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.

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. Massimo Mastrangeli

dr. J. Martinez Castañeda

Abstract

Something about soundzones, something about perceptual models, something about perceptual sound zones.

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

An introduction to the work and the problem it seeks to solve.

- An introduction to sound zones as a concept.
- Where sound zones come short, motivation to use perceptual models.
- Statement of the goal of the project: creating a perceptual multi-zone algorithm.
- A description of what the rest of the document will contain.

Literature Review 2

Skeleton of Chapter

In order to build a perceptual sound zone algorithm, we review literature for both sound zones and perceptual models. This chapter will also motivate what sound zone techniques and perceptual models will be used in the design of the perceptual sound zone algorithm.

- I will discuss my literature review into sound zones.
  - Pressure Matching Approaches
  - Acoustic Contrast Approaches
  - Mode Matching Approaches
- I will discuss my literature review into perceptual models.
  - Dau Model
  - Detectibility Models, i.e. Par and Taal
  - Distraction Model
  - Speech Intelligibility Based, i.e. SIIB and STOI
- I will briefly discuss the prior work done on perceptual sound zones.
  - The work by Tae-Woong Lee.
- I will discuss and motivate the chosen sound zone approach (pressure matching) and the chosen perceptual model (detectibility).

3 Perceptual Models

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 mention that there are two implementations in literature: Par and Taal. I mention the trade-offs between the two.
- I describe the Par Detectibility.
  - The underlying perceptual ideas
  - How its implemented
  - How its calibrated
- I describe the Taal Detectibility.
  - The underlying perceptual ideas
  - How its implemented
  - How its calibrated
- I compare the Par and Taal Detectibility.
- I show that my implementations of the Par and Taal Detectibility are valid. This is done by comparing the masking curve predictions to a reference implementation of the Dau model.

4 Perceptual Sound Zones

Skeleton of Chapter

Describes the design of the perceptual sound zone algorithm.

- Introduction of the data model
- Derivation of the frame-based processing framework
- 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 0: Reference Algorithm
  - 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 Algorithm 0 and Algorithm 1 in terms of Distraction and PEAQ.
  - Compare Algorithm 0 and Algorithm 2 when varying the constraint on the detectibility of the interference in terms of Distraction and PEAQ.
  - Compare Algorithm 0 and Algorithm 3 when varying the constraint on the detectibility of the reproduction error in terms of Distraction and PEAQ.

4.1 Data Model

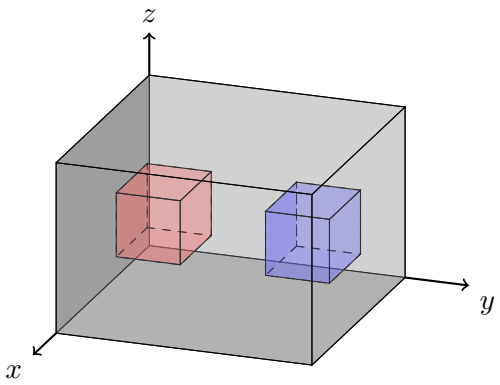
TODO: CITATION NEEDED, I have not cited anything. Probably should? In this section the base data model will be introduced. This data model will be later used in the derivation of the sound zone algorithms.

Firstly, in subsection 4.1.1 a spatial description of a room will be given. This description will include the the contents of the room, namely the zones and loudspeakers contained within. Finally, after introducing the room, the sound zone problem can be more formally defined.

4.1.1 Room Model and Sound Zone Problem Statement

TODO: I am not great at topology, so I would love some tips on how to write this down in a better way... 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 work can however be extended to a larger number of zones). This description can then be used to pose the sound zone problem in a more formal way.

In general, 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 goal of the sound zone algorithm is to use the loudspeakers to realise a specified target sound pressure in the space defined 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 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. First, a way of modeling the target sound pressure will be discussed. Afterwards, a way of realizing said target sound pressure by controlling the loudspeaker inputs will be given

mathematically. Finally, the data model is used to state the goal of the sound zone algorithm more formally.

4.1.1.1 Defining Target Sound 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 continious 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.

In this work, the latter approach will be taken. **TODO: Why does discretizing make sense? Add more reasons...** 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.

TODO: Update the previously introduced graphic to include the control points

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.

4.1.1.2 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 realised sound pressure in control point m due to playing from loudspeaker l.

$$p^{(l,m)}[n] = \left(h^{(l,m)} * x_l\right)[n] \tag{4.1}$$

The realised 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:

$$p^{(m)}[n] = \sum_{l=0}^{N_L} p^{(l,m)}[n] \tag{4.2}$$

$$= \sum_{l=0}^{N_L} \left(h^{(l,m)} * x_l\right)[n] \tag{4.3}$$

4.1.1.3 Sound Zone Problem Statement

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 realised sound pressure $p^{(m)}[n]$ attains the target sound pressure $t^{(m)}[n]$ for all control points $m \in A \cup B$.

