Block-Based Perceptually Adaptive Sound Zones With Reproduction Error Constraints

De Koeijer, Niels; Moller, Martin Bo; Martinez, Jorge; Martinez-Nuevo, Pablo; Hendriks, Richard C.

DOI
10.1109/TASLP.2024.3407487

Publication date
2024

Document Version
Final published version

Published in
IEEE/ACM Transactions on Audio Speech and Language Processing

Block-Based Perceptually Adaptive Sound Zones With Reproduction Error Constraints

Niels de Koeijer, Martin Bo Møller, Jorge Martinez, Pablo Martínez-Nuevo, and Richard C. Hendriks, *Senior Member, IEEE*

***Abstract*—Sound zone algorithms control the inputs to a loudspeaker array such that spatially distinct zones, each with separate audio content, are created. This work proposes a sound zone approach which includes a model of human auditory perception in the optimization problem designing the loudspeaker control filters. The control filters are therefore optimized directly for human experience, rather than by proxy through sound pressure, as is done in typical approaches. The proposed optimization problem features a perceptually weighted constraint on the bright zone reproduction error, which allows the user of the algorithm to specify the desired bright zone quality. The proposed method achieves 2 to 4 dB of additional acoustic contrast and is expected to yield less distracting dark-zone interference for the same perceived quality when compared to a traditional approach.**

***Index Terms*—Sound zones, sound field control, perceptual masking models, adaptive control.**

I. Introduction

SOUND field control algorithms aim to control an array of loudspeakers such that spatially distinct zones each with their own audio content are created. These zones are commonly referred to as "sound zones". In the ideal case, this is done in such a way that there is minimal interference between zones, allowing listeners to enjoy the audio content of one zone without perceiving content reproduced in the other zones. Creating sound zones is typically valuable for situations within which multiple people occupy a shared space without necessarily wanting to share sound. Such situations include homes and car cabins [1], but also reducing the sound leakage of an outdoor concert to the surrounding environment [2].

The creation of sound zones is often approached as an optimization problem, where loudspeaker control filters are designed based on predictions of the resulting sound pressure at control locations in space [3], [4]. These predictions are commonly made based on simulated or measured impulse responses, which capture the relationship between loudspeaker input and resulting sound pressure at the control locations. Typically, an optimization problem is solved separately per zone. For each zone, one defines a set of control points in space forming a "bright zone", in which one desires to reproduce specific audio content, and another set forming a "dark zone", in which one wishes to limit the amount of sound pressure leaking from the bright zone [3], [4], [5], [6]. By overlapping multiple bright-dark zone pairs, the desired effect of multiple zones with minimal interference is achieved.

Both time-invariant and adaptive approaches have been proposed for the design of the control filters. In the former case, a fixed set of loudspeaker control filters is designed without considering temporal changes in e.g., audio content or room impulse responses [7], [8]. In the latter case, the control filters are instead updated regularly based on changes in the environment. Adaptive approaches include both adaptive-filtering based approaches [9] as well as approaches where a complete optimization problem is solved for each control filter [5]. Such approaches have been shown to outperform time-invariant ones. For instance, in [5] it was shown that applying the moving horizon adaptive control technique to sound zones yields an additional 4 dB of acoustic contrast compared to static time-invariant filters.

Recent work explores the addition of models of human auditory perception to adaptive sound zone algorithms [10], [11], [12], [13], [14]. These perceptual models quantify how one sound can mask another as perceived by a human observer. One motivation for including perception in the algorithms is that it allows the control filters to be optimized directly for the experience of the listener, rather than by proxy through sound pressure, as is done for typical approaches.

In work by Donley et al. [12] an existing sound zone approach [15] was augmented with perceptual information. Instead of aiming for a minimal sound pressure level in the dark zone in the optimization problem, the proposed method aims for a dark zone sound pressure level equal to the masking threshold. This

Manuscript received 8 August 2023; revised 15 January 2024 and 25 March 2024; accepted 6 May 2024. Date of publication 30 May 2024; date of current version 14 June 2024. This work was supported by the Innovation Fund Denmark through Project Interactive Sound Zones for Better Living under Grant 9069-00038B. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Enzo De Sena. *(Corresponding author: Niels de Koeijer.)*

Niels de Koeijer is with the Audio Technology Department, Bang & Olufsen A/S, DK7600 Struer, Denmark (e-mail: nemk@bang-olufsen.dk).

Martin Bo Møller is with the Audio Technology Department, Bang & Olufsen A/S, DK7600 Struer, Denmark, and also with the Section on Artificial Intelligence and Sound, Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: mim@bang-olufsen.dk).

Jorge Martinez is with the Multimedia Computing Group, Intelligent Systems Department, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 XE Delft, The Netherlands (e-mail: J.A.MartinezCastaneda@tudelft.nl).

Pablo Martínez-Nuevo is with the Artificial Intelligence Department, Bang & Olufsen A/S, DK7600 Struer, Denmark (e-mail: pmnuevo@alum.mit.edu).

Richard C. Hendriks is with the Signal Processing Systems Group, Microelectronics Department, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: r.c.hendriks@tudelft.nl).

Digital Object Identifier 10.1109/TASLP.2024.3407487

relaxation thus allows some sound pressure in the dark zone, as long as it is inaudible to humans. The authors show that this dark zone relaxation in turn allows for more of the optimization effort to go into reducing the spatial reproduction error in the bright zone.

Lee et al. proposed a perceptually weighted time-invariant (P-VAST) and a perceptually weighted adaptive (AP-VAST) [10], [11] sound zone algorithms. Both algorithms extend the variable span trade-off (VAST) sound zone framework by including a time-domain weighting filter in the design of the loudspeaker control filters [6]. The perceptual weighting is introduced as the reciprocal of the masking threshold. Doing so adds a higher weight to more perceptually relevant frequencies during the filter design. Both P-VAST and AP-VAST were found to outperform the non-perceptual reference approaches in terms of perceptual speech quality measures and through a listening test.

The presented work further explores the benefits of incorporating perceptual information in sound zone algorithms. A block-based sound zone framework is proposed where perceptual models are leveraged to explicitly quantify and constrain the perceived quality of the reproduced sound pressure in the bright zone. This is in contrast with existing approaches in which the dark zone audibility is either constrained, or the problem is entirely unconstrained, thus controlling the bright zone performance implicitly [3], [16]. The proposed framework yields both traditional and perceptual sound zone algorithms, which allows isolated study of the benefits of the perceptual modelling. The perceptual model considered in this work is the "distortion detectability", which was originally proposed by van de Par et al. [17], and later refined by Taal et al. [18], [19]. These models have two beneficial properties in relation to the sound zones. First, they can be expressed as convex functions which can be included directly in the optimization framework. Secondly, the output of the model has the interpretation that a value of 1 corresponds to an error signal which is just noticeable. In this work the baseline model by van de Par et al. is compared to the refinement by Taal et al. as well as no perceptual model.

The rest of this document is structured as follows. First, in Section II, a block-based adaptive sound zone approach with reproduction error constraints is proposed. This method serves as the foundation for the perceptually adaptive approaches. Next, Section III introduces the two implementations of the distortion detectability perceptual model, by van de Par et al. and Taal et al. respectively. These are then, in Section IV, combined with the block-based adaptive sound zone approach to form two perceptually adaptive sound zone approaches. Subsequently, Section V investigates the benefits of including perception in sound zone algorithms by comparing the performance of simulations of the adaptive and perceptually adaptive sound zone approaches. Additionally, the proposed approaches are compared to the AP-VAST method [10], [11].

II. Block-Based Adaptive Sound Zone Approach

Sound zone algorithms attempt to control the sound pressure in space by controlling the output of a loudspeaker array. In this work, the proposed sound zone algorithms do this by adaptively altering the control filters for each loudspeaker based on the audio content to be reproduced. In this section, an adaptive sound zoning method with reproduction error constraints is presented, in which the control filters are designed separately for each block of input audio. This approach serves both as the foundation and as a comparison case for the perceptually adaptive sound zone algorithms proposed in Section IV. The data model used in this approach is inspired by the single-channel overlap-save block convolution scheme originally proposed by Moulines et al. [20]. This scheme can operate on blocks of arbitrary size, which is critical for the perceptually adaptive sound zone algorithm, as the timescale used affects the prediction performance of the perceptual models used in this work (discussed in more detail in Section III).

