Signal-Adaptive and Perceptually Optimized Sound Zones with Variable Span Trade-Off Filters

Lee, Taewoong; Nielsen, Jesper Kjær; Christensen, Mads Græsbøll

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

DOI (link to publication from Publisher):
10.1109/TASLP.2020.3013397

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

Signal-Adaptive and Perceptually Optimized Sound Zones with Variable Span Trade-Off Filters

Taewoong Lee, *Student Member, IEEE,* Jesper Kjær Nielsen, *Member, IEEE,* and Mads Græsbøll Christensen, *Senior Member, IEEE*

***Abstract*—Creating sound zones has been an active research field since the idea was first proposed. So far, most sound zone control methods rely on either an optimization of physical metrics such as acoustic contrast and signal distortion or a mode decomposition of the desired sound field. By using these types of methods, approximately 15 dB of acoustic contrast between the reproduced sound field in the target zone and its leakage to other zone(s) has been reported in practical set-ups, but this is typically not high enough to satisfy the people inside the zones. In this paper, we propose a sound zone control method shaping the leakage errors so that they are as inaudible as possible for a given acoustic contrast. The shaping of the leakage errors is performed by taking the time-varying input signal characteristics and the human auditory system into account when the loudspeaker control filters are calculated. We show how this shaping can be performed using variable span trade-off filters, and we show theoretically how these filters can be used for trading signal distortion in the target zone for acoustic contrast. The proposed method is evaluated based on physical metrics such as acoustic contrast and perceptual metrics such as STOI. The computational complexity and processing time of the proposed method for different system set-ups are also investigated. Lastly, the results of a MUSHRA listening test are reported. The test results show that the proposed method provides more than 20% perceptual improvement compared to existing sound zone control methods.**

***Index Terms*—Adaptive control, human auditory system, masking effect, sound zones, variable span trade-off filters.**

I. Introduction

SOUND zones are different listening areas in the same acoustic environment for different audio contents, and these zones are created by controlling a set of loudspeakers. Typically, two types of sound zones are considered: a bright zone and a dark zone. The bright zone is a confined region in which the desired sound field is reproduced as faithfully as possible, whereas the dark zone is a confined region in which the energy of a reproduced sound field is suppressed as much as possible. These two zones are created by filtering the signals fed into the loudspeakers, and multiple bright zones can be obtained by superimposing the individual bright and dark zones for every input signal. Many different applications of sound zones have been studied, including outdoor concerts [1], automobile cabins [2]–[6], pedestrian alert systems [7], mobile devices [8], personal computers [9], and other applications [10], [11].

Many different methods for designing the loudspeaker control filters have been proposed over the last two decades since the concept was first introduced in [12]. Generally, these control methods seek to reproduce the desired sound field in the bright zone as faithfully as possible while also suppressing its leakage to the dark zone as much as possible. The proposed methods can be largely divided into three categories: mode matching methods, acoustic contrast control (ACC) methods, and pressure matching (PM) methods. Mode matching methods are based on that any sound field can be decomposed as an infinite sum of spatial harmonics. In practice, however, the sum is truncated up to a finite number of spatial harmonics often referred to as modes. The fundamental idea is based on [13], and several subsequent mode matching methods have been proposed [14]–[17].

The ACC methods are designed to maximize the acoustic contrast, defined as the ratio of the acoustic potential energies between the bright and dark zones, and this is achieved by solving a generalized eigenvalue problem [18]. Since ACC only optimizes the acoustic contrast, it will in general not maintain the spatial characteristics of the desired sound field. Consequently, the ACC methods are most useful in situations where the spatial characteristics are either not important or very hard to reproduce due to complicated, dynamic acoustics environments such as in car cabins [2]–[6]. Various variations of the ACC method have been proposed. These include the energy difference maximization [19], the planarity control [20], subband optimization [3], multiple constraints on the acoustic contrast for different frequency bands [6], and the broadband ACC (BACC) method [21]. The BACC method is different from the other methods in that it operates in the time-domain instead of the frequency-domain. Since the BACC method typically will produce control filters that will filter out most of the energy in the input signal, except for the few frequencies where the maximum acoustic contrast can be obtained, the reproduced sound field will typically be severely distorted. Various ways of mitigating this problem have been proposed in [22]–[24].

The PM methods produce control filters that minimize the reproduction error, defined as the difference between the reproduced and desired sound fields in the bright and dark zones. The original method was proposed in [25], [26]. Compared to the ACC methods, the signal distortion is much smaller, but so is the acoustic contrast. To allow the user to trade-off these two, largely two types of combination method have been studied. In [27], a method which combines the energy difference maximization method [19] and PM [26] in order to control the acoustic contrast and the reproduction error has been proposed. In [28], a method sometimes referred to as

T. Lee, J. K. Nielsen, and M. G. Christensen are with the Audio Analysis Lab, CREATE, Aalborg University, 9000 Aalborg, Denmark (e-mail: tlee, jkn, mgc@create.aau.dk)

ACC-PM has been proposed, and it is a more flexible PM method where the user can control the relative importance of reproducing the desired sound field and minimizing acoustic potential energy in the dark zone. The ACC-PM method has also been proposed in a broadband version in [29]. We note in passing that, despite its name, ACC-PM is actually not a combination of the ACC and PM methods and is also referred to by different names, e.g., in [30], [31]. A true combination of the BACC and PM methods in the time-domain has recently been proposed in [32].

Until now, an acoustic contrast of more than approximately 15 dB has only been reported in highly idealized experiments where, e.g., an impractical number of loudspeakers are used, the acoustic environment is time-invariant, or the performance is evaluated using oracle knowledge of the acoustic environment [15]. Unfortunately, however, a much higher contrast than 15 dB is needed, as reported in [33]. In [34], [35], it was also found that a target-to-interferer ratio (TIR) of at least 25 dB is needed. TIR is a metric closely related to the acoustic contrast, but it measures the ratio of either the acoustic potential energy or loudness between the reproduced and interfering sound fields in a given zone (see [36] for more on this).

Except for [17] where the sound zones were optimized for preserving speech privacy and for [37] where the pre-echoes were controlled over the attenuation of reflections in a reverberant environment, existing sound zone control methods design the control filters by minimizing physical metrics. A problem of quantifying the performance using physical metrics such as acoustic contrast and signal distortion is that they do not directly relate to the human auditory system. Moreover, the loudspeaker control filters are typically designed assuming input signals with flat spectra. The main advantage of this is that the control filters can be designed offline, but the disadvantage is that array effort[1] is wasted on controlling input signal frequency components which might not be present in the input signal or are inaudible. This is a general disadvantage of the frequency-domain methods in which the control filters are designed independently for every frequency bin. With the exception of [39], sound zone control methods in the frequency-domain do not trade-off the reproduction error in one frequency bin for the reproduction error in another frequency bin.

In this paper, we propose a perceptually optimized sound zone control method in the time-domain, which takes both the input signal characteristics and the human auditory system into account on a segment-by-segment level and gives explicit control of the trade-off between (weighted) acoustic contrast and signal distortion. This approach is inspired by perceptual audio coding, where quantization errors have been successfully hidden by exploiting the characteristics of the human auditory system. Famously in the early 1990s [40], the so-called 13 dB miracle [41, Ch. 10] demonstrated that this approach drastically lowered the requirements to the signal-to-quantization noise level without impacting the perceived quality, and these principles have later been standardized in, e.g., MPEG-1/2 Layer-3 (MP3) [42], [43]. In the sound zones application, we have reproduction errors instead of quantization errors. By using masking curves for designing weighting filters that shape the reproduction errors in a perceptually meaningful way, we can, therefore, ensure that the largest control effort is spent on maximizing the contrast and/or minimizing the reproduction error in the perceptually most important frequency regions. The proposed sound zone control method will be based on the variable span linear filter, which is a subspace approach initially proposed for signal enhancement [44]–[48] (see [49] for more on the relation between these problems). An interesting feature of the proposed method is that it reduces to existing sound zone control methods, such as broadband PM, BACC, and broadband ACC-PM, in special cases. We remark that this paper is an extension of our preliminary work reported in [32], which considered only the un-weighted case, and [50], which considered the weighted but non-adaptive case. Moreover, we here also report more elaborate experimental analyses and results, including a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) listening test [51].

The paper is organized as follows: in Sec. II, the sound zone control method with an arbitrary weighting of the reproduction error is explained, and it is shown how the input signal characteristics can be taken into account. In Sec. III, we discuss how the weighting filters are designed to take the characteristics of the human auditory system into account. Furthermore, it is extended to explain how the input signals are segmented in blocks and how the loudspeaker control filters are updated. In Sec. IV, the performance of the proposed method is evaluated via not only typical physical metrics such as the acoustic contrast (AC), the normalized signal distortion (nSDP), and the TIR, but also perceptual metrics, including the short-time objective intelligibility (STOI) [52] and the instantaneous perceptual similarity measure (PSMt) from the perception model based audio quality assessment method (PEMO-Q)[2] [54]. In addition to this, the results of a MUSHRA listening test are reported. Finally, in Sec. V, the paper is concluded.

II. A Weighted VAST Framework

In this section, the proposed weighted variable span trade-off (VAST) framework is described. To do this, we consider the simple system setup depicted in Fig. 1. The figure shows the bright and dark zones as spatially confined regions sampled by M_{B} and M_{D} microphone positions, respectively. Moreover, the figure shows L loudspeakers, with the lth loudspeaker having the finite impulse response (FIR) control filter with filter coefficients $\boldsymbol{q}_l$ and the input signal $x[n]$. As alluded to in the introduction, we can design multiple bright zones by superimposing the solutions to the individual bright and dark zones for each input signal. Throughout the theoretical part of this paper, we, therefore, consider the problem of creating a bright zone and a dark zone, and we use subscripts $_{\mathrm{B}}$ and $_{\mathrm{D}}$ to represent the bright and dark zones, respectively.

[1]The array effort is defined as the sum of mean squared control filters [38].

[2]PEMO-Q was chosen because it shows a higher prediction accuracy in known data and a more robust prediction performance on completely new data over the perceptual evaluation of audio quality (PEAQ) in [53].

