Description of Pressure Matching

The "pressure matching" (PM) approach is widely used in the literature to solve the sound zone problem [1, 2]. In this section, a multi-zone pressure matching algorithm is derived using the data model given in Section 3.1.

In pressure matching approaches, one attempts to design suitable loudspeaker input signals in such a way that the resulting sound pressure in the zone matches the specified target sound pressure for that zone. Simultaneously the sound pressure that results in other zones as interference or cross-talk due to reproducing the target in the bright zone is minimized [1, 30].

This goal can be stated formally as choosing $x_{\mathcal{A}}^{(l)}[n]$ such that the resulting achieved pressure $p_{\mathcal{A}}^{(m)}[n]$ attains the target sound pressure $t^{(m)}[n]$ in all control points $m \in A$. At the same time, however, $p_{\mathcal{A}}^{(m)}[n]$ should result in minimal sound pressure in all control points $m \in B$. Any sound pressure resulting from $x_{\mathcal{A}}^{(l)}[n]$ in zone $\mathcal{B}$ can be understood as leakage or cross-talk between the zones. Similar arguments can be given for $x_{\mathcal{B}}^{(l)}[n]$.

An optimization problem that achieves this goal is formulated as follows:

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

where the $||\cdot||_2^2$ operator denotes the squared L^2-norm. In this case, the norm is taken over the time-domain sequence indexed by n.

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

$$\mathrm{RE}_{\mathcal{Z}}^{(m)} = \left|\left|p_{\mathcal{Z}}^{(m)}[n] - t^{(m)}[n]\right|\right|_2^2 \qquad \forall\, m \in Z, \tag{3.8}$$

$$\mathrm{LE}_{\mathcal{Z}}^{(m)} = \left|\left|p_{\mathcal{Z}}^{(m)}[n]\right|\right|_2^2 \qquad \forall\, m \notin Z. \tag{3.9}$$

With the following interpretations:

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

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

Using these definition we can rewrite the optimization problem:

$$\underset{x_{\mathcal{A}}^{(l)}[n],x_{\mathcal{B}}^{(l)}[n]\,\forall\,l}{\arg\min} \quad \sum_{m\in A}\mathrm{RE}_{\mathcal{A}}^{(m)}+\sum_{m\in B}\mathrm{LE}_{\mathcal{A}}^{(m)}+\sum_{m\in B}\mathrm{RE}_{\mathcal{B}}^{(m)}+\sum_{m\in A}\mathrm{LE}_{\mathcal{B}}^{(m)}. \tag{3.10}$$

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

Some pressure matching approaches attempt to control this trade-off by introducing weights for the different error terms, or by adding constraints. Choosing constraints can, however, be challenging as the squared L^2 pressure error does not always correlate well with how the error is perceived by humans.

3.2.2 Description of Acoustic Contrast Control

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

In this section, a multi-zone ACC algorithm is described. As the previously described data model is in the time domain, this approach will take inspiration from a time-domain approach found in literature known as the broadband acoustic contrast control (BACC) approach [31, 32, 2].

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

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

$$x_{\mathcal{Z}}^{(l)}[n] = \left(w_{\mathcal{Z}}^{(l)} * s_{\mathcal{Z}}\right)[n]. \tag{3.11}$$

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

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

$$\mathrm{AC}_{\mathcal{Z}} = \frac{\sum_{m\in Z}\left|\left|p_{\mathcal{Z}}^{(m)}[n]\right|\right|_2^2}{\sum_{m\notin Z}\left|\left|p_{\mathcal{Z}}^{(m)}[n]\right|\right|_2^2}. \tag{3.12}$$

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

$$\underset{w_{\mathcal{A}}^{(l)}[n],\, w_{\mathcal{B}}^{(l)}[n]\, \forall\, l}{\arg\max} \quad \mathrm{AC}_{\mathcal{A}} + \mathrm{AC}_{\mathcal{B}}. \tag{3.13}$$

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

Acoustic contrast control is discussed mainly for completion as an alternative to the pressure matching approach. As is discussed in Section 3.3, pressure matching is the main inspiration for the proposed perceptual sound zone approach used in the proposed perceptual sound zone algorithm.

3.3 Proposal of Perceptual Sound Zone Framework

This section proposes and motivates a perceptual sound zone framework that makes use of the "Par distortion detectability" discussed in Chapter 2 to construct sound zones in a manner inspired by the "pressure matching" approach discussed in Section 3.2. This perceptual sound zone framework is used to propose perceptual sound zone algorithms in Chapter 4.

As is motived by Section 2.2, the Par distortion detectability is the perceptual model to be used in the perceptual sound zone algorithm. Recall from Section 2.3 that the detectability $D(x[n], \varepsilon[n])$ quantifies how detectable a disturbance $\varepsilon[n] \in \mathbb{R}^{N_x}$ is in presence of a masking signal $x[n] \in \mathbb{R}^{N_x}$. Note that the Par distortion detectability assumes that the time-scale of its inputs are short, in the order of 20 to 200 ms.

It is noted in Section 2.3.3 that the Par distortion detectability measure is a convex function of the disturbance signal $\varepsilon[n]$ when the masking signal is held constant. As such, one approach is to specify a sound zone algorithm optimizes over this disturbance signal in some way. This is be done by adopting a model for the disturbance $\varepsilon[n]$ and the masking signal $x[n]$.

One natural choice for the disturbance signal are the sound pressure errors from the pressure matching approach.

As discussed in Section 3.2.1, pressure matching constructs sound zones by minimizing the sum of the reproduction error in the bright zone $\mathrm{RE}_{\mathcal{Z}}^{(m)}$ and the leakage to the dark zone $\mathrm{LE}_{\mathcal{Z}}^{(m)}$. The original definitions of these equations are given by Equations (3.8) and (3.9). Their definition is repeated for the convenience of the reader:

$$\mathrm{RE}_{\mathcal{Z}}^{(m)} = \left|\left|p_{\mathcal{Z}}^{(m)}[n] - t^{(m)}[n]\right|\right|_2^2 \qquad \forall\ m \in Z, \tag{3.14}$$

$$\mathrm{LE}_{\mathcal{Z}}^{(m)} = \left|\left|p_{\mathcal{Z}}^{(m)}[n]\right|\right|_2^2 \qquad \forall\ m \notin Z. \tag{3.15}$$

Consider modeling the errors from the pressure matching approach as the disturbances $\varepsilon[n]$. To this end, define the reproduction error detectability $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ and the leakage error detectability $\mathrm{LED}_{\mathcal{Z}}^{(m)}$:

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

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

The reproduction error detectability and the leakage error detectability are building blocks that form a framework with which perceptual sound zone algorithms can be created. In these definitions, both the masking signal $x[n]$ and the disturbance signal $\varepsilon[n]$ of the disturbance detectability $D(x[n], \varepsilon[n])$ are modeled:

- The reproduction error detectability $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ models the distortion signal as the reproduction error, which is defined as the the deviation of the achieved sound pressure in the bright zone from the target sound pressure, i.e $p_{\mathcal{Z}}^{(m)}[n] - t^{(m)}[n]$.

As such, the reproduction error detectability can be understood as the detectability of the reproduction error in the presence of the target sound pressure for that control point m.

- The leakage error detectability $\mathrm{LED}_{\mathcal{Z}}^{(m)}$ models the distortion signals as the leakage error, which is defined as the achieved sound pressure in the dark zone, i.e., $p_{\mathcal{Z}}^{(m)}[n]$.

  The leakage error detectability can thus be understood as the detectability of the achieved dark zone pressure, or interference, in the presence of the target sound pressure for that control point m.

- For both reproduction error detectability $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ and the leakage error detectability $\mathrm{LED}_{\mathcal{Z}}^{(m)}$ the masking signal $x[n]$ is modeled as the target sound pressure for the given control point m, i.e $t^{(m)}[n]$.

  As a result, the masking properties of the target sound pressures are used for both the reproduction error detectability and the leakage error detectability.

The expected behavior of minimizing the reproduction error detectability and the leakage error detectability is that the reconstruction and leakage errors are shaped in such a way that they are masked to a degree by the target sound pressure. As a result, the errors should become minimally detectable.