A. Pressure Prediction Model

Assume an array of L loudspeakers is presented an input audio sequence denoted by $u[n]$. The input signal $u[n]$ is partitioned into overlapping blocks of size N_b. Here, a hop size of N_r samples is used, resulting in an overlap of $N_b - N_r$ samples between subsequent blocks. To index these blocks, let the s^{th} block of the input signal be denoted by $\mathbf{u}(s) \in \mathbb{R}^{N_b}$, which can be formed as follows:

$$\mathbf{u}(s) := \begin{bmatrix} u[sN_r] & \dots & u[sN_r + N_b - 1] \end{bmatrix}^T. \tag{1}$$

In this work, the design of the control filters is based on predictions of the pressure arising at spatially sampled control locations m. The resulting pressure prediction from the loudspeaker l to a point in space m is modelled through the linear convolution between the input block $\mathbf{u}(s)$, the control filter $\mathbf{w}^{(l)}(s) \in \mathbb{R}^{N_w}$, and the room impulse response (RIR) $\mathbf{h}^{(l,m)} \in \mathbb{R}^{N_h}$. The total sound pressure prediction $\mathbf{p}^{(m)}(s) \in \mathbb{R}^{N_v}$ is defined as follows:

$$\mathbf{p}^{(m)}(s) := \sum_l \mathbf{h}^{(l,m)} \circledast_{N_b} \mathbf{w}^{(l)}(s) \circledast_{N_b} \mathbf{u}(s). \tag{2}$$

Here, $\circledast_{N_b}$ denotes the modulo-N_b circular convolution of two sequences. Due to the wrap-around of the modulo operator, the circular convolution will only coincide with only the last $N_v = N_b - N_w - N_h + 2$ samples of the corresponding linear convolution. The proposed block-based approach thus discards the first $N_a = N_b - N_v$ samples. The motivation for using the circular convolution is that it can be performed efficiently through element-wise multiplication in the frequency domain. To leverage this, let $\tilde{\mathbf{u}}(s) \in \mathbb{C}^{N_b}$, $\tilde{\mathbf{h}}^{(l,m)} \in \mathbb{C}^{N_b}$, and $\tilde{\mathbf{w}}^{(l)}(s) \in \mathbb{C}^{N_w}$ denote the frequency domain representation of the input signal, acoustic transfer functions (ATFs) and control filters respectively. Note that the ATFs have been zero-padded to block size N_b before applying the discrete Fourier transform (DFT).

The frequency domain representation of the truncated sound pressure for the block s at control point m is denoted by $\tilde{\mathbf{p}}^{(m)}(s) \in \mathbb{C}^{N_v}$. It is defined as follows:

$$\tilde{\mathbf{p}}^{(m)}(s) := \mathbf{G}_{01}\tilde{\mathbf{U}}(s)\tilde{\mathbf{H}}^{(m)} \left(\mathbf{I}_{N_l} \otimes \mathbf{G}_{10}\right) \tilde{\mathbf{w}}(s). \tag{3}$$

Here, $\otimes$ denotes the Kronecker product. The vector $\tilde{\mathbf{w}}(s) \in \mathbb{C}^{N_l N_w}$ contains the concatenation of frequency domain control

filters for block s:

$$\tilde{\mathbf{w}}(s) := \begin{bmatrix} \tilde{\mathbf{w}}^{(1)}(s) \\ \vdots \\ \tilde{\mathbf{w}}^{(L)}(s) \end{bmatrix}. \tag{4}$$

The matrix $\mathbf{G}_{10} \in \mathbb{C}^{N_b \times N_w}$ denotes a matrix that is designed to zero-pad the frequency-domain filter $\tilde{\mathbf{w}}^{(l)}(s)$ from original length N_w to block length N_b, and is defined as:

$$\mathbf{G}_{10} := \mathbf{F}_{N_b} \mathbf{Z}_{N_b, N_w} \mathbf{F}_{N_w}^{-1}. \tag{5}$$

Here, $\mathbf{F}_{N_w}^{-1} \in \mathbb{C}^{N_w \times N_w}$ and $\mathbf{F}_{N_b} \in \mathbb{C}^{N_b \times N_b}$ are an N_w-point IDFT matrix and N_b-point DFT matrix respectively [21], and are thus used for time-frequency conversion. The matrix $\mathbf{Z}_{N_b,N_w} \in \mathbb{R}^{N_b \times N_w}$ implements time-domain zero padding of a signal from length N_w to a length N_b samples, which can be defined as:

$$\mathbf{Z}_{N_b,\, N_w} := \begin{bmatrix} \mathbf{I}_{N_w} \\ \mathbf{0}_{N_b - N_w,\, N_w} \end{bmatrix}. \tag{6}$$

Here, $\mathbf{0}_{N_b - N_w, N_w} \in \mathbb{R}^{N_b - N_w \times N_w}$ denotes a matrix of all zeros, and $\mathbf{I}_{N_w} \in \mathbb{R}^{N_w \times N_w}$ denotes an identity matrix.

Subsequently, after zero padding, (3) describes multiplication with matrices $\tilde{\mathbf{H}}^{(m)} \in \mathbb{C}^{N_b \times N_w N_l}$ and $\tilde{\mathbf{U}}(s) \in \mathbb{C}^{N_b \times N_b}$, which implement the circular convolution through frequency domain inner products between the input block and ATFs as described in (2). These can be expressed as:

$$\tilde{\mathbf{H}}^{(m)} := \begin{bmatrix} \text{diag}\left(\tilde{\mathbf{h}}^{(1,m)}\right) & \dots & \text{diag}\left(\tilde{\mathbf{h}}^{(L,m)}\right) \end{bmatrix}, \tag{7}$$

$$\tilde{\mathbf{U}}(s) := \text{diag}\left(\tilde{\mathbf{u}}(s)\right). \tag{8}$$

As mentioned, due to the modulo operator, the first N_a samples of (2) will contain time-domain aliasing. Therefore, only the last N_v valid samples are kept. To this end, multiplication with matrix $\mathbf{G}_{01} \in \mathbb{C}^{N_v \times N_b}$ in (3) performs time-domain truncation to length N_v on frequency domain representation of length N_b, which can be written as:

$$\mathbf{G}_{01} := \mathbf{F}_{N_v} \mathbf{T}_{N_v,\, N_b} \mathbf{F}_{N_b}^{-1}. \tag{9}$$

Here, matrix $\mathbf{T}_{N_v,\, N_b} \in \mathbb{R}^{N_v,\, N_b}$ corresponds to discarding the first N_a samples of a sequence of length N_b, resulting in a sequence of length N_v. This can be defined in matrix form as:

$$\mathbf{T}_{N_v,\, N_b} := \begin{bmatrix} \mathbf{0}_{N_v,\, N_a} & \mathbf{I}_{N_v} \end{bmatrix}. \tag{10}$$

B. Sound Zones as an Optimization Problem

In this section, it is shown how the previously introduced pressure prediction model given by (3) can be used in an optimization problem to find sets of control filters $\mathbf{w}^{(l)}(s)$ such that sound zones are created. Without loss of generality, this section presents a situation with only two zones. Consider the situation shown in Fig. 1, where the space has been partitioned into two zones. Each zone has been sampled by a set of control points m, forming two disjoint sets: one for zone A with points $m \in \mathcal{A}$, and one for zone B with points $m \in \mathcal{B}$. Present alongside the control points is a set of N_l loudspeakers. For each zone $\text{C} \in \{\text{A}, \text{B}\}$, a target sound pressure $\tilde{\mathbf{t}}_{\text{C}}^{(m)}(s) \in \mathbb{C}^{N_v}$ is defined for all points

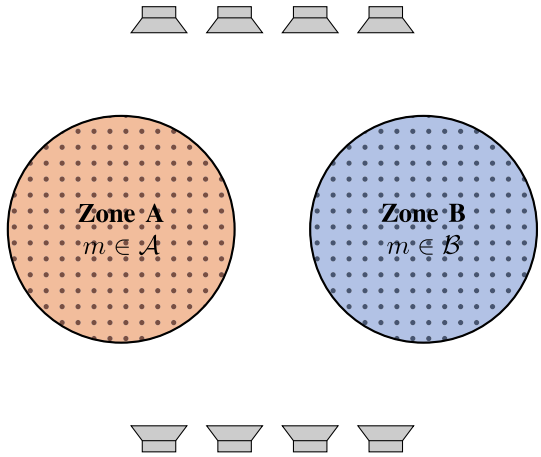

Fig. 1. A room with loudspeakers and two zones. The zones are sampled in space, defining control points $m \in \mathcal{A}$ and $m \in \mathcal{B}$ for zone A and zone B, respectively.

