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

CAS-2021-??

M.Sc. Thesis

Sound Zones with a Cost Function based on Human Hearing

Niels Evert Marinus de Koeijer B.Sc.

Abstract

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

TU Delft

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

Sound Zones with a Cost Function based on Human Hearing

Subtitle Compulsory?

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Electrical Engineering

by

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

This work was performed in:

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

Delft University of Technology

Delft University of Technology
Department of
Microelectronics & Computer Engineering

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

Dated: September 15, 2021

Chairman: ________________________
dr.ir. R.C. Hendriks

Advisors: ________________________
dr. M. Bo Møller

____
dr. P. Martinez Nuevo

Committee Members: ________________________
dr. M. Mastrangeli

____
dr. J. Martinez Castañeda

Abstract

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

Acknowledgments

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

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

Contents

List of Figures

List of Tables

1 Introduction

Skeleton of Chapter

In the introduction I will discuss the following:

- Introduction to sound zones:
  - Explain what sound zones are, and why we want them.
  - Explain where the state of the art comes short.
- Introduction to perceptual approach.
  - Motivate the perceptual approach.
    - Optimizes for perceptual experience, rather than sound pressure
    - Show how it could solve the short-coming of the state of the art
  - Briefly discuss prior work in this approach.
    - Work done by Taewoong Lee [1][2].
    - Work done by Jacob Donley [3][4]
  - Introduce goal of thesis:
    - Furthering perceptual sound zones state of the art?
- Give structure of the document

2 Review of Perceptual Model Literature

Skeleton of Chapter

In order to build a perceptual sound zone algorithm, we review literature for perceptual models to find a suitable perceptual model.

- I will discuss the criteria which will determine which perceptual model is chosen.
  - Complexity
  - Feasibility to optimize
- I will discuss my literature review into perceptual models to find a model that best fits the criteria.
  - Dau Model
  - Detectability Models, i.e. Par and Taal
  - Distraction Model
  - Audio quality models, PEAQ, VISQOL
  - Speech Intelligibility Based, i.e. SIIB and STOI
- I will discuss and motivate what perceptual model will be used in the optimization problem.
  - This is done by means of summarizing the findings, and then reflecting on the previously introduced criteria. From this, I will conclude that the **Par Detectability** is best suited.
- I will discuss and motivate what perceptual models will be used for evaluation. From this I will conclude that the speech intelligibility metrics are suitable.

3 Implementation of Perceptual Model

Skeleton of Chapter

Here, I describe the implementation of the chosen perceptual models, i.e. the detectability.

- I give a high-level description of detectability.
- I describe the Par Detectability.
  - The underlying perceptual ideas
  - How its implemented
  - How its calibrated
- I show that my implementations of the Par Detectability is valid.
  - This is done by comparing the masking curve predictions to a reference implementation of the Dau model.

Review of Sound Zone Algorithms Literature 4

Skeleton of Chapter

At this point the Par Detectability as been selected as the perceptual model of choice. In this chapter, the literature will be reviewed in order to find a suitable sound zone approach for integrating the perceptual model into.

- I will define the criteria that are required for a sound zone algorithm.
  - Par Detectability is a distortion measure which can be expressed as an L2 error.
- I will discuss my literature review into sound zones.
  - Pressure Matching (PM) Approaches.
  - Acoustic Contrast Control (ACC) Approaches.
  - Mode Matching Approaches.
- I will summarize the results and reflect on the requirements. From this I will conclude that the pressure matching approach is the most suited as it uses an L2 error as well.

5 Implementation of Reference Sound Zone Algorithm

Skeleton of Chapter

In the preceding chapter it was concluded that a pressure matching approach was best suited for building a perceptual sound zone algorithm. In this chapter, a reference pressure matching implementation will be given that will function as a the basis for the perceptual algorithm to be introduced in later chapters.

- Introduction of the data model
  - Mathematical description of the room, loudspeakers, zones, room impulse responses.
  - Definition of the sound pressure in their room, how it relates to the loudspeaker input signals.
- Introduction of the Multi-Zone Pressure-Matching (MZ-PM) approach.
  - Contains the complete derivation of the approach, starting from the previously introduced data model.
- Extension of the MZ-PM approach to work on a short-time scale
  - The motivation for this is that the previously-derived approach requires knowledge of the full time domain signal, which is unrealistic in practice.
  - In addition to this, the perceptual model works on a short time-scale
- Extension of the short-time MZ-PM approach to work in the frequency domain.
  - This is a requirement for the perceptual model, as it functions in the STFT domain.

5.1 Introduction

5.2 Data Model

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

The contents of this section are as follows.

First, subsection 5.2.1 introduces a spatial description of a room containing two zones and a loudspeaker array. Then, subsection 5.2.2 defines the objective of the sound zone algorithm as realizing target sound pressure at discrete points in the room.

The relation between the sound pressure in the room and loudspeaker input signals will then be given in subsection 5.2.3, completing the mathematical framework. This is then used in subsection 5.2.4 to select a suitable target sound pressure which will be used in the remainder of this thesis.