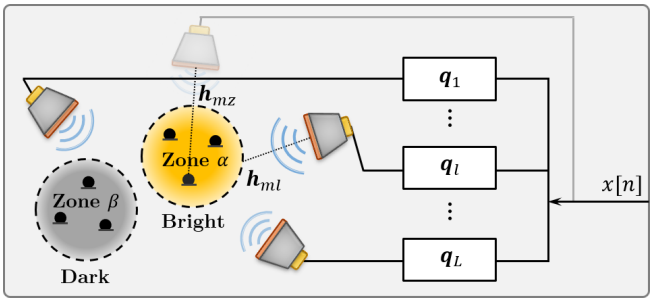

Fig. 1. An illustration of a system geometry of sound zones. The input signal $x[n]$ is fed into L loudspeakers after being filtered by the corresponding control filter $\{\boldsymbol{q}_l\}_{l=1}^{L}$. The RIR from loudspeaker l to control point m and the impulse response from virtual source z to control point m is represented as $\boldsymbol{h}_{ml}$ and $\boldsymbol{h}_{mz}$, respectively.

The reproduced sound pressure $p_m[n]$ at microphone position or control point m is represented by the linear convolution between input signal $x[n]$, the L control filters $\{\boldsymbol{q}_l\}_{l=1}^{L}$ of length J, and the L room impulse responses (RIRs) $\{\boldsymbol{h}_{ml}\}_{l=1}^{L}$ of length K, i.e.,

$$\begin{aligned} p_m[n] &= \sum_{l=1}^{L}\sum_{k=0}^{K-1}\sum_{j=0}^{J-1} x[n-k-j]h_{ml}[k]q_l[j] \\ &= \sum_{l=1}^{L} \boldsymbol{y}_{ml}^T[n]\boldsymbol{q}_l = \boldsymbol{y}_m^T[n]\boldsymbol{q}, \end{aligned} \tag{1}$$

where

$$\boldsymbol{y}_{ml}[n] = \boldsymbol{X}[n]\boldsymbol{h}_{ml}, \tag{2}$$

$$\boldsymbol{h}_{ml} = \begin{bmatrix} h_{ml}[0] & \cdots & h_{ml}[K-1] \end{bmatrix}^T, \tag{3}$$

$$\boldsymbol{X}[n] = \begin{bmatrix} x[n] & \cdots & x[n-K+1] \\ \vdots & \ddots & \vdots \\ x[n-J+1] & \cdots & x[n-K-J+2] \end{bmatrix}, \tag{4}$$

$$\boldsymbol{y}_m[n] = \begin{bmatrix} \boldsymbol{y}_{m1}^T[n] & \cdots & \boldsymbol{y}_{mL}^T[n] \end{bmatrix}^T, \tag{5}$$

$$\boldsymbol{q} = \begin{bmatrix} \boldsymbol{q}_1^T & \cdots & \boldsymbol{q}_L^T \end{bmatrix}^T, \tag{6}$$

$$\boldsymbol{q}_l = \begin{bmatrix} q_l[0] & \cdots & q_l[J-1] \end{bmatrix}^T. \tag{7}$$

The known signal vector $\boldsymbol{y}_{ml}[n]$ is the uncontrolled reproduced sound pressure at control point m originating from loudspeaker l, as this is what we have when there is no control over the zones, i.e., the control filters are all equal to the Kronecker delta function. The goal is then to design the control filters $\boldsymbol{q}$ so that the reproduced sound pressure $p_m[n]$ matches a desired sound pressure $d_m[n]$ across all control points as well as possible. Typically, the desired pressures are all 0 for the control points in the dark zone, whereas those in the bright zone are defined as part of a sound field generated by a virtual source z emitting $x[n]$. Thus, the desired sound pressure at control point m is defined as

$$d_m[n] = \begin{cases} (h_{mz} * x)[n] & m \in \mathcal{M}_{\mathrm{B}} \\ 0 & m \in \mathcal{M}_{\mathrm{D}} \end{cases}, \tag{8}$$

where $*$ denotes the linear convolution operator, $\mathcal{M}_{\mathrm{B}}$ and $\mathcal{M}_{\mathrm{D}}$ are the set of control point indices for the bright and dark zones, respectively, and $h_{mz}[n]$ is the impulse response from the virtual source z to control point m, as depicted in Fig. 1. Note that sound zone control methods have to implicitly perform dereverberation in order to match the desired and reproduced sound fields if the desired sound field is defined in an anechoic environment.

In sound zone control, two zones labeled α and β are typically considered as illustrated in Fig. 1. If we consider two zones, each having their own desired sound field, then the bright and dark zones for audio input signal $x^{(\alpha)}[n]$ are zone α and zone β, respectively, and those for audio input signal $x^{(\beta)}[n]$ are zone β and zone α, respectively. To this end, multiple bright zones can be obtained when the two reproduced sound fields are superposed.

How close the reproduced sound field is to the desired sound field can be quantified by the reproduction error, defined as the difference between the desired and reproduced sound pressures across all control points in a given zone such that

$$\varepsilon_m[n] = d_m[n] - p_m[n]\ . \tag{9}$$

More generally, we can filter the reproduction error by a weighting filter $w_m[n]$ so that

$$\tilde{\varepsilon}_m[n] = (w_m * \varepsilon_m)[n] = \tilde{d}_m[n] - \tilde{p}_m[n]\ , \tag{10}$$

where, e.g., $\tilde{p}_m[n]$ means that $p_m[n]$ has been filtered with the weighting filter $w_m[n]$. If we plug this into (1), we obtain the weighted and reproduced sound pressure at control point m as

$$\tilde{p}_m[n] = (w_m * p_m)[n] = \sum_{l=1}^{L} \tilde{\boldsymbol{y}}_{ml}^T[n]\boldsymbol{q}_l = \tilde{\boldsymbol{y}}_m^T[n]\boldsymbol{q} \tag{11}$$

where $\tilde{\boldsymbol{y}}_{ml}^T[n]$ is defined as in (2), except for that the source signal $x[n]$ is pre-filtered with the weighting filter. Note that the weighting filter is assumed to be known and is used to shape the reproduction error according to some design criterion. This will be elaborated upon in the next section.

We are now able to measure the distance between $\tilde{p}_m[n]$ and $\tilde{d}_m[n]$ for each of the zones. This can describe how much distortion is present in the bright zone and how much power is remaining in the dark zone. This allows us to define the weighted signal distortion power (SDP) $\tilde{\mathcal{S}}_B(\boldsymbol{q})$ and the weighted residual error power $\tilde{\mathcal{S}}_D(\boldsymbol{q})$, respectively, as

$$\begin{aligned} \tilde{\mathcal{S}}_{\mathrm{B}}(\boldsymbol{q}) &= \frac{1}{|\mathcal{M}_{\mathrm{B}}|N}\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_{\mathrm{B}}} |\tilde{\varepsilon}_m[n]|^2 \\ &= \tilde{\sigma}_d^2 - 2\boldsymbol{q}^T\tilde{\boldsymbol{r}}_{\mathrm{B}} + \boldsymbol{q}^T\tilde{\boldsymbol{R}}_{\mathrm{B}}\boldsymbol{q}, \end{aligned} \tag{12}$$

$$\tilde{\mathcal{S}}_{\mathrm{D}}(\boldsymbol{q}) = \frac{1}{|\mathcal{M}_{\mathrm{D}}|N}\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_{\mathrm{D}}} |\tilde{\varepsilon}_m[n]|^2 = \boldsymbol{q}^T\tilde{\boldsymbol{R}}_{\mathrm{D}}\boldsymbol{q}, \tag{13}$$

where N is the number of observations and

$$\tilde{\sigma}_d^2 = \frac{1}{|\mathcal{M}_\mathrm{B}|N} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_\mathrm{B}} |\tilde{d}_m[n]|^2, \tag{14}$$

$$\tilde{\boldsymbol{r}}_\mathrm{B} = \frac{1}{|\mathcal{M}_\mathrm{B}|N} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_\mathrm{B}} \tilde{\boldsymbol{y}}_m[n] \tilde{d}_m[n], \tag{15}$$

$$\tilde{\boldsymbol{R}}_\mathrm{C} = \frac{1}{|\mathcal{M}_\mathrm{C}|N} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_\mathrm{C}} \tilde{\boldsymbol{y}}_m[n] \tilde{\boldsymbol{y}}_m^T[n], \tag{16}$$

with $\tilde{\boldsymbol{R}}_\mathrm{C}$ for $\mathrm{C} \in \{\mathrm{B}, \mathrm{D}\}$ being the spatial correlation matrix for the corresponding zone, $\tilde{\boldsymbol{r}}_\mathrm{B}$ being the spatial correlation vector for the bright zone, and $\tilde{\sigma}_d^2$ being the variance of the desired sound field.

Using these definitions, we can pose the convex optimization problem

$$\text{minimize } \tilde{\mathcal{S}}_\mathrm{B}(\boldsymbol{q}) \text{ subject to } \tilde{\mathcal{S}}_\mathrm{D}(\boldsymbol{q}) \leq \epsilon, \tag{17}$$

where ϵ is a nonnegative scalar representing the power allowed in the dark zone. The Lagrangian function corresponding to the problem in (17) is

$$\mathcal{L}(\boldsymbol{q}) = \tilde{\mathcal{S}}_\mathrm{B}(\boldsymbol{q}) + \mu(\tilde{\mathcal{S}}_\mathrm{D}(\boldsymbol{q}) - \epsilon), \tag{18}$$

where $\mu \geq 0$ is the Lagrange multiplier. As also assumed in the signal enhancement literature [47], this Lagrange multiplier is here treated as a user-defined parameter that controls the trade-off between minimizing $\tilde{\mathcal{S}}_\mathrm{B}(\boldsymbol{q})$ and suppressing $\tilde{\mathcal{S}}_\mathrm{D}(\boldsymbol{q})$. We note in passing that minimizing (18) for $\mu = 1$ and μ equal to a constant produces the broadband PM solution and the broadband ACC-PM solution, respectively, when no weighting is applied, i.e., $w[n]$ is the Kronecker delta function, and the input signal is assumed to have a flat spectrum.