$m \in \mathcal{C}$ as:

$$\tilde{\mathbf{t}}_{\text{C}}^{(m)}(s) := \mathbf{G}_{01} \tilde{\mathbf{U}}_{\text{C}}(s) \tilde{\mathbf{H}}_t^{(m)} \mathbf{G}_{10} \tilde{\mathbf{w}}_t \quad \forall\, m \in \mathcal{C}. \tag{11}$$

Here, $\tilde{\mathbf{w}}_t \in \mathbb{C}^{N_w}$ and $\tilde{\mathbf{H}}_t^{(m)} \in \mathbb{C}^{N_b \times N_w}$ respectively describe the filter and ATF of a virtual source to point m. This source can be placed at any arbitrary point in space. In this work, the virtual source is chosen equal to one of the loudspeakers in the control array, which ensures that there is always a trivial set of control filters that attain the target sound pressure. The fixed control filter $\tilde{\mathbf{w}}_t$ implements a modelling delay, ensuring that loudspeakers further from the control points than the target source can contribute to sound pressure with causal control filters. The input matrix $\tilde{\mathbf{U}}_{\text{C}}(s) \in \mathbb{C}^{N_b \times N_b}$ contains block s of desired audio for zone $\text{C} \in \{\text{A}, \text{B}\}$.

The goal of the proposed sound zone algorithm is to design two sets of control filters $\mathbf{w}_{\text{A}}^{(l)}(s)$ and $\mathbf{w}_{\text{B}}^{(l)}(s)$ for all loudspeakers l in the array that attempt to recreate the target sound pressures in all control points in their respective zones, whilst minimizing the leakage to the other zone. Briefly, consider only the filters for zone A. The previously introduced pressure model given by (3) can be used used to define the sound pressure due to the filters $\mathbf{w}_{\text{A}}^{(l)}(s)$ and input $\mathbf{u}_{\text{A}}(s)$:

$$\tilde{\mathbf{p}}_{\text{A}}^{(m)}(s) := \mathbf{G}_{01} \tilde{\mathbf{U}}_{\text{A}}(s) \tilde{\mathbf{H}}^{(m)} \left(\mathbf{I}_{N_l} \otimes \mathbf{G}_{10}\right) \tilde{\mathbf{w}}_{\text{A}}(s). \tag{12}$$

The definition above is valid for all $m \in \mathcal{A} \cup \mathcal{B}$. For convenience, the following notation is introduced:

$$\tilde{\mathbf{p}}_{\text{A}\to\text{A}}^{(m)}(s) = \tilde{\mathbf{p}}_{\text{A}}^{(m)}(s) \quad m \in \mathcal{A}, \tag{13}$$

$$\tilde{\mathbf{p}}_{\text{A}\to\text{B}}^{(m)}(s) = \tilde{\mathbf{p}}_{\text{A}}^{(m)}(s) \quad m \in \mathcal{B}. \tag{14}$$

Here, $\tilde{\mathbf{p}}_{\text{A}\to\text{A}}^{(m)}(s)$ can be understood as the sound pressure intended for zone A going to zone A, and $\tilde{\mathbf{p}}_{\text{A}\to\text{B}}^{(m)}(s)$ can be understood as the sound pressure intended for zone A leaking into zone B. Similar definitions can be made for the sound pressure intended for zone B.

With these definitions in place, the sound zone algorithm can be introduced. In this work, the control filters $\mathbf{w}_{\mathrm{A}}^{(l)}(s)$ and $\mathbf{w}_{\mathrm{B}}^{(l)}(s)$ are designed in isolation of one another, and then combined afterwards. When designing the control filters $\mathbf{w}_{\mathrm{A}}^{(l)}(s)$, zone A plays the role of a *bright zone*, in which a certain target sound pressure is desired. Conversely, zone B plays the role of a *dark zone*, where minimal sound pressure is desired. When designing $\mathbf{w}_{\mathrm{B}}^{(l)}(s)$, the roles of zones A and B are reversed. The combination of the optimal control filters then yields the desired result. The combination of filters is explained in more detail in Section II-C.

Thus, the optimization can be given in terms of the zone A control filters $\mathbf{w}_{\mathrm{A}}^{(l)}(s)$ from this point onwards. The optimization problem considered in this work is given as:

$$\begin{aligned}\tilde{\mathbf{w}}_{\mathrm{A}}^{(l)}(s) = \underset{\tilde{\mathbf{w}}_{\mathrm{A}}^{(l)}(s)}{\arg\min} \sum_{m\in\mathcal{B}} \|\tilde{\mathbf{p}}_{\mathrm{A}\to\mathrm{B}}^{(m)}(s)\|_2^2 \\ \text{subject to} \sum_{l} \|\tilde{\mathbf{w}}_{\mathrm{A}}^{(l)}\|_2^2 \le E_0 \\ \|\tilde{\mathbf{p}}_{\mathrm{A}\to\mathrm{A}}^{(m)}(s) - \tilde{\mathbf{t}}_{\mathrm{A}}^{(m)}(s)\|_2^2 \le Q_0 \|\tilde{\mathbf{t}}_{\mathrm{A}}^{(m)}(s)\|_2^2 \quad \forall\, m \in \mathcal{A}. \end{aligned} \tag{15}$$

The problem can be understood as follows. In the cost function, the total sound pressure energy in zone B leaking from zone A is minimized. Meanwhile, the normalized mean squared error (NMSE) between target sound pressure $\tilde{\mathbf{t}}_{\mathrm{A}}^{(m)}$ and achieved sound pressure $\tilde{\mathbf{p}}_{\mathrm{A}\to\mathrm{A}}^{(m)}(s)$ in zone A is kept below a certain level Q_0. Hence, by solving (15), a set of control filters is obtained where dark zone sound pressure energy is minimal for a given quality level (in terms of NMSE) in the bright zone. The filter energy constraint serves to limit the total array gain of the loudspeaker array, limiting the amount of sound pressure that can arise outside the controlled region.

The problem studied in this work can be understood as a variation on the typical pressure matching (PM) approach [3]. This variation was first studied in the work by Hu et al. [22]. In traditional PM approaches, the constraints and cost are flipped relative to (15): the leakage of zone A to zone B is constrained, whilst the reproduction error in zone A is minimized. One difficulty of the proposed method is in finding a suitable value for the constraint value Q_0 on the NMSE, as there is no clear and consistent perceptual or physical interpretation of how it relates to perceived quality. This highlights another motivation for replacing the NMSE quality measure with a perceptually informed one, as perceptual models are designed to have a consistent perceptual interpretation, where similar output values correspond to similar perceived experiences.

The problem described in (15) is a convex quadratically constrained quadratic program (QCQP) [23], and thus can readily be solved by off-the-shelf solvers. Solving a QCQP for each frame of audio is computationally expensive, which may currently limit potential of the proposed method to run in real-time. Alternative computationally efficient implementations of this algorithm may be derived in the future.

C. Transmitting Solutions Through a Loudspeaker Array

After finding the sets of control filters by solving (15) for zone $\mathrm{C} \in \{\mathrm{A}, \mathrm{B}\}$ for all blocks s, an alternative loudspeaker input sequence $v_{\mathrm{C}}^{(l)}[n] \in \mathbb{R}^{N_x}$ is created. This alternative input sequence will synthesize the corresponding bright-dark zone pair when played through the array. This is done by applying the filters $\mathbf{w}_{\mathrm{C}}^{(l)}(s)$ to their respective input blocks $\mathbf{u}_{\mathrm{C}}(s)$ for each loudspeaker l. To this end, let $\mathbf{v}_{\mathrm{C}}^{(l)}(s) \in \mathbb{R}^{N_v}$ denote the s^{th} block of $v_{\mathrm{C}}^{(l)}[n]$, with the following definition:

$$\mathbf{v}_{\mathrm{C}}^{(l)}(s) = \mathbf{T}_{N_v,N_b}\mathbf{F}_{N_b}^{-1}\tilde{\mathbf{U}}(s)\mathbf{G}_{10}\tilde{\mathbf{w}}_{\mathrm{C}}^{(l)}(s). \tag{16}$$

Using the existing definitions, it can be seen that $\mathbf{v}^{(l)}(s)$ is given as the truncated circular convolution between the filter designed for transducer l with the input block $\mathbf{u}(s)$, subsequently truncated to length N_v.

The loudspeaker input sequence $v_{\mathrm{C}}^{(l)}[n]$ can then be obtained through the combination of the individual blocks $\mathbf{v}_{\mathrm{C}}^{(l)}(s)$. Each block is delayed sN_r samples and subsequently multiplied by a causal window $m[n] \in \mathbb{R}^{N_w}$, which is assumed to be constant overlap add (COLA) compliant [24] for the given hop size N_r. The sequence can thus be expressed as follows:

$$v_{\mathrm{C}}^{(l)}[n] = \sum_{s} v_{\mathrm{C}}^{(l)}[n - sN_r;\, s]m[n - sN_r]. \tag{17}$$

Here, the causal sequences $v_{\mathrm{C}}^{(l)}[n;\, s] \in \mathbb{R}^{N_v}$ correspond to the values contained within the block $\mathbf{v}_{\mathrm{C}}^{(l)}(s)$.