5.2.1 Room Topology

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

The room R can be modeled as a closed subset of three dimensional space, $\mathcal{R} \subset \mathbb{R}^3$. The two non-overlapping zones $\mathcal{A}$ and $\mathcal{B}$ are contained within the room R, i.e. $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{R}$ where $\mathcal{A} \cap \mathcal{B} = \emptyset$. In addition to the zones, the room $\mathcal{R}$ also contains N_L loudspeakers, which can be modeled as discrete points. The room, loudspeakers and zones are visualized in Figure 5.1.

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

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

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

5.2.2 Defining Target Pressure

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

Currently, the zones are given as continuous regions in space. Sound zone approaches will attempt to recreate a specified pressure in the entire region of space defined by $\mathcal{A}$ and $\mathcal{B}$. Other approaches will instead discretize the zones by sampling the continuous zones $\mathcal{A}$ and $\mathcal{B}$ into so-called control points. The sound pressure is then controlled only in these control points.

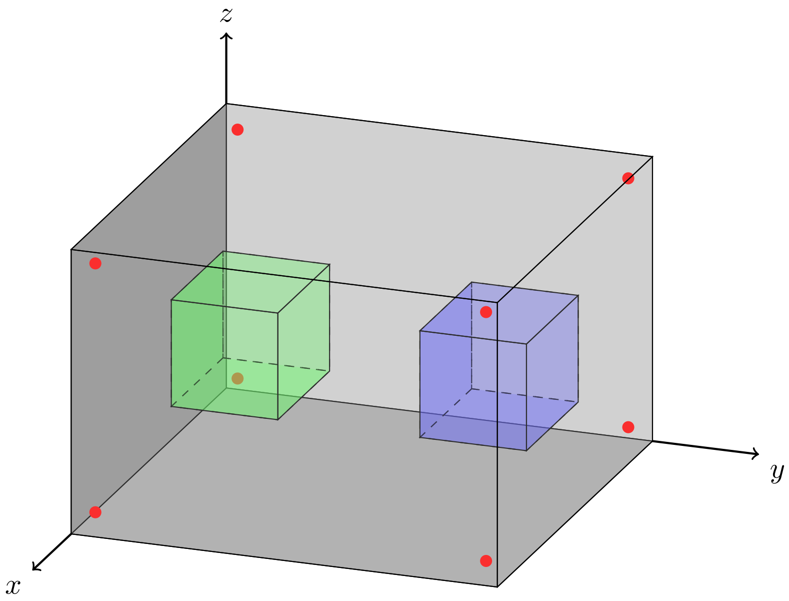

Figure 5.1: The room $\mathcal{R} \subset \mathbb{R}^3$ containing the zones $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{R}$ depicted in green and blue respectively. The room contains $N_L = 8$ loudspeakers, which are denoted by the red dots in the corners of the room.

In this work, a pressure matching approach is used, and thus the latter approach will be taken. Thus, we discretize zones $\mathcal{A}$ and $\mathcal{B}$ into a total of N_a and N_b control points respectively. Let A and B denote the sets of the resulting control points points contained within zones $\mathcal{A}$ and $\mathcal{B}$ respectively.

Now let $t^m[n]$ denote the target sound pressure at control point m in either A or B, i.e. $m \in A \cup B$. Our goal is thus to realize $t^m[n]$ in all control points $m \in A \cup B$ using the loudspeakers present in the room. The relationship between the loudspeaker input signals and the sound pressure is the topic of the next section.

5.2.3 Realizing Sound Pressure through the Loudspeaker

The sound pressure produced by the loudspeakers can be controlled by specifying their input signals. Mathematically speaking, let $x^{(l)}[n] \in \mathbb{R}^{N_x}$ denote the loudspeaker input signal for the l^{th} loudspeaker. As such, the goal of the sound zone algorithm is to find loudspeaker inputs $x^{(l)}[n]$ such that the target sound pressure $t^m[n]$ is realized for all $m \in A \cup B$.

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

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

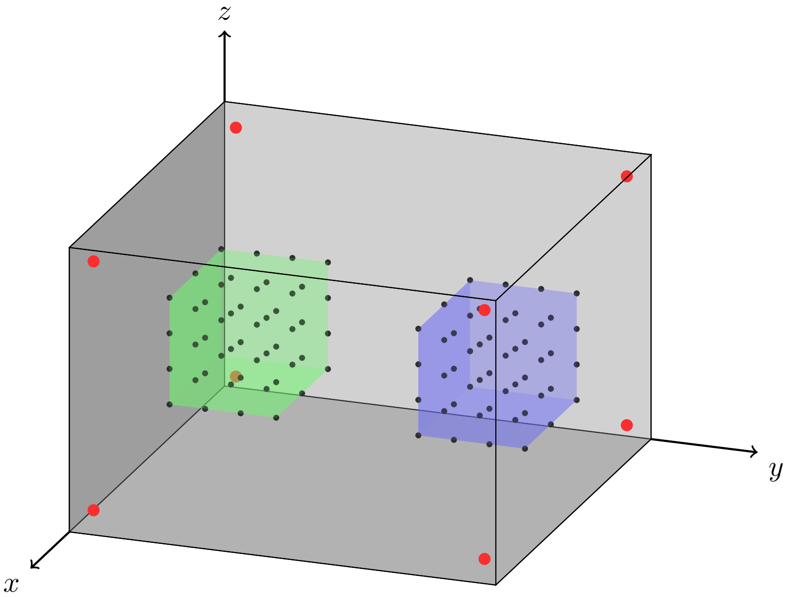

Figure 5.2: The previously introduced room $\mathcal{R}$ with zones $\mathcal{A}$ and $\mathcal{B}$ discretized.

