



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!



# 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**

Copyright © 2021 Circuits and Systems Group  
All rights reserved.

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

---

<b>Abstract</b>	<b>v</b>
<b>Acknowledgments</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Objectives and Organization . . . . .	3
1.2.1 Search and Implementation of a Perceptual Model . . . . .	3
1.2.2 Search and Implementation of a Sound Zone Approach . . . . .	3
1.2.3 Implementation of a Perceptual Sound Zone Algorithm . . . . .	4
<b>2 Perceptual Model Review and Implementation</b>	<b>5</b>
2.1 Introduction . . . . .	5
2.2 Perceptual Background . . . . .	6
2.2.1 Workings of the Ear . . . . .	6
2.2.2 Auditory Masking . . . . .	6
2.2.3 Threshold of Quiet . . . . .	6
2.3 Review of Perceptual Models . . . . .	7
2.3.1 Objective Measures . . . . .	7
2.3.2 Audio Coding Models . . . . .	10
2.4 Selection of Perceptual Model . . . . .	13
2.4.1 Criteria for Selecting Perceptual Model . . . . .	13
2.4.2 Selection of Perceptual Model . . . . .	13
2.5 Implementation of Selected Perceptual Model . . . . .	15
2.5.1 Introduction to Par Detectability . . . . .	15
2.5.2 Computation Details of the Par Detectability . . . . .	15
2.5.3 Calibration of the the Par Detectability . . . . .	16
2.6 Conclusion . . . . .	18
<b>3 Sound Zone Approach Review and Implementation</b>	<b>19</b>
3.1 Introduction . . . . .	19
3.2 Review of Sound Zone Approaches . . . . .	20
3.3 Selection of Sound Zone Approach . . . . .	21
3.4 Data Model . . . . .	22
3.4.1 Room Topology . . . . .	22
3.4.2 Defining Target Pressure . . . . .	23
3.4.3 Realizing Sound Pressure through the Loudspeaker . . . . .	23
3.4.4 Choice of Target Pressure . . . . .	24
3.5 Implementation of Selected Sound Zone Algorithm . . . . .	26
3.6 Block Based Approach . . . . .	28
3.6.1 Modeling Block Based Knowledge Limitations . . . . .	28
3.6.2 Block Based Loudspeaker Input Signal Computation . . . . .	30

3.6.3	Block Based Sound Pressure . . . . .	30
3.6.4	Derivation of Block-Based Multi-Zone Pressure-Matching . .	31
3.7	Conclusion . . . . .	33
<b>4</b>	<b>Implementation and Evaluation of a Perceptual Sound Zone Algorithm</b>	<b>35</b>
4.1	Introduction . . . . .	35
4.2	Frequency Domain Conversion . . . . .	36
4.2.1	Frequency Domain Transformation . . . . .	36
4.2.2	Quantities the Frequency Domain . . . . .	37
4.2.3	Proposed Frequency Domain Approach . . . . .	37
4.3	Unconstrained Perceptual Pressure Matching . . . . .	39
4.4	Perceptual Minimization Approach . . . . .	39
4.5	Constrained Perceptual Pressure Matching . . . . .	40
4.6	Perceptual Constraining Approach . . . . .	40
4.7	Results . . . . .	41
4.8	Conclusion . . . . .	42
<b>5</b>	<b>Conclusion</b>	<b>43</b>

# Introduction

---

## 1.1 Motivation

Sound systems are used world wide to fill rooms with enjoyable audio content. Problems arise however when multiple people in the same room want to enjoy different audio content at the same time.

For example, one person may want to enjoy a show, while an other may want to listen to their music. If they are in the same room, their desires clash: neither person can fully enjoy their chosen activity without disturbing the other. In short, the interference of multiple source of audio leads to a situation where both individual experiences are diminished.

In recent years, attempts have been made to solve this problem by controlling the spatial reproduction of sound in such a way that different areas in a room have different sound content.

One class of algorithms that attempt to do so are known as sound zone algorithms [1]. Sound zone algorithms partition the space of the room into multiple so-called sound zones. Each sound zone is assigned different audio content.

The sound zone algorithms decides how to use the sound system to reproduce audio content in each zone. Using the principals of constructive and destructive interference, this is done in such a way that minimal content of one zone is audible in the others.

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

In practice however, sound zone algorithms do not always do a perfect job.

The performance of algorithms depends on the environment and the available sound system. Depending on the situation, the interference between zones can typically only be reduced by so much. As such, audio content of one zone is often still audible in other zones. In addition to this, reducing interference between zones can also come at the cost of reducing quality of the reproduced audio in the zones.

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

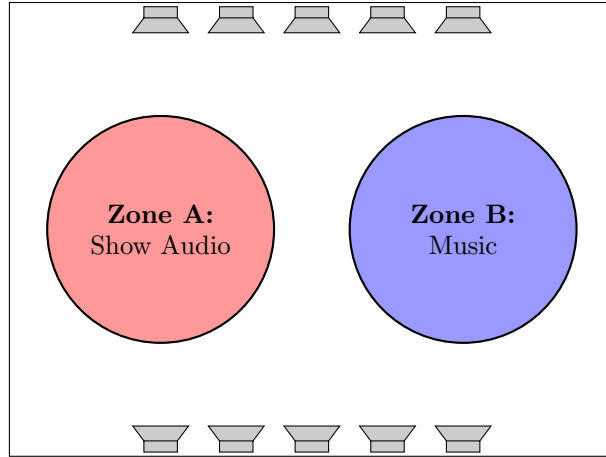


Figure 1.1: A room containing a sound system consisting of an array of loudspeakers and two zones. The goal of the sound zone algorithm is to control the sound system in such a way that the red zone contains the audio of a show, and the blue zone contains the music.

sound pressure, which doesn't always capture what is important for the perception of sound. As such, including a perceptual model may allow the algorithm to focus on the parts of the audio content that matter perceptually.

