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

Something about soundzones, something about perceptual models, something about perceptual sound zones. MIM: 1) At which point do you introduce the Data model? 2) It's OK to push things to the future work. PMN: 1) What is the story? In two sentances. Elevator Pitch. Helps you prioritize. 2) What is it that I'm doing differently. 3) Problems with Tae-Woongs work

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

Something about soundzones, something about perceptual models, something about perceptual sound zones. MIM: 1) At which point do you introduce the Data model? 2) It's OK to push things to the future work. PMN: 1) What is the story? In two sentances. Elevator Pitch. Helps you prioritize. 2) What is it that I'm doing differently. 3) Problems with Tae-Woongs work

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

TODO: I am considering merging some chapters (e.g. merging the literature review and the implementation chapters).

List of Figures

List of Tables

1 Introduction

Thesis Outline

I envision the outline of my thesis as follows:

- **Introduction:**
  - Introduction to sound zones:
    - Explain sound zones, 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
    - 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:
  - Give structure of the document
- **Review of Sound Zones**
  - Formal introduction to the sound zone problem.
    - Introduction of room model:
      - Room
      - Zones
      - Loudspeakers
    - Goal of sound zone algorithm:
      - Attaining Target Sound Pressure
    - Way of attaining goal:
      - Controlling loudspeaker inputs
      - Relation sound pressure and loudspeaker inputs

  - Short review of prior work, typical approaches.
    * Pressure Matching
    * Acoustic Contrast Control
    * Mode Matching
- **Literature Review of Perceptual Models for use in Sound Zones**
  - Criteria for perceptual models for sound zones
    * Easy to optimize for
    * Feasible to compute in real time
  - Masking Models
    * Par Distraction Models
    * Taal Distraction Models
    * MPEG Model I and II
    * Dau Model
  - Objective Measures
    * Distraction Model
    * Speech Metrics
      · STOI
      · SIIB
      · PESQ
    * Audio Metrics
      · ViSQOL
      · PEAQ
  - Selection of perceptual model
    * Par Distraction selected.
      · Convex, easy to optimize for.
      · Feasible to compute in real time.
    * Note that other models can be used for evaluation.
- **Analysis of Detectability for Sound Zone Algorithms**
  - Detailed discussion of Detectability
    * Psycho-acoustical background
    * Implementation details

        - Constructing frequency domain weights
        - Calibration of the model
    - Discussion of which Sound Zone Approach best suited for Detectability
        - Review of discussed sound zone algorithms with respect to Detectability
            - Mode Matching, Pressure Matching in a different domain.
            - Acoustic Contrast Control, no content control.
            - Pressure Matching, Detectability is an alternative to squared error.
        - Conclusion that Pressure Matching is the best
- **Implementation of Detectability Minimization Algorithms**
    - Derivation of the Reference Pressure Matching Approach
    - Derivation of the Detectability Minimization Approach
    - Derivation of the Detectability Constraining Approach

Review of Perceptual Model Literature

2

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
  - Detectibility 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 the chosen sound zone approach (pressure matching) and the chosen perceptual model (detectibility). This is done by means of summarizing the findings, and then reflecting on the criteria. From this, I will conclude that the **Par Detectibility** is best suited.
- I will discuss what perceptual models will be used for evaluation. From this I will conclude that PEAQ / VISQOL are useful for quality evaluation, and that the distraction model is useful for leakage evaluation.

2.1 Introduction

Implementation of Perceptual Model

3

Skeleton of Chapter

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

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

3.1 Introduction

Review of Sound Zone Algorithms Literature

4

Skeleton of Chapter

At this point the Par Detectibility 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 Par Detectibility.

- I will define the criteria that are required for a sound zone algorithm.
  - I am yet to think of any.
- I will discuss my literature review into sound zones.
  - Pressure Matching (PM) Approaches.
  - Acoustic Contrast Control (ACC) Approaches.
  - Mode Matching Approaches.
  - The work by Tae-Woong Lee.
- I will summarize the results and reflect on the requirements. From this I will conclude that the pressure matching approach is the most suited.

4.1 Introduction

Implementation of Reference Sound Zone Algorithm

5

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.