To play multiple solutions of (15) simultaneously, one can sum over their respective loudspeaker input signals. For instance, in the previously introduced two-zone case, one could play $v_{\mathrm{A}}^{(l)}[n]$ and $v_{\mathrm{B}}^{(l)}[n]$ simultaneously.

III. Distortion Detectability Perceptual Model

In this work, the adaptive sound zone approach introduced in Section II is augmented with perceptual information. The perceptual model used is "distortion detectability", a perceptual model originally proposed for sinusoidal audio coding. The distortion detectability, denoted by $D(\tilde{\mathbf{x}}, \tilde{\mathbf{e}}) : \mathbb{C}^N \times \mathbb{C}^N \to \mathbb{R}^+$, estimates how detectable a sinusoidal disturbance signal $\tilde{\mathbf{e}} \in \mathbb{C}^N$ is to a human observer who is simultaneously listening to a masking signal $\tilde{\mathbf{x}} \in \mathbb{C}^N$. Note that $D(\tilde{\mathbf{x}}, \tilde{\mathbf{e}})$ is a function of the frequency domain representation of the disturbance and masking signals. Here, a higher value of distortion detectability implies that the distortion $\tilde{\mathbf{e}}$ is more detectable in the presence of $\tilde{\mathbf{x}}$.

Distortion detectability was originally introduced by van de Par et al. as a perceptual spectral masking model with low computational complexity [17]. The model uses results from psycho-acoustics, to model how an audio signal is perceived by a human observer. The complexity is kept low by excluding certain stages of the human auditory system that are computationally intensive to model. Alongside the original, an alternative implementation of the distortion detectability was later proposed Taal et al. [18], [19]. This implementation added temporal masking to the model whilst keeping its computational complexity low.

The following is a brief description of both models, focusing mainly on the computational aspects relevant to this work.

A. Original Distortion Detectability by Van De Par et al.

The van de Par distortion detectability $D_{\text{Par}}(\tilde{\mathbf{x}}, \tilde{\mathbf{e}})$ is expressed as a perceptually weighted L_2-norm which can be given as follows:

$$D_{\text{Par}}(\tilde{\mathbf{x}}, \tilde{\mathbf{e}}) = \|\mathbf{G}_{\text{Par}}(\tilde{\mathbf{x}})\tilde{\mathbf{e}}\|_2^2. \tag{18}$$

Here, the frequency-domain perceptual weighting matrix $\mathbf{G}_{\text{Par}}(\tilde{\mathbf{x}}) : \mathbb{C}^N \to \mathbb{R}^{N\times N}$ is determined based on the masking properties of the masking signal $\tilde{\mathbf{x}}$ through:

$$\mathbf{G}_{\text{Par}}(\tilde{\mathbf{x}}) = \text{diag}\left(\sqrt{\sum\nolimits_{i=1} \frac{C_s \tilde{\mathbf{h}}_i^2}{\|\tilde{\mathbf{h}}_i \odot \tilde{\mathbf{x}}\|_2^2 + C_a}}\right). \tag{19}$$

Here, $\odot$ represents the Hadamard product between two vectors. The vector $\tilde{\mathbf{h}}_i \in \mathbb{C}^N$ is the joint transfer function of an outer-middle ear filter and the i^{th} filter in a gammatone filter bank, partially modelling the processing that occurs in the human auditory system. Scalars $C_s \in \mathbb{R}^+$ and $C_a \in \mathbb{R}^+$ are calibration constants that calibrate the model such that the distortion detectability is equal to 1 when the disturbance signal $\tilde{\mathbf{e}}$ is "just noticeable" to a human observer in the presence of the masking signal $\tilde{\mathbf{x}}$. The calibration achieves this by using distortion-masking signal pairs that are known to be "just noticable" from perceptual literature. This calibration thus provides a consistent perceptual interpretation of distortion detectability. The calibration procedure is followed exactly as is detailed in the original work [17].

Note that (18) is convex as a function of the disturbance signal $\tilde{\mathbf{e}}$ when $\tilde{\mathbf{x}}$ is held constant, as it is a composition of an affine function and the squared L_2-norm [23].

B. Time-Domain Distortion Detectability by Taal et al.

Par's distortion detectability is a spectral masking model operating on frequency domain representations of its inputs. As such, it assumes that the masking properties of the masking signal are stationary in time. When this is not the case, e.g., when the time domain masking signal $\mathbf{x}$ consists of a period of silence followed by a burst of audio, distortion detectability $D_{\text{Par}}(\tilde{\mathbf{x}}, \tilde{\mathbf{e}})$ may make an incorrect prediction of the detectability of the distortion signal [18], [19]. Hence, the prediction performance of the Par distortion detectability is better on shorter time scales. In examples from the original paper, time scales between 23.2 and 100.0 ms are used [17].

To alleviate this drawback, Taal's version of distortion detectability includes both temporal and spectral masking, hence considering temporal variations in the input signals. Taal's distortion detectability can be expressed as:

$$D_{\text{Taal}}(\tilde{\mathbf{x}}, \tilde{\mathbf{e}}) = \sum_i \|\mathbf{G}_{\text{Taal}}^{(i)}(\mathbf{x})\,(\mathbf{h}_i \circledast_N \mathbf{e})\|_2^2. \tag{20}$$

Note that Taal's distortion detectability is given in terms of time domain representations of the disturbance $\mathbf{e} \in \mathbb{R}^N$ and masking signals $\mathbf{x} \in \mathbb{R}^N$. Here, $\mathbf{G}_{\text{Taal}}^{(i)}(\mathbf{x}) : \mathbb{R}^N \to \mathbb{R}^{N\times N}$ represents the time-domain perceptual weighting matrix. In contrast to the Par distortion detectability, in the implementation by Taal et al. perceptual weighting matrices are defined separately for each gammatone filter tap i, defined as:

$$\mathbf{G}_{\text{Taal}}^{(i)}(\mathbf{x}) = \text{diag}\left(\sqrt{\left(\frac{C_2}{|\mathbf{h}_i \circledast_N \mathbf{x}|^2 \circledast_N \mathbf{h}_s + C_1}\right) \circledast_N \mathbf{h_s}}\right). \tag{21}$$

Here, $\mathbf{h}_i \in \mathbb{R}^N$ denotes a time-domain version of the filter $\tilde{\mathbf{h}}_i$ introduced in Section III-A. The vector $\mathbf{h}_s \in \mathbb{R}^N$ contains a low-pass filter to model the envelope extraction stage in the human auditory system. The scalars $C_1 \in \mathbb{R}^+$ and $C_2 \in \mathbb{R}^+$ perform a similar calibration as in Par's distortion detectability.

Note from (20) that $D_{\text{Taal}}(\tilde{\mathbf{x}}, \tilde{\mathbf{e}})$ is defined as a sum of squared L_2-norms, each corresponding to the i^{th} channel of the gammatone filter bank. To simplify notation of subsequent algorithms, this work will rewrite $D_{\text{Taal}}(\tilde{\mathbf{x}}, \tilde{\mathbf{e}})$ to a form similar to (18). To achieve this, consider the following equivalent expression for Taal's distortion detectability:

$$D_{\text{Taal}}(\tilde{\mathbf{x}}, \tilde{\mathbf{e}}) = \sum_i \|\mathbf{G}_{\text{Taal}}^{(i)}(\mathbf{x})\mathbf{F}_N^{-1}\tilde{\mathbf{H}}_i\tilde{\mathbf{e}}\|_2^2, \tag{22}$$

$$= \tilde{\mathbf{e}}^H \mathbf{Q}_{\text{Taal}}(\mathbf{x})\tilde{\mathbf{e}}. \tag{23}$$

Here, $\tilde{\mathbf{H}}_i$ is a diagonal matrix containing the entries of $\tilde{\mathbf{h}}_i$. As circular convolution in time coincides with entry-wise multiplication in frequency, (22) is identical to (20). The matrix $\mathbf{Q}_{\text{Taal}}(\mathbf{x}) \in \mathbb{C}^{N\times N}$ is then obtained by expanding the squared L_2-norm into a quadratic form and collecting terms, resulting in the following definition:

$$\mathbf{Q}_{\text{Taal}}(\mathbf{x}) = \sum_i \tilde{\mathbf{H}}_i^H \mathbf{F}_N^{-H} \mathbf{G}_{\text{Taal}}^{(i)}(\mathbf{x})^H \mathbf{G}_{\text{Taal}}^{(i)}(\mathbf{x}) \mathbf{F}_N^{-1} \tilde{\mathbf{H}}_i. \tag{24}$$

By taking a matrix square root, e.g. through Cholesky factorization, Taal's distortion detectability can be defined as:

$$D_{\text{Taal}}(\tilde{\mathbf{x}}, \tilde{\mathbf{e}}) = \|\mathbf{G}_{\text{Taal}}(\mathbf{x})\tilde{\mathbf{e}}\|_2^2, \tag{25}$$

where $\mathbf{G}_{\text{Taal}}(\mathbf{x})^H \mathbf{G}_{\text{Taal}}(\mathbf{x}) = \mathbf{Q}_{\text{Taal}}(\mathbf{x})$. Note that this can always be done, as $\mathbf{Q}_{\text{Taal}}(\mathbf{x})$ is positive semi-definite because it is a sum of Gram matrices, which are positive semi-definite by definition. The representation given by (25) is thus a mathematically equivalent representation to the one given by the original definition (20).