Early results show that the perceptual sound zone approach is promising. Recent work by Donley et al. explored including the absolute threshold of hearing, which models the lowest sound pressure humans can hear, into sound zone algorithms. In this pursuit it was found that the quality of the reproduced audio in the zones [2]. Later, Lee et al. showed that including a perceptually-motivated weighting in the sound zone algorithm outperforms traditional algorithms [3, 4].

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

## 1.2 Objectives and Organization

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

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

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

### 1.2.1 Search and Implementation of a Perceptual Model

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

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

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

### 1.2.2 Search and Implementation of a Sound Zone Approach

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

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

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

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

reference implementation with which the perceptual sound zone algorithm can be compared.

### **1.2.3 Implementation of a Perceptual Sound Zone Algorithm**



# Perceptual Model Review and Implementation

---

# 2

## 2.1 Introduction

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

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

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

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

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

## **2.2 Perceptual Background**

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

### **2.2.1 Workings of the Ear**

### **2.2.2 Auditory Masking**

### **2.2.3 Threshold of Quiet**

## 2.3 Review of Perceptual Models

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

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

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

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

### 2.3.1 Objective Measures

In order to determine the perceived quality of audio one approach is to use listening tests. These are tests in which subjects are asked to rate a property (or properties) of a set of audio stimuli. One example where these tests are performed is for the evaluation of listening aids, where they are used determine the speech intelligibility [5]. Other examples include determining which loudspeaker has the best perceived sound quality.

Performing listening tests is however often cumbersome due to the large amount of human labour involved. This motivates the use of objective quality measures, which attempt to predict the outcomes of these listening tests. This is very useful for algorithm developers for example, as they can get an indication of how well they are doing without having to perform a cumbersome test. Note however that a objective quality measure does not replace a listening test: it can only be used to give an indication.

The objective measure that will be considered take a reference and degraded audio stimuli as inputs. The quality of the degraded audio stimuli is then determined by the measure. For example, a speech intelligibility measure will determine the intelligibility of a degraded audio file.

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

correlates with audio quality, optimizing over such a measure could improve sound zone algorithms.

As such, this section will explore various objective measures. This will be done by considering various different objective measure classes.

#### 2.3.1.1 Objective Speech Quality

There have been a number of attempts to create objective measures to quantify the quality of speech. In this section three objective speech quality measures will be discussed. Namely the Perceptual Evaluation of Speech Quality (PESQ) [6] measure, Perceptual Objective Listening Quality Assessment (POLQA) [7] measure, and Virtual Speech Quality Objective Listener (ViSQOL) [8, 9] measure.

PESQ is a metric which attempts to determine the perceived quality of speech [6], and was standardized by the International Telecommunication Union (ITU-T) in 2001. PESQ is computed by first applying an auditory transform that maps the reference and degraded speech into a time-frequency representation of the perceived loudness. So-called symmetric and So-called symmetric and asymmetric disturbances are determined between the time-frequency bins of the reference and degraded speech. A non-linear average is then taken over frequency bins the resulting in an average disturbances per time bin. This is then mapped by a linear output mapping to correspond to listening test outcomes.

POLQA is another speech quality metric, and can be considered a successor to PESQ. It was standardized by the International Telecommunication Union (ITU-T) in 2011. It is meant to be the successor of PESQ, with the intention of having more accurate predictions on a wider range of distortions. POLQA works similarly to PESQ in that it also determines an internal representation of human auditory perception of the clean reference speech signal and the distorted speech signal. POLQA differs from PESQ in that it is designed to be capable of handling global temporal compression and expansions.

ViSQOL is a metric developed in collaboration with Google. In contrast to the previous methods, which relied on internal human auditory system representations of the reference and degraded speech inputs, ViSQOL primarily uses the Neurogram Similarly Index Measure (NSIM) to make its predictions. Neurograms contain the neural firing activity of the auditory nerve in time-frequency bins. NSIM determines how similar the firing patterns of two neurograms are. This similarity is then related to the outcomes of listening tests through a laplacian fit [8].

In general, PESQ, POLQA and ViSQOL require many steps to compute and are not easy to optimize for. Some attempts have been made to reformulate PESQ in order to make it more tractable for optimization by approximating the disturbances by other functions [10].

#### 2.3.1.2 Objective Speech Intelligibility

Intelligibility of speech can be understood as the percentage of words identified correctly given a degraded speech signal. In this section, two objective speech intel-

ligibility metrics will be discussed. Namely, the Short Time Objective Intelligibility (STOI) [5] measure and the Speech Intelligibility In Bits (SIIB) [11] measure.

STOI takes as input the clean and degraded speech signals. The clean and degraded speech signals are then converted into 1/3 octave bands, and then segmented into short time frames. The final output value of STOI then the average correlation coefficient between the clean and degraded segmented input signals, averaged over all bands and all segments.

SIIB is computed through the mutual information between a clean speech signal and the speech signal received by a listener. As such, the idea behind SIIB is that the intelligibility of speech is related to the information shared between intended and degraded speech [11]. In order to compute the mutual information, the paper models the transmission of an intended message from speaker to listener as a communication channel. Among other aspects, this transmission channel includes a model of the human auditory system.

Both STOI and SIIB are difficult to optimize for directly. For STOI, this is due to the removal of silent regions and the clipping operator are non-differentiable operations. Furthermore, the computation of the correlation coefficient is a non-convex function of the degraded speech.

SIIB is in general non-convex and non-differentiable as it uses a K-nearest neighbor estimator to compute the mutual information. However, if the communication channel is approximated as gaussian, the mutual information can be computed in closed form, and SIIB becomes a differentiable measure.

### 2.3.1.3 Objective Audio Quality Measures

The previous objective quality metrics were designed mainly for speech purposes. In this section, a number of objective quality metrics will be discussed that are designed for perceived audio quality. Namely, the Perceptual Evaluation of Audio Quality (PEAQ) [12], POLQA Music [13] and ViSQOLAudio [14]. The latter two are adapted versions of the previously discussed POLQA and ViSQOL speech quality measures. It was found that with some adjustments, the speech quality models could be used to determine audio quality.