represent said sound pressure. It can be defined as follows:

$$p^{(l,m)}[n] = \left(h^{(l,m)} * x^{(l)}\right)[n] \tag{5.1}$$

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

$$p^{(m)}[n] = \sum_{l=0}^{N_L} p^{(l,m)}[n] \tag{5.2}$$

$$= \sum_{l=0}^{N_L} \left(h^{(l,m)} * x^{(l)}\right)[n] \tag{5.3}$$

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

5.2.4 Choice of Target Pressure

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

Assume that the user of the sound zone system has selected desired playback audio signals $s_{\mathcal{A}}[n] \in \mathbb{R}^{N_x}$ and $s_{\mathcal{B}}[n] \in \mathbb{R}^{N_x}$ that they wish to hear in zone $\mathcal{A}$ and $\mathcal{B}$ respectively.

In order to accommodate the wishes of the user, the target sound pressure is chosen as follows:

$$t^{(m)}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * s_{\mathcal{A}} \right)[n] \qquad \forall m \in A \tag{5.4}$$

$$t^{(m)}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * s_{\mathcal{B}} \right)[n] \qquad \forall m \in B \tag{5.5}$$

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

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

5.3 Multi-Zone Pressure-Matching Solution Approach

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

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

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

$$x^{(l)}[n] = x_{\mathcal{A}}^{(l)}[n] + x_{\mathcal{B}}^{(l)}[n] \tag{5.6}$$

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

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

$$p_{\mathcal{Z}}^{(m)}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * x_{\mathcal{Z}}^{(l)} \right)[n] \tag{5.7}$$

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

$$p^{(m)}[n] = p_{\mathcal{A}}^{(m)}[n] + p_{\mathcal{B}}^{(m)}[n] \tag{5.8}$$

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

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

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

$$\underset{x_{\mathcal{A}}^{(l)}[n],\, x_{\mathcal{B}}^{(l)}[n]\,\forall\, l}{\arg\min} \quad \sum_{m\in A} \left|\left|p_{\mathcal{A}}^{(m)}[n] - t^{(m)}[n]\right|\right|_2^2 + \sum_{m\in A} \left|\left|p_{\mathcal{B}}^{(m)}[n]\right|\right|_2^2 + \tag{5.9}$$

$$\sum_{m\in B} \left|\left|p_{\mathcal{B}}^{(m)}[n] - t^{(m)}[n]\right|\right|_2^2 + \sum_{m\in B} \left|\left|p_{\mathcal{A}}^{(m)}[n]\right|\right|_2^2 \tag{5.10}$$

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

$$\mathrm{RE}_{\mathcal{Z}} = \sum_{m\in A} \left|\left|p_{\mathcal{A}}^{(m)}[n] - t^{(m)}[n]\right|\right|_2^2 \tag{5.11}$$

$$\mathrm{LE}_{\mathcal{Z}} = \sum_{m\in A} \left|\left|p_{\mathcal{B}}^{(m)}[n]\right|\right|_2^2 \tag{5.12}$$

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

$$\underset{x_{\mathcal{A}}^{(l)}[n],\, x_{\mathcal{B}}^{(l)}[n]\,\forall\, l}{\arg\min} \quad \mathrm{RE}_{\mathcal{A}} + \mathrm{LE}_{\mathcal{A}} + \mathrm{RE}_{\mathcal{B}} + \mathrm{LE}_{\mathcal{B}} \tag{5.13}$$

From this it becomes clear that this approach results in trade-off between minimizing the reproduction errors $\mathrm{RE}_{\mathcal{Z}}$ and leakages $\mathrm{LE}_{\mathcal{Z}}$. Some pressure matching approaches attempt to control this trade-off by introducing weights for the different error terms, or by adding constraints. Choosing constraints can however be challenging as the mean square pressure error is difficult to interpret.

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

5.4 Block-Based Multi-Zone Pressure-Matching

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

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

Instead, a block-based approach can be used where the incoming samples of the desired playback signals are used in real-time as they become available. The system buffers a block of H incoming samples samples, and then solves the sound zone problem for the newest block. The buffering results in a latency of H samples, which could be acceptable assuming H is chosen sufficiently small.

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

For these reasons, this section will adapt existing Multi-Zone Pressure Matching approach introduced in **??** to accommodate for block based processing.

5.4.1 Mathematical Block Model

For the block-processing based sound zone approach, the incoming samples of the desired playback signals for both zones $s_{\mathcal{A}}[m]$ and $s_{\mathcal{B}}[m]$ are buffered into blocks. As such, the sound zone system only has knowledge of the most recent block, denoted by block index μ. The relation between the global time index n and block index μ is given as follows:

$$\mu = \lfloor n/H \rfloor \tag{5.14}$$

Thus at a time n, up to and including the μ^{th} blocks of desired playback signals $s_{\mathcal{A}}[n]$ and $s_{\mathcal{B}}[n]$ are known, i.e. $0 \leq n \leq \mu H$, assuming $s_{\mathcal{A}}[n]$ and $s_{\mathcal{B}}[n]$ are causal.