1. To form the basis for the perceptual algorithm.
2. To function as a reference implementation to evaluate performance.

The chapter will discuss the following

- Introduction of a data model from which the sound zone problem can be stated mathematically.
- Derivation of a multi-zone pressure matching approach from the previously introduced mathematical framework.
- Extension of the multi-zone pressure matching to perform on a frame by frame basis. This is done in order to allow for real-time sound zones.

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. This is followed by subsection 5.2.3, which discusses how the loudspeaker array can be used to realize the target sound pressure.

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 can be reframed as finding loudspeaker input signals such that a 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.

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

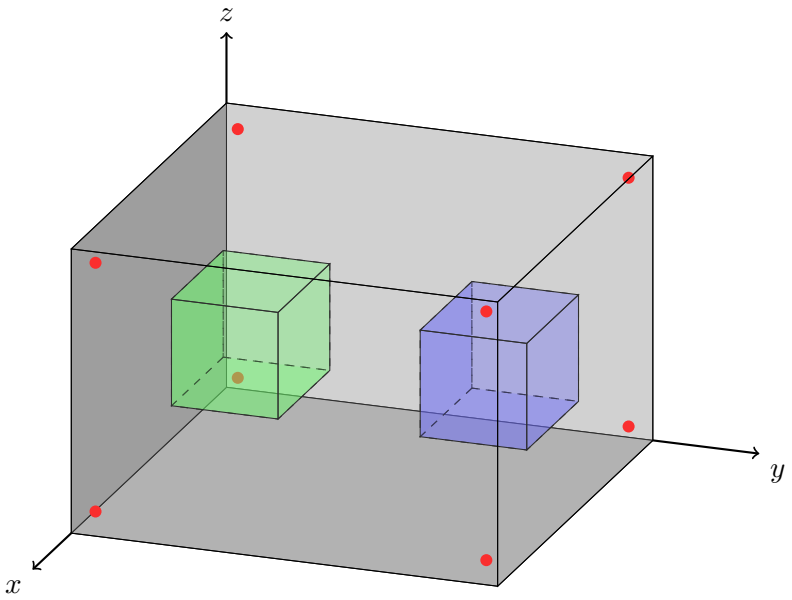

Figure 5.1: The room $\mathcal{R} \subset \mathbb{R}^3$ containing the zones $\mathcal{A} \subset \mathcal{R}$ and $\mathcal{B} \subset \mathcal{R}$ depicted in green and blue respectively. The room contains $N_L = 8$ loudspeakers, which are denoted by the red dots in the corners of the room.

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

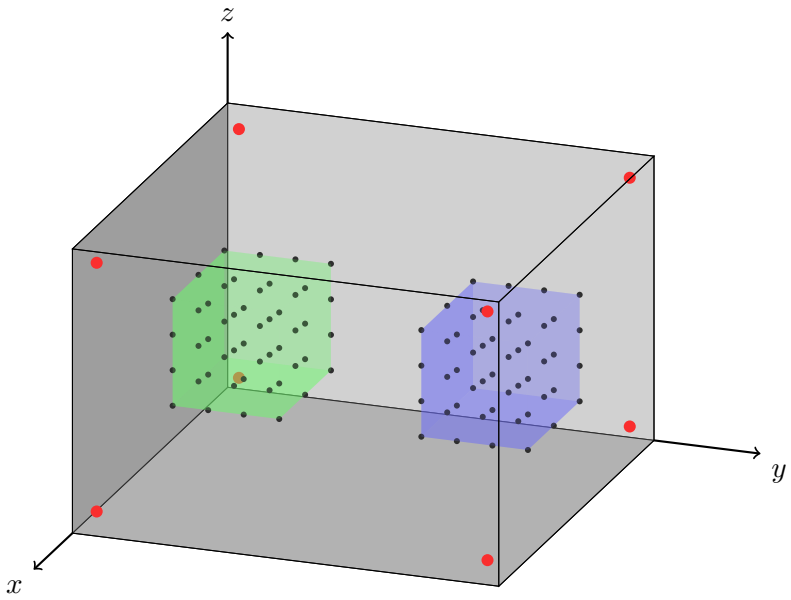

Figure 5.2: The previously introduced room $\mathcal{R}$ with zones $\mathcal{A}$ and $\mathcal{B}$ discretized.