PEAQ is a audio quality metric standardized by the International Telecommunication Union (ITU-T) [12]. PEAQ estimates a quality grade by first computing an internal representation based on the human auditory system of the reference and degraded audio signals. The differences between the two internal representations then results in a number of perceptually relevant features. PEAQ calls these features Model Output Variables (MOVs). These MOVs are then mapped to the final audio quality grade through a neural network.

PEAQ, POLQA Music and ViSQOLAudio are all difficult to optimize for.

#### 2.3.1.4 Distraction Model

One especially promising objective measure is the distraction proposed by Francome et al. in 2015 [?].

The distraction was determined to be the attribute that best describes the perceptual experience of interfering audio programs through an elicitation study performed in 2014 [?]. This prompted the creation of the model.

To create the model, a listening test was performed where the participants were subjected to audio-on-audio interference. The subjects were played a target signal to focus listening to [?]. At the same time, an interferer was played to distract the participant. The participants were given a scale between 0 and 100 on which they were asked to rate how distracting the interference was when listening to the target program. Here, a score of 100 is being very distracted.

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

Computing said features cannot not be performed in real time, as the original distraction model is too computationally complex. To this end, Rämö et al. proposed a version of the distraction model that could be run in real-time. This was done by approximating the features of the original distraction model by computationally less complex alternatives.

While easy to compute, the real-time distraction model by Rämö et al. is non-differentiable as the model uses piecewise functions and non-convex due to taking the logarithm of the square of the input signals.

### 2.3.2 Audio Coding Models

Audio coding algorithms attempt to find an low-bitrate representation of an audio input signal, as a form of compression. This process is usually lossy, as reducing the bitrate introduces errors.

These errors can be a detriment to the listening experience. As such, most audio coding algorithms use a perceptual model. The perceptual model is used to introduce encoding errors in such a way that the audio output signal is perceptually indistinguishable from the audio input signal [15].

The perceptual model typically takes form of a distortion function which determines how audible the difference between a reference input audio signal and a distorted output audio signal is. This function is used to encode an input audio signal such that it has minimal distortion for a specified bitrate.

The perceptual models used in audio coding are promising for integration into a sound zone algorithm, as they are often mathematically tractable. As such, this section will explore a number perceptual models from audio coding.

### 2.3.2.1 ISO MPEG Models

The ISO/IEC 11172-3 standard specifies a coded representation for audio files [16], and a decoder for said representation. An encoder said representation is not part of the standard. This is done deliberately, to allow for future improvements to the encoder, without having to change the standard [17].

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

The psycho-acoustical models work by subdividing the input audio signal into different frequency bands, modeling the frequency bands in the human auditory system. Separately per band, the model then determines how much quantization noise can be added without it becoming audible. As such, the model assumes that the distortion signal is noise-like [18].

The output of the psycho-acoustical model is thus the amount of noise that can be added per band. In the case of audio coding, this can then be used to control quantization noise. This technique has however also been used for various signal processing purposes, such as audio watermarking [15].

### 2.3.2.2 Par Detectability

In 2005, van der Par et al. proposed a novel perceptual model [18]. The perceptual model defines a distortion measure which determines the “detectability” of a distortion signal in presence of a masking signal. That is to say, how likely is a human to detect the distortion signal.

The proposed method differentiates itself from the previously discussed ISO MPEG models in two ways.

Firstly, the paper uses newer findings from psycho-acoustic literature, namely spectral integration. In spectral integration, the masking effects from neighboring bands are taken into account when computing the masking effects. The psycho-acoustical models defined in the ISO MPEG standard does not do this, and effectively works independently [15].

Secondly, it assumes that the distortion signal is sinusoidal, rather than noise-like. As such, it is more effective in hiding sinusoidal distortion.

On top of this, the proposed distortion measure can be expressed as an L2-norm. This mathematical tractability makes for easy integration into existing least-square problems [15].

The van der Par model has been used in many signal processing applications, examples ranging from speech enhancement to removing perceptually irrelevant sinusoidal components [19, 20].

### 2.3.2.3 Taal Detectability

A paper from 2012 by Taal et al. proposed a novel perceptual model [15]. The perceptual model also defines a distortion measure which determines the detectability, with identical interpretation from the Par Detectability model.

In contrast to the Par Detectability, the Taal Detectability measure takes temporal information into account. The inclusion of temporal information allows for the suppression of pre-echoes, an artifact introduced by assuming that the masking effects are stationary over the short time frame over which it operates.

In contrast to other temporal perceptual models, the Taal Detectability has a relatively low computational complexity. In addition to this, it can also be expressed as an L2-norm, which makes it a good candidate for optimization. The computational demand was however shown to be higher than the Par Detectability [15], especially for larger number of input samples.



## 2.4 Selection of Perceptual Model

In section 2.3 a literature review of various models of the human auditory system were given. In this section, one perceptual model will be selected for use in a sound-zone algorithm.

The other perceptual models will still be of use however during the evaluation of the results of the perceptual sound zone algorithm.

The approach to select a perceptual model is as follows. First, subsection 2.4.1 will discuss the criteria that will inform the decision. Next, subsection 2.4.2 will use these criteria to evaluate the perceptual models discussed in the literature review in section 2.3. From this evaluation the most suited perceptual model will be selected for use in a perceptual sound zone algorithm.

### 2.4.1 Criteria for Selecting Perceptual Model

This section will define desirable criteria for the perceptual model for integration with a sound zone algorithm. These criteria will be used to select a perceptual model from the models discussed in the literature review.

1. **Mathematical Tractability.**

It is a desirable property for the perceptual model to be easy to include in optimization problems. For example, perceptual models that can be differentiated are therefore favoured over optimization problems that are non-differentiable.

2. **Computational Overhead:**

It is desirable for the inclusion of the perceptual model to add minimal computational load to the sound zone algorithm. The additional overhead of the perceptual model should not increase the run time of the sound zone algorithm by many orders of magnitude.

### 2.4.2 Selection of Perceptual Model

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

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

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

can be added per frequency band, but does not provide an immediate means of using that information in optimization.