The notation for Taal's distortion detectability given in (25) is now similar to that of Par's distortion detectability given by (18): both are given as the squared norm of an affine transformation of the frequency-domain disturbance signal $\tilde{\mathbf{e}}$. The Taal distortion detectability is thus also a convex function in $\tilde{\mathbf{e}}$ when $\tilde{\mathbf{x}}$ is held fixed. Note that in general, the perceptual weighting matrix for the Taal distortion detectability is dense rather than diagonal, leading to additional computational complexity relative to the Par implementation.

In the original derivation of the Taal model, it is assumed that the distortion signal is relatively small compared to the masking signal [18], [19]. As such, it is to be expected that the model may yield inaccurate results when this assumption is violated.

This notation unifies Par's and Taal's distortion detectability under a common framework. Hence, in further discussions, the "Par" and "Taal" subscripts of the distortion detectability functions $D(\tilde{\mathbf{x}}, \tilde{\mathbf{e}})$ will be dropped.

IV. Perceptually Adaptive Sound Zone Approach

In this section, a perceptually adaptive sound zone algorithm is proposed by constructing an algorithm based on the block-based adaptive sound zone approach from Section II using the definitions of the distortion detectability introduced in Section III. Doing so, the optimization problem from (15) can be adapted to form a new optimization problem:

$$\begin{aligned} \tilde{\mathbf{w}}_{\mathrm{A}}^{(l)}(s) = \underset{\tilde{\mathbf{w}}_{\mathrm{A}}^{(l)}(s)}{\arg\min} \quad & \sum_{m\in\mathcal{B}} D\left(\tilde{\mathbf{l}}_{\mathrm{B}}^{(m)}(s),\, \tilde{\mathbf{p}}_{\mathrm{A}\to\mathrm{B}}^{(m)}(s)\right) \\ \text{subject to} \quad & \sum_{l} \|\tilde{\mathbf{w}}_{\mathrm{A}}^{(l)}\|_2^2 \le E_0 \\ D\left(\tilde{\mathbf{t}}_{\mathrm{A}}^{(m)}(s),\, \tilde{\mathbf{t}}_{\mathrm{A}}^{(m)}(s) - \tilde{\mathbf{p}}_{\mathrm{A}\to\mathrm{A}}^{(m)}(s)\right) & \le Q_0 \quad \forall\, m \in \mathcal{A}. \end{aligned} \tag{26}$$

Comparing (26) with (15), it is clear that the cost function and the reproduction error constraint of (26) have been specified in terms of distortion detectability.

For the distortion detectability in the cost function, a new quantity is introduced as masking signal, namely the latent sound pressure in zone B, denoted by $\tilde{\mathbf{l}}_{\mathrm{B}}^{(m)}(s) \in \mathbb{C}^{N_v}$. This sound pressure is assumed to be passively present at the control points $m \in \mathcal{B}$ during playback of block s. This sound pressure could for instance arise due to the playback of another zone, or due to some background noise known to be present in zone B. The sinusoidal distortion signal is taken to be the achieved sound pressure leaking from zone A to zone B, denoted by $\tilde{\mathbf{p}}_{\mathrm{A}\to\mathrm{B}}^{(m)}(s)$. The effect of minimizing the cost function is that the sound pressure in zone B, due to leakage from the sound pressure meant for zone A, is made to be minimally detectable in the presence of the predicted latent sound pressure. Implicitly, the cost function in (26) thus leverages the masking properties of the latent sound pressure. Thus, (26) allows one to account for additional information about the environment within which the sound zones exist. If the latent sound pressure is not known, it is assumed to be zeros, thus taking only the threshold in quiet into account.

For the reproduction error constraints, the masking signal is taken to be the target sound pressure for zone A, namely $\tilde{\mathbf{t}}_{\mathrm{A}}^{(m)}(s)$. The distortion signal is taken to be the deviation of the achieved zone A sound pressure $\tilde{\mathbf{p}}_{\mathrm{A}\to\mathrm{A}}^{(m)}(s)$ from the corresponding target sound pressure. The resulting distortion detectability captures how detectable the deviation from the target sound pressure is. It is therefore a perceptual measure of the perceived quality in zone A. As mentioned in Section III, distortion detectability of 1 corresponds to a just noticeable distortion. This provides a perceptual interpretation, which can be used to make an informed selection of Q_0: values below 1 correspond to inaudible differences with the target sound pressure, and values above 1 become audible.

Note that when applying the method from (26) to design filters for zone B after doing so for zone A, a recursive relationship appears. Designing filters for zone B changes latent sound pressure in zone B, which was assumed to be static when designing filters for zone A. As such, there exists a dependence between filters. In this work it is assumed that these changes to the latent sound pressure is minimal, and are not considered further. Exploring this recursion may prove interesting future work.

As discussed in Section III, distortion detectability is convex as a function of the disturbance signal if the masking signal is held constant. As both the latent sound pressure and the target sound pressure are constant in (26), the problem is a convex quadratically constrained quadratic program (QCQP), which can readily be solved by off-the-shelf solvers. Note also that (26) provides a valid algorithm for both the Par and Taal versions of the distortion detectability from Section III.

V. Simulations & Results

To evaluate the benefits of including perceptual information in sound zone algorithms, this section describes the simulation and subsequent evaluation of the non-perceptual and perceptual proposed sound zone approaches. To contextualize this work, the perceptually adaptive sound zone approach AP-VAST is included in the simulations and comparisons [10], [11]. First, in Section V-A the simulation setup is described. Next, Section V-B introduces a number of perceptual and physical evaluation measures. Finally, Section V-C discusses the evaluation of the simulation.

A. Simulation Setup

To obtain sufficient data to evaluate the performance of the algorithms, extensive simulations are performed of all three proposed algorithms: the reference adaptive sound zone algorithm discussed in Section II, and the two perceptually adaptive sound algorithms discussed in Section IV using Par's and Taal's distortion detectability implementations respectively[1]. To compare the algorithms effectively, various simulations were made for each algorithm, each for a wide range of the reproduction error constraints Q_0 given in (15) and (26). The values for Q_0 were chosen such that the evaluation measures yield a range of results that allow for meaningful analysis and conclusions. For the perceptual algorithms, special care was taken to include the calibration point for a "just noticeable" distortion, i.e. the point of calibration around "1", and the values slightly above and below it. A range for the constraint Q_0 of 0.1 to 0.35 is used for the reference adaptive sound zone algorithm. Similarly, ranges of 0.2 to 15.0 and 0.5 to 15.0 for the Par and Taal versions of the perceptually adaptive sound zone algorithm, respectively.

The Habets implementation of the well-known image-source room impulse response method by Allen et al. [25], [26] is used to simulate a $4.3 \times 6.0 \times 2.7$ meter rectangular room with a reverberation time of 200 ms. A birds-eye overview of this room is depicted in Fig. 2. Two zones, each containing 9 control points

[1]An implementation of the proposed method and simulations methodology provided at github.com/nielsdekoeijer/perceptually-adaptive-sound-zones.

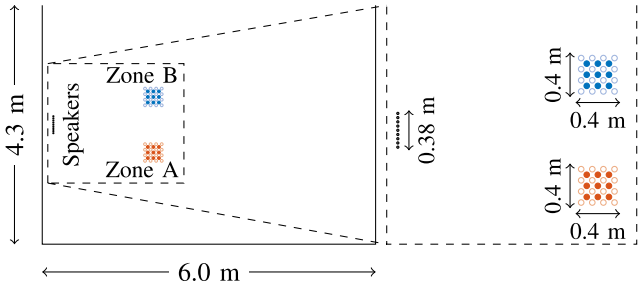

Fig. 2. Depiction of the simulated room used in the simulations. The room has $N_l = 10$ loudspeakers and two zones. The zones are sampled in space, depicting control points and validation points $m \in \mathcal{A}$ and $m \in \mathcal{B}$. The control points are depicted by the solid circles and the validation points are depicted by the hollow circles.

are defined in the room. To minimize the effects of over-fitting during the evaluation, each zone is sampled by an additional 16 validation points, which do not coincide with any of the existing control points. The sound pressure at the validation points is used for all subsequently presented results.