The matrices $\tilde{\boldsymbol{R}}_\mathrm{B}$ and $\tilde{\boldsymbol{R}}_\mathrm{D}$ are real, symmetric, and at least semi-positive definite matrices. Provided that $\tilde{\boldsymbol{R}}_\mathrm{D}$ is positive definite, these properties allow us to compute a joint diagonalization for those two matrices [49], [55, Ch. 8.7]. As exploited for signal enhancement [47], we can use this diagonalization to obtain more control over the trade-off between the SDP and the acoustic contrast. Specifically, we obtain this control by solving (18) with a low-rank approximation to the control filters $\boldsymbol{q}$. The two matrices $\tilde{\boldsymbol{R}}_\mathrm{B}$ and $\tilde{\boldsymbol{R}}_\mathrm{D}$ can be jointly diagonalized as

$$\boldsymbol{U}_{LJ}^T \tilde{\boldsymbol{R}}_\mathrm{B} \boldsymbol{U}_{LJ} = \boldsymbol{\Lambda}_{LJ}, \qquad \boldsymbol{U}_{LJ}^T \tilde{\boldsymbol{R}}_\mathrm{D} \boldsymbol{U}_{LJ} = \boldsymbol{I}_{LJ}, \tag{19}$$

where $\boldsymbol{\Lambda}_{LJ} = \mathrm{diag}(\lambda_1, \cdots, \lambda_{LJ})$ is a diagonal matrix containing the generalized eigenvalues in descending order, i.e., $\lambda_1 \geq \cdots \geq \lambda_{LJ} \geq 0$, $\boldsymbol{I}_{LJ}$ is the $LJ \times LJ$ identity matrix, and $\boldsymbol{U}_{LJ}$ is a nonsingular matrix containing the generalized eigenvectors sorted according to the eigenvalues. The matrices $\boldsymbol{U}_{LJ}$ and $\boldsymbol{\Lambda}_{LJ}$ are computed by solving the eigenvalue problem

$$\tilde{\boldsymbol{R}}_\mathrm{D}^{-1} \tilde{\boldsymbol{R}}_\mathrm{B} \boldsymbol{U}_{LJ} = \boldsymbol{U}_{LJ} \boldsymbol{\Lambda}_{LJ} \ . \tag{20}$$

It is worth noting that $\tilde{\boldsymbol{R}}_\mathrm{D}$ is typically positive definite when $M_D \mathrm{min}(N, K + J - 1) \geq LJ$.

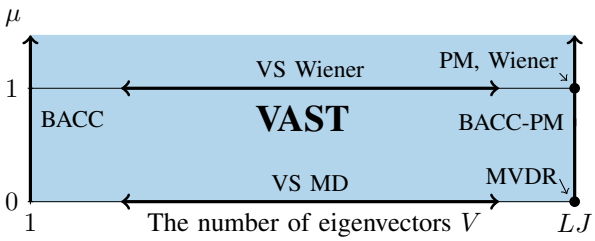

Fig. 2. VAST plane that illustrates how the various special cases of the VAST solutions are related in terms of a function of the user parameters $1 \leq V \leq LJ$ and $\mu \geq 0$.

Since any vector can be represented as a linear combination of the columns of a nonsingular matrix, $\boldsymbol{q}$ can be written as

$$\boldsymbol{q} = \boldsymbol{U}_{LJ} \boldsymbol{a}_{LJ}, \tag{21}$$

where $\boldsymbol{a}_{LJ}$ is an LJ coefficient vector. If we plug (21) into (12) and (13), we obtain

$$\tilde{\mathcal{S}}_\mathrm{B}(\boldsymbol{U}_{LJ}\boldsymbol{a}_{LJ}) = \tilde{\sigma}_d^2 - 2\boldsymbol{a}_{LJ}^T \boldsymbol{U}_{LJ}^T \tilde{\boldsymbol{r}}_\mathrm{B} + \boldsymbol{a}_{LJ}^T \boldsymbol{\Lambda}_{LJ} \boldsymbol{a}_{LJ}, \tag{22}$$

$$\tilde{\mathcal{S}}_\mathrm{D}(\boldsymbol{U}_{LJ}\boldsymbol{a}_{LJ}) = \boldsymbol{a}_{LJ}^T \boldsymbol{a}_{LJ}. \tag{23}$$

Interestingly, we can observe from (23) that $\tilde{\mathcal{S}}_\mathrm{D}$ is only represented by $\boldsymbol{a}_{LJ}$. Hence, this joint diagonalization leads us to analyze how $\tilde{\mathcal{S}}_\mathrm{B}$ and $\tilde{\mathcal{S}}_\mathrm{D}$ behave in terms of the eigen information. Furthermore, we benefit from introducing a $V(\leq LJ)$-rank approximation by forcing the $LJ - V$ smallest eigenvalues to 0, which directly reduces $\tilde{\mathcal{S}}_\mathrm{D}$. How this affects $\tilde{\mathcal{S}}_\mathrm{B}$ is explained later. Now we can approximate $\boldsymbol{q}$ by using the first V eigenvectors such that

$$\boldsymbol{q} \approx \boldsymbol{U}_V \boldsymbol{a}_V, \tag{24}$$

where $1 \leq V \leq LJ$ and optimize $\mathcal{L}$ over $\boldsymbol{a}_V$ instead of $\boldsymbol{q}$ directly. The cost function (18) is then

$$\begin{aligned} \mathcal{L}(\boldsymbol{U}_V \boldsymbol{a}_V) = \tilde{\sigma}_d^2 - 2\boldsymbol{a}_V^T \boldsymbol{U}_V^T \tilde{\boldsymbol{r}}_\mathrm{B} + \boldsymbol{a}_V^T \boldsymbol{\Lambda}_V \boldsymbol{a}_V \\ + \mu(\boldsymbol{a}_V^T \boldsymbol{a}_V - \epsilon). \end{aligned} \tag{25}$$

The solution to this is analytically derived and given by

$$\begin{aligned} \boldsymbol{a}_{\text{P-VAST}}(V, \mu) &= \arg\min_{\boldsymbol{a}_\mathrm{V}} \mathcal{L}(\boldsymbol{U}_V \boldsymbol{a}_\mathrm{V}) \\ &= \left[\boldsymbol{\Lambda}_V + \mu \boldsymbol{I}_V\right]^{-1} \boldsymbol{U}_V^T \tilde{\boldsymbol{r}}_\mathrm{B} \ . \end{aligned} \tag{26}$$

Finally, we plug (26) into (24) and obtain the control filter as

$$\boldsymbol{q}_{\text{P-VAST}}(V, \mu) = \boldsymbol{U}_V \boldsymbol{a}_{\text{P-VAST}}(V, \mu) = \sum_{v=1}^{V} \frac{\boldsymbol{u}_v^T \tilde{\boldsymbol{r}}_\mathrm{B}}{\lambda_v + \mu} \boldsymbol{u}_v, \tag{27}$$

where λ_v and $\boldsymbol{u}_v$ are the vth generalized eigenvalue and eigenvector, respectively.

Interestingly, we can obtain different solutions by varying V and μ, including many existing solutions as shown in Fig. 2 assuming no weighting of the reproduction error and an input signal with a flat spectrum. For example, the BACC solution is obtained when $V = 1$, the broadband PM (or Wiener) solution is obtained when $V = LJ$ and $\mu = 1$, the broadband ACC-PM solution is obtained when $V = LJ$, and the MVDR solution is obtained when $V = LJ$ and $\mu = 0$. As we have shown in App. A, the maximum acoustic contrast but also

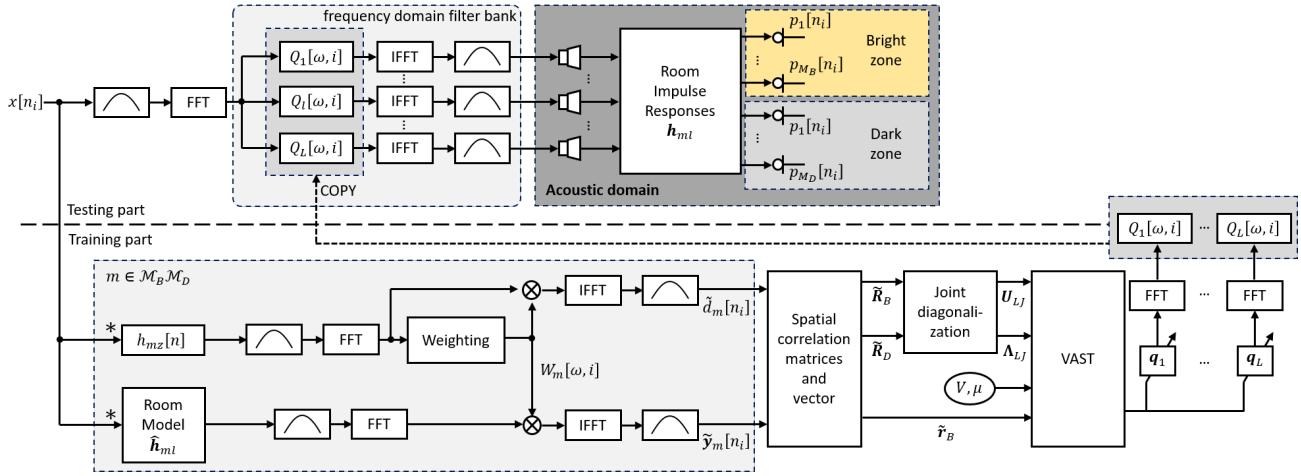

Fig. 3. The block diagram to get the reproduced sound field at the given time segment i for AP-VAST. Note that * denotes the linear convolution in the time-domain, $W_m[\omega, i]$ denotes the weighting filter in the frequency-domain at frequency ω and time segment i at control point m, FFT and IFFT denote the fast Fourier transform and its inverse, respectively, $\mathcal{M}_\mathrm{B}\mathcal{M}_\mathrm{D}$ denotes a concatenated index set of $\mathcal{M}_\mathrm{B}$ and $\mathcal{M}_\mathrm{D}$, and $Q_l[\omega, i]$ denotes the control filters in the frequency-domain at frequency ω and time segment i for loudspeaker l.

the maximum SDP is obtained for $V = 1$. Increasing V and keeping μ fixed decrease both, and we obtain the minimum SDP but also the minimum acoustic contrast for the maximum number of eigenvectors, i.e., $V = LJ$. Thus, V is also a user parameter that can be used for controlling the trade-off between the acoustic contrast and the SDP. Clearly, μ also controls aspects of this trade-off and can, e.g., be set so that the acoustic contrast is completely ignored (the MVDR solution).