As the desired playback signals $s_{\mathcal{A}}[m]$ and $s_{\mathcal{B}}[m]$ are revealed in a block-wise fashion, the sound zone system cannot compute the entirety of loudspeaker input signals $x_{\mathcal{A}}^{(l)}[n]$ and $x_{\mathcal{B}}^{(l)}[n]$.

Instead, one approach is to compute the loudspeaker input signals in the same block-wise fashion, by finding the H newest samples of the loudspeaker input signals as the H newest samples of desired playback signals $s_{\mathcal{A}}[m]$ and $s_{\mathcal{B}}[m]$ are revealed to the system.

5.4.2 Block Based Multi-Zone Pressure-Matching

As discussed previously, the Multi-Zone Pressure-Matching (MZ-PM) algorithm attempts to control the loudspeaker input signals $x_{\mathcal{A}}^{(l)}[n]$ and $x_{\mathcal{B}}^{(l)}[n]$ such that a specified target sound pressure $t^{(m)}[n]$ is attained at all control points m. Here, target sound pressure $t^{(m)}[n]$ is determined by the sound pressure that arises due to the desired playback signals $s_{\mathcal{A}}[n]$ and $s_{\mathcal{B}}[n]$.

The introduction of the block-based approach limits the knowledge of $s_{\mathcal{A}}[n]$ and $s_{\mathcal{B}}[n]$ up to and including the most recent block μ. As such, the desired playback signals are only known up for $0 \leq n \leq \mu H$, assuming causal desired playback signals.

This limitation has implications for the computation of the loudspeaker input signals and the target sound pressure. Neither quantities can be computed in their entirety due to the limitation in knowledge of the desired playback signals. Because the desired playback signals are revealed to the system in blocks of size H, the system will instead compute the loudspeaker input signals and target sound pressure at the same rate.

As such, after a new block of desired playback signals is revealed, a new block of loudspeaker input signals will be computed such that a new block of target sound pressure is best attained. Adapting the existing MZ-PM algorithm to operate on a block-by-block basis is the topic of this section.

Defining Block-Based Loudspeaker Input Signals

First, consider the implications of the block based processing on the loudspeaker input signals. The goal is to compute $x_{\mathcal{Z}}^{(l)}[n]$ in a blocks of size H. To do so, consider the segmentation of the sound pressure that is realized due to loudspeaker input signal $x_{\mathcal{Z}}^{(l)}[n]$:

$$\begin{aligned}
p_{\mathcal{Z}}^{(m)}[n] &= \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * x_{\mathcal{Z}}^{(l)}\right)[n] && (5.15)\\
&= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] x_{\mathcal{Z}}^{(l)}[b] && (5.16)\\
&= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] x_{\mathcal{Z}}^{(l)}[b] \sum_{k=-\infty}^{\infty} w[b-kH] && (5.17)\\
&= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\infty} x_{\mathcal{Z}}^{(l)}[b] w[b-kH] && (5.18)\\
& && (5.19)
\end{aligned}$$

Here, $w[n] \in \mathbb{R}^{N_w}$ is a window that is defined to be non-zero for $-N_w + 1 \leq n \leq 0$. As such, it is a non-causal window. It is chosen such that it satisfies the constant overlap add (COLA) condition for a hop size H, which is given as follows:

$$\sum_{k=-\infty}^{\infty} w[n-kH] = 1 \quad \forall\, n \tag{5.20}$$

The interpretation of the rewrite of $p_{\mathcal{Z}}^{(m)}[n]$ above can be understood as a projection of the loudspeaker input signals $x_{\mathcal{Z}}^{(l)}[n]$ onto a basis frames, formed by a basis of overlapping windows $w[n]$.

The hop-size is chosen to be equal to the block size H, as such the overlap is equal to $N_w - H$. Note also that the windowed blocks need not be overlapping in general, i.e. one possible choice of window is the rectangular window with $N_w = H$.

This forms individual segments $x_{\mathcal{Z}}^{(l)}[n]w[n-\mu H]$, which have support $-N_w + 1 + \mu H \leq n \leq \mu H$. Due to the properties of the COLA condition, the individual segments can be recombined to form the complete loudspeaker input signal:

$$x_{\mathcal{Z}}^{(l)}[n] = \sum_{k=-\infty}^{\infty} x_{\mathcal{Z}}^{(l)}[n]w[n-kH] \tag{5.21}$$

This model can be used order to compute the complete loudspeaker input signal $x_{\mathcal{Z}}^{(l)}[n]$ block-by-block by solving the sound zone problem per segment $x_{\mathcal{Z}}^{(l)}[n]w[n-kH]$.

The idea is then to compute the $x_{\mathcal{Z}}^{(l)}[n]w[n-\mu H]$ segment of $x_{\mathcal{Z}}^{(l)}[n]$ such that said target pressure is best attained. In order to do so, let $x_{\mathcal{Z},\mu}^{(l)}[n] \in \mathbb{R}^{N_w}$ represent the content of the μ^{th} frame.