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

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 loudspeaker input 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 target sound pressure can be understood as the sound pressure that results when separately playing the selected loudspeaker input signals $s_{\mathcal{A}}[n]$ and $s_{\mathcal{B}}[n]$ through the loudspeakers in the room.

In other words, the sound pressure in a zone is chosen such that it matches the sound pressure arises in said zone when playing only the desired input signal. For example, when in zone A, the target sound pressure is set equal to the sound pressure corresponding to playing only $s_{\mathcal{A}}[n]$ from the loudspeakers. 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 discussing it 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 information 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. 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 are decomposed into two parts as follows:

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

Here, $x^{(l)}_{\mathcal{A}}[n]$ and $x^{(l)}_{\mathcal{B}}[n]$ are the parts of the loudspeaker input signal responsible for reproducing the target sound pressure in zone $\mathcal{A}$ and $\mathcal{B}$ respectively. Now, it is possible to consider the sound pressure that arises due to the separate loudspeaker input signals:

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

$$\tag{5.8}$$

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

The idea in this approach is to chose $x^{(l)}_{\mathcal{A}}[n]$ and such that the resulting pressure $p^{(m)}_{\mathcal{A}}[n]$ attains the target sound pressure $t^{(m)}[n]$ in all $m \in A$. At the same time however, $p^{(m)}_{\mathcal{A}}[n]$ should not result in any sound pressure in all $m \in B$. Any sound pressure resulting from $x^{(l)}_{\mathcal{A}}[n]$ in zone $\mathcal{B}$ is essentially leakage, or cross-talk between zones. Similar arguments can be given for $x^{(l)}_{\mathcal{B}}[n]$: it should reproduce the target sound pressure for $m \in B$ but no sound pressure for $m \in A$.

In the MZ-PM approach, the loudspeaker weights $x^{(l)}_{\mathcal{A}}[n]$ and $x^{(l)}_{\mathcal{B}}[n]$ that achieve this goal are found by minimizing the difference between the intended pressure and the realized pressure as follows:

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

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

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 and leakages. Some pressure matching approaches attempt to control this trade-off by introducing weights for the different error terms, or constraints. Choosing constraints can however be challenging as the mean square pressure error is difficult to interpret.

The optimization problem can be solved in the time and the frequency domain. In frequency domain approaches, the convolutions become inner products, which typically results in a lower computational complexity.

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. This results in a latency of at least H, which could be acceptable assuming H is chosen sufficiently small.

In addition to the benefit of block-based processing, 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 processing 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 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 $s_{\mathcal{A}}[m]$ and $s_{\mathcal{B}}[m]$ are known.

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 to attain 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 $x_{\mathcal{Z}}^{(l)}[n]$:

$$p_{\mathcal{Z}}^{(m)}[n] = \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * x_{\mathcal{Z}}^{(l)} \right)[n] \tag{5.15}$$

$$= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] x_{\mathcal{Z}}^{(l)}[b] \tag{5.16}$$

$$= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] x_{\mathcal{Z}}^{(l)}[b] \sum_{k=-\infty}^{\infty} w[b-kH] \tag{5.17}$$

$$= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\infty} x_{\mathcal{Z}}^{(l)}[b] w[b-kH] \tag{5.18}$$

$$\tag{5.19}$$

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 non-causal. It is chosen such that 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 overlapping windows $w[n]$. The hop-size is chosen to be equal to the block size H, as such the overlap is equal to $N_w - H$. This forms individual frames $x_{\mathcal{Z}}^{(l)}[n]w[n-\mu H]$, which have support $-N_w+1+\mu H \leq n \leq \mu H$. Due to the properties of the COLA condition, the individual frames 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 frame $x_{\mathcal{Z}}^{(l)}[n]w[n-kH]$. When the μ^{th} block of $s_{\mathcal{Z}}[n]$ is revealed, this allows for an update to target sound pressure (how this is done is discussed later). The idea is then to compute the $x_{\mathcal{Z}}^{(l)}[n]w[n-\mu H]$ frame 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:

$$t^{(m)}[n] = \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * s_{\mathcal{Z}}\right)[n] \tag{5.22}$$

$$= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{Z}}[b] \tag{5.23}$$

$$= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{Z}}[b] \sum_{k=-\infty}^{\infty} w[b-kH] \tag{5.24}$$

$$= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{\mu=-\infty}^{\infty} s_{\mathcal{Z}}[b] w[b-\mu H] \tag{5.25}$$

$$\tag{5.26}$$

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 blocks. Note also that the windowed blocks need not be overlapping in general, i.e. one possible choice of window is the rectangular window H with $N_w = H$.