A second order conic problem (SOCP) solver is used to solve the optimization problems. As such, the time and space complexity of the simulations are determined by the dimensions of the matrices in the optimization problems. As the perceptual models do not alter the dimensions of the matrices involved, the execution time was found to be approximately identical for all methods. To limit the complexity of the problem, the frequency range of the audio content for each zone is limited to between 300 Hz and 4000 Hz, with a corresponding sampling rate of 8000 Hz. To effectively reproduce the frequency range, the wavelength of the lowest and highest frequency inform the design of the loudspeaker array used to form the sound zones. A line array is used with a length of 38 cm, corresponding to $1/3$ of the wavelength of the lowest frequency. Loudspeakers are placed $1/2$ of the wavelength of the highest frequency apart, resulting in 10 loudspeakers placed 4.2 cm apart.

For the experiments, zones A and B are assigned distinct audio content from a set of five songs from the pop, electronic, and rock genres, which are all loudness matched in accordance with the EBU R-128 recommendation [27]. All 20 possible combinations of bright- and dark zone are run for each algorithm for all aforementioned Q_0 constraint values. A 5-second excerpt of each song is used for each simulation.

Based on the room dimensions, 50 ms control filters are used to allow for each speaker to effectively control each control point. With the chosen sampling rate of 8000 Hz, this results in a control filter length of $N_w = 400$ samples and a room impulse response length of $N_h = 1600$ samples. As the distortion detectability by Par et al. assumes that the masking properties of its inputs are stationary, a length of 80 ms is deemed appropriate. This corresponds to $N_v = 640$ non-aliased samples. These requirements result in a total block size of $N_b = 2638$ samples. Trapezoidal windows satisfying the COLA conditions for an $1/8$th overlap between subsequent blocks are used for reconstruction.

In the simulations, only one bright-dark zone pair is considered. Thus, the problems (15) and (26) are only solved for zone A. Thus, control filters are found such that the zone A target sound pressure is reproduced up to quality level Q_0, whilst minimizing the sound pressure that leaks towards zone B. For the perceptual approach given by (26), the zone B latent sound pressure prediction, denoted by $\tilde{\mathbf{l}}_{\mathrm{B}}^{(m)}(s)$, is chosen to be equal to the zone B target sound pressure $\tilde{\mathbf{t}}_{\mathrm{B}}^{(m)}(s)$ for all points $m \in \mathcal{B}$. This corresponds to the situation where there is another sound zone being reproduced in zone B. The optimization problem is thus incentivized to shape the leakage from zone A to B to be masked by the zone B target sound pressure. The motivation for considering only a single bright-dark zone pair is to obtain as simple simulations as possible, while remaining sufficient for gaining insight into the performance of the proposed algorithms.

The results obtained using the proposed methods are compared to results obtained using AP-VAST[2] [10], [11]. The simulations utilize the same RIRs, hop size, and filter lengths as the proposed method, while the block size of AP-VAST is chosen as twice the hop size to comply with the 50% overlap for the processing scheme suggested in [11].

Two user-defined parameters must be chosen to adjust the performance of AP-VAST, μ and the number of eigenvectors V. By choosing $\mu = 1$, as done for the results in [11], it is possible to interpret the AP-VAST solution as ranging between a perceptually weighted version of ACC (a single eigenvector) and PM (all eigenvectors). AP-VAST relies on a joint diagonalization of two spatial correlation matrices representing the perceptually weighted natural loudspeaker responses in the bright and dark zone, respectively. To ensure the positive definiteness of the dark-zone spatial correlation matrix, a diagonal loading was introduced by adding an identity matrix scaled by $10^{-8}\sigma_0$, where σ_0 denotes the largest singular value of the dark-zone spatial correlation matrix. Furthermore, all weighting vectors describing inverse masking curves were normalized to unit vectors as this yielded a substantial improvement to algorithm performance in the considered scenario.

B. Evaluation Measures

To effectively evaluate the performance of the algorithm, both "physically based" and "perceptually based" evaluation measures are used. Physically based measures depend wholly on properties of the reproduced sound, whilst perceptual measures also consider how a human would perceive said sound. It should also be noted that perceptual measures can only give an indication. Hence, for conclusive results objective evaluation through listening tests is required.

1) Mean Normalized Mean Squared Error (NMSE): The normalized mean square error (NMSE) is a physical measure which quantifies how well the target pressure has been attained in a zone. Let $\mathbf{p}_{\mathrm{A}\to\mathrm{A}}^{(m)}$ and $\mathbf{t}_{\mathrm{A}}^{(m)}$ denote the time-domain sound pressure for points $m \in \mathcal{A}$ in zone A. The average NMSE over

[2]The AP-VAST method was implemented based on the original manuscript. The implementation can be found at github.com/macoustics/ap-vast-unofficial.

all N_A points in zone A can then be given as:

$$\overline{\mathrm{NMSE}}_\mathrm{A} = \frac{1}{N_\mathrm{A}} \sum_{m \in \mathcal{A}} \frac{\|\mathbf{p}^{(m)}_{\mathrm{A}\to\mathrm{A}} - \mathbf{t}^{(m)}_\mathrm{A}\|_2^2}{\|\mathbf{t}^{(m)}_\mathrm{A}\|_2^2}. \tag{27}$$

2) Acoustic Contrast: The acoustic contrast (AC) is a physical measure that quantifies the relative sound pressure energy difference between the sound pressure in bright and dark zones. Using the previously defined notation, it can be defined for zone A as:

$$\mathrm{AC}_\mathrm{A} = 10 \log_{10} \left(\frac{N_\mathrm{B} \sum_{m \in \mathcal{A}} \|\mathbf{p}^{(m)}_{\mathrm{A}\to\mathrm{A}}\|_2^2}{N_\mathrm{A} \sum_{m \in \mathcal{B}} \|\mathbf{p}^{(m)}_{\mathrm{A}\to\mathrm{B}}\|_2^2} \right). \tag{28}$$

3) Mean Perceptual Evaluation of Audio Quality (PEAQ): A perceptual measure tailored especially to evaluate the audio quality of a sound zone does not exist. Instead, in this paper, the Perceptual Evaluation of Audio Quality (PEAQ) model is used. PEAQ is a well-known perceptual measure standardized by the ITU-R [28]. PEAQ is designed for the evaluation of distortion due to audio coding. Given a reference and a distorted audio signal, PEAQ : $\mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ aims to predict the objective difference grade (ODG) between the two signals on a scale from -4 (bad) to 0 (excellent). In this work, the mean PEAQ rating for a zone A is defined as:

$$\overline{\mathrm{PEAQ}}_\mathrm{A} = \frac{1}{N_\mathrm{A}} \sum_{m \in \mathcal{A}} \mathrm{PEAQ}\left(\mathbf{t}^{(m)}_\mathrm{A}, \mathbf{p}^{(m)}_{\mathrm{A}\to\mathrm{A}}\right). \tag{29}$$

Note that possible interference from zone B is not considered.

4) Mean Distraction: In work by Francombe et al., "distraction" was found to be the most pertinent attribute for characterizing the listener experience in an audio-on-audio interference scenario [29]. In later work, a perceptual model was proposed predicting how distracted a listener will be when exposed to a given target and interfering audio signals [30]. Distraction has been shown to be correlated with the overall experienced quality within a sound zone [31]. The distraction model, denoted as function "Distraction : $\mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^+$", makes a prediction of how distracted the listener will be from the target signal as a result of the interfering signal on a scale from 0 (corresponding to not at all distracting) to 100 (corresponding to overpowering). A real-time viable version of the distraction model was later proposed by Rämö et al. [32], which simplified stages of the original model to reduce computational complexity. This real-time version is the implementation used in this work. To evaluate the performance of the algorithm, a prediction of the average experienced distraction for listeners in zone B due to the sound pressure leaking in from zone A is made. It is defined as follows:

$$\overline{\mathrm{Distraction}}_\mathrm{B} = \frac{1}{N_B} \sum_{m \in \mathcal{B}} \mathrm{Distraction}\left(\mathbf{l}^{(m)}_\mathrm{B}, \mathbf{p}^{(m)}_{\mathrm{A}\to\mathrm{B}}\right). \tag{30}$$

In this case, the interferer is taken to be the sound leaking from zone A to zone B, denoted by $\mathbf{p}^{(m)}_{\mathrm{A}\to\mathrm{B}}$. The target audio is taken to be the latent audio in zone B, denoted by $\mathbf{l}^{(m)}_\mathrm{B}$.