III. Signal-Adaptive and Perceptually Optimized Sound Zones

As alluded to in the introduction, audio coding was revolutionized by exploiting simple mathematical models for the human auditory system. These models encode the principle that a certain sound also known as maskee becomes less audible or inaudible in the presence of a stronger masker close to the maskee in the time- and/or frequency-domain [56]. This phenomenon is generally referred to as the masking effect, and it allows us to make large modifications to audio signals without changing how they are perceived by humans. In the sound zones application, we can only find a set of control filters that renders the reproduction errors to exactly zero provided that the multiple-input/multiple-output inverse theorem (MINT) conditions are satisfied [57], [58], i.e., it is necessary (but not sufficient) to have more loudspeakers than control points, something which is seldom satisfied in practice. We thus cannot avoid making a reproduction error, but we can seek to shape this error to be as inaudible as possible. In the proposed sound zone control method, the reproduction error is shaped in the following way. For control point m, we compute a masking curve from a given input signal based on a psychoacoustic model, e.g., [59]. This masking curve is defined as an amplitude spectrum describing the sound pressure level below which any sound is modeled to be inaudible. The weighting filter is calculated as the reciprocal of the masking curve. In other words, we apply a small weight to those part of the spectrum where the masker has a high power, whereas we penalize reproduction errors more by applying a larger weight in those part of the spectrum where the masker has a low or no power.

TABLE I
Desired signal and masker for control point m and given time segment i

Zone	$\alpha \quad (m \in \mathcal{M}_\alpha)$	$\beta \quad (m \in \mathcal{M}_\beta)$
Desired signal	$d_m^{(\alpha)}[n_i]$	$d_m^{(\beta)}[n_i]$
Masker	$d_m^{(\alpha)}[n_i]$	$d_m^{(\beta)}[n_i]$

To compute the masking curve at control point m, we must first figure out where the control point is located. If it is in zone α, say, the masking curve is computed from the desired signal $d_m^{(\alpha)}[n]$ at this control point when zone α is the bright zone. Note that we have used the superscript $(\cdot)^{(\alpha)}$ on the desired signal to stress that this signal does not change with which zones are considered as bright or dark zones. Thus, when zone α is considered to be the bright zone, the masking curves for the control points in the dark zone, i.e., zone β, are calculated not from $d_m[n] = 0$ but from $d_m^{(\beta)}[n]$. Note that the masking curves are calculated from the desired signal to avoid an iterative procedure for computing the control filters, although the actual masker is the reproduced signal. For precisely this reason, the masking curves used in audio coding are also calculated from the unquantized signal, although the actual masker is the quantized signal [60]. The above discussion is summarized in Table I. If a zone is desired to be a dark zone, the masking curve for the zone will simply be the threshold in quiet.

Although average masking curves can be computed from audio signals, we can expect to obtain the best performance if the masking curves are updated on a segment basis so that the control filters are adapted to the current input signal segment. To do this, we divide $x[n]$ into I time segments, and $\boldsymbol{q}$ is calculated at each of these time segments. For the segment-

wise approach, the observation index n can be considered as a local time-index, and this is related to the global time-index n_i as

$$n_i = (N-\eta)(i-1)+n, \qquad i \in \mathcal{I}, \tag{28}$$

where $\mathcal{I}$ denotes the set of the segment indices, $\eta \in \{0,1,\ldots,N-1\}$ is the number of overlapping samples between segments, and $n = 0,1,\cdots,N-1$. This indexing is used in Fig. 3, which shows the implementation of the proposed sound zone control method, referred to as signal-adaptive and perceptually optimized variable span trade-off (AP-VAST). Fig. 3 shows that the weighting and the filtering with the control filters are implemented in the short-time Fourier-transform (STFT) domain with a 50% overlap and with identical analysis and synthesis windows given by [61]

$$g[n] = \sin\left\{\frac{\pi}{N}\left(n+\frac{1}{2}\right)\right\}. \tag{29}$$

This implementation of the time-varying filtering is adopted since it has proven successful in many speech and audio processing applications, including audio coding. The room model $\hat{\boldsymbol{h}}_{ml}$ indicates that the RIRs have been measured or modeled in advance. It should be noted that AP-VAST has a special case, which we refer to as perceptually optimized sound zones (P-VAST), that uses averaged input signal statistics and masking curves.

IV. Experimental Validation and Discussion

This section presents an evaluation of the proposed method and comparisons to the reference methods in both anechoic and reverberant environments. We here consider the case of two bright zones. In other words, two input signals are considered, and each of them becomes the desired signal in the corresponding zone. Individual problems of sound zones for each input signal are solved, then the reproduced sound fields are superimposed.

A. Set-up

As depicted in Fig. 4, a circular array with a radius of $r_c = 2$ m with eight omnidirectional loudspeakers evenly placed on the circumference was considered. The two virtual sources were located outside of the array at the same location, which was 0.5 m away from the 7th loudspeaker. It is depicted as a dashed line loudspeaker in Fig. 4. The zones were located in the interior of the loudspeaker array and spatially sampled by 25 control points on a 2D grid with $l_a = 5$ cm spacing between the control points to cover the size of a human head. The control points are shown as black dots in Fig. 4 and were used to calculate the control filters. Besides, 16 monitor points were used to evaluate the performance. These points are shown as gray crosses in Fig. 4 and located in between the control points. The centers of the two zones were $l_c = 2$ m apart from each other. All loudspeakers, control points, and the virtual sources were assumed to be located in the same plane at the height of 1.5 m. All the parameters that are common in all experiments are summarized in Table II.

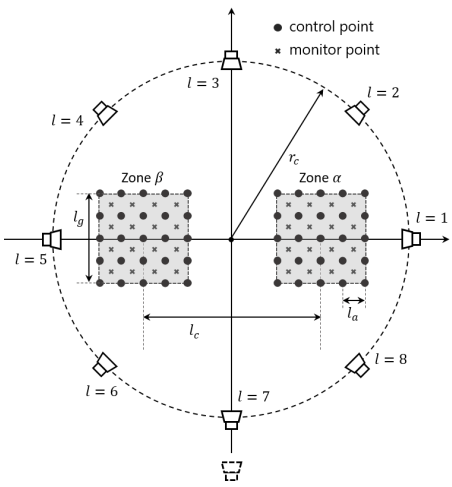

Fig. 4. The system geometry used in the validations. Note that the illustration does not follow the actual scale but is magnified for better visualization.

For the performance evaluation, the typical physical metrics – AC, nSDP, and TIR – as well as the perceptual metrics – STOI [52] and PSMt[3] [54] – were used. AC, nSDP, and TIR are here defined as

$$\mathrm{AC} = 10\log_{10}\left(\frac{M_D}{M_B}\frac{\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_B}|p_m[n]|^2}{\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_D}|p_m[n]|^2}\right), \tag{30}$$

$$\mathrm{nSDP}_m = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1}|p_m[n]-d_m[n]|^2}{\sum_{n=0}^{N-1}|d_m[n]|^2}\right), \tag{31}$$

$$\mathrm{TIR}_m^{(\alpha;\beta)} = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1}|p_m^{(\alpha)}[n]|^2}{\sum_{n=0}^{N-1}|p_m^{(\beta)}[n]|^2}\right). \tag{32}$$

Although these metrics are typically defined for an input signal with a flat spectrum, we calculated the metrics from the desired and reproduced sound fields because they should depend on those sound fields. Since AC is defined as a zone-wise metric, AC was calculated zone-wise. On the other hand, nSDP, TIR, and the perceptual metrics were calculated point-wise. Note that the metrics calculated point-wise can be easily converted to the zone-wise metrics and vice versa. The mean values and the error bars with the 95% confidence intervals are shown for the metrics calculated point-wise[4].

It is also important to note that AC and nSDP are the metrics per input signal, whereas TIR is the metric per zone. Besides, the reference signal for STOI and PSMt is the desired signal at each point $d_m^{(\alpha)}[n]$, and the processed signal for STOI and PSMt is the observed signal that is the sum of the reproduced signal $p_m^{(\alpha)}[n]$ and $p_m^{(\beta)}[n]$ at the corresponding point. We used the freely available MATLAB toolboxes of STOI and PSMt for

[3]PSM and PSMt showed a similar trend throughout all experiments, so only the results of PSMt are displayed.

[4]For the case of multiple bright zones, a family of perceptual source separation metrics in [62] was also reviewed. Unlike STOI and PSMt, however, these metrics did not correlate very well with an informal listening test, so we did not include the metrics in this paper.

TABLE II
THE PARAMETER DETAILS FOR THE SIMULATIONS

Variable	Value	Variable	Value
M_B, M_D (control)	25	M'_B, M'_D (monitor)	16
L	8	r_c	2 m
l_c	2 m	l_g	0.2 m
l_a	0.05 m		
sampling frequency	16 kHz	speed of sound, c	343 m/s
K	3200	J	240

obtaining the results [52], [54], respectively, and the default values in each toolbox were used. Finally, the RIRs were calculated using the RIR generator toolbox [63], which is a MATLAB implementation of the image source method [64], for both the anechoic environment ($T_{60} = 0$ s) and the reverberant environment ($T_{60} = 200$ ms). Lastly, the 48 kHz RIRs were generated and then downsampled to 16 kHz.