The rest of section 4.1 will describe how this problem can be solved in greater detail. First, subsection 4.1.2 discusses how the target sound pressure will be chosen. Afterwards, subsection 4.1.3 will discuss the Pressure Matching (PM) approach that can be used to solve the stated sound zone problem. This approach will form the foundation on which the perceptual sound zone algorithms will be constructed.

4.1.2 Choice of Target Pressure

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.

The target sound pressure $t^{(m)}[n]$ describes the desired content for a specific control point m. 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 intend to hear in zone $\mathcal{A}$ and $\mathcal{B}$ respectively.

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

$$t^{(m)}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * s_{\mathcal{A}}\right)[n] \qquad \forall\, m \in A \tag{4.4}$$

$$t^{(m)}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * s_{\mathcal{B}}\right)[n] \qquad \forall\, m \in B \tag{4.5}$$

This choice for target sound pressure can be understood as the sound pressure that results when playing the selected loudspeaker input signals $s_{\mathcal{A}}[n]$ and $s_{\mathcal{B}}[n]$ seperately using the loudspeakers. 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. Similarly, when in zone B, the target sound pressure is set to the pressure that arises from playing only $s_{\mathcal{B}}[n]$.

The motivation for choosing this target is that it physically attainable with the given loudspeakers and room. **TODO: Expand this motivation with a couple more arguments.**

4.1.3 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. **TODO: Expand this motivation with a couple more arguments...?**

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:

$$x_l[n] = x_{l,\mathcal{A}}[n] + x_{l,\mathcal{B}}[n] \tag{4.6}$$

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 seperate loudspeaker input signals:

$$p_{\mathcal{A}}^{(m)}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * x_{l,\mathcal{A}}\right)[n] \tag{4.7}$$

$$p_{\mathcal{B}}^{(m)}[n] = \sum_{l=0}^{N_L} \left(h^{(l,m)} * x_{l,\mathcal{B}}\right)[n] \tag{4.8}$$

Here, $p_{\mathcal{A}}^{(m)}[n]$ and $p_{\mathcal{B}}^{(m)}[n]$ can be understood to be the pressure that arrises 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 $p_{\mathcal{A}}^{(m)}[n]$ 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 attain 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:

$$\underset{x_{l,\mathcal{A}}[n],\, x_{l,\mathcal{B}}[n]\,\forall\, l}{\arg\min} \quad \sum_{m\in A} \left|\left| p_{\mathcal{A}}^{(m)}[n] - t^{(m)}[n] \right|\right|_2^2 + \sum_{m\in A} \left|\left| p_{\mathcal{B}}^{(m)}[n] \right|\right|_2^2 + \tag{4.9}$$

$$\sum_{m\in B} \left|\left| p_{\mathcal{B}}^{(m)}[n] - t^{(m)}[n] \right|\right|_2^2 + \sum_{m\in B} \left|\left| p_{\mathcal{A}}^{(m)}[n] \right|\right|_2^2 \tag{4.10}$$

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 reproduciton error and leakage for zone $\mathcal{B}$. Typically, 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.

The 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 introdued in the following sections.

4.2 Frame-Based Processing Framework

It is assumed that the desired playback signals $s_{\mathcal{A}}[n]$ and $s_{\mathcal{B}}[n]$ are not known in their entirety at a given time n. The motivation for this is that a user can change the desired playback content in real-time. As such, the system should be able to accomodate for this. **TODO: This really needs better motivation...**

It is assumed that $s_{\mathcal{A}}[m]$ and $s_{\mathcal{B}}[m]$ are known from $\infty \leq m \leq \mu H$, where $\mu = \lfloor n/H \rfloor$. This can be interpreted as the desired playback signals is revealed in blocks of size H, and that at a time $n = \mu H$, the μ^{th} block is known.

This assumption as a number of implications on the previously derived equations.

4.2.0.1 Implications for Computing Target Pressure

As the desired playback signals are not known entirely, the target can also not be computed enitrely as before. As the playback signals are revealed in blocks of size H, it makes sense to update the target signal in blocks of H.

Consider the following rewrite of the target signal for the desired signal of zone $\mathcal{A}$:

$$t^{(m)}[n] = \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * s_{\mathcal{A}}\right)[n] \tag{4.11}$$

$$= \sum_{l=0}^{N_L-1} t^{(l,m)}[n] \tag{4.12}$$

Here, $t^{(m,l)}[n]$ is the contribution of the l^{th} loudspeaker to the target sound pressure at reproduction point m. Consider the following rewrite:

$$t^{(m,l)}[n] = \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{A}}[b] \tag{4.13}$$

$$= \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{A}}[b] \sum_{k=-\infty}^{\infty} w[b-kH] \tag{4.14}$$

$$= \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\infty} s_{\mathcal{A}}[b] w[b-kH] \tag{4.15}$$

$$= \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\infty} s_{\mathcal{A},k}[b] \tag{4.16}$$

Here, $w[n] \in \mathbb{R}^{N_w}$ is a window that satisfies the COLA condition for a hopsize H. The window is defined to be non-zero for $-N_w + 1 \leq n \leq 0$, as such it is non-causal. Furthermore, the windows are overlapping, thus $N_w > H$.