Note that using the target sound pressure as a masking signal is an approximation: Generally, it cannot be assumed that the achieved sound algorithm exactly matches the target sound pressure perfectly, as the target is not always attainable for the given room, zones, and set of loudspeakers. In the ideal case, the masking properties of the total achieved sound pressure would be used instead. However, this quantity depends on the optimizer. Adopting the total sound pressure in place of the target sound pressure for the masking signal results in a non-convex problem. As stated in Section 2.3.3, the detectability is only convex if the masking signal is constant. As such, the masking effects of the achieved pressure are approximated by those of the target sound pressure.

Note also that the detectability is proposed to operate on short time-frequency domain segments with a time resolution of 20 to 200 milliseconds. As such, the existing data model and pressure matching approach must be changed to operate in a short time-frequency domain fashion. This is done in Section 4.1.

This framework of error detectabilities is found to be a promising and natural way of creating sound zone algorithms directly through the Par disturbance detectability and is used to state two perceptual sound zone algorithms in Chapter 4.

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

Chapter 4

Perceptual Sound Zone Algorithm Proposal and Implementation

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

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

Chapter Structure

This chapter is structured as follows.

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

4.1 Proposal of Short-Time Frequency-Domain Pressure Matching

In Section 3.3 a perceptual sound zone framework is proposed based on the pressure matching approach discussed in Section 3.2. In this framework, the Par detectability measure quantifies the perceptual cost of sound pressure errors. In doing so, Section 3.2 introduces the concepts of reproduction error detectability $\text{RED}_{\mathcal{Z}}^{(m)}$ and the leakage error detectability $\text{LED}_{\mathcal{Z}}^{(m)}$ per control point m.

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

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

4.1.1 Short-Time Pressure Matching

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

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

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

In order to formulate their short-time representations, $s_{\mathcal{Z}}[n]$ and $x_{\mathcal{Z}}^{(l)}[n]$ are split up into multiple overlapping blocks by using shifted windows $w[n - kH]$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}\tilde{t}^{(m)}[n, \mu] &= \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * \tilde{s}_{\mathcal{Z}}[\mu]\right)[n] \\ &= \sum_{l=0}^{N_L-1} \sum_{k=-\infty}^{\mu} \left(h^{(l,m)} * \tilde{s}_{\mathcal{Z},k} w_k\right)[n] \\ &= \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * \tilde{s}_{\mathcal{Z},\mu} w_\mu\right)[n] + \tilde{t}^{(m)}[n, \mu - 1].\end{aligned} \tag{4.7}$$

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

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

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

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

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

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

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

It is shown in Equation (4.8) and Equation (4.7) that all quantities can be computed recursively.

This is used in the block-based pressure matching approach by computing the blocks of the loudspeaker input signal $x_{\mathcal{Z}}^{(l)}[n]$ one by one. As such, the k^{th} loudspeaker input signal coefficient $\tilde{x}_{\mathcal{Z},\mu}^{(l)}[n]$ is computed such that the resulting resulting sound pressure $\tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu]$ best matches the target sound pressure $\tilde{t}^{(m)}[n,\mu]$.

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

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

$$\begin{aligned}\underset{\tilde{x}_{\mathcal{A},\mu}^{(l)}[n],\,\tilde{x}_{\mathcal{B},\mu}^{(l)}[n]\,\forall\, l}{\arg\min} \quad & \sum_{m\in A}\left|\left|\tilde{p}_{\mathcal{A}}^{(m)}[n,\mu]-\tilde{t}_{\mu}^{(m)}[n,\mu]\right|\right|_2^2+\sum_{m\in A}\left|\left|\tilde{p}_{\mathcal{B}}^{(m)}[n,\mu]\right|\right|_2^2+\\ & \sum_{m\in B}\left|\left|\tilde{p}_{\mathcal{B}}^{(m)}[n,\mu]-\tilde{t}_{\mu}^{(m)}[n,\mu]\right|\right|_2^2+\sum_{m\in B}\left|\left|\tilde{p}_{\mathcal{A}}^{(m)}[n,\mu]\right|\right|_2^2.\end{aligned} \tag{4.9}$$

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

This is beneficial, as due to overlap, this allows block μ to potentially improve the results of previous blocks. However, the loudspeaker input signals of the current block μ can only affect so many previous blocks (depending on the overlap). As such, to reduce the complexity of the optimization without affecting the results, one may choose to truncate the number of previous blocks $-\infty \leq k \leq \mu-1$ in the history of the optimization.

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

4.1.2 Short-Time Frequency-Domain Pressure Matching

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

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

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

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

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

$$X[k] = \sum_{n=0}^{N_w-1} x[n] \exp\left(\frac{-j2\pi kn}{N_w + N_h - 1}\right). \tag{4.10}$$

Converting the previously introduced block-based pressure matching to a frequency domain equivalent version essentially involves converting the sound pressures $\tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu]$ and $\tilde{t}^{(m)}[n,\mu]$ to their frequency domain counterparts, which are denoted by $\tilde{P}_{\mathcal{Z},\mu}^{(m)}[k] \in \mathbb{C}^{N_w+N_h-1}$ and $\tilde{T}_{\mu}^{(m)}[k] \in \mathbb{C}^{N_w+N_h-1}$ respectively.

This is done as follows:

$$\tilde{T}^{(m)}[k,\mu] = \sum_{l=0}^{N_L} H^{(l,m)}[k]\tilde{S}_{\mathcal{Z},\mu}[k], \tag{4.11}$$

$$\tilde{P}_{\mathcal{Z}}^{(m)}[k,\mu] = \sum_{l=0}^{N_L} H^{(l,m)}[k]\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k]. \tag{4.12}$$

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 signals, which are defined as follows:

$$\tilde{S}_{\mathcal{Z},\mu}[k] = \sum_{n=\mu H-N_w+1}^{\mu H} \left[\sum_{k=-\infty}^{\mu} \tilde{s}_{\mathcal{Z},\mu}[n]w[n-kH]\right] \exp\left(\tfrac{-j2\pi k(n-\mu H+N_w-1)}{N_w+N_h-1}\right), \tag{4.13}$$

$$\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k] = \sum_{n=\mu H-N_w+1}^{\mu H} \left[\sum_{k=-\infty}^{\mu} \tilde{x}_{\mathcal{Z},\mu}^{(l)}[n]w[n-kH]\right] \exp\left(\tfrac{-j2\pi k(n-\mu H+N_w-1)}{N_w+N_h-1}\right). \tag{4.14}$$

This definition takes the short-time Fourier transformation of block μ of all contributions to the desired playback signal and the loudspeaker input signals up to and including block μ. As such, the history formed by the previous blocks is also taken into account. Note that the window is implicitly included in the transformed quantities. This is done for ease of notation.

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

$$\begin{aligned}\underset{\tilde{x}^{(l)}_{\mathcal{A},\mu}[n],\, \tilde{x}^{(l)}_{\mathcal{B},\mu}[n]\, \forall\, l}{\arg\min} \quad & \sum_{m\in A} \left|\left| \tilde{P}^{(m)}_{\mathcal{A}}[k,\mu] - \tilde{T}^{(m)}[k,\mu] \right|\right|_2^2 + \sum_{m\in A} \left|\left| \tilde{P}^{(m)}_{\mathcal{B}}[k,\mu] \right|\right|_2^2 + \\ & \sum_{m\in B} \left|\left| \tilde{P}^{(m)}_{\mathcal{B}}[k,\mu] - \tilde{T}^{(m)}[k,\mu] \right|\right|_2^2 + \sum_{m\in B} \left|\left| \tilde{P}^{(m)}_{\mathcal{A}}[k,\mu] \right|\right|_2^2 \end{aligned} \tag{4.15}$$

Note also how the optimization is still performed over the time domain loudspeaker input signals $\tilde{x}^{(l)}_{\mathcal{A},\mu}[n]$ and $\tilde{x}^{(l)}_{\mathcal{B},\mu}[n]$. 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 principle, this introduces more complexity than solving directly over the frequency domain loudspeaker input coefficient $\tilde{X}^{(l)}_{\mathcal{Z},\mu}[k]$. This, however, introduces issues as it requires the truncation of the time-domain version to the first N_w samples.

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

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

4.2 Proposal of Perceptual Pressure Matching Algorithms

Previously, Section 3.3 noted that the pressure matching approach could be formulated using the detectability. In this section, this approach is used to propose two sound zone algorithms.