As such, the decision is between the Par and Taal detectability. As discussed, the Taal detectability takes into account more perceptual. This is however this has been shown to be at the cost of computational complexity. As such, the Par detectability will be selected for use in the sound zone algorithm.

## 2.5 Implementation of Selected Perceptual Model

In section 2.4, it was determined that the Par detectability was the perceptual model most suited for integration in a sound zone algorithm from the perceptual models considered in the literature review given in section 2.3. In this section, the perceptual model is considered in greater detail.

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

### 2.5.1 Introduction to Par Detectability

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

For the model to be accurate, the signals  $x[n]$  and  $\varepsilon[n]$  should be short-time signals. The paper uses a signal length  $N_x$  corresponding to between 20 to 200 milliseconds. This is important, as the model assumes that the psycho-acoustical properties of  $x[n]$  and  $\varepsilon[n]$  are stationary. Note that the Taal detectability does not make this assumption [15].

It is assumed that a human is listening to both the masking signal and the disturbance signal at the same time. The detectability can then be understood as the probability that a human listener can detect the disturbance signal  $\varepsilon[n]$  in presence of the masking signal  $x[n]$  [18]. The signal  $x[n] \in \mathbb{R}^N$  is referred to as the masking signal because it masks the disturbance signal  $\varepsilon[n]$  to a degree.

The metric is normalized in such a way that the detectability  $D(x[n], \varepsilon[n])$  is equal to 1 when the disturbance signal  $\varepsilon[n]$  is just noticeable in presence of masking signal  $x[n]$ . This can be understood as the probability of detecting  $\varepsilon[n]$  in presence of  $x[n]$  being equal to 1, or a 100%.

The detectability  $D(x[n], \varepsilon[n])$  can also attain a value larger than 1, which clashes with the interpretation of detectability as a probability. In fact, a detectability of larger than 1 will often be used in this work. The larger values of the detectability correspond with a more noticeable disturbance signal  $\varepsilon[n]$  in presence of masking signal  $x[n]$ .

### 2.5.2 Computation Details of the Par Detectability

This section will explore the Par detectability. Here, it will be shown how human auditory system is modeled and how the detectability is computed.

The first thing to note about the Par detectability is that it operates in the frequency domain version. To this end, let  $X[k]$  and  $\mathcal{E}[k]$  denote the frequency domain representations of the masking signal  $x[n]$  and the disturbance signal  $\varepsilon[n]$  respectively.

The Par detectability starts by computing an internal representation of the input signals  $X[k]$  and  $\mathcal{E}[k]$ . This internal representation models the filtering that occurs in the human auditory system. Note that the exact same filtering is applied to both  $X[k]$  and  $\mathcal{E}[k]$ .

The Par detectability models this filtering in two steps. The first step models how parts of the ear filter the incoming sound with an outer- and middle-ear filter  $H_{\text{om}}[k]$ . Next, a 4<sup>th</sup> order Gammatone filter bank is applied, modeling the filtering of the basilar membrane inside the ear [18]. The Gammatone filter bank consists of  $N_g$  filters, which are denoted by  $\Gamma_i[k]$

After filtering, the power per Gammatone filter tap is computed. Let  $M_i$  and  $S_i$  denote the output power of the  $i^{\text{th}}$  filter tap of the for  $X[k]$  and  $\mathcal{E}[k]$  respectively. The relationship between the input quantities and the output power of the filter taps can be given as follows:

$$M_i = \frac{1}{N_x} \sum_{k=0}^{N_x} |H_{\text{om}}[k]|^2 |\Gamma_i[k]|^2 |X[k]|^2 \quad (2.1)$$

$$S_i = \frac{1}{N_x} \sum_{k=0}^{N_x} |H_{\text{om}}[k]|^2 |\Gamma_i[k]|^2 |\mathcal{E}[k]|^2 \quad (2.2)$$

The output powers can then be used to define the within-channel detectability  $D_i$ . This can be thought of the detectability per filter tap. The within-channel detectability is defined as follows:

$$D_i = \frac{N_x S_i}{N_x M_i + C_a} \quad (2.3)$$

Here,  $C_a$  is a calibration constant that models the absolute threshold of hearing. This can be understood by considering the case where  $M_i = 0$ .

The total detectability  $D(x[n], \epsilon[n])$  can then be computed by combining the within channel detectabilities. It is defined as follows:

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

$$= C_s L_{\text{eff}} \sum_{i=0}^{N_g} \frac{\sum_{k=0}^{N_x} |H_{\text{om}}[k]|^2 |\Gamma_i[k]|^2 |\mathcal{E}[k]|^2}{\sum_{k=0}^{N_x} |H_{\text{om}}[k]|^2 |\Gamma_i[k]|^2 |X[k]|^2 + C_a} \quad (2.5)$$

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

### 2.5.3 Calibration of the the Par Detectability

This section will describe the calibration of the Par Detectability. As mentioned, a just noticeable disturbance signal must result in a detectability  $D(x[n], \epsilon[n])$  of 1. In order to do so, calibration constants  $C_a$  and  $C_s$  are chosen in a certain way.

This can be done in different ways, but the paper [\[18\]](#)