As the reference methods for the performance comparison to the proposed method, AP-VAST, we used broadband PM (i.e., equivalent to broadband ACC-PM in [29] with $\xi = \mu/(1+\mu) = 0.5$), the frequency-domain ACC[5] [18], VAST [32], and P-VAST. As a baseline, we also evaluated the performance without any control, i.e., the control filters were all set equal to the Kronecker delta function. For AP-VAST, the time-varying weighting filters were obtained from masking curves computed from each 60 ms time segment with a 50% overlap, whereas the weighting filter for P-VAST was calculated from an averaged masking curve computed from all 60 ms time segments. The psychoacoustic model in [59] was used to calculate masking curves. AP-VAST followed the implementation described in Fig. 3. Lastly, a control filter length of $J = 240$ samples (i.e., 15 ms at a sampling frequency of 16 kHz which corresponds to a frequency grid of 66.67 Hz in the ACC method) for all methods, and a segment length of $N = 960$ samples (i.e., 60 ms at a sampling frequency of 16 kHz) for AP-VAST were used[6]. Therefore, the number of eigenvectors V was in the range from 1 to 1920 since $1 \leq V \leq LJ$. Note that the same (V, μ) for the two input signals were used for each data point in the following figures.

B. Performance evaluation on the proposed method AP-VAST

In the first experiment, we considered two input signals as six seconds of dialogues excerpt from the Disney movie "Zootopia" in two different languages (English and Danish). These were used for the desired signal in zone α and zone β, respectively. The energy of the two signals was calibrated to be identical, and they were downsampled from 44.1 kHz to 16 kHz. The number of segments for AP-VAST was equal to $I = 201$.

[5]Note that a regularization based on the truncated singular value decomposition (TSVD) in [28] was used and that the magnitude normalization factor of the control filter was calculated as described in [28] for the entire frequency range, except at the first frequency component (i.e., the DC-component) and the Nyquist frequency, which were both set to 0.

[6]To compute the joint diagonalization, $\tilde{\boldsymbol{R}}_D$ has to be full rank, which requires $M_D \min(N, K+J-1) \geq LJ$.

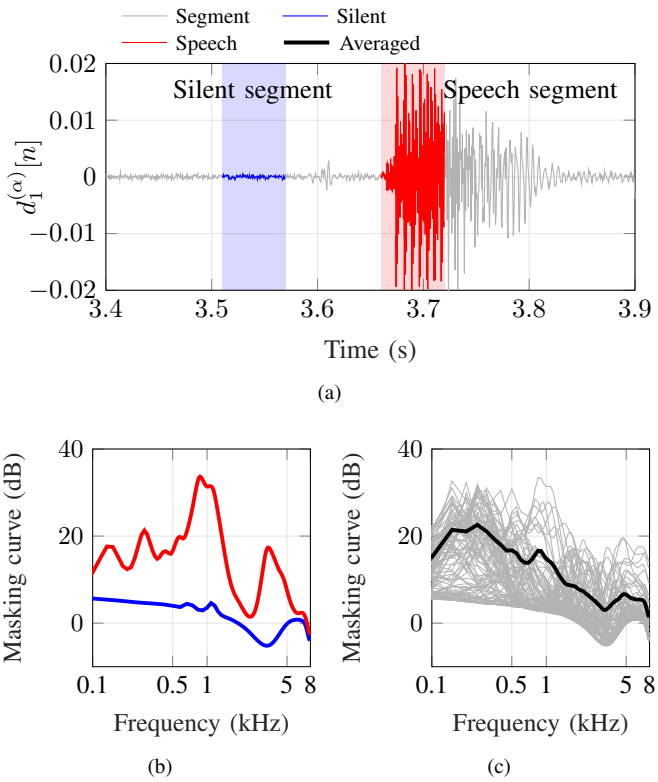

Fig. 5. (a) A subset of the speech signal for zone α in the first experiment, one of the speech segments and one of the silent segments are shown in red and in blue, respectively, (b) The masking curve (red) and the masking curve (blue) are computed from the speech segment and the silent segment, respectively, (c) Masking curves from each segment (gray) and the averaged masking curve across all the masking curves (black).

Masking curves of the speech signal for zone α are shown in Fig. 5. Note that the desired signal at the 1st control point $d_1^{(\alpha)}[n]$ is considered. As an example of the masking curves, one of the silent segments and one of the speech segments are depicted in blue and in red, respectively in panel (a). From these segments, the masking curves corresponding to each of the segments are calculated and plotted in Fig. 5 (b) with the same color as the segments. In other words, if the input segment barely contains any signal characteristics, a masking curve (blue) close to the threshold-in-quiet with a shallow slope is obtained. Otherwise, the masking curve (red) is calculated based on the corresponding input segment. Lastly, the masking curve (gray) from each segment and the averaged masking curve (black) across all the segments are shown in Fig. 5 (c). The averaged masking curve is used in P-VAST.

First, AC and nSDP obtained by AP-VAST from the speech signal for zone α for five different μ's are shown in Fig. 6. Regardless of μ, a clear trend is that AC and nSDP decrease with increasing V. For $V = 1$, in any case of μ, the highest AC but also the largest nSDP is obtained. The smallest nSDP along with the lowest AC is obtained when $V = LJ$ with a fixed μ. This trend certainly shows the trade-off between AC and SDP, which is the core property of AP-VAST.

As μ increases from $\mu = 0$, both AC and nSDP increase for a fixed V. The lower bound of AC and nSDP can be observed when $\mu = 0$. In this case, AP-VAST searches for the solution to minimize the signal distortion in the bright

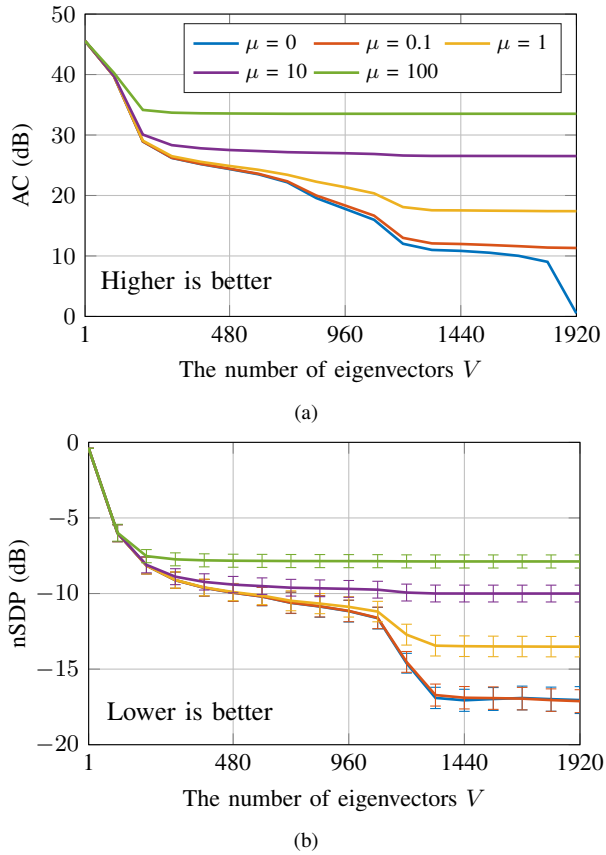

Fig. 6. (a) AC in dB and (b) nSDP in dB as a function of V for five different μ: $\mu = 0$ (blue), $\mu = 0.1$ (red), $\mu = 1$ (yellow), $\mu = 10$ (purple), and $\mu = 100$ (green). Note that nSDP is represented with the 95% confidence interval (error bars).

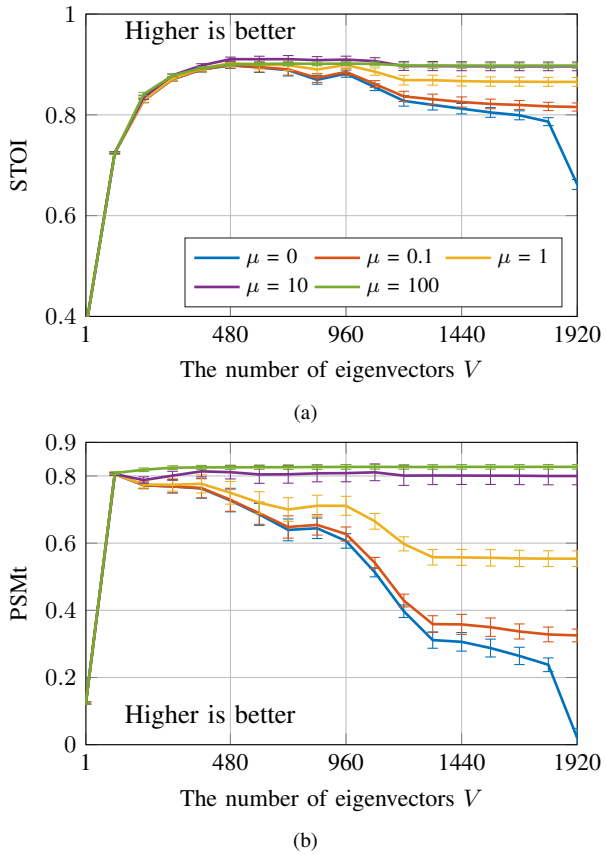

Fig. 7. (a) STOI in a range between 0.0 and 1.0 and (b) PSMt in a range of between -1.0 to 1.0 as a function of V for five different μ: $\mu = 0$ (blue), $\mu = 0.1$ (red), $\mu = 1$ (yellow), $\mu = 10$ (purple), and $\mu = 100$ (green). Note that both are plotted with the 95% confidence interval (error bars).

zone instead of reducing the residual power in the dark zone. Hence, the case of $V = 1920$ and $\mu = 0$ guarantees the least nSDP, but so is AC, which is close to 0 dB. Interestingly, AC and nSDP become less sensitive for V as μ increases. If μ is large enough, e.g., $\mu = 100$, the degradation on AC and nSDP with increasing V is smaller than when μ is small, e.g., $\mu = 1$. This trend can be interpreted as AP-VAST seeks the solution to reduce the residual power in the dark zone more than to minimize the signal distortion in the bright zone as μ increases. Even though TIR is not plotted in this experiment, TIR follows the same trend as AC.