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

In Section 3.3 the description of reproduction error detectability $\mathrm{RED}_{\mathcal{Z}}^{(m)}$ and the leakage error detectability $\mathrm{LED}_{\mathcal{Z}}^{(m)}$ is given in terms of full-length input sequences. As noted, this is inaccurate as the detectability operates on a short-time scale and is only done this way to convey the concept.

Henceforth, using the concepts introduced in Section 4.1, let $\mathrm{RED}_{\mathcal{Z}}^{(m)}[\mu]$ and $\mathrm{LED}_{\mathcal{Z}}^{(m)}[\mu]$ denote the reproduction error detectability and leakage error detectability corresponding to block μ in control point m.

To define these error detectability quantities, recall that detectability is defined as follows:

$$D(x[n], \varepsilon[n]) = ||W_x[k]\mathcal{E}[k]||_2^2 \tag{4.16}$$

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

$$\begin{aligned}\mathrm{RED}_{\mathcal{Z}}^{(m)}[\mu] &= D(\tilde{t}^{(m)}[n,\mu], \tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu] - \tilde{t}^{(m)}[n,\mu]) \\ &= \left|\left|W_{\tilde{t}^{(m)}[\mu]}[k]\left(\tilde{P}_{\mathcal{Z}}^{(m)}[k,\mu] - \tilde{T}^{(m)}[k,\mu]\right)\right|\right|_2^2,\end{aligned} \tag{4.17}$$

$$\begin{aligned}\mathrm{LED}_{\mathcal{Z}}^{(m)}[\mu] &= D(\tilde{t}^{(m)}[n,\mu], \tilde{p}_{\mathcal{Z}}^{(m)}[n,\mu]) \\ &= \left|\left|W_{\tilde{t}^{(m)}[\mu]}[k]\left(\tilde{P}_{\mathcal{Z}}^{(m)}[k,\mu]\right)\right|\right|_2^2.\end{aligned} \tag{4.18}$$

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

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

4.2.1 Proposal of Unconstrained Perceptual Pressure Matching

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

Consider the following optimization problem:

$$\begin{aligned} \underset{\tilde{x}_{\mathcal{A}}^{(l)}[n,\mu],\, \tilde{x}_{\mathcal{B}}^{(l)}[n,\mu]\,\forall\, l}{\arg\min} \quad & \sum_{m\in A} \mathrm{RED}_{\mathcal{A}}^{(m)}[\mu] + \sum_{m\in B} \mathrm{LED}_{\mathcal{A}}^{(m)}[\mu] + \\ & \sum_{m\in B} \mathrm{RED}_{\mathcal{B}}^{(m)}[\mu] + \sum_{m\in A} \mathrm{LED}_{\mathcal{B}}^{(m)}[\mu]. \end{aligned} \tag{4.19}$$

The total detectability of the reproduction errors and the leakage errors is minimized by optimizing over the block-based representations of the loudspeaker input signals $\tilde{x}_{\mathcal{A}}^{(l)}[n,\mu]$ and $\tilde{x}_{\mathcal{B}}^{(l)}[n,\mu]$. The expected behavior of this optimization problem is that the sound pressure errors will be shaped in such a way that they are masked by the target sound pressure.

4.2.2 Proposal of Constrained Perceptual Pressure Matching

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

The motivation for this approach is that it is hypothesized that the Par distortion detectability has a consistent perceptual interpretation. As mentioned in Section 2.3, a detectability of 1 will consistently imply “just noticeable” [7].

This makes choosing constraints for detectability easier than typical non-perceptual pressure matching approaches. These approaches typically directly constrain the sound pressure error, for which it is difficult to determine constraints [33] as the sound pressure error does not have a consistent perceptual interpretation. As a result, a sound pressure error constraint can lead to widely varying results perceptually.

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

$$\begin{aligned} \underset{\tilde{x}_{\mathcal{A}}^{(l)}[n,\mu],\, \tilde{x}_{\mathcal{B}}^{(l)}[n,\mu]\,\forall\, l}{\arg\min} \quad & \sum_{m\in B} \mathrm{LED}_{\mathcal{A}}^{(m)}[\mu] + \sum_{m\in A} \mathrm{LED}_{\mathcal{B}}^{(m)}[\mu], \\ \text{subject to} \quad & \mathrm{RED}_{\mathcal{A}}^{(m)}[\mu] \leq D_0 \quad \forall\, m \in A \\ & \mathrm{RED}_{\mathcal{B}}^{(m)}[\mu] \leq D_0 \quad \forall\, m \in B. \end{aligned} \tag{4.20}$$

Here, D_0 is the maximum allowed detectability of the reproduction error per control point m. The intended effect of this constraint is limiting how detectable the deviation of the achieved sound pressure is from the specified target. Effectively, this allows for controlling the quality of the achieved sound pressure per control point m. To the knowledge of the authors, this is a novelty.

It is also hypothesized that the constraint D_0 allows for more control over the tradeoff between reproduction error detectability and leakage error detectability.

It should be noted that constraining the leakage error detectability is also a good choice. In doing so the algorithm is limited in the amount of interference that it allows. A possible optimization problem that achieves this is defined as follows:

$$\begin{aligned} \underset{\tilde{x}_{\mathcal{A}}^{(l)}[n,\mu],\, \tilde{x}_{\mathcal{B}}^{(l)}[n,\mu]\,\forall\, l}{\arg\min} \quad & \sum_{m\in B} \mathrm{RED}_{\mathcal{A}}^{(m)}[\mu] + \sum_{m\in A} \mathrm{RED}_{\mathcal{B}}^{(m)}[\mu], \\ \text{subject to} \quad & \mathrm{LED}_{\mathcal{B}}^{(m)}[\mu] \le D_0 \quad \forall\, m \in A \\ & \mathrm{LED}_{\mathcal{A}}^{(m)}[\mu] \le D_0 \quad \forall\, m \in B. \end{aligned} \tag{4.21}$$

This problem is not pursued any further in this thesis. Constraining the detectability of the leakage error is, however, found to be promising future work, which is discussed further in Chapter 6.

Chapter 5

Evaluation of Proposed Perceptual Sound Zone Algorithms

In the Chapter 4, two perceptual sound zone algorithms are proposed based on the proposed perceptual sound zone framework given in Section 3.3. This chapter performs an evaluation of the proposed perceptual algorithms in order to determine the benefits, if any, of including perceptual information in the proposed fashion.

This is done by comparing the proposed algorithms with a reference algorithm.

Chapter Structure

In order to achieve this goal, this chapter is structured as follows:

- This chapter begins with a description of the evaluation methodology in Section 5.1, describing the experiments and the measures used in the subsequent evaluation.
- Section 5.2 then presents and discusses the results of the described experiments. The proposed algorithms are compared to a reference algorithm through various perceptual measures in order to determine their relative performance.

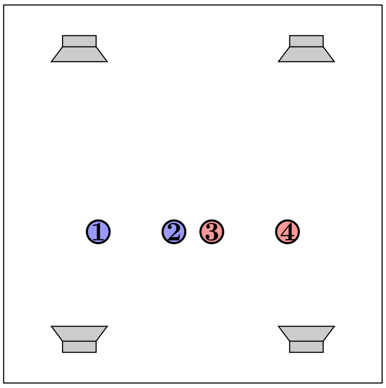


Figure 5.1: A birds-eye view of a square 5 by 5 meter room used in the simulations for the evaluation of the algorithms. Two zones each consisting of two control points are depicted in blue and red. The control points are numbered for ease of reference. The four loudspeakers are placed in the corners of the room.

5.1 Evaluation Methodology

This section discusses the general approach for evaluation of the perceptual sound zone algorithms proposed in Section 4.2.1 and Section 4.2.2. All algorithms will be evaluated on the basis of simulations. As all derived algorithms are computationally intensive, many of the parameter considerations for the simulations are chosen such that the computational load is kept feasible.

For the evaluation, the constrained- and unconstrained perceptual pressure matching approaches are compared to the short-time frequency-domain pressure matching approach given in Section 4.1 by Equation (4.15).

5.1.1 Simulation Configuration

All sound zone algorithms are evaluated in a simulated square room of 5 by 5 meters, with a ceiling height of 3.4 meters. There are two zones, each consisting of two control points. In order to obtain a sufficiently challenging problem, the zones are placed in close proximity to each other, with the two closest points being 0.5 meters apart. This is important, as a trivial problem makes it difficult to highlight the differences in performance between the perceptual and non-perceptual approaches.