## 2.6 Conclusion

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

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

# Sound Zone Approach Review and Implementation

---

# 3

## **3.1 Introduction**

## **3.2 Review of Sound Zone Approaches**



### **3.3 Selection of Sound Zone Approach**

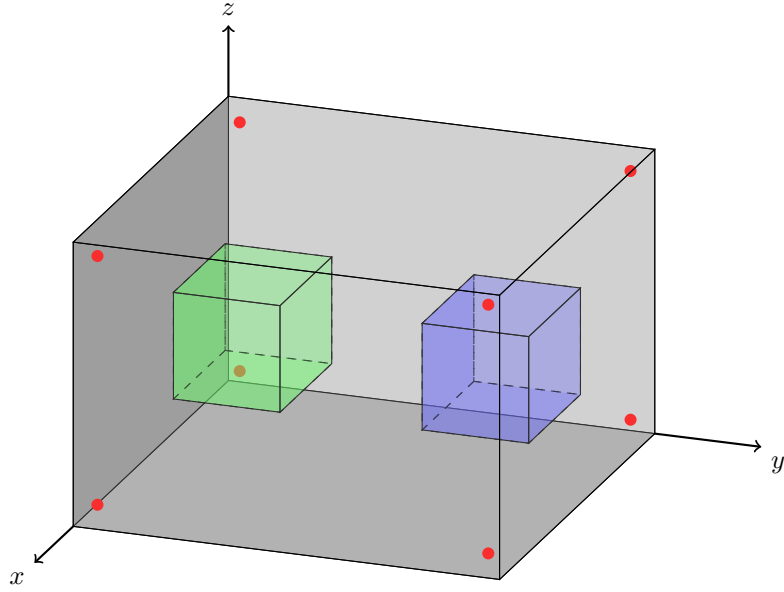


Figure 3.1: The room  $\mathcal{R} \subset \mathbb{R}^3$  containing the zones  $\mathcal{A} \subset \mathcal{R}$  and  $\mathcal{B} \subset \mathcal{R}$  depicted in green and blue respectively. The room contains  $N_L = 8$  loudspeakers, which are denoted by the red dots in the corners of the room.

### 3.4 Data Model

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

The contents of this section are as follows.

First, ?? introduces a spatial description of a room containing two zones and a loudspeaker array. Then, subsection 3.4.2 defines the objective of the sound zone algorithm as realizing target sound pressure at discrete points in the room.

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

#### 3.4.1 Room Topology

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

The room  $R$  can be modeled as a closed subset of three dimensional space,  $\mathcal{R} \subset \mathbb{R}^3$ . The two non-overlapping zones  $\mathcal{A}$  and  $\mathcal{B}$  are contained within the room  $R$ , i.e.  $\mathcal{A} \subset \mathcal{R}$  and  $\mathcal{B} \subset \mathcal{R}$  where  $\mathcal{A} \cap \mathcal{B} = \emptyset$ . In addition to the zones, the room  $\mathcal{R}$  also contains  $N_L$  loudspeakers, which can be modeled as discrete points. The room, loudspeakers and zones are visualized in Figure 3.1.

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.

### 3.4.2 Defining Target Pressure

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

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

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

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

### 3.4.3 Realizing Sound Pressure through the Loudspeaker

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

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

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

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

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

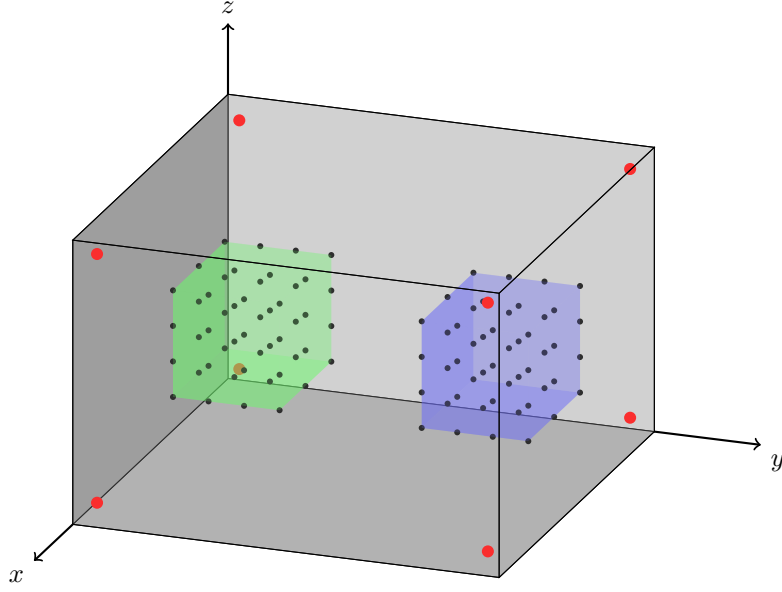


Figure 3.2: The previously introduced room  $\mathcal{R}$  with zones  $\mathcal{A}$  and  $\mathcal{B}$  discretized.

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

$$p^{(m)}[n] = \sum_{l=0}^{N_L} p^{(l,m)}[n] \quad (3.2)$$

$$= \sum_{l=0}^{N_L} (h^{(l,m)} * x^{(l)})[n] \quad (3.3)$$

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  $m \in A \cup B$ .

#### 3.4.4 Choice of Target Pressure

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

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

pressure is chosen as follows:

$$t^{(m)}[n] = \sum_{l=0}^{N_L} (h^{(l,m)} * s_{\mathcal{A}})[n] \quad \forall m \in A \quad (3.4)$$

$$t^{(m)}[n] = \sum_{l=0}^{N_L} (h^{(l,m)} * s_{\mathcal{B}})[n] \quad \forall m \in B \quad (3.5)$$

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 is physically attainable with the given loudspeakers and room.

### 3.5 Implementation of Selected Sound Zone Algorithm

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

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

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

$$x^{(l)}[n] = x_{\mathcal{A}}^{(l)}[n] + x_{\mathcal{B}}^{(l)}[n] \quad (3.6)$$

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:

$$p_{\mathcal{Z}}^{(m)}[n] = \sum_{l=0}^{N_L} \left( h^{(l,m)} * x_{\mathcal{Z}}^{(l)} \right) [n] \quad (3.7)$$

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:

$$p^{(m)}[n] = p_{\mathcal{A}}^{(m)}[n] + p_{\mathcal{B}}^{(m)}[n] \quad (3.8)$$

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:

$$\arg \min_{x_{\mathcal{A}}^{(l)}[n], x_{\mathcal{B}}^{(l)}[n] \forall l} \sum_{m \in A} \left\| p_{\mathcal{A}}^{(m)}[n] - t^{(m)}[n] \right\|_2^2 + \sum_{m \in A} \left\| p_{\mathcal{B}}^{(m)}[n] \right\|_2^2 + \quad (3.9)$$

$$\sum_{m \in B} \left\| p_{\mathcal{B}}^{(m)}[n] - t^{(m)}[n] \right\|_2^2 + \sum_{m \in B} \left\| p_{\mathcal{A}}^{(m)}[n] \right\|_2^2 \quad (3.10)$$