When block μ of $s_{\mathcal{Z}}[n]$ is revealed, $x_{\mathcal{Z},\mu}^{(l)}[n]$ can be computed such that the target pressure defined by the new desired playback content is best attained. The μ^{th} frame can then be added in an overlap-add like fashion to compute the loudspeaker input signal in real-time.

Before deriving how the loudspeaker input signal frame content $x_{\mathcal{Z},\mu}^{(l)}[n]$ can be computed the block-wise computation the target sound pressure must be discussed. This is the topic of the next paragraph.

5.4.3 Block-Based Target Pressure Computation

As mentioned, the playback signals $s_{\mathcal{A}}[n]$ and $s_{\mathcal{B}}[n]$ are revealed in blocks of size H in the block based framework. As such, for block μ, the knowledge of the desired playback signals is limited, which results in a limited knowledge in the target sound pressure $t^{(m)}[n]$.

The target sound pressure can be segmented in a way analogous to the segmentation of the loudspeaker input signals:

$$
\begin{aligned}
t^{(m)}[n] &= \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * s_{\mathcal{Z}}\right)[n] && (5.22)\\
&= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{Z}}[b] && (5.23)\\
&= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{Z}}[b] \sum_{k=-\infty}^{\infty} w[b-kH] && (5.24)\\
&= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{\mu=-\infty}^{\infty} s_{\mathcal{Z}}[b] w[b-\mu H] && (5.25)\\
& && (5.26)
\end{aligned}
$$

In the rewrite above, the desired playback signal $s_{\mathcal{Z}}[n]$ is projected onto a basis spanned by windows $w[n]$ of size N_w. The equation above essentially sums the contributions individual contributions of windowed segments.

This allows for a formulation of $t_{\mu}^{(m)}[n]$ in which we only consider the contribution up to and including the μ^{th} segment. Such a formulation is given as follows:

$$
\begin{aligned}
t_{\mu}^{(m)}[n] &= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\mu} s_{\mathcal{Z}}[b] w[b-\mu H] && (5.27)\\
&= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{Z}}[b] \left(w[b-\mu H] + \sum_{k=-\infty}^{\mu-1} w[b-\mu H] \right) && (5.28)\\
&= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{Z}}[b] w[b-\mu H] + t_{\mu-1}^{(m,l)}[n] && (5.29)
\end{aligned}
$$

As can be seen, $t_{\mu}^{(m)}[n]$ is expressed as the contribution of the windowed segment μ and the contribution of all previous segments $-\infty \leq k \leq \mu - 1$.

The computation can be performed recursively: to compute $t_{\mu}^{(m)}[n]$, we compute the convolution of the current windowed block $s_{\mathcal{Z},\mu}[n] w[n-\mu H]$ with the RIRs, and then add the history of previous blocks defined by $t_{\mu-1}^{(m,l)}[n]$.

Thus, $t_{\mu}^{(m)}[n]$ can be understood to be the target pressure given blocks up to μ. As new blocks are revealed, the target target sound pressure can be updated. Note that this definition converges to the "true" target sound pressure:

$$
t_{\infty}^{(m)}[n] = t^{(m)}[n] \qquad (5.30)
$$

One interpretation of was done so far is that the target sound pressure can be performed by breaking the convolution of the desired playback signal $s_{\mathcal{Z}}[n]$ with the room impulse responses $h^{(l,m)}[n]$ into a sum of convolutions of windowed blocks of the desired playback signal. In doing so, the target sound pressure can be computed in real-time as new samples of $s_{\mathcal{Z}}[n]$ come available.

The target sound pressure $t_{\mu}^{(m)}[n]$ will be used in the computation of $x_{\mathcal{Z},\mu}^{(l)}[n]$. The idea here is to choose $x_{\mathcal{Z},\mu}^{(l)}[n]$ such that the resulting sound pressure best matches $t_{\mu}^{(m)}[n]$.

Derivation of Block-Based Multi-Zone Pressure-Matching

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

To begin, let the sound pressure realized after playing the first μ loudspeaker input signal segments $x_{\mathcal{Z},\mu}^{(l)}[n]$ be denoted by $p_{\mathcal{Z},\mu}^{(m)}[n]$. This sound pressure can be expressed as follows:

$$p_{\mathcal{Z},\mu}^{(m)}[n] = \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\mu} x_{\mathcal{Z},k}^{(l)}[b]w[b-kH] \tag{5.31}$$

$$= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] x_{\mathcal{Z},\mu}^{(l)}[b]w[b-\mu H] + p_{\mathcal{Z},\mu-1}^{(m)}[n] \tag{5.32}$$

Note that $p_{\mathcal{Z},\mu}^{(m)}[n]$ can be computed recursively just as was the case with the target sound pressure: the sound pressure at a given moment is equal to the sound pressure due to the most recent loudspeaker input signal frame $k = \mu$ and the previous loudspeaker input frames $-\infty \leq k \leq \mu - 1$.

The solution approach: computing the loudspeaker input signal frames $x_{\mu}^{(l)}[n]$ such that the resulting sound pressure $p_{\mathcal{Z},\mu}^{(m)}[n]$ best attains $t_{\mu}^{(l)}[n]$.