The room contains four loudspeakers placed in the corners of the room at the height of 1.2 meters. Omnidirectional loudspeakers, which radiate energy equally in all directions [29], are used. An image depicting the entire setup is given in Figure 5.1.

In order to synthesize room impulse responses used to relate the loudspeaker inputs and the resulting sound pressure in the control points, the image source method [34] is used, specifically, the Habets implementation [29].

For reasons of computational complexity, the simulated room impulse responses are limited to reverberation times of at most 200 milliseconds. The definition of the reverberation time is the time required for the intensity of reflections to reduce 60 dB relative to the direct path sound pressure [29]. To put the reverberation time used in the experiment into context, it has been found in an investigation into the reverberation time of furnished rooms that similarly sized rooms have an average maximal reverberation time of 720 milliseconds [35].

To define the target sound pressure, content must be selected for the two zones. For this evaluation, English speech content is used. The motivation for this is that the objective speech quality measures described in Chapter 2 are found to be more robust than the discussed general audio quality measures. Another practical reason is that many general audio quality measures have little to no free and openly available implementations, whereas the speech measures do.

The speech signals are downsampled to 8000 Hz. The motivation for this is that it is computationally intensive to run the algorithms at a higher sampling rate. Another motivation is that the majority of the speech energy is contained between 150 and 4000 Hz [36].

In total, a dataset of 4 loudness-matched speech signals is used for evaluation. For each experiment, one speech signal is assigned to each zone. All possible combinations of the speech signals are formed, resulting in a total of 12 possible configurations. However, due to the symmetry of the loudspeaker, room, and zones given by Figure 5.1 this simplifies to a total of 6 configurations.

5.1.2 Simulation Evaluation Measures

This section introduces the measures that are used to evaluate the simulation results of the reference and perceptual algorithms.

The perceptual measures PESQ, STOI, SIIB, and Distraction from the literature review discussed in Section 2.1 is used for the evaluation. In addition to these perceptual measures, two traditional physical measures, namely the "normalized mean square error" (NMSE) and the "acoustic contrast" (AC) in terms of sound pressure is used in the evaluation of sound zone algorithms [37] are also used.

The motivation for including physical measures is to test the hypothesis that, while typically outperforming in terms of perceptual measures, the perceptual sound zone algorithms likely do not outperform the reference sound zone algorithm in terms of physical measures. The hypothesis for this is that the reference pressure matching approach optimizes over a physical error measure.

A brief summary is given of the outputs and inputs of the sound zone algorithm is given below. These serve as the inputs to the evaluation measures.

- **Target Sound Pressure $t^{(m)}$:**
  The desired sound pressure per control point m, defined by the speech signals corresponding to the zone $\mathcal{Z}$ the control point is in.

- **Achieved Sound Pressure $p_{\mathcal{Z}}^{(m)}$:**
  The sound pressure is achieved by the algorithm at control point m for the

zone $\mathcal{Z}$. This sound pressure has two different interpretations, depending on the control point under consideration.

  - **Achieved Bright Zone Sound Pressure for Zone $\mathcal{Z}$:**
    When $m \in Z$, the achieved sound pressure is the achieved approximation of the target sound pressure $t^{(m)}$ for control point m. This can be understood as the approximation of the target sound pressure for a control point m, sans leakage or interference.
  - **Achieved Dark Zone Sound Pressure for Zone $\mathcal{Z}$:**
    When $m \notin Z$, the achieved sound pressure represents the sound pressure that arises in other zones due to reproducing the bright zone for zone $\mathcal{Z}$. It can be understood as the leakage or interference in other zones due to reproducing zone $\mathcal{Z}$.

- **Total Achieved Sound Pressure $\sum_{\mathcal{Z}} p_{\mathcal{Z}}^{(m)}$:**
  The sound pressure in a control point m due to contributions of all zones $\mathcal{Z}$. This represents the sound pressure that a user of the sound zone system would experience, and thus it contains both the approximation of the target sound pressure and the interference.

For more information on these quantities, the reader is referred to Section 3.1, where they are introduced. What follows is a description of various categories of evaluation measures.

5.1.2.1 Perceptual Quality Measures

The first category of measures that are discussed are perceptual measures that estimate the perceived quality of speech audio.

One of the metrics that is used is the Perceptual Evaluation of Speech Quality (PESQ). As described in Section 2.1.1, PESQ is a metric that grades the quality of a degraded speech signal with respect to a reference speech signal. The resulting quality grade will be between 0 and 5, where 5 is the highest obtainable grade.

Another set of metrics that will be used are two speech intelligibility metrics: the Short-Time Objective Intelligibility (STOI) and the Speech Intelligibility in Bits (SIIB). STOI provides an intelligibility score between 0 and 1, where 1 is the highest score. SIIB instead scores the intelligibility with an information rate given in bits/s, lower-bounded by 0 bits/s. Maximum intelligibility corresponds to a rate of about 150 bits/s [17].

All perceptual measures evaluate the quality of a “degraded” input stimuli with respect to a ”reference” stimuli. Using the previously introduced quantities, the measures are used to evaluate sound zone performance as follows:

- **PESQ, STOI and SIIB of the Total Achieved Sound Pressure with respect to the Target Sound Pressure:**
  Corresponds to the quality/intelligibility of the achieved sound pressure, including interference, and will be referred to as the **“total PESQ”**, **“total STOI”** and the **“total SIIB”** per control point m.

- **PESQ, STOI and SIIB of the Achieved Bright Zone Sound Pressure with respect to the Target Sound Pressure:**
  Corresponds to the quality/intelligibility of the achieved sound pressure sans interference. This quantity is referred to as the **"bright zone PESQ"**, **"bright zone STOI"** and **"bright zone SIIB"** per control point m for zone $\mathcal{Z}$.

5.1.2.2 Perceptual Interference Measure

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

The distraction will be used as follows:

- **Distraction of achieved Dark Zone Sound Pressure with respect to the Achieved Bright Zone Sound Pressure:**
  This quantifies how distracting the dark zone sound pressures are when listening to the bright zone sound pressure per control point m. This will simply be referred to as the **"Distraction"** per control point m.

5.1.2.3 Physical Measures

One physical measure is the acoustic contrast (AC) between the achieved bright zone sound pressure, and the achieved dark zone sound pressure can be used as a non-perceptual measure of interference. Initially introduced in Section 3.2, the acoustic contrast between two time-domain sequences $x[n] \in \mathbb{R}^N$ and $y[n] \in \mathbb{R}^N$ is given as follows:

$$\mathrm{AC}(x, y) = 10 \log_{10} \left(\frac{||x[n]||_2^2}{||y[n]||_2^2} \right). \tag{5.1}$$

Another non-perceptual metric that can be used to evaluate the quality of the result is the normalized mean square error (NMSE). The NMSE of the deviation of time sequence $x[n]$ from reference time sequence $y[n]$ is given as:

$$\mathrm{NMSE}(x, y) = 10 \log_{10} \left(\frac{||y[n] - x[n]||_2^2}{||y[n]||_2^2} \right). \tag{5.2}$$

Note that both physical quantities are given in decibels. These physical measures are used to evaluate sound zone performance as follows:

- **NMSE of the Total Achieved Sound Pressure with respect to the Target Sound Pressure:**
  Describes the NMSE between the target and achieved sound pressure and will be referred to as the **"total NMSE"**.

- **NMSE of the Achieved Bright Zone Sound Pressure with respect to the Target Sound Pressure:**
  This is the NMSE between the target and the achieved sound pressure sans

interference and will henceforth be referred to as the **“bright zone NMSE”** per control point m.

- **Acoustic contrast between the Achieved Bright Zone Sound Pressure and the Achieved Dark Zone Sound Pressure:**
  This measure quantifies the ratio of the acoustic potential energy of the bright zone and of the dark zone. From here on referred to as the **“Acoustic Contrast”** per control point m.