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:

$$\text{RE}_{\mathcal{Z}} = \sum_{m \in A} \left\| p_{\mathcal{A}}^{(m)}[n] - t^{(m)}[n] \right\|_2^2 \quad (3.11)$$

$$\text{LE}_{\mathcal{Z}} = \sum_{m \in A} \left\| p_{\mathcal{B}}^{(m)}[n] \right\|_2^2 \quad (3.12)$$

Here,  $\text{RE}_{\mathcal{Z}}$  is the reproduction error and  $\text{LE}_{\mathcal{Z}}$  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:

$$\arg \min_{x_{\mathcal{A}}^{(l)}[n], x_{\mathcal{B}}^{(l)}[n] \forall l} \text{RE}_{\mathcal{A}} + \text{LE}_{\mathcal{A}} + \text{RE}_{\mathcal{B}} + \text{LE}_{\mathcal{B}} \quad (3.13)$$

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

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

### 3.6 Block Based Approach

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

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

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

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

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

#### 3.6.1 Modeling Block Based Knowledge Limitations

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

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

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

$$\mu(n) = \lfloor n/H \rfloor \quad (3.14)$$

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

Let the available knowledge of the desired playback signal with knowledge of blocks



up to  $\mu$  be denoted as  $\tilde{s}_{\mathcal{Z}}^R[n, \mu]$ . It can be defined as follows:

$$\tilde{s}_{\mathcal{Z}}^R[n, \mu] = \sum_{k=-\infty}^{\mu} \tilde{s}_{\mathcal{Z},k}^R[n] w_H^R[n - kH] \quad (3.15)$$

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

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

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

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

$$\sum_{k=-\infty}^{\infty} w[n - kH] = 1 \quad \forall n \quad (3.16)$$

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

$$s_{\mathcal{Z}}[n] = \sum_{k=-\infty}^{\infty} s_{\mathcal{Z}}[n] w[n - kH] \quad (3.17)$$

$$= \sum_{k=-\infty}^{\infty} \tilde{s}_{\mathcal{Z},k}[n] w[n - kH] \quad (3.18)$$

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

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

$$\tilde{s}_{\mathcal{Z}}[n, \mu] = \sum_{k=-\infty}^{\mu} \tilde{s}_{\mathcal{Z},k}[n] w[n - kH] \quad (3.19)$$

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

### 3.6.2 Block Based Loudspeaker Input Signal Computation

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

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

$$\tilde{x}_{\mathcal{Z}}^{(l)}[n, \mu] = \sum_{k=-\infty}^{\mu} \tilde{x}_{\mathcal{Z},k}^{(l)}[n]w[n - kH] \quad (3.20)$$

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

$$= \tilde{x}_{\mathcal{Z}}^{(l)}[n, \mu - 1] + \tilde{x}_{\mathcal{Z},\mu}^{(l)}[n]w[n - \mu H] \quad (3.22)$$

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

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

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

### 3.6.3 Block Based Sound Pressure

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

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

The block-based target sound pressure will be denoted by  $\tilde{t}^{(m)}[n, \mu]$ , and can be defined by simply substituting the definition for the block-based desired playback

signal  $\tilde{s}_{\mathcal{Z}}[n, \mu]$  into the definition of the target pressure ??:

$$\tilde{t}^{(m)}[n, \mu] = \sum_{l=0}^{N_L} \left( h^{(l,m)} * \tilde{s}_{\mathcal{Z}} \right) [n] \quad (3.23)$$

$$= \sum_{l=0}^{N_L} \sum_{k=-\infty}^{\mu} \sum_{m=0}^{N_h-1} h^{(l,m)}[m] \tilde{s}_{\mathcal{Z},k}[n-m] w[n-kH-m] \quad (3.24)$$

$$= \tilde{t}^{(m)}[n, \mu-1] + \sum_{l=0}^{N_L} \sum_{m=0}^{N_h-1} h^{(l,m)}[m] \tilde{s}_{\mathcal{Z},\mu}[n-m] w[n-\mu H-m] \quad (3.25)$$

The definition above holds for all points  $m \in \mathcal{Z}$ , the points contained in zone  $\mathcal{Z}$ . As can be seen, the block based target sound pressure for the block  $\mu$  can be computed recursively by adding the contribution of the newest block  $\tilde{s}_{\mathcal{Z},\mu}[n]$  the target sound pressure of the previous block.

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

$$\tilde{p}_{\mathcal{Z}}^{(m)}[n, \mu] = \sum_{l=0}^{N_L} \left( h^{(l,m)} * \tilde{x}_{\mathcal{Z}}^{(l)} \right) [n] \quad (3.26)$$

$$= \sum_{l=0}^{N_L} \sum_{k=-\infty}^{\mu} \sum_{m=0}^{N_h-1} h^{(l,m)}[m] \tilde{x}_{\mathcal{Z},k}^{(l)}[n-m] w[n-kH-m] \quad (3.27)$$

$$= \tilde{p}_{\mathcal{Z}}^{(m)}[n, \mu-1] + \sum_{l=0}^{N_L} \sum_{m=0}^{N_h-1} h^{(l,m)}[m] \tilde{x}_{\mathcal{Z},\mu}^{(l)}[n-m] w[n-\mu H-m] \quad (3.28)$$

The definition above again holds for all points  $m \in \mathcal{Z}$ . As can be seen, the block based resulting sound pressure for the block  $\mu$  can also be computed recursively.

With this, all quantities required for the block-based formulation of the Multi-Zone Pressure-Matching (MZ-PM) approach are known. The next section will use these quantities to state the block-based MZ-PM algorithm.