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

$$t_\mu^{(m)}[n] = \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\mu} s_{\mathcal{Z}}[b] w[b-\mu H] \tag{5.27}$$

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

$$= \sum_{l=0}^{N_L-1} \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{Z}}[b] w[b-\mu H] + t_{\mu-1}^{(m,l)}[n] \tag{5.29}$$

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

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:

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

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 $t_{\mu}^{(m)}[n]$.

As $t_{\mu}^{(m)}[n]$ is shown to converge to the true sound pressure, the sum of the all contributions of $x_{\mathcal{Z},\mu}^{(l)}[n]$ is thus chosen to best attain it. How this is done exactly is discussed in the next section.

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 frames $x_{\mathcal{Z},\mu}^{(l)}[n]$ be denoted by $p_{\mathcal{Z},\mu}^{(m)}[n]$, which 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 at a time, all frames $k \leq \mu - 1$ are assumed constant. Note however that $x_{\mu}^{(l)}[n]$ can only influence samples $\mu H - N_w + 1$ due to its finite support.

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. As such, the block based approach had a dual purpose.

However, there is 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 give 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).

TODO: Something about zero padding

5.5.1 Proposed Frequency Domain Approach

TODO: Add zero padding... Kinda tricky but can be done elegantly.

$$\underset{x_{\mathcal{A},\mu}^{(l)}[n],\, x_{\mathcal{B},\mu}^{(l)}[n]\,\forall\, l}{\arg\min} \quad \sum_{m\in A}\left|\left|\hat{p}_{\mathcal{A},\mu}^{(m)}[\omega]-\hat{t}_{\mu}^{(m)}[\omega]\right|\right|_2^2+\sum_{m\in A}\left|\left|\hat{p}_{\mathcal{B},\mu}^{(m)}[\omega]\right|\right|_2^2+ \tag{5.36}$$

$$\sum_{m\in B}\left|\left|\hat{p}_{\mathcal{B},\mu}^{(m)}[\omega]-\hat{t}_{\mu}^{(m)}[\omega]\right|\right|_2^2+\sum_{m\in B}\left|\left|\hat{p}_{\mathcal{A},\mu}^{(m)}[\omega]\right|\right|_2^2 \tag{5.37}$$

$$\text{subject to} \quad \hat{p}_{\mathcal{A},\mu}^{(m)}[\omega]=\sum_{l=0}^{N_L-1}\hat{h}^{(l,m)}[\omega] \tag{5.38}$$

Perceptual Sound Zone Algorithm

6

Skeleton of Chapter

Describes the design of the perceptual sound zone algorithm.

- Introduction of detectibility to form perceptual cost function building blocks:
  - Detectibility of reproduction error
  - Detectibility of interference
- Show how building blocks can be used to form different perceptual algorithms:
  - Algorithm 1: Minimization of detectibility of reproduction error and detectibility of interference.
  - Algorithm 2: Minimization of detectibility of reproduction error subject to constraint on detectibility of interference.
  - Algorithm 3: Minimization of detectibility of interference subject to constraint on detectibility of reproduction error.
- Evaluate different algorithms:
  - Compare Reference and Algorithm 1 in terms of Distraction and PEAQ.
  - Compare Reference and Algorithm 2 when varying the constraint on the detectibility of the interference in terms of Distraction and PEAQ.
  - Compare Reference and Algorithm 3 when varying the constraint on the detectibility of the reproduction error in terms of Distraction and PEAQ.

6.1 Introduction

7 Conclusion

Skeleton of Chapter

A conclusion about the work.

8 Bibliography

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.