Secondly, STOI and PSMt obtained by AP-VAST from the speech signal for zone α for five different μ's are shown in Fig. 7. As alluded to earlier in this section, the observed signal is the sum of the reproduced signal by input signal α and the interference signal by input signal β. This affects STOI and PSMt to decrease, especially as μ decreases and V increases. The lower bound of STOI and PSMt can be found for $\mu = 0$ as in AC and nSDP, and they decrease as V increases. PSMt drops sharply for $V \geq 960$ and $\mu < 1$ particularly because the interference is more dominant than the reproduced signal. Therefore, we can see this as AC and TIR are more important than nSDP in order to have higher STOI and/or PSMt. We can expect from the STOI metric that the reproduced sound is intelligible if $V \geq 120$ is used in this experiment[7]. On top of this, we can expect from PSMt that it has a better perception if $\mu \geq 1$. We can expect that STOI and PSMt will not decrease in the case of one bright zone and one dark zone as V increases or μ decreases because nSDP decreases. A similar trend in these metrics is also observed from zone β.

C. Performance comparison

In the previous experiment, we investigated AP-VAST with respect to the physical and perceptual metrics as a function of V and μ in Figs 6 and 7, respectively. In this experiment, a comparison between AP-VAST and the reference methods is carried out. Specifically, how the signal-adaptive approach improves the performance is investigated by comparing VAST, P-VAST, and AP-VAST. The same input signals as in the previous experiment are used, and μ is set to $\mu = 1$.

AC, nSDP, and TIR from the speech signal for zone α performed by five different methods are illustrated in Fig. 8. As seen in Fig. 8 (a), AC from all methods is improved compared to the initial AC, which is 0 dB due to the symmetry of the system. ACC and PM provide around 15.7 dB and 14.3 dB of AC, respectively, whereas VAST and P-VAST vary depending on V, but equal to or higher than 15 dB. PM gives the lower

[7]According to [52], a STOI score of more than 0.80 maps to approximately 100 % speech intelligibility.

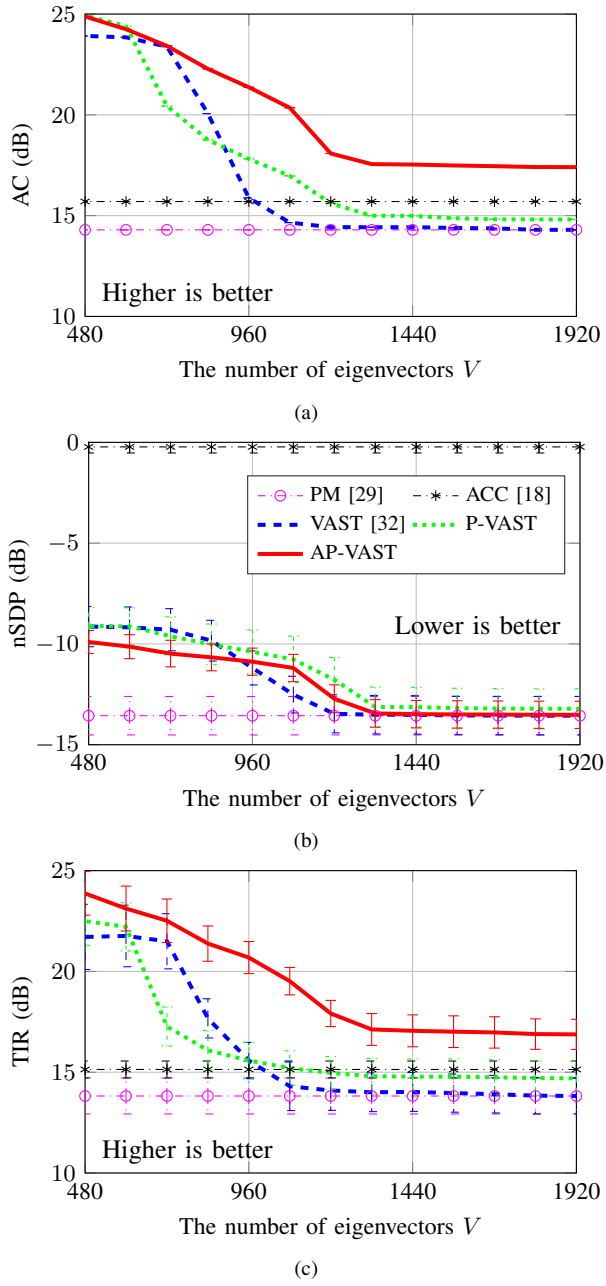

Fig. 8. (a) AC in dB, (b) nSDP in dB, and (c) TIR in dB as a function of V for five different methods: PM (purple dash-dot) [29], ACC (black dash-dot) [18], VAST (blue dash) [32], P-VAST (green dot), and AP-VAST (red solid). Note that nSDP and TIR are represented with the 95% confidence interval (error bars).

bound of AC, and we can observe that AC by VAST converges to the lower bound as V increases. AP-VAST provides the highest AC across V in this experiment and follows the same trend, which can be found in AC and nSDP, as in the previous experiment.

As depicted in Fig. 8 (b), ACC and PM provide around 0 dB and -13.5 dB of nSDP, respectively. Note that the initial nSDP is about 11.9 dB, but this is excluded in Fig. 8 (b) for better visualization. Interestingly, nSDP of not only VAST but also P-VAST and AP-VAST seems to be upper- and lower-bounded by ACC and PM, respectively. By comparing Figs. 8 (a) and

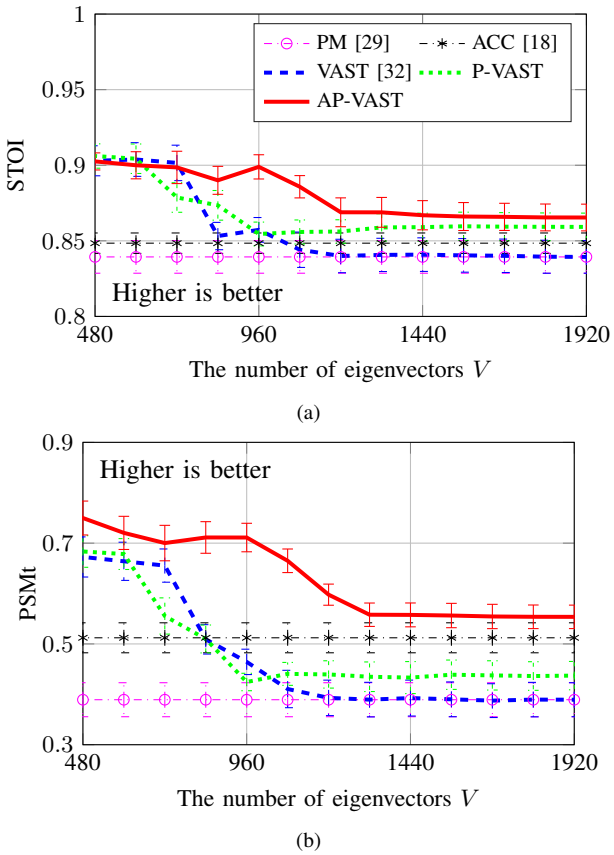

Fig. 9. (a) STOI in a range between 0.0 and 1.0 and (b) PSMt in a range between -1.0 to 1.0 as a function of V for five different methods: PM (purple dash-dot) [29], ACC (black dash-dot) [18], VAST (blue dash) [32], P-VAST (green dot), and AP-VAST (red solid). Note that both metrics are represented with the 95% confidence interval (error bars).

8 (b), we can observe that higher AC yields larger nSDP from ACC, PM, and VAST, which do not have perceptual weighting. However, this is not the case for P-VAST and AP-VAST since higher AC can be observed even at lower nSDP by comparing P-VAST and AP-VAST, e.g., when $V = 960$. As alluded to previously in this section, TIR follows the same trend in AC, as seen in Figs. 8 (a) and (c).

In Fig. 9, we observe STOI and PSMt. When there is no control, a STOI of 0.6 and a PSMt of -0.01 are observed but excluded in Fig. 9 for better visualization. We can observe that STOI and PSMt by VAST converge to that of PM as V increases. P-VAST gives higher scores than VAST, and finally AP-VAST provides the highest STOI and PSMt amongst the methods across any V.

From these experiments, we can observe that AP-VAST does not have the best performance across the physical metrics but does have it for the perceptual metrics.

D. Performance comparison in a reverberant environment

In the third and last experiment, we considered two input signals, which were 4.5 seconds of track 49 (English female speech) and track 50 (English male speech) excerpt from the EBU SQAM database [65]. They were the desired signal in zone α and zone β, respectively. As in the signals used in the previous experiments, the energy of the signals was

TABLE III
THE MEAN AND THE 95% CONFIDENCE INTERVAL OF THE PERFORMANCE METRICS OF THE SPEECH SIGNAL FOR ZONE β IN THE REVERBERANT ENVIRONMENT

Method	Parameter		Performance metric				
	$1 \leq V \leq 1920$	$\mu \geq 0$	AC (dB)	nSDP (dB)	TIR (dB)	STOI	PSMt
No control	NA[a]	NA	-1.3	14.2 ± 0.5	-0.1 ± 0.3	0.64 ± 0.004	0.14 ± 0.025
PM [29]	1920	1.0	10.5	$\mathbf{-9.9 \pm 0.7}$	11.5 ± 0.5	0.80 ± 0.003	0.39 ± 0.009
ACC [18]	NA	NA	9.2	0.4 ± 0.8	9.3 ± 0.4	0.77 ± 0.013	0.40 ± 0.014
VAST [32]	1080	1.0	12.2	-7.8 ± 0.7	$\mathbf{12.8 \pm 0.7}$	0.80 ± 0.004	0.43 ± 0.016
P-VAST	1080	1.0	**14.9**	-7.2 ± 0.6	11.8 ± 0.6	0.76 ± 0.004	0.40 ± 0.022
AP-VAST	1080	1.0	12.0	-8.4 ± 0.4	12.2 ± 0.3	$\mathbf{0.82 \pm 0.004}$	$\mathbf{0.54 \pm 0.010}$

[a] NA: Not applicable

TABLE IV
THE COMPUTATIONAL COMPLEXITY FOR THE CALCULATION OF SPATIAL STATISTICS AND CONTROL FILTERS

Method	Spatial statistics	Control filter
PM [29]	$\mathcal{O}(ML^2J^2)$	$\mathcal{O}(L^3J^3)$ for solving the least squares
ACC [18]	$\mathcal{O}(L^2J)$	$\mathcal{O}(L^3J)$ for solving J GEPs
VAST [32]	$\mathcal{O}(MNL^2J^2)$	$\mathcal{O}(L^3J^3)$ for computing JD
P-VAST	$\mathcal{O}(MNL^2J^2)$	$\mathcal{O}(L^3J^3)$ for computing JD
AP-VAST	$\mathcal{O}(IMNL^2J^2)$	$\mathcal{O}(IL^3J^3)$ for computing I JDs

GEP: generalized eigenvalue problem
JD: joint diagonalization

calibrated to be identical, and the signals were downsampled from 44.1 kHz to 16 kHz. The number of segments for AP-VAST was equal to $I = 135$.