In the rewrite above, the desired playback signal $s_{\mathcal{A}}[n]$ is projected onto a basis of overlapping frames of size N_w. The projection results in a sum of frames $s_{\mathcal{A},k}[n]$. The

support of $s_{\mathcal{A},k}[n]$ is deifned by the shifted window that is used to synthesise it, i.e. it is non-zero for $-N_w+1+kH \leq n \leq kH$. As the COLA condition is met for the chosen window, the sum over all frames reconstructs $s_{\mathcal{A}}[n]$ perfectly.

At a time $n = \mu H$, the frames up to $k = \mu$ can be computed. Let $t_{\mu}^{(m,l)}[n]$ represent the target using frames up to $k = \mu$:

$$t_{\mu}^{(m,l)}[n] = \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\mu} s_{\mathcal{A},k}[b] \tag{4.17}$$

$$= \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{A},\mu}[b] + \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\mu-1} s_{\mathcal{A},k}[b] \tag{4.18}$$

$$= \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] s_{\mathcal{A},\mu}[b] + t_{\mu-1}^{(m,l)}[n] \tag{4.19}$$

As can be seen, $t_{\mu}^{(m,l)}[n]$ can expressed as the contribution of the current frames and the contribution of all previous frames. In addition, note how the computation can be performed recursively. To compute $t_{\mu}^{(m,l)}[n]$, we compute the convolution of the current frame $s_{\mathcal{A},\mu}[n]$ with the RIRs, and then add the history of previous frames.

Thus, $t_{\mu}^{(m,l)}[n]$ can be considered an estimation of the target given frames up to μ. As new frames are revealed, the target estimation can be updated. Note that this definition converges to the "true" target estimation: $t_{\infty}^{(m,l)}[n] = t^{(m,l)}[n]$.

Essentially, this approach allows for the real-time computation for the target signal.

4.2.0.2 Implications for Computing Loudspeaker Inputs

Just as the target is comuted as new frames are revealed, the loudspeaker input should also be computed this way. **TODO: I need to motivate this?** Consider the following rewrite of the realized sound pressure $p^{(l)}[n]$:

$$p^{(m)}[n] = \sum_{l=0}^{N_L-1} \left(h^{(l,m)} * x^{(l)} \right)[n] \tag{4.20}$$

$$= \sum_{l=0}^{N_L-1} p^{(l,m)}[n] \tag{4.21}$$

$p^{(m,l)}[n]$ is the contribution of the l^{th} loudspeaker to the realized sound pressure at reproduction point m. Consider the following:

$$p^{(m,l)}[n] = \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b]x^{(l)}[b] \tag{4.22}$$

$$= \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b]x^{(l)}[b] \sum_{k=-\infty}^{\infty} w[b-kH] \tag{4.23}$$

$$= \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\infty} x^{(l)}[b]w[b-kH] \tag{4.24}$$

$$= \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\infty} x_k^{(l)}[b] \tag{4.25}$$

Analagous to the derivation for the target sound pressure $t^{(m,l)}[n]$, the loudspeaker input signal is project onto a basis consisting of windows $w[n]$. This results in frames $x_k^{(l)}[n]$. Just as with the target sound pressure, let $p_\mu^{(m,l)}[n]$ represent the realised sound pressure using frames up to $k = \mu$:

$$p_\mu^{(m,l)}[n] = \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\mu} x_k^{(l)}[b] \tag{4.26}$$

$$= \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b]x_\mu^{(l)}[b] + \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b] \sum_{k=-\infty}^{\mu-1} x_k^{(l)}[b] \tag{4.27}$$

$$= \sum_{b=n-N_h+1}^{n} h^{(l,m)}[n-b]x_\mu^{(l)}[b] + p_{\mu-1}^{(m,l)}[n] \tag{4.28}$$

As such, the realised pressure can be computed recursively just like the target sound pressure. Using this data model also allows for the recursive computation of loudspeaker frames $x_\mu^{(l)}[n]$. The loudspeaker input signals must be chosen such that sound pressure realized by them must approximate the target sound pressure.

This reveales one possible solution approach: computing the loudspeaker frames $x_\mu^{(l)}[n]$ such that the target sound pressure $t_\mu^{(l)}[n]$ is attained. **TODO: Expand on this. Why does this approach make sense? There are probably other ways of doing it aswell...**

5 Conclusion

Skeleton of Chapter

A conclusion about the work.

6 Bibliography