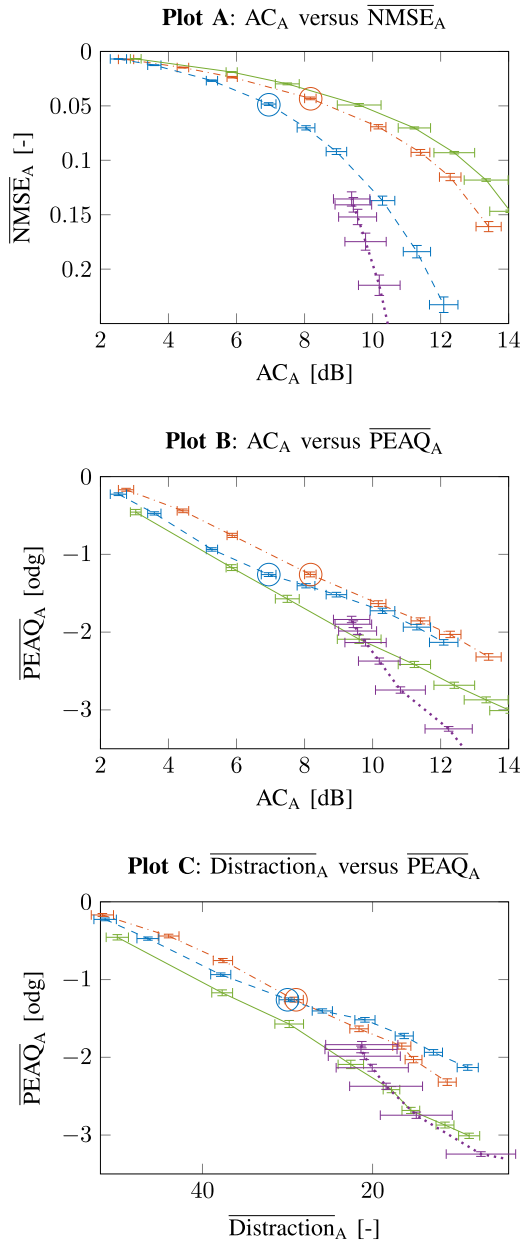

Fig. 3. Evaluation of the three proposed algorithms in the simulated room described in Section V-A: Par (-·-·-), Taal (- - -), Reference (——) and AP-VAST (·····), with measures described in Section V-B for various constraint values Q_0 with results averaged over all evaluation points for all media items. A circle is placed on the point of calibration where $Q_0 = 1$ for Par and Taal versions of the algorithm. Error bars depict corresponding 95% confidence intervals.

C. Simulation Results

In Fig. 3, the simulations described in Section V-A are shown evaluating the sound present in zone A using the measures defined in Section V-B. For the proposed methods, each data point represents a specific constraint values Q_0 from (15) and (26), averaged over all validation points and all media items. For AP-VAST, each data point corresponds to a different number of eigenvectors taken into account when constructing the control filters. This parameter is varied between utilizing only the largest eigenvector $V = 1$ and utilizing all eigenvectors $V = 4000$, in

increments of 500 for each data point. Sound zone algorithms typically trade-off between bright zone sound quality and dark zone sound pressure leaking into the bright-zone. To show this trade-off, all plots have a measure on the x-axis which captures an aspect of the dark zone performance (acoustic contrast or mean distraction), whilst having a measure on the y-axis which captures an aspect of the bright zone performance (mean NMSE or mean PEAQ in zone A). The axes have been given such that moving upwards along the y-axis or left to right along the x-axis corresponds to an improvement in the respective measure. As the constraint value Q_0 aims to constraint the (perceived) bright zone quality, it decreases going from left to right along the x-axis. Special attention is given to the constraint value $Q_0 = 1$ for the Par and Taal versions of the algorithm. As discussed in Section III, this is the point of calibration where the distortion signal is "just noticeable" in presence of the masking signal. For AP-VAST, increasing the number of eigenvectors V adds more detail to the resulting control filter with each additional eigenvector included, yielding improved (perceived) bright zone quality. As such, moving downwards on along the y-axis corresponds to a decreasing number of eigenvectors.

First, in Fig. 3(a), the two physical measures, namely mean NMSE and acoustic contrast, are shown for zone A for different values of Q_0 and V. It can be seen that the reference adaptive sound zone approach attains a lower NMSE than both perceptually adaptive approaches and AP-VAST for the same acoustic contrast for all values of constraint Q_0. The reference approach having the best performance for this measure is to be expected, as it optimizes directly over the NMSE. Note that, as described in Section V-B, NMSE does not model perception.

Accordingly, in Fig. 3(b), the mean PEAQ ODG bright zone quality measure is used in place of the mean NMSE. This time, for the same contrast, both the Par- and Taal-based perceptually adaptive approaches attain a lower PEAQ ODG score for all constraint values. AP-VAST outperforms the reference approach when $V \geq 2000$, but is in all cases outperformed by the proposed, perceptual approaches. The Par model-based approach achieves the best performance. For a fixed PEAQ ODG, it can be seen that the Par approach attains an additional 2 to 4 dB in acoustic contrasts for the same bright zone quality when compared to the reference.

Furthermore, in Fig. 3(c) the mean distraction is used in place of the acoustic contrast. Similarly, it can be seen that for a fixed mean PEAQ ODG, the proposed perceptually adaptive approaches yield lower distraction scores, corresponding to less distracting sound zones. Hence, it seems that when considering perceptual measures, the perceptually adaptive approaches outperform the non-perceptual reference approaches. Due to the large confidence intervals for distraction, it cannot be concluded how AP-VAST compares to the other methods.

From the preceding comparison between AP-VAST and the proposed perceptual approaches, it can be seen that it is outperformed for both NMSE and PEAQ versus AC as depicted by Fig. 3(b). Notably, even when including all $V = 4000$ eigenvectors are included in the solution, AP-VAST does not attain near-perfect PEAQ nor NMSE scores as the proposed approaches do. This limitation can be explained by the particular choice of $\mu = 1$ or the regularization required in the joint diagonalization of the spatial correlation matrices.

When comparing the Par- and Taal-based implementations, the mean PEAQ and contrast scores suggest that the two perceptual approaches perform similarly, with Par typically performing slightly better in terms of acoustic contrast. When considering the mean PEAQ and mean distraction, it can be seen that the Taal and Par based methods perform similarly at the point of calibration $Q_0 = 1$, with the Taal based method overtaking Par for higher constraint values (right of the point denoting $Q_0 = 1$). Despite being a more complicated perceptual model, the Taal based approach shows lackluster performance for low constraint values (left of the point denoting $Q_0 = 1$). One possible explanation for this is that, as described in Section III-B, the Taal model assumes that the distortion signal is small relative to the masking signal. It is possible that the leakage at lower constraint values is large, leading to inaccurate estimates of the distortion detectability, subsequently leading to poor results. Interestingly, it can also be seen that the "just noticeable" constraint value $Q_0 = 1$ corresponds to a PEAQ ODG close to -1.0, which corresponds to "Perceptible, but not annoying" [28], which is in line with the perceptual interpretation of distortion detectability.

The simulation results summarized in Fig. 3 show that the proposed algorithms perform competitively when compared to the state of the art. Note that the potential solution space for each frame spanned by the parameters μ and V for AP-VAST likely includes the solutions of the proposed methods. However, a fixed choice of μ and V might not consistently satisfy quality constraints as given by (26).

The quality constraint is shown to be perceptually consistent, with the constraint value of $Q_0 = 1$ yielding similar results for both perceptual approaches with respect to both PEAQ and distraction. This property provides the operator of the algorithm a consistent point of reference when tuning.