For the reverberant environment, a room with $T_{60} = 200$ ms and a volume of 140 m^3 was considered. In order to compare AP-VAST to the reference methods, the user parameters V and μ are selected as $V = 1080$ and $\mu = 1$, respectively. Note that we can expect AP-VAST to have a lower nSDP as well as higher AC, STOI, and PSMt if a high μ is selected, which can be explained in Figs. 6 and 7. However, here $\mu = 1$ is specifically chosen in order to compare AP-VAST to PM directly. Although the dereverberation might not be performed well because the length of the control filter is shorter than that of the reverberation, the performance comparison is still fair since this applies to all the candidate methods.

The performance of all the metrics as a function of mean and 95% confidence interval is summarized in Table III. In general, compared to the performance in the anechoic environment depicted in Figs. 8 and 9 for $V = 1080$, performance degradation on all metrics is observed from all methods due to the reverberation. The highest AC, the minimum nSDP, and the largest TIR are obtained by P-VAST, PM, and VAST, respectively, but none of them provides the highest STOI or PSMt. Although AP-VAST provides neither the highest AC nor the lowest nSDP in this experiment, AP-VAST provides the highest STOI and PSMt.

E. Computational complexity and processing time

The experiments showed that AP-VAST outperformed the reference methods in terms of the perceptual metrics. Thus, there is a clear benefit to making a sound zone control algorithm signal-adaptive and perceptually optimized. The price for this, however, is a higher computational complexity since we must update the control filters for every segment instead of just once for all segments. To quantify this, we here use the big-O notation $\mathcal{O}$ from [55] to denote the computational complexity of an algorithm. By using this notation, the joint diagonalization in AP-VAST, P-VAST, and VAST has a computational complexity of order $\mathcal{O}(L^3J^3)$ where L and J are the number of loudspeakers and the control filter length, respectively. Since the joint diagonalization has to be performed for every segment in AP-VAST, the resulting complexity is $\mathcal{O}(IL^3J^3)$ where I is the number of segments. The joint diagonalization is performed from the spatial statistics. If the spatial statistics are computed naïvely, the computational complexity is $\mathcal{O}(MNL^2J^2)$ for every segment where M and N are the number of control points and the segment length, respectively. Note that the complexity becomes high for P-VAST if N becomes large. The broadband PM and ACC-PM, on the other hand, demand the same order of complexity as the joint diagonalization, i.e., $\mathcal{O}(L^3J^3)$ to solve a large least-squares problem. Since the ACC method is solved in the frequency-domain, we get many smaller problems rather than one big problem, one for every frequency bin, which results in a complexity of $\mathcal{O}(L^3)$ for solving the generalized eigenvalue problem and a complexity of $\mathcal{O}(L^2)$ for forming the spatial statistics in one of J frequency bins. The above discussion is summarized in Table IV.

Lastly, the mean processing times for calculating the spatial correlation matrix and the joint diagonalization by AP-VAST are shown in Fig. 10. Note that the 95% confidence interval was negligible compared to the processing time. The same setup and the input signals used in Sec. IV-B except for the number of loudspeakers were used for computing the processing times. Four different numbers of loudspeakers were chosen, $L = \{4, 8, 12, 16\}$; therefore, the corresponding dimensions of the spatial correlation matrices are $LJ = \{960, 1920, 2880, 3840\}$, respectively, because the length of control filters is $J = 240$. All timings were computed on a Windows 10 desktop PC (Dell OptiPlex 5040) with a 3.4 GHz Intel(R) Core(TM) i7-6700 CPU and 8 GB RAM using the function in MATLAB 2019a called `timeit`. As can be seen from Fig. 10, when $LJ = 960$ for a 60 ms time segment,

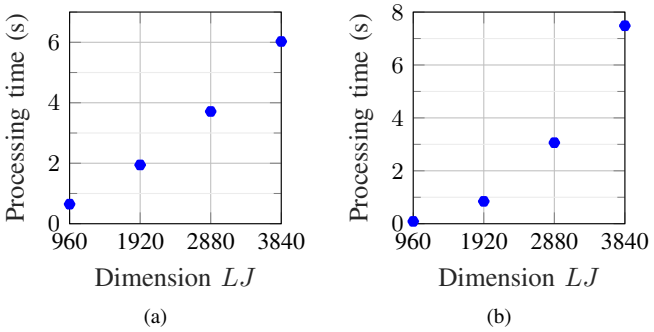

Fig. 10. The processing time of AP-VAST for calculating (a) The spatial correlation matrix $\boldsymbol{R}_{\mathrm{C}}$ and (b) The joint diagonalization. Four different numbers of loudspeakers $L = \{4, 8, 12, 16\}$ with a fixed control filter length $J = 240$ lead us to have four different dimensions LJ of $\boldsymbol{R}_{\mathrm{C}}$: $LJ = \{960, 1920, 2880, 3840\}$.

TABLE V
LIST OF EXCERPTS USED IN THE MUSHRA TESTS

Scenario	Data set	Zone	Excerpt
S1	D1	α	Female speech, Track 49[a]
	D2	β	Male speech, Track 50[a]
S2	D3	α	Pop music, Track 69[a]
	D4	β	Pop music, Track 70[a]
S3	D5	α	Piano solo
	D6	β	Orchestra, Track 66[a]
S4	D7	α	Guitar solo
	D8	β	Male speech
S5	D9	α	Zootopia dialogue in Danish
	D10	β	Zootopia dialogue in English

[a] EBU SQAM in [65]

the mean processing times are approximately 647 ms and 88 ms, respectively, and they are nearly 6027 ms and 7484 ms when $LJ = 3840$, respectively. The processing time for computing the joint diagonalization grows approximately eight times every time LJ doubles due to the cubic complexity, as summarized in Table IV. It should be noted that real-time in a practical setup would be challenging due to its substantial computational complexity.

F. Formal listening Test

In order to quantify the perceived performance of the candidate methods, we conducted a subjective listening test. The case of two bright zones was considered for five different scenarios. This led us to have ten different data sets[8], as summarized in Table V. In other words, the two reproduced sound fields were superposed; therefore, interference is present in the processed signal. In the listening test, the signal at the center of each zone was played back. Through this listening test, we evaluated the overall preference, i.e., the quality and the attenuation of the leakage to the other zone, of the candidate methods.

1) Set-up: As illustrated in Fig. 11, a uniform linear array of 16 equally-spaced loudspeakers was considered. Each zone,

[8]The audio examples of the reproduced signals are available online at the following link: https://tinyurl.com/apvast2020

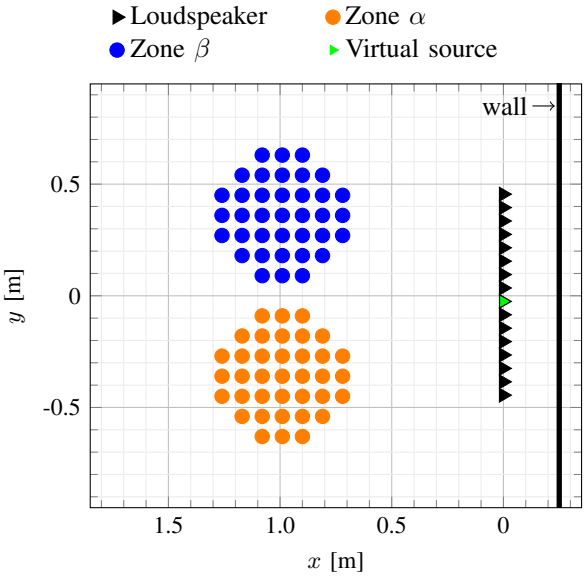

Fig. 11. The system geometry for the MUSHRA listening test. The number of loudspeakers (black triangle) $L = 16$, the number of control points (orange and blue circles) $M = 37$ for each zone, and the virtual source (green triangle) are shown.

which consists of 37 control points, was located in front of the loudspeaker array. The distance between the loudspeakers was 9 cm. The same distance was used for the space between the control points in each of the zones. A room with the dimensions 4.5 m $\times$ 4.5 m $\times$ 2.2 m and a reverberation time $T_{60} = 300$ ms was considered. The measured RIRs used in [39], [66] were used for the listening tests. The impulse response of the desired sound field $h_{mz}[n]$ for the bright zone was chosen from the RIR of the 8th loudspeaker, after being shortened to contain only the direct path component. The RIRs were resampled from 48 kHz to 16 kHz. Except for the above modifications, the same information described in Sec. IV-A was used. Therefore, V is in a range of $1 \leq V \leq 3840 = LJ$.

The user parameters V and μ were selected as $V = 3840$ and $\mu = 1$, respectively, for both AP-VAST and P-VAST. This choice allows us to directly compare the perceived performance of the perceptually optimized sound zones (AP-VAST and P-VAST) to that of the physically optimized sound zones (PM), as well as the perceived performance between AP-VAST and P-VAST.