As such, the optimization is over one frame $k = \mu$ at a time, all other frames $k \leq \mu - 1$ are assumed constant. Note however that $x_{\mu}^{(l)}[n]$ can only influence samples $-N_w + 1 + \mu H \leq n \leq N_h + \mu H$ of $p_{\mathcal{Z},\mu}^{(m)}[n]$ due to its finite support. Therefore, the optimization only considers the sound pressure that can be controlled by $x_{\mu}^{(l)}[n]$.

The final optimization problem can be thus stated as follows:

$$\underset{x_{\mathcal{A},\mu}^{(l)}[n],\, x_{\mathcal{B},\mu}^{(l)}[n]\,\forall\, l}{\arg\min} \quad \sum_{m\in A} \left|\left|p_{\mathcal{A},\mu}^{(m)}[n] - t_{\mu}^{(m)}[n]\right|\right|_2^2 + \sum_{m\in A} \left|\left|p_{\mathcal{B},\mu}^{(m)}[n]\right|\right|_2^2 + \tag{5.33}$$

$$\sum_{m\in B} \left|\left|p_{\mathcal{B},\mu}^{(m)}[n] - t_{\mu}^{(m)}[n]\right|\right|_2^2 + \sum_{m\in B} \left|\left|p_{\mathcal{A},\mu}^{(m)}[n]\right|\right|_2^2 \tag{5.34}$$

The problem above is solved recursively for loudspeaker input signal frames $x_{\mathcal{A},\mu}^{(l)}[n]$ and $x_{\mathcal{B},\mu}^{(l)}[n]$ as new samples $s_{\mathcal{A}}[n]$ and $s_{\mathcal{B}}[n]$ are revealed. The final loudspeaker input signals can then be found in real-time as follows:

$$x_{\mathcal{Z}}^{(l)}[n] = \sum_{k=-\infty}^{\mu} x_{\mathcal{Z},\mu}^{(l)}[n]w[n-kH] \quad \forall\, n \leq \mu H - N_w + H \tag{5.35}$$

The expression above is only valid up to $n \leq \mu H - N_w + H$ due to missing overlapping frames. The resulting $x_{\mathcal{Z}}^{(l)}[n]$ can then be played in real-time as the loudspeaker input signal frames $x_{\mathcal{Z},\mu}^{(l)}[n]$ are being computed.

5.5 Frequency Domain Block Based Multi-Zone Pressure Matching

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

Firstly, one advantage of using this algorithm over its non block-based counterpart is that it can work in real-time.

Secondly, the block-based approach works on a variable time-scale determine by the block size H. As a result, it can operate on short time-scales. This is useful, as the perceptual model that we wish to integrate operates on short time-scales of the order of 20 to 200 milliseconds.

There is however an additional adjustment that needs to be made before the perceptual model can be integrated. Currently, the BB-MZ-PM algorithm operates in the time domain, whereas the perceptual model operates in the frequency domain.

For this reason, this section will convert the existing time domain BB-MZ-PM algorithm to an equivalent frequency domain formulation. By equivalent it is meant that the algorithms has the same resulting loudspeaker input signals $x_{\mathcal{Z}}^{(l)}$.

In order to relate the frequency domain and the time domain, a natural choice is are discrete fourier transform (DFT) and inverse discrete fourier transform (IDFT).

5.5.1 Quantities the Frequency Domain

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

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

One approach is to evaluate the sound pressures $p_{\mathcal{Z},\mu}^{(m)}[n]$ and $t_{\mu}^{(m)}[n]$ first in the time domain, and then take the DFT. However, the computation of the time-domain versions involves convolutions with the room impulse responses. Another approach is to instead and compute the time-domain convolutions in frequency domain.

Here, use can be made of a property of the DFT: inner product in the frequency domain corresponds to a circular convolution in the time domain. Circular convolution can be made to correspond to linear convolution by means of zero-padding to convolution length.

As such, imagine the convolution between a RIR $h \in \mathbb{R}^{N_h}$ and a loudspeaker input signal $x \in \mathbb{R}^{N_x}$. Here, the convolution length is equal to $N_c = N_h + N_x - 1$. The following holds:

$$\sum_{n=0}^{N_c-1} (h * x)[n] \exp(\omega k n) = \left[\sum_{n=0}^{N_c-1} h[n] \exp(\omega k n)\right] \circ \left[\sum_{n=0}^{N_c-1} x[n] \exp(\omega k n)\right] \tag{5.36}$$

$$= \hat{h}[k] \circ \hat{x}[k] \tag{5.37}$$

Here, $\omega = 2\pi/N_c$. The expression above relates the DFT of the linear convolution between RIR and loudspeaker input signal to the inner product between the DFT of the RIR and loudspeaker input signals The DFT as defined here implicitly zero-pads the input signals to convolution length N_c.