Measure	Unconstrained Perceptual PM Mean (± 95% CI)	Reference PM Mean (± 95% CI)
Total PESQ	3.154 ± 0.081	2.609 ± 0.084
Bright Zone PESQ	3.345 ± 0.087	4.107 ± 0.051
Total STOI	0.943 ± 0.003	0.940 ± 0.006
Bright Zone STOI	0.950 ± 0.003	0.989 ± 0.001
Total SIIB	1114.306 ± 23.762	893.225 ± 63.815
Bright Zone SIIB	1260.117 ± 14.290	1311.041 ± 12.333
Total NMSE	-4.929 ± 0.235 dB	-13.529 ± 0.856 dB
Bright Zone NMSE	-5.241 ± 0.248 dB	-16.600 ± 0.875 dB
Distraction	7.828 ± 1.868	12.693 ± 3.405
Acoustic Contrast	13.258 ± 0.379 dB	16.075 ± 0.936 dB

Table 5.1: Summary of the results of the evaluation of the unconstrained perceptual pressure matching approach and the reference pressure matching approach using the evaluation metrics defined in Section 5.1.

5.2 Evaluation of Proposed Algorithms

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

This section will evaluate the results of the performed experiments for both proposed approaches. To this end, Section 5.2.1 the unconstrained perceptual pressure matching approach is evaluated, and in Section 5.2.2 the constrained perceptual pressure matching approach is evaluated.

In order to effectively describe various points in the room from the simulations, the control points numbering given by Figure 5.1 is used.

5.2.1 Evaluating Unconstrained Perceptual Pressure Matching

In this section, the unconstrained perceptual pressure matching algorithm will be evaluated in accordance to the approach discussed in Section 5.1. This is done by first evaluating the results of the simulations of the various setups quantitatively through the proposed measures. From this, conclusions are drawn, which are then motivated qualitatively by reasoning about algorithm behavior by considering waveforms generated by the investigated algorithms.

Quantitative Analysis of Simulation Results

In order to quantify the performance of the unconstrained perceptual pressure matching approach, the various measures introduced in Section 5.1 are determined for all 12 simulations. The measures are averaged over all simulations and each control point. For comparison purposes, the reference pressure matching approach is simulated and evaluated in an identical fashion.

The results of this experiment is summarized in Table 5.1. This table depicts the mean and 95% confidence interval of the measures taken over all four control points and all six unique experiments. As discussed in Section 2.1, it is important to note that the perceptual measures can only be used as an indication, and further listening tests must be performed to draw any real conclusions.

From the results in the table, the following observations are made:

1. The perceptual approach outperforms the reference in two perceptual measures evaluating the total experience: total PESQ and SIIB attain higher values for the perceptual approach. The perceptual approach also attains a higher value for STOI, however, the values are too close to draw any real conclusions.

   Note that these measures evaluate the total experience, taking into consideration the total sound pressure per control point, including interference. This implies that the perceptual approach may result in an overall better perceptual experience.

2. The perceptual approach outperforms the reference in terms of the perceptual distraction measure, implying that the interference in the perceptual approach may be less distracting

3. The reference approach outperforms the perceptual approach in perceptual measures evaluating the bright-zone quantities: bright zone PESQ, STOI, and SIIB are all higher for the reference approach. Note that these measures are sans interference: they only evaluate how well the achieved sound pressure attains the target, ignoring interference.

   This implies that disregarding interference, the reference approximates the target more effectively perceptually. However, from Item 1 it is known that the total experience results in a better quality of experience.

   This implies that, although the reference algorithm approximates the target better perceptually, the interference that it introduces a sufficient disturbance to be outperformed by the reference.

4. The reference approach outperforms the perceptual approach for all physical measures: total and bright zone NMSE and acoustic contrast. This is to be expected, as the reference approach optimizes the NMSE directly.

   Interestingly, although the total NMSE is lower, the reference is outperformed in terms of all total perceptual measures, as discussed in Item 1.

   In addition to this, the acoustic contrast between intended and interfering sound pressure for the reference approach is over twice as large as the perceptual approach. Nevertheless, the perceptual approach is less distracting according to the distraction model as discussed in Item 2.

   These results imply that NMSE or AC may not be optimal measures for the evaluation of the perceptual experience of sound zones.

In summary, from the observations above, it is concluded that the perceptual sound zone algorithm may outperform the reference sound zone algorithm in terms of perceptual experience. This seems to be due to the perceptually disturbing interference introduced by the reference algorithm. As such, the perceptual algorithm seems to make a better perceptual trade-off between reproduction of the target sound pressure and suppression of the interference than the reference.

This is because, when disregarding noise, the reference algorithm has a better reproduction of the target perceptually. However, when the noise is added, the reference algorithm gets outperformed by the perceptual approach. The distraction ratings also indicate that the noise introduced by the reference algorithm is more distracting.

Analyzing Algorithm Behavior

In the preceding section, the reference and perceptual algorithms are compared quantitatively. Results indicate that the perceptual algorithm outperforms the reference in terms of the total experience. This section considers the behavior of the algorithm in an attempt to explain these results.

To this end, consider Figure 5.2. This figure contains a plot of the waveforms of the target sound pressure and achieved dark-zone sound pressure for both the perceptual and non-perceptual variants of pressure matching for control point $m = 2$ (see: Figure 5.1) from the experiments.

The selected control point is in zone $\mathcal{A}$. As explained in Section 5.1, the achieved dark-zone sound pressure can be understood as the interference due to another zone, in this case, zone $\mathcal{B}$.

Consider the highlighted region for the perceptual algorithm. From this, it can be seen that the magnitude of the interference is correlated to the magnitude of the target sound pressure for zone $\mathcal{A}$. Contrast this to the highlighted region for the reference algorithm, where the interference is at a relatively constant level.

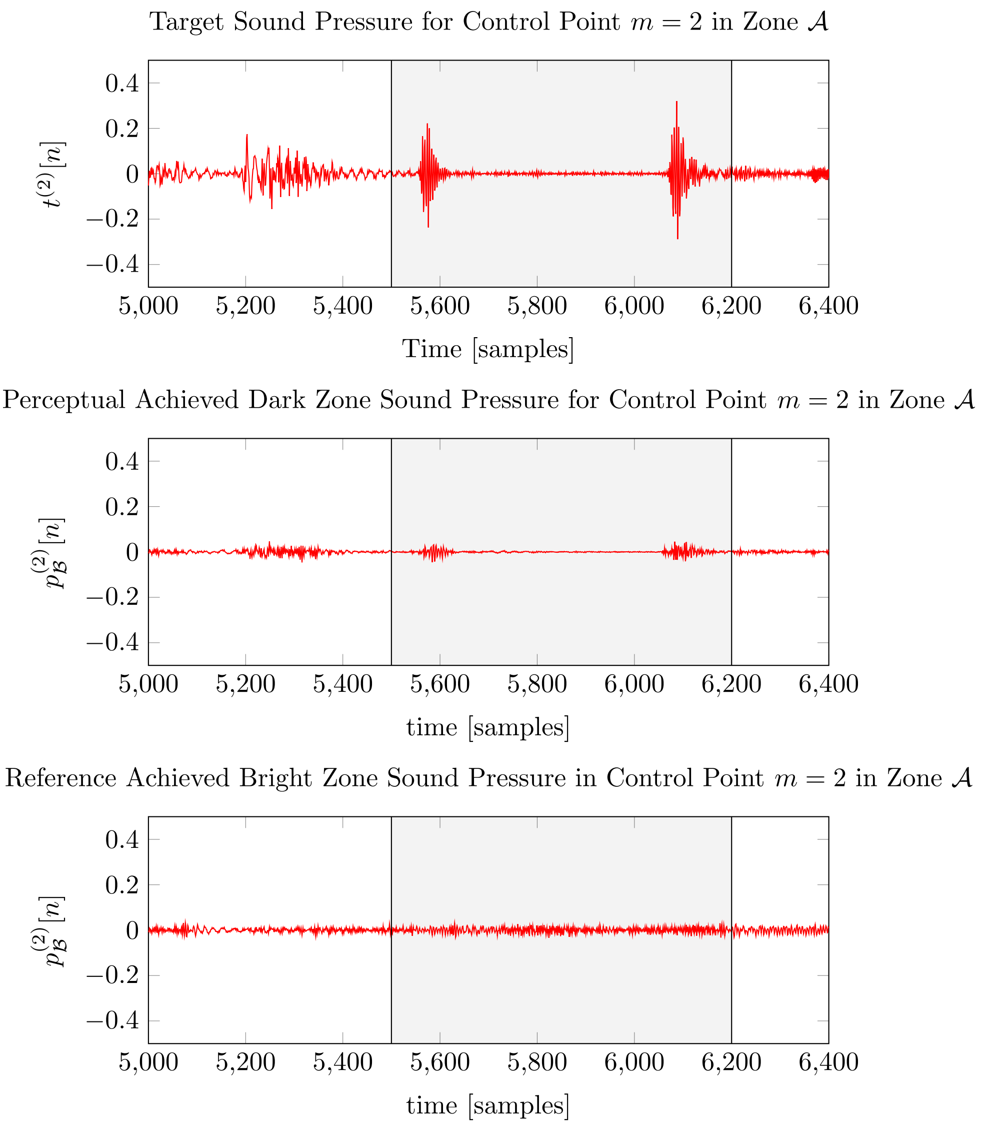
This may explain why, while having lower overall contrast, the perceptual approach outperforms the reference approach in terms of distraction and overall perceptual experience. When determining the interference for control point $m = 2$, the perceptual algorithm takes the target sound pressure for $\mathcal{A}$ into account. Effectively, when the target sound pressure is relatively loud, more interference is allowed as it is masked to a degree by the target sound pressure.