D. Understanding Differences Between Proposed Methods

To better understand how the proposed perceptual sound zone approaches are attaining better contrast than the reference approach for the same PEAQ ODG, an additional investigation is conducted to sketch the differences between the behaviour of the resulting control filters. To this end, a new measure is introduced, namely the mean target attenuation for zone A:

$$\overline{\mathrm{TA}}_{\mathrm{A}} = \frac{1}{N_{\mathrm{A}}} \sum_{m \in \mathcal{A}} 20 \log_{10} \left(\frac{\|\mathbf{p}_{\mathrm{A}\to\mathrm{A}}^{(m)}\|_2}{\|\mathbf{t}_{\mathrm{A}}^{(m)}\|_2} \right). \tag{31}$$

The target attenuation captures the average reduction in sound pressure energy between achieved and target sound pressures in zone A. This measure is of interest as it was found that the algorithms have a tendency to reduce the bright zone energy to minimize the dark zone leakage error. One explanation is that, after hitting the physical limits of the leakage reduction achievable with the line array, one can still reduce the dark zone leakage by simply "turning down" the bright zone. Naturally, the quality constraint Q_0 limits the degree to which this is possible. However, the various versions of the algorithm may make this trade-off differently.

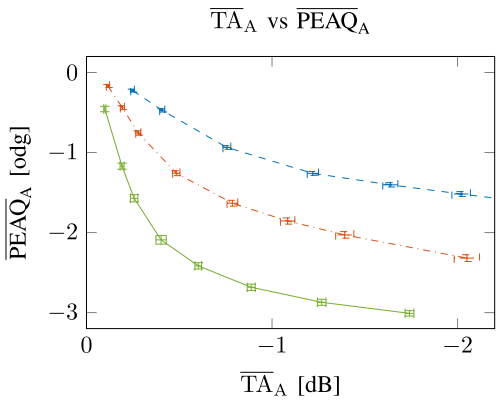

Fig. 4. Mean Target Attenuation for zone A ($\overline{\text{TA}}_\text{A}$), versus mean PEAQ score for zone A ($\overline{\text{PEAQ}}_\text{A}$) for Par (-·-·-), Taal (- - -), and Reference (——). The error bars depict 95% confidence intervals over the respective mean.

To investigate this, in Fig. 4, the mean target attenuation is plotted against the mean PEAQ ODG score for various constraint values. It can be seen that the perceptual approaches attain the same attenuation as the reference at lower PEAQ ODG scores, with the Taal-based approach attaining the lowest PEAQ ODG scores for all constraint values. This hints at the perceptually-based approaches using energy in a more efficient way: they achieve a certain bright zone quality with the same or less sound pressure energy by prioritizing perceptually relevant components. This serves as a possible explanation as to why the perceptual approaches can achieve an additional 2 to 4 dB of contrast at the same level of bright zone quality: when less sound pressure is used in the bright zone, less sound pressure will leak to the dark zone.

VI. Conclusion

This paper discusses three sound zone algorithms. First, in Section II a traditional sound zone approach based on sound pressure prediction was presented, featuring constraints on the reproduced sound pressure in the bright zone, allowing one to specify a desired quality of reproduced audio. This approach was then augmented in Section IV with two implementations of the distortion detectability model; one version by van de Par et al. and another by Taal et al. To investigate the benefit of including perceptual information, the traditional approach was compared to the two perceptual approaches in Section V. It was shown that the traditional approach outperforms the perceptual approaches in terms of physical measures such as Normalized Mean Square Error (NMSE) and Acoustic Contrast (AC). However, when using perceptual measures such as the Perceptual Evaluation of Audio Quality (PEAQ) score or distraction, both perceptual approaches outperformed the reference. The algorithm using the distortion detectability proposed by van de Par et al. achieves 2 to 4 dB of AC at the same PEAQ quality level when compared to the reference, which is significant enough to motivate potential listening tests to corroborate these results in the future. It is shown that one possible explanation for this result is that the perceptual approaches achieve a certain level of bright zone quality with less energy, resulting in less leakage to the dark zone. Additional future work includes algorithm optimization for real-time applications.

Acknowledgment

The authors would also like to thank Jon Francombe for his helpful comments and feedback.

References

[1] J. Cheer, S. J. Elliott, and M. F. S. Gálvez, "Design and implementation of a car cabin personal audio system," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 412–424, 2013.
[2] F. Heuchel, D. Caviedes Nozal, and F. T. Agerkvist, "Sound field control for reduction of noise from outdoor concerts," *J. Audio Eng. Soc.*, no. 10107, Oct. 2018.
[3] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 81–91, Mar. 2015.
[4] J.-W. Choi and Y.-H. Kim, "Generation of an acoustically bright zone with an illuminated region using multiple sources," *J. Acoust. Soc. America*, vol. 111, no. 4, pp. 1695–1700, 2002.
[5] M. B. Møller and J. Østergaard, "A moving horizon framework for sound zones," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 256–265, 2020.
[6] J. K. Nielsen, T. Lee, J. R. Jensen, and M. G. Christensen, "Sound zones as an optimal filtering problem," in *Proc. IEEE 52nd Asilomar Conf. Signals Syst. Comput.*, 2018, pp. 1075–1079.
[7] Y. Cai, M. Wu, and J. Yang, "Design of a time-domain acoustic contrast control for broadband input signals in personal audio systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 341–345.
[8] M. F. S. Gálvez, S. J. Elliott, and J. Cheer, "Time domain optimization of filters used in a loudspeaker array for personal audio," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 11, pp. 1869–1878, Nov. 2015.
[9] L. Vindrola, M. Melon, J.-C. Chamard, and B. Gazengel, "Use of the filtered-x least-mean-squares algorithm to adapt personal sound zones in a car cabin," *J. Acoust. Soc. America*, vol. 150, no. 3, pp. 1779–1793, 2021.
[10] T. Lee, J. K. Nielsen, and M. G. Christensen, "Towards perceptually optimized sound zones: A proof-of-concept study," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 136–140.
[11] T. Lee, J. K. Nielsen, and M. G. Christensen, "Signal-adaptive and perceptually optimized sound zones with variable span trade-off filters," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2412–2426, 2020.
[12] J. Donley and C. Ritz, "Multizone reproduction of speech soundfields: A perceptually weighted approach," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2015, pp. 342–345.
[13] J. Francombe et al., "Perceptually optimized loudspeaker selection for the creation of personal sound zones," *J. Audio Eng. Soc.*, no. 4-1, Sep. 2013.
[14] D. Wallace and J. Cheer, "Design and evaluation of personal audio systems based on speech privacy constraints," *J. Acoust. Soc. America*, vol. 147, no. 4, pp. 2271–2282, 2020.
[15] J. Donley and C. Ritz, "An efficient approach to dynamically weighted multizone wideband reproduction of speech soundfields," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, 2015, pp. 60–64.
[16] F. Olivieri, F. M. Fazi, S. Fontana, D. Menzies, and P. A. Nelson, "Generation of private sound with a circular loudspeaker array and the weighted pressure matching method," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 8, pp. 1579–1591, Aug. 2017.
[17] 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 J. Adv. Signal Process.*, vol. 2005, no. 9, pp. 1–13, 2005.

[18] C. Taal and R. Heusdens, "A low-complexity spectro-temporal based perceptual model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2009, pp. 153–156.

[19] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Trans. Audio Speech, Lang. Process.*, vol. 20, no. 5, pp. 1553–1564, Jul. 2012.

[20] E. Moulines, O. A. Amrane, and Y. Grenier, "The generalized multidelay adaptive filter: Structure and convergence analysis," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 14–28, Jan. 1995.

[21] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: JHU Press, 2013.

[22] M. Hu, H. Zou, J. Lu, and M. G. Christensen, "Maximizing the acoustic contrast with constrained reconstruction error under a generalized pressure matching framework in sound zone control," *J. Acoust. Soc. America*, vol. 151, no. 4, pp. 2751–2759, 2022.

[23] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[24] J. O. Smith, "Spectral audio signal processing," 2011, Accessed: Dec. 30, 2022, [Online]. Available: http://ccrma.stanford.edu/jos/sasp/

[25] E. A. Habets, "Room impulse response generator," vol. 1, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, Tech. Rep. 2.2.4, 2006.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. America*, vol. 65, no. 4, pp. 943–950, 1979.

[27] Geneva: European Broadcasting Union, Revision 4, Recommendation number EBU R 128, Aug. 2020. [Online]. Available: tech.ebu.ch

[28] T. Thiede et al., "PEQA-the ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29, 2000.

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

[30] J. Francombe, R. Mason, and M. D. S. Bech, "A model of distraction in an audio-on-audio interference situation with music program material," *J. Audio Eng. Soc.*, vol. 63, no. 1/2, pp. 63–77, 2015.

[31] K. Baykaner et al., "The relationship between target quality and interference in sound zone," *J. Audio Eng. Soc.*, vol. 63, no. 1/2, pp. 78–89, 2015.

[32] J. Rämö, S. Bech, and S. H. Jensen, "Real-time perceptual model for distraction in interfering audio-on-audio scenarios," *IEEE Signal Process. Lett.*, vol. 24, no. 10, pp. 1448–1452, Oct. 2017.