This property can be used to find the frequency domian version of the pressure $p_{\mathcal{Z},\mu}^{(m)}[n]$ as follows: **TODO: Spontaneously introduced some funky notation... Also, this may not make sense. Consider the time shift required for the DFT.**

$$\hat{p}_{\mathcal{Z},\mu}^{(m)}[k] = \left\{\sum_{n=0}^{N_c-1}\left[\sum_{l=0}^{N_L-1}\left(h^{(l,m)} * x_{\mathcal{Z},\mu}^{(l)} w_\mu\right)[n]\right] \exp(\omega k n)\right\} + \hat{p}_{\mathcal{Z},\mu-1}^{(m)}[k] \tag{5.38}$$

$$= \left\{\sum_{l=0}^{N_L-1} \hat{h}^{(l,m)}[k] \circ \left[\sum_{n=0}^{N_c-1}\left(x_{\mathcal{Z},\mu}^{(l)}[n-\mu H] w_\mu[n]\right) \exp(\omega k n)\right]\right\} + \hat{p}_{\mathcal{Z},\mu-1}^{(m)}[k] \tag{5.39}$$

TODO: Similar arguments for the target pressure

5.5.2 Proposed Frequency Domain Approach

TODO: Essentially, we zero-pad and go to the frequency domain for all quantities. Then all convolutions become inner products. Zero-padding is done to convolution length. We optimize still over the time domain signal.

$$\underset{x_{\mathcal{A},\mu}^{(l)}[n],\, x_{\mathcal{B},\mu}^{(l)}[n]\, \forall\, l}{\arg\min} \quad \sum_{m \in A} \left|\left|\hat{p}_{\mathcal{A},\mu}^{(m)}[k] - \hat{t}_\mu^{(m)}[k]\right|\right|_2^2 + \sum_{m \in A} \left|\left|\hat{p}_{\mathcal{B},\mu}^{(m)}[k]\right|\right|_2^2 + \tag{5.40}$$

$$\sum_{m \in B} \left|\left|\hat{p}_{\mathcal{B},\mu}^{(m)}[k] - \hat{t}_\mu^{(m)}[k]\right|\right|_2^2 + \sum_{m \in B} \left|\left|\hat{p}_{\mathcal{A},\mu}^{(m)}[k]\right|\right|_2^2 \tag{5.41}$$

$$\text{subject to} \quad \hat{p}_{\mathcal{A},\mu}^{(m)}[k] = \text{Windowing, Zero-pad, DFT of } x_{\mathcal{A},\mu}^{(l)}[n] \quad \forall\, m \in \mathcal{A} \tag{5.42}$$

$$\hat{p}_{\mathcal{B},\mu}^{(m)}[k] = \text{Windowing, Zero-pad, DFT of } x_{\mathcal{B},\mu}^{(l)}[n] \quad \forall\, m \in \mathcal{B} \tag{5.43}$$

Implementation of Perceptual Sound Zone Algorithm 6

Skeleton of Chapter

Describes the design of the perceptual sound zone algorithm.

- Introduction of detectability into the various terms of the existing STFT-based Multi-Zone Pressure-Matching approach.
- Show how these terms can be used to form multiple algorithms
  - Unconstrained minimization of detectability
  - Constrained minimization of detectability
    - Motivate by discussing shortcomings of unconstrained approach

6.1 Introduction

6.2 Perceptual Minimization Approach

Essentially, argue that Detectability is defined as:

$$D(\nu[n], \epsilon[n]) = ||\mathbf{w}(\hat{\nu}[\omega]) \circ \hat{\boldsymbol{\epsilon}}||_2^2 \tag{6.1}$$

The detectability quantifies how detectable masked signal $\epsilon[n]$ is in presence of masker signal $\nu[n]$. It is calibrated such that $D(\nu[n], \epsilon[n]) = 1$ for a masker-masked signal pair Where $\hat{x}[\omega]$ is the DFT of the masker $\nu[n]$ and $\hat{\boldsymbol{\epsilon}}$ a vector containing the DFT of the masked signal $\epsilon[n]$.

As can be seen, the masker defines the detectability weighting vector $\mathbf{w}(\hat{x}[\omega])$, which is a weighting informed by the masking effects of $\hat{x}[\omega]$. The computation for this masking vector is done in the frequency domain.

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

$$\begin{aligned}
\underset{x_{\mathcal{A},\mu}^{(l)}[n],\, x_{\mathcal{B},\mu}^{(l)}[n]\,\forall\, l}{\arg\min} \quad & \sum_{m \in A} \left|\left| \hat{p}_{\mathcal{A},\mu}^{(m)}[\omega] - \hat{t}_{\mu}^{(m)}[\omega] \right|\right|_2^2 + \sum_{m \in A} \left|\left| \hat{p}_{\mathcal{B},\mu}^{(m)}[\omega] \right|\right|_2^2 + \\
& \sum_{m \in B} \left|\left| \hat{p}_{\mathcal{B},\mu}^{(m)}[\omega] - \hat{t}_{\mu}^{(m)}[\omega] \right|\right|_2^2 + \sum_{m \in B} \left|\left| \hat{p}_{\mathcal{A},\mu}^{(m)}[\omega] \right|\right|_2^2 \qquad (6.2) \\
\text{subject to} \quad & \hat{p}_{\mathcal{A},\mu}^{(m)}[\omega] = \text{Windowing, Zero-pad, DFT of } x_{\mathcal{A},\mu}^{(l)}[n] \quad \forall\; m \in \mathcal{A} \\
& \hat{p}_{\mathcal{B},\mu}^{(m)}[\omega] = \text{Windowing, Zero-pad, DFT of } x_{\mathcal{B},\mu}^{(l)}[n] \quad \forall\; m \in \mathcal{B}
\end{aligned}$$