In doing so, the interference is less detectable and thus perceptually less disturbing, which serves as a possible explanation to the results given in Section 5.2.1

5.2.2 Evaluating Constrained Perceptual Pressure Matching

This section details the evaluation of the constrained perceptual pressure matching algorithm discussed in Section 4.2.2. As discussed, this algorithm minimizes the detectability of the leakage error whilst constraining the detectability of the reproduction error.

This leverages the fact that the Par detectability has a perceptual interpretation. That is to say, if two disturbances result in the same detectability, this should mean



Target Sound Pressure for Control Point $m = 2$ in Zone $\mathcal{A}$

Perceptual Achieved Dark Zone Sound Pressure for Control Point $m = 2$ in Zone $\mathcal{A}$

Reference Achieved Bright Zone Sound Pressure in Control Point $m = 2$ in Zone $\mathcal{A}$

Figure 5.2: Depiction of the wave forms of the target sound pressure and achieved dark zone sound pressure for the unconstrained perceptual pressure matching approach and the reference pressure matching approach.

that they are equally detectable perceptually. As such, using the detectability of the reproduction errors in constraints could allow for more precise control over the quality of the reproduced audio.

Traditional pressure matching approaches such as the reference pressure matching algorithm can also attempt to constrain the reproduction error. However, the reference pressure matching algorithm uses the mean square sound pressure error rather than detectability, which does not always correlate well with perception. That is to say; two mean square pressure errors can vary widely perceptually.

This section seeks to explore the degree of control that the reproduction error detectability constraints provide. In this case, no comparison is made to the reference.

Instead, the measures defined in Section 5.1 are used to determine the performance of the constrained perceptual pressure matching approach for varying constraint values. Afterward, a qualitative analysis of the algorithm is given through waveforms in order to motivate the quantitative results.

Quantifying Algorithm Performance

As mentioned, the goal is to analyse the behavior the constrained perceptual pressure matching algorithm introduced in Section 4.2.2 for various values of the constraint D_0 shown in Equation (4.20). To quantify the performance of the algorithm, the measures introduced in Section 5.1 are used.

Consider Figure 5.3, where the results of the experiments for the constrained perceptual pressure matching algorithm are depicted for different values of D_0. The measures as depicted are averaged over all six unique experiments and over each of the four points in the room. The error bars indicate the 95% confidence intervals.

The following observations are made:

1. It can be seen that the bright zone PESQ, NMSE, STOI, and SIIB are all strictly decreasing or increasing as a function of D_0. Recall that the bright zone quantities refer to the reproduction of the intended target sound pressures sans interference. Thus, increasing the constraints correlates with a lowering of the quality of the reproduced target perceptually. This makes sense, as higher D_0 allows for a more detectable reproduction error.

2. It can be seen that the total PESQ, STOI, and SIIB are not strictly decreasing functions of D_0. The total quantities refer to the perceptual quality of all the sound pressure in the control points, including interference. Interestingly, the measures peak at a constraint value D_0 of about 3 or 5.

   This effect is likely due to the interference introducing considerable perceptual distortion, as the achieved sound pressure sans interference is strictly decreasing. Low values of the constraint limit the deviation from the target sound pressure and thus limit how much interference suppression can take place.

   For high values of D_0, the total and bright zone quantities converge to one another. This corresponds to the interference being so small that it is perceptually irrelevant.

3. From the distraction plot, it can be seen that for lower constraint values, the distraction decreases as a function of D_0. At a constraint value of about 15, the distraction starts increasing again.

   The rise in distraction can potentially be explained by a decrease in achieved bright zone sound pressure energy. One way of reducing the amount of interference is by decreasing the bright zone sound energy. As such, the algorithm has an incentive to do so. Increasing constraint values D_0 allows for larger reproduction errors, which allow for a lower-energy representation of the achieved bright zone sound pressure. Thus, the increase in distraction may be due to a decrease in achieved bright zone sound pressure energy without a meaningful decrease of interference.

This effect can also be observed through the mean acoustic contrast, however, due to the size of the error bars are it is difficult to draw a conclusion based on these results.

From the observations above, it is concluded that perceptually constrained pressure matching allows for a degree of control of the perceived quality of the bright zone sound pressure. This can be seen in the small confidence intervals and from the observation in Item 1 where it is shown that the bright zone measures are all a strictly decreasing functions of the constraint value D_0. However, as given by Item 2, the total perceived quality (including interference) is not a strictly decreasing function of D_0, so the constraint must be chosen carefully.

Finally, it is shown in Item 3 that the distraction can be controlled to a degree through the constraint values D_0, as the distraction is a strictly decreasing function of D_0 for low constraint values. Higher constraint values seem to increase the distraction. It is theorized that this is due to there being diminishing returns in increasing the constraint D_0, as the interference is not decreased in a perceptually meaningful way. This can be seen from the convergence of the bright zone and the total perceptual measures as discussed in Item 2. Furthermore, the bright zone sound pressure energy is theorized to be decreased due to the further relaxation of the constraints.

Analyzing Algorithm Behavior

In the previous section, it is hypothesized that the perceptually constrained pressure matching approach allows for accurate control of the perceived sound pressure. This section explores the effects that increasing the value of D_0 has on the waveforms of the bright zone sound pressure and the dark zone sound pressure.

To this end, consider Figure 5.4. This plot depicts wave forms for control point $m = 2$ in zone $\mathcal{A}$ for one of the experiments for different values of the constraint D_0.

As can be seen from the plots on the right-hand side, increasing the constraint value D_0 seems to reduce the interference. This makes sense, as increasing the constraint relaxes the required reproduction error detectability and allows for more interference suppression.

Interestingly, one can see the frequency-weighting that occurs in Par detectability. As discussed in Section 2.3.3, the Par detectability has a low perceptual weighting for lower frequencies due to the threshold of hearing. As can be seen, for $D_0 = 1$, many high frequencies are still present in the achieved dark zone sound pressure. For $D_0 = 21$, only lower frequencies remain.

The achieved bright zone sound pressures on the left-hand side provide evidence for the claim that the achieved bright zone sound pressure decreases with increasing constraints. When compared to the target for that zone, the total energy present seems to decrease greatly between constraint values $D_0 = 1$ and $D_0 = 21$.

5.3 Comparison with Approaches from Literature

In this section, the results obtained in this work are contrasted with other perceptual sound zone approaches from the literature.