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

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

As mentioned before, the approach that will be taken is to compute the  $\mu^{\text{th}}$  coefficient of the loudspeaker input signal  $\tilde{x}_{\mathcal{Z},\mu}^{(l)}$  such that the resulting sound pressure  $\tilde{p}_{\mathcal{Z}}^{(m)}[n, \mu]$  best matches the target sound pressure  $\tilde{t}^{(m)}[n, \mu]$ .

Note that in this approach only the most recent loudspeaker coefficients  $\tilde{x}_{\mathcal{Z},\mu}^{(l)}$  are being controlled. The previous coefficients are assumed to have already been played. As such, they are held fixed.

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

$$\arg \min_{\tilde{x}_{\mathcal{A},\mu}^{(l)}, \tilde{x}_{\mathcal{B},\mu}^{(l)} \forall l} \sum_{m \in \mathcal{A}} \left\| \tilde{p}_{\mathcal{A}}^{(m)}[n, \mu] - \tilde{t}_{\mu}^{(m)}[n, \mu] \right\|_2^2 + \sum_{m \in \mathcal{A}} \left\| \tilde{p}_{\mathcal{B}}^{(m)}[n, \mu] \right\|_2^2 + \quad (3.29)$$

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

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

The problem above is solved recursively for all loudspeaker input signal coefficients  $\tilde{x}_{\mathcal{A},\mu}^{(l)}[n]$  and  $\tilde{x}_{\mathcal{B},\mu}^{(l)}[n]$  as new blocks  $\tilde{s}_{\mathcal{A},\mu}[n]$  and  $\tilde{s}_{\mathcal{B},\mu}[n]$  are revealed. The final loudspeaker input signals  $\tilde{x}_{\mathcal{Z}}^{(l)}[n, \infty]$  can then be found by means of ??.

### 3.7 Conclusion



# Implementation and Evaluation of a Perceptual Sound Zone Algorithm

---

# 4

## 4.1 Introduction

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

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

This chapter is structured as follows.

## 4.2 Frequency Domain Conversion

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

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

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

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

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

### 4.2.1 Frequency Domain Transformation

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

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

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



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

$$X[k] = \sum_{n=0}^{N_w+N_h-2} x[n] \exp\left(\frac{-j2\pi kn}{N_w + N_h - 1}\right) \quad (4.1)$$

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

#### 4.2.2 Quantities the Frequency Domain

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

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

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

$$\tilde{T}^{(m)}[k, \mu] = \tilde{T}^{(m)}[k, \mu - 1] + \sum_{l=0}^{N_L} H^{(l,m)}[k] \tilde{S}_{\mathcal{Z},\mu}[k] \quad (4.2)$$

$$\tilde{P}_{\mathcal{Z}}^{(m)}[k, \mu] = \tilde{P}_{\mathcal{Z}}^{(m)}[k, \mu - 1] + \sum_{l=0}^{N_L} H^{(l,m)}[k] \tilde{X}_{\mathcal{Z},\mu}^{(l)}[k] \quad (4.3)$$

Here,  $H^{(l,m)}[k] \in \mathbb{C}^{N_w+N_h-1}$  is the transformed version of the room impulse responses. Furthermore,  $\tilde{S}_{\mathcal{Z},\mu}[k] \in \mathbb{C}^{N_w+N_h-1}$  and  $\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k] \in \mathbb{C}^{N_w+N_h-1}$  are the frequency domain versions of the desired playback signal and the loudspeaker input signal, which are defined as follows:

$$\tilde{S}_{\mathcal{Z},\mu}[k] = \sum_{n=0}^{N_w+N_h-2} \tilde{s}_{\mathcal{Z},\mu}[n] w[n - \mu H] \exp\left(\frac{-j2\pi kn}{N_w + N_h - 1}\right) \quad (4.4)$$

$$\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k] = \sum_{n=0}^{N_w+N_h-2} \tilde{x}_{\mathcal{Z},\mu}^{(l)}[n] w[n - \mu H] \exp\left(\frac{-j2\pi kn}{N_w + N_h - 1}\right) \quad (4.5)$$

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

#### 4.2.3 Proposed Frequency Domain Approach

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

$$\arg \min_{\tilde{x}_{\mathcal{A},\mu}^{(l)}[n], \tilde{x}_{\mathcal{B},\mu}^{(l)}[n] \forall l} \sum_{m \in \mathcal{A}} \left\| \tilde{P}_{\mathcal{A}}^{(m)}[k, \mu] - \tilde{T}^{(m)}[k, \mu] \right\|_2^2 + \sum_{m \in \mathcal{A}} \left\| \tilde{P}_{\mathcal{B}}^{(m)}[k, \mu] \right\|_2^2 + \quad (4.6)$$

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

Note how the optimization is still performed over the time domain signal. This was done to constrain the loudspeaker input signal coefficient to size  $N_w$ , as that is an assumption made by the frame-based processing.

In principal, this introduces more complexity than solving directly over the frequency domain loudspeaker input coefficient  $\tilde{X}_{\mathcal{Z},\mu}^{(l)}[k]$ . This however introduces issues as it requires the truncation of the time-domain version to  $N_w$  samples, which was found to introduce artifacts.

### 4.3 Unconstrained Perceptual Pressure Matching

### 4.4 Perceptual Minimization Approach

Detectability is defined as:

$$D(x[n], e[n]) = \|\mathbf{w}(X[k]) \circ \mathbf{E}[k]\|_2^2 \quad (4.8)$$

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

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

$$\arg \min_{\tilde{x}_{\mathcal{A},\mu}^{(l)}[n], \tilde{x}_{\mathcal{B},\mu}^{(l)}[n] \forall l} \sum_{m \in \mathcal{A}} \left\| \tilde{P}_{\mathcal{A}}^{(m)}[k, \mu] - \tilde{T}^{(m)}[k, \mu] \right\|_2^2 + \sum_{m \in \mathcal{A}} \left\| \tilde{P}_{\mathcal{B}}^{(m)}[k, \mu] \right\|_2^2 + \quad (4.9)$$

$$\sum_{m \in \mathcal{B}} \left\| \tilde{P}_{\mathcal{B}}^{(m)}[k, \mu] - \tilde{T}^{(m)}[k, \mu] \right\|_2^2 + \sum_{m \in \mathcal{B}} \left\| \tilde{P}_{\mathcal{A}}^{(m)}[k, \mu] \right\|_2^2 \quad (4.10)$$