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

$$\begin{aligned}
\underset{x_{\mathcal{A},\mu}^{(l)}[n],\, x_{\mathcal{B},\mu}^{(l)}[n]\,\forall\, l}{\arg\min} \quad & \sum_{m \in A} \left|\left| \mathbf{w}\left(\hat{t}_{\mu}^{(m)}[\omega]\right) \circ \left[\hat{p}_{\mathcal{A},\mu}^{(m)}[\omega] - \hat{t}_{\mu}^{(m)}[\omega] \right] \right|\right|_2^2 + \sum_{m \in A} \left|\left| \mathbf{w}\left(\hat{t}_{\mu}^{(m)}[\omega]\right) \circ \left[\hat{p}_{\mathcal{B},\mu}^{(m)}[\omega] \right] \right|\right|_2^2 + \\
& \sum_{m \in B} \left|\left| \mathbf{w}\left(\hat{t}_{\mu}^{(m)}[\omega]\right) \circ \left[\hat{p}_{\mathcal{B},\mu}^{(m)}[\omega] - \hat{t}_{\mu}^{(m)}[\omega] \right] \right|\right|_2^2 + \sum_{m \in B} \left|\left| \mathbf{w}\left(\hat{t}_{\mu}^{(m)}[\omega]\right) \circ \left[\hat{p}_{\mathcal{A},\mu}^{(m)}[\omega] \right] \right|\right|_2^2 \\
\text{subject to} \quad & \hat{p}_{\mathcal{A},\mu}^{(m)}[\omega] = \text{Windowing, Zero-pad, DFT of } x_{\mathcal{A},\mu}^{(l)}[n] \quad \forall\; m \in \mathcal{A} \\
& \hat{p}_{\mathcal{B},\mu}^{(m)}[\omega] = \text{Windowing, Zero-pad, DFT of } x_{\mathcal{B},\mu}^{(l)}[n] \quad \forall\; m \in \mathcal{B}
\end{aligned} \tag{6.3}$$

6.3 Perceptual Constraining Approach

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

$$
\begin{array}{rl}
\underset{x^{(l)}_{\mathcal{A},\mu}[n],\, x^{(l)}_{\mathcal{B},\mu}[n]\,\forall\, l}{\arg\min} & \displaystyle\sum_{m\in A}\left|\left|\mathbf{w}\left(\hat{t}^{(m)}_{\mu}[\omega]\right)\circ\left[\hat{p}^{(m)}_{\mathcal{B},\mu}[\omega]\right]\right|\right|_2^2+\sum_{m\in B}\left|\left|\mathbf{w}\left(\hat{t}^{(m)}_{\mu}[\omega]\right)\circ\left[\hat{p}^{(m)}_{\mathcal{A},\mu}[\omega]\right]\right|\right|_2^2 \\
\text{subject to} & \displaystyle\sum_{m\in A}\left|\left|\mathbf{w}\left(\hat{t}^{(m)}_{\mu}[\omega]\right)\circ\left[\hat{p}^{(m)}_{\mathcal{A},\mu}[\omega]-\hat{t}^{(m)}_{\mu}[\omega]\right]\right|\right|_2^2\leq Q \quad \forall\, m\in\mathcal{A} \\
& \hat{p}^{(m)}_{\mathcal{A},\mu}[\omega]=\text{Windowing, Zero-pad, DFT of } x^{(l)}_{\mathcal{A},\mu}[n] \quad \forall\, m\in\mathcal{A} \\
& \displaystyle\sum_{m\in B}\left|\left|\mathbf{w}\left(\hat{t}^{(m)}_{\mu}[\omega]\right)\circ\left[\hat{p}^{(m)}_{\mathcal{B},\mu}[\omega]-\hat{t}^{(m)}_{\mu}[\omega]\right]\right|\right|_2^2\leq Q \quad \forall\, m\in\mathcal{B} \\
& \hat{p}^{(m)}_{\mathcal{B},\mu}[\omega]=\text{Windowing, Zero-pad, DFT of } x^{(l)}_{\mathcal{B},\mu}[n] \quad \forall\, m\in\mathcal{B}
\end{array}
\tag{6.4}
$$

Here, we are minimizing the perceptually-weighted leakage, subject to constraints limiting the detectability of the reproduction error per point m. This allows for this trade-off.

7 Evaluation of Perceptual Sound Zone Algorithm

Skeleton of Chapter

This chapter aims to evaluate the previously derived algorithms by comparing the perceptual algorithms with the reference algorithm.

- Introduction to the comparison methodology
  - Use of STOI
  - Use of PESQ
  - Use of Distraction
- Evaluation of reference vs unconstrained problem
- Evaluation of reference vs constrained problem for different values of the constraint

8 Conclusion

Skeleton of Chapter

A conclusion about the work.

Bibliography

[1] T. Lee, J. K. Nielsen, and M. G. Christensen, “Signal-adaptive and perceptually optimized sound zones with variable span trade-off filters,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2412–2426, 2020.

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

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

[4] J. Donley, C. Ritz, and W. B. Kleijn, “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.