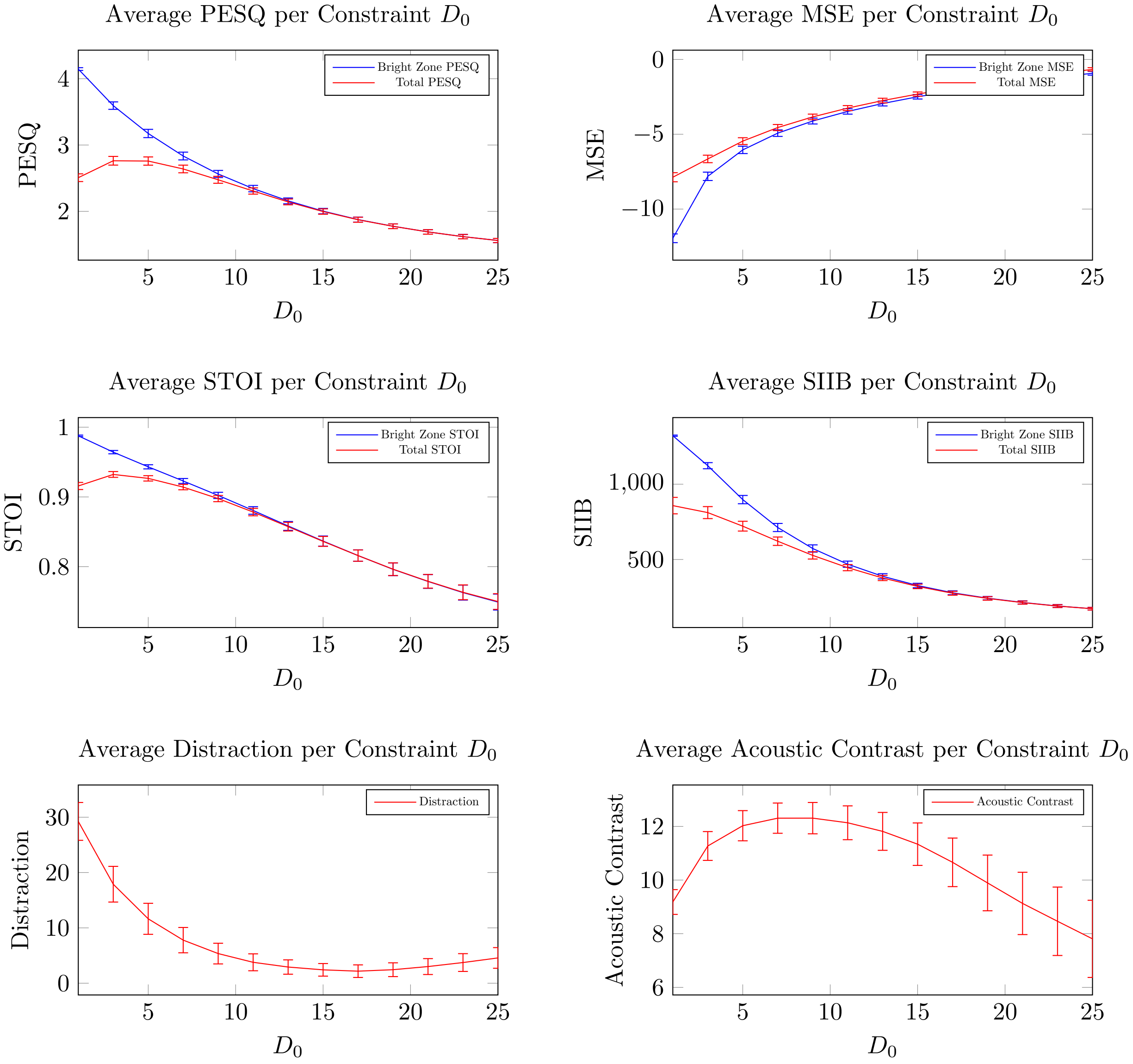
In prior work by Lee et al., the signal-adaptive perceptual variable span trade-off (AP-VAST) perceptual sound zone approach was proposed [4, 5]. Here, the existing variable span trade-off (VAST) sound zone framework is extended with a time-domain perceptual weighting filter. The perceptual weighting is determined by the reciprocal of the masking curves of the target sound pressure [5].

As such, the approach by Lee et al. and the proposed approaches are similar. The approach by Lee et al., however, does not directly optimize over a perceptual model as is done in the proposed approach. Therefore, the perceptual interpretation of the cost function, which enables the perceptually motivated constraints proposed in this work may not be preserved when using AP-VAST.

Similarly to this work, the AP-VAST framework was shown to outperform a reference pressure matching and acoustic contrast control approach in terms of PESQ and STOI [4]. In addition to this, Lee et al. showed through a MUSHRA listening test that the perceptual approach had a 20% better performance than existing non-perceptual approaches [5]. Due to differences in setups, the approach by Lee et al. and the approaches proposed in this paper cannot be directly compared. One way to effectively compare the two approaches is through listening tests.

In other work, Donley et al. showed how sound zones could be constructed by optimizing over the speech intelligibility contrast (SIC) between zones to improve the speech privacy [3, 38, 39]. This is similar to the approach done in this work, as the sound zones are constructed by direct optimization of a perceptual model.

Optimizing for speech privacy, however, also allows for the addition of white noise in order to increase privacy. As such, while the proposed approach and the approach by Donley et al. are similar mathematically, the outcomes are quite different and, therefore, difficult to compare.


Average PESQ per Constraint D_0

Average MSE per Constraint D_0

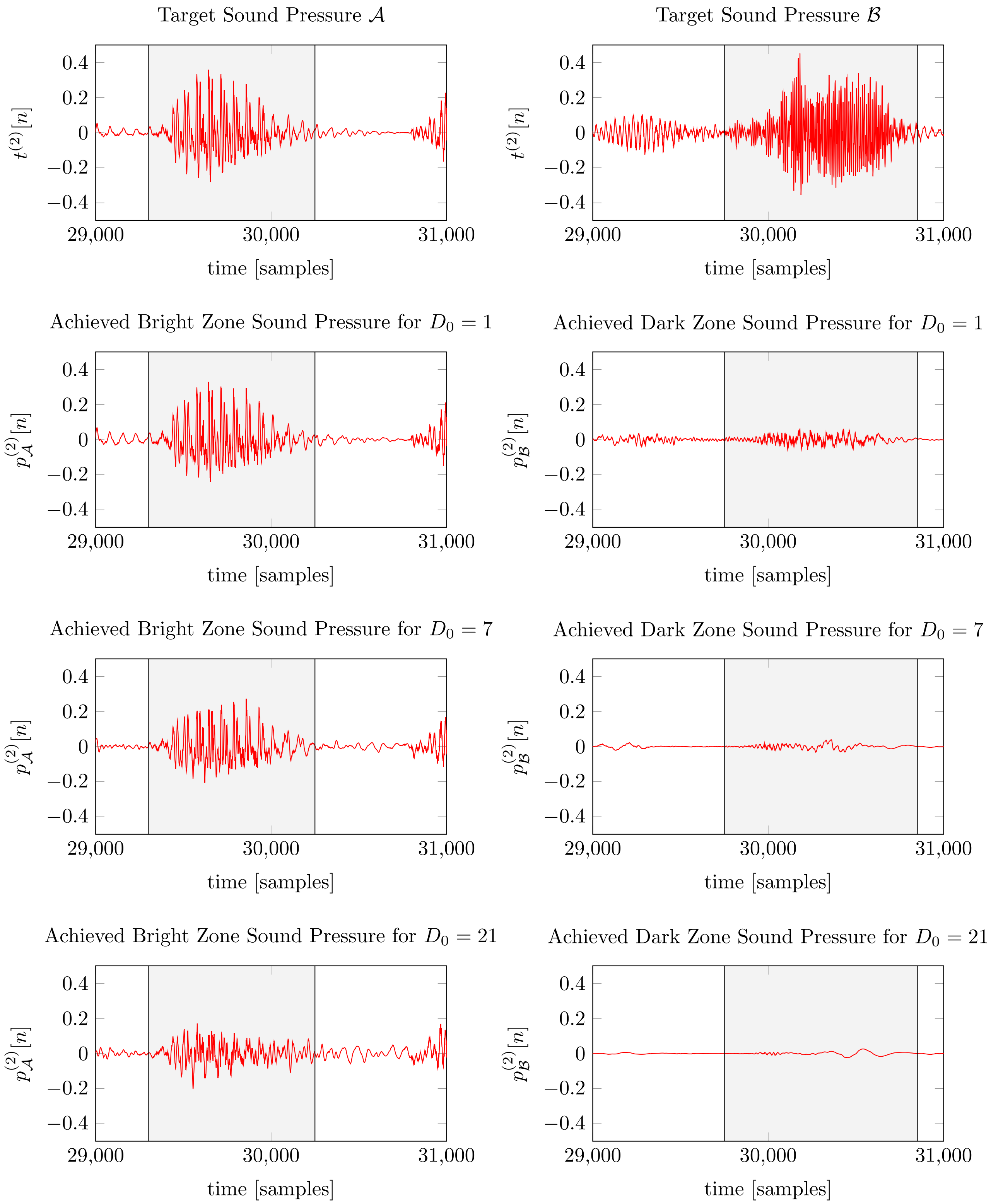
Average STOI per Constraint D_0

Average SIIB per Constraint D_0

Average Distraction per Constraint D_0

Average Acoustic Contrast per Constraint D_0

Figure 5.3: Plots depicting the various perceptual and physical measures introduced in Section 5.1 for various values of the constraint D_0 of the constrained perceptual pressure matching algorithm given by Equation (4.20). All measures all averaged over all 4 control points (see Figure 5.3) and all 12 simulations. The error bars show the first standard deviation in the data.