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

$$\arg \min_{\tilde{x}_{\mathcal{A},\mu}^{(l)}[n], \tilde{x}_{\mathcal{B},\mu}^{(l)}[n] \forall l} \sum_{m \in \mathcal{A}} \left\| \mathbf{w} \left( \tilde{T}^{(m)}[k, \mu] \right) \left[ \tilde{P}_{\mathcal{A}}^{(m)}[k, \mu] - \tilde{T}^{(m)}[k, \mu] \right] \right\|_2^2 + \quad (4.11)$$

$$\sum_{m \in \mathcal{A}} \left\| \mathbf{w} \left( \tilde{T}^{(m)}[k, \mu] \right) \left[ \tilde{P}_{\mathcal{B}}^{(m)}[k, \mu] \right] \right\|_2^2 + \quad (4.12)$$

$$\sum_{m \in \mathcal{B}} \left\| \mathbf{w} \left( \tilde{T}^{(m)}[k, \mu] \right) \left[ \tilde{P}_{\mathcal{B}}^{(m)}[k, \mu] - \tilde{T}^{(m)}[k, \mu] \right] \right\|_2^2 + \quad (4.13)$$

$$\sum_{m \in \mathcal{B}} \left\| \mathbf{w} \left( \tilde{T}^{(m)}[k, \mu] \right) \left[ \tilde{P}_{\mathcal{A}}^{(m)}[k, \mu] \right] \right\|_2^2 \quad (4.14)$$

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

## 4.5 Constrained Perceptual Pressure Matching

## 4.6 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:

$$\arg \min_{\tilde{x}_{\mathcal{A},\mu}^{(l)}[n], \tilde{x}_{\mathcal{B},\mu}^{(l)}[n] \forall l} \sum_{m \in \mathcal{B}} \left\| \mathbf{w} \left( \tilde{T}^{(m)}[k, \mu] \right) \left[ \tilde{P}_{\mathcal{A}}^{(m)}[k, \mu] \right] \right\|_2^2 + \quad (4.15)$$

$$\sum_{m \in \mathcal{A}} \left\| \mathbf{w} \left( \tilde{T}^{(m)}[k, \mu] \right) \left[ \tilde{P}_{\mathcal{B}}^{(m)}[k, \mu] \right] \right\|_2^2 \quad (4.16)$$

$$\text{subject to} \quad \left\| \mathbf{w} \left( \tilde{T}^{(m)}[k, \mu] \right) \left[ \tilde{P}_{\mathcal{A}}^{(m)}[k, \mu] - \tilde{T}^{(m)}[k, \mu] \right] \right\|_2^2 \leq Q \quad \forall m \in \mathcal{B} \quad (4.17)$$

$$\left\| \mathbf{w} \left( \tilde{T}^{(m)}[k, \mu] \right) \left[ \tilde{P}_{\mathcal{B}}^{(m)}[k, \mu] - \tilde{T}^{(m)}[k, \mu] \right] \right\|_2^2 \leq Q \quad \forall m \in \mathcal{A} \quad (4.18)$$

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

## 4.7 Results

## 4.8 Conclusion







# Bibliography

---

- [1] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, “Personal sound zones: Delivering interface-free audio to multiple listeners,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 81–91, 2015.
- [2] J. Donley and C. H. Ritz, “Multizone reproduction of speech soundfields: A perceptually weighted approach,” 2015.
- [3] T. Lee, J. K. Nielsen, and M. G. Christensen, “Towards perceptually optimized sound zones: A proof-of-concept study,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 136–140, IEEE, 2019.
- [4] T. Lee, J. K. Nielsen, and M. G. Christensen, “Signal-adaptive and perceptually optimized sound zones with variable span trade-off filters,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2412–2426, 2020.
- [5] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [6] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, pp. 749–752, IEEE, 2001.
- [7] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, “Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment,” *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.
- [8] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, “Visqol: The virtual speech quality objective listener,” in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, pp. 1–4, VDE, 2012.
- [9] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O’Gorman, and A. Hines, “Visqol v3: An open source production ready objective speech and audio metric,” pp. 1–6, 2020.
- [10] J. Kim, M. El-Kharmy, and J. Lee, “End-to-end multi-task denoising for joint sdr and pesq optimization,” *arXiv preprint arXiv:1901.09146*, 2019.
- [11] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, “An instrumental intelligibility

- metric based on information theory,” *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2017.
- [12] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, “Peaq-the itu standard for objective measurement of perceived audio quality,” *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.
  - [13] P. Pořta and J. G. Beerends, “Subjective and objective assessment of perceived audio quality of current digital audio broadcasting systems and web-casting applications,” *IEEE Transactions on Broadcasting*, vol. 61, no. 3, pp. 407–415, 2015.
  - [14] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, “Visqo-laudio: An objective audio quality metric for low bitrate codecs,” *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL449–EL455, 2015.
  - [15] C. H. Taal, R. C. Hendriks, and R. Heusdens, “A low-complexity spectro-temporal distortion measure for audio processing applications,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1553–1564, 2012.
  - [16] I. J. S. 29, “Information technology — coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s — part 3: Audio,” techreport 3, International Organization for Standardization, Geneva, CH, Oct. 1993.
  - [17] D. Pan, “A tutorial on mpeg/audio compression,” *IEEE multimedia*, vol. 2, no. 2, pp. 60–74, 1995.
  - [18] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, “A perceptual model for sinusoidal audio coding based on spectral integration,” *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1–13, 2005.
  - [19] P. Balazs, B. Laback, G. Eckel, and W. A. Deutsch, “Time–frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking,” *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 1, pp. 34–49, 2009.
  - [20] C. H. Taal, J. Jensen, and A. Leijon, “On optimal linear filtering of speech for near-end listening enhancement,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, 2013.