2) MUSHRA test: A MUSHRA listening test was conducted according to the recommendation in [51] using a webMUSHRA software [67]. In total, 20 listeners with self-reported normal-hearing have participated in the tests. The listeners were asked to be located in a quiet place with wearing a pair of headphones[9]. The listening test was divided into a training phase and an evaluation phase. In the training phase, the listeners were asked to get familiar with the interface and all the processed signals. In the evaluation phase, the listeners had to rate seven differently processed signals, the so-called stimuli, per data set in a range of 0 to 100 according to the

[9]We conducted the listening tests using an online platform [67] due to COVID-19 lockdown.

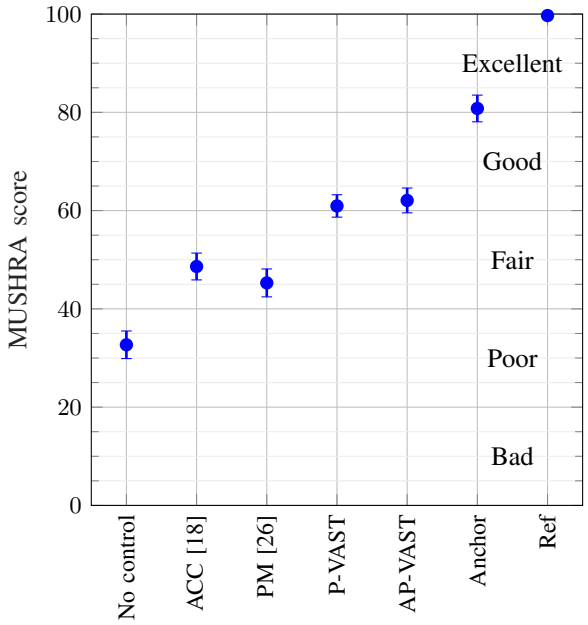

Fig. 12. The mean values and the 95% confidence intervals of all MUSHRA scores for four different methods and a hidden reference and the two anchors. In total, 1400 ratings, specifically, 200 ratings (20 participants for 10 different data sets) per method, were used. Note that the standard anchor and the hidden reference are denoted as Anchor and Ref, respectively.

TABLE VI
THE WILCOXON SIGNED-RANK TEST STATISTICS FOR DIFFERENT PAIRS OF THE SOUND ZONE CONTROL METHODS

Pair	PM P-VAST	PM AP-VAST	ACC P-VAST	ACC AP-PVAST	P-VAST AP-VAST
Z	-9.536	-10.066	-9.018	-9.580	-1.145
p	< 0.001	< 0.001	< 0.001	< 0.001	0.252

Z denotes the Z satistic.

quality with respect to the known reference. These stimuli included the hidden reference and a standard anchor (the low-pass filtered version of the reference at 3.5 kHz). The hidden reference was used for examining the consistency of the listener's responses. The no control version was used as an additional anchor. The remaining methods were ACC, PM, P-VAST, and AP-VAST. In the test, they were displayed in random order. Therefore, one listener had to give 70 ratings (10 data sets, 7 stimuli) in total.

The statistical analyses at a significance level α of 0.05 were conducted in SPSS 25 and MATLAB 2019a. A Friedman test [68] was performed[10], as suggested in [69], because a Shapiro-Wilk test [70], [71] found that the MUSHRA scores were not normally distributed. It should be noted that the reference and the two anchors are excluded in the analyses. From the results of the Friedman test, a statistically significant difference was found in the perceived performance across all data sets between the related groups (ACC, PM, P-VAST, and AP-VAST), $N_{\mathrm{F}} = 200, \chi^2(3) = 183.799$, $p < 0.001$

[10]Initially, a one-way analysis of variance (ANOVA) with repeated measures was considered to compare the mean MUSHRA scores between ACC, PM, P-VAST, and AP-VAST because all the ratings by the same listener are on the same continuous, dependent variable, i.e., the MUSHRA scores.

TABLE VII
THE WILCOXON SIGNED-RANK TEST STATISTICS BETWEEN AP-VAST AND P-VAST FOR DIFFERENT DATA SETS

Scenario	S1	S2	S3	S4	S5
Dataset	D1	D3	D5	D7	D9
Z	-1.269	-1.188	-3.511	-1.572	-2.073
p	0.205	0.235	$\underline{< 0.001}$	0.116	$\mathbf{0.038}$
Dataset	D2	D4	D6	D8	D10
Z	-0.763	-0.982	-3.287	-2.833	-2.199
p	0.445	0.326	$\mathbf{0.001}$	$\mathbf{0.005}$	$\mathbf{0.028}$

(**boldface**): H_0 is rejected, and AP-VAST has a higher mean MUSHRA score than P-VAST. (underlined): H_0 is rejected, and P-VAST has a higher mean MUSHRA score than AP-VAST. Otherwise, H_0 cannot be rejected due to insufficient evidence to reject it.

where N_{F} is the number of ratings per method, $\chi^2(d)$ is the Friedman's Chi-square statistic with d degree of freedom, and p is the significance of the result. It can be interpreted as at least one pairwise difference is present, which is not surprising at all, considering Fig. 12. Therefore, post hoc analysis with Wilcoxon signed-rank tests [72], [73] was separately conducted. A Bonferroni correction was applied, resulting in a corrected $\alpha = 0.05/5 = 0.01$. We can observe the statistically significant pairwise difference ($p < 0.001$) between all the combinations, as summarized in Table VI, except for the pair of P-VAST and AP-VAST.

We can visually see this difference by using a plot that shows the mean MUSHRA scores with the 95% confidence intervals, as shown in Fig. 12. First, we can observe that the lowest mean MUSHRA score (32.7) when no control is considered. One of the significant observations is that the perceptual approaches (AP-VAST and P-VAST) outperform the existing methods (ACC and PM) by at least 10 points in general, which is more than a 20% improvement. Specifically, a significant improvement, which is more than 15 points, is observed by comparing the scores between AP-VAST and PM. The mean MUSHRA scores for the four methods were as follows: 48.63 for ACC, 45.29 for PM, 60.95 for P-VAST, and 62.07 for AP-VAST. We emphasize that such a perceived difference was achieved even in the worst case of AC, when $V = LJ$ for a fixed μ, for AP-VAST and P-VAST. Secondly, the standard anchor obtained about the mean MUSHRA score of 80.8, which is the second-highest score. This observation can be interpreted that most listeners preferred the situation in which no interference is present, i.e., TIR is infinite, even though the processed signal is low-pass filtered. In other words, higher TIR and/or AC is preferred even the distortion (SDP) is present. This tendency seems more noticeable when speech is the desired signal than the case of music. Thirdly, the mean MUSHRA score by AP-VAST is slightly higher than that by P-VAST.

We conducted Wilcoxon signed-rank tests for all the data sets separately to identify the statistically significant pairwise difference between AP-VAST and P-VAST. A two-tailed paired t-test [74, Ch. 8.4] could not be applied because the pairwise difference was not normally distributed, as in the previous analysis. The null hypothesis H_0 is as follows: the

perceived performance between AP-VAST and P-VAST is the same. The null hypothesis is rejected if $p < 0.05$; otherwise, the null hypothesis cannot be rejected due to a lack of evidence to reject it at the significant level $\alpha = 0.05$. The test statistics of all the Wilcoxon tests are summarized in Table VII. The null hypothesis H_0 is rejected for the five data sets (D5, D6, D8, D9, and D10), which show a statistically significant difference in the perceived performance between AP-VAST and P-VAST. Specifically, the higher MUSHRA scores were obtained by AP-VAST over P-VAST from the four data sets: D6, D8, D9, and D10. We believe that segment-dependent V and μ based on certain design criteria for constraints rather than fixed V and μ could lead the optimal performance of AP-VAST, which will be dependent on the statistics of the input signals and different acoustic environments.

V. Conclusion

In this paper, we proposed a signal-adaptive method for creating perceptually optimized sound zones by using variable span trade-off filters in the time-domain. This method was achieved by taking the characteristics of input signals and the human auditory system into account segment-wise. The characteristics of input signals were integrated into the spatial correlation matrices, and the human auditory system was quantified mathematically by using a psychoacoustic model. Masking thresholds were calculated by using this model from the given input signal and used as perceptual weighting filters applied to the input signals. To this end, it allowed us to shape the reproduction error perceptually so that the interference becomes less or ideally inaudible to the listener in a given zone according to the human auditory system. Exploiting the joint diagonalization of the spatial correlation matrices allowed us to have a flexible control filter that trades-off the acoustic contrast and the signal distortion. Through validations in both anechoic and reverberant environments, the performance in terms of the physical metrics – AC, TIR, and nSDP – as well as the perceptual metrics – STOI and PSMt – was measured. The performance across all metrics, zones, and input signals was reasonably consistent, all indicating that the proposed method provides a perceptually better reproduction of the desired sound field, even though the physical metrics are not necessarily better. Lastly, through a MUSHRA listening test, it was verified that the perceptually optimized sound zones provide more than 20% better perceived performance in terms of the mean MUSHRA score compared to the existing sound zone control methods.

Appendix A

A. Acoustic contrast

Since the acoustic contrast $\gamma(\boldsymbol{q})$ is the ratio between the power in the bright and dark zones, it can be written as

$$\gamma(\boldsymbol{q}) = \frac{M_{\mathrm{D}}}{M_{\mathrm{B}}} \frac{\boldsymbol{q}^T \boldsymbol{R}_{\mathrm{B}} \boldsymbol{q}}{\boldsymbol{q}^T \boldsymbol{R}_{\mathrm{D}} \boldsymbol{q}}, \tag{33}$$

and if we plug $\boldsymbol{q}_{\text{P-VAST}}(V,\mu)$ from (27) into $\gamma(\boldsymbol{q})$, then it yields

$$\gamma\left(\boldsymbol{q}_{\text{P-VAST}}(V,\mu)\right) = \frac{M_{\mathrm{D}}}{M_{\mathrm{B}}} \frac{\boldsymbol{a}_{\text{P-VAST}}^T(V,\mu)\boldsymbol{\Lambda}_V \boldsymbol{a}_{\text{P-VAST}}(V,