Target Sound Pressure $\mathcal{A}$

Target Sound Pressure $\mathcal{B}$

Achieved Bright Zone Sound Pressure for $D_0 = 1$

Achieved Dark Zone Sound Pressure for $D_0 = 1$

Achieved Bright Zone Sound Pressure for $D_0 = 7$

Achieved Dark Zone Sound Pressure for $D_0 = 7$

Achieved Bright Zone Sound Pressure for $D_0 = 21$

Achieved Dark Zone Sound Pressure for $D_0 = 21$

Figure 5.4: Plots depicting waveforms of the perceptually constrained pressure matching approach introduced in Section 4.2.2 for varying constraint values of D_0. The wave forms are taken from the experiment detailed in Section 5.1, and depict control point $m = 2$ in zone $\mathcal{A}$ from Figure 5.1. The left plots depict the achieved bright zone sound pressure for $m = 2$ and the right plots the achieved dark zone sound pressure, or interference, for that same control point.

Chapter 6

Conclusion

Sound zone algorithms attempt to control the spatial distribution of sound in order to create zones with distinct audio content in a room. This work aims to explore how a perceptual model of the human auditory system can be integrated into the cost function of sound zone algorithms and what benefits this may have.

In the first research question RQ1, we sought to answer:

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

In this work, it is shown that a perceptual sound zone algorithm can be stated directly using a perceptual sound zone framework based on the Par detectability distortion perceptual model and pressure matching sound zone approach. This sound zone framework uses the perceptual model to determine how perceptually detectable the errors in sound pressure are.

The framework is used in this work to propose two sound zone algorithms. The first algorithm, "unconstrained perceptual pressure matching", in which the total detectability of the sound pressure errors is minimized. The other algorithm, "constrained perceptual pressure matching", in which the detectability of the interference between zones is minimized while constraining the detectability of error in the reproduced audio.

The second research question RQ2 is posed as follows:

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

This work sought to answer this question by investigating the properties of the two proposed perceptual sound zone algorithms, and comparing it with a reference non-perceptual pressure matching approach. Findings suggest that the benefits of including perceptual models in sound zone algorithms are twofold:

- The work indicates that the proposed unconstrained perceptual pressure matching outperforms the non-perceptual pressure matching in terms of perceptual speech measures PESQ, STOI, SIIB, and Distraction. Investigation indicates that one possible reason for this is that the interference introduced by the reference algorithm is more perceptually disturbing.

- The work shows that proposed constrained perceptual pressure matching allows for control over the perceived quality of the sound in the zones. By leveraging the perceptual interpretation of the Par distortion detectability, one can specify the desired minimum level of quality. This work shows that the perceptual constraint correlates directly with the quality that is reproduced.

  This is a challenge for non-perceptual sound zone approaches as the cost functions are typically constructed using physical measures. These physical measures have no consistent perceptual interpretation, meaning that the same constraints can lead to widely varying perceptual results.

While this work successfully answers its research questions, there are still many promising directions of perceptual sound zones research. The following future work is found to be of interest:

- As discussed, it is shown that the unconstrained perceptual pressure matching algorithm outperforms the reference non-perceptual pressure matching algorithm in terms of various objective perceptual measures. However, as discussed in Section 2.1, these objective measures can only be used to give an indication of performance.

  To objectively determine if the perceptual approach does indeed outperform traditional approaches, formal listening tests must be conducted.

- As shown, the constrained perceptual pressure matching approach can be used to control the reproduced audio quality through the detectability of the reproduction error. However, the degree to which this is possible with a non-perceptual pressure matching approach is not explored in this work.

  It is of interest to compare the performance of perceptual and non-perceptual constraints to obtain a complete understanding of the differences.

- The proposed perceptual sound zone framework can be used to formulate more perceptual sound zone algorithms. One algorithm that is of particular interest is an algorithm that constrains the detectability of the interference rather than the reproduction error. This can then be readily compared to non-perceptual pressure matching approaches, which often include a similar, non-perceptual constraint.

- Currently, the proposed perceptual sound zone algorithms are posed as optimization problems that use both time and frequency domain representations of the optimizers. The translation between domains is suspected to greatly increase the computational complexity of the algorithm.

  As such, it is of interest to obtain a version of the algorithm that operates in a single domain to reduce the computational complexity.

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

[3] J. Donley and C. Ritz, "Multizone reproduction of speech soundfields: A perceptually weighted approach," in *Proceedings of APSIPA Annual Summit and Conference*, vol. 16, no. 19, 2015.

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

[5] ——, "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.

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

[7] 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.

[8] M. Torcoli, T. Kastner, and J. Herre, "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1530–1541, 2021.

[9] J. Herre and S. Dick, "Psychoacoustic models for perceptual audio coding—a tutorial review," *Applied Sciences*, vol. 9, no. 14, p. 2854, 2019.

[10] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Transac-*

tions on Audio, Speech, and Language Processing, vol. 20, no. 5, pp. 1553–1564, 2012.

[11] 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.

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

[13] 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.

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

[15] 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," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX).* IEEE, 2020, pp. 1–6.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] J. Francombe, R. Mason, M. Dewhirst, and S. Bech, "Elicitation of attributes for the evaluation of audio-on-audio interference," *The Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2630–2641, 2014.

[21] ——, "A model of distraction in an audio-on-audio interference situation with music program material," *Journal of the Audio Engineering Society*, vol. 63, no. 1/2, pp. 63–77, 2015.

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

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

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

[25] 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.

[26] 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.

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

[28] L. Vindrola, M. Melon, J.-C. Chamard, B. Gazengel, and G. Plantier, “Personal sound zones: A comparison between frequency and time domain formulations in a transportation context,” in *Audio Engineering Society Convention 147.* Audio Engineering Society, 2019.

[29] E. A. Habets, “Room impulse response generator,” *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[30] 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.

[31] S. J. Elliott and J. Cheer, “Regularisation and robustness of personal audio systems,” *ISVR Technical Memorandum 995*, 2011.

[32] 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.

[33] L. Shi, T. Lee, L. Zhang, J. K. Nielsen, and M. G. Christensen, “Generation of personal sound zones with physical meaningful constraints and conjugate gradient method,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 823–837, 2021.

[34] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-

room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[35] C. Díaz and A. Pedrero, "The reverberation time of furnished rooms in dwellings," *Applied Acoustics*, vol. 66, no. 8, pp. 945–956, 2005.

[36] ETSI, "Etsi ts 103 737 v1.1.2 speech and multimedia transmission quality (stq)," European Telecommunications Standards Institute 2010, 650 Route des Lucioles F-06921 Sophia Antipolis Cedex - FRANCE, techreport V1.1.2, Aug. 2010-08.

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

[38] J. Donley, C. Ritz, and W. B. Kleijn, "Improving speech privacy in personal sound zones," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 311–315.

[39] ——, "Multizone soundfield reproduction with privacy-and quality-based speech masking filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1041–1055, 2018.

Appendix A

Calibration of the the Par Detectability Measure

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

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

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

A.1 Relating Digital Representation and Sound Pressure Level

One difficulty of taking the threshold of hearing into account is that it is typically given in terms of sound pressure level (SPL), measured in dB. The one-sided spectrum of the threshold of hearing in dB SPL can be approximated by the following function [6]:

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

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

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

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

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

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

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

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

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

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

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

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

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

A.2 Determining Calibration Constants

There are various ways of calibrating this model, but this section will discuss the method of calibrating that is given in the original paper [7]. 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 the presence of an in-phase sinusoidal masking signal that is 18 dB SPL louder [7]. To model this, consider the following masking and disturbance signals.

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

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

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

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

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

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

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

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

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

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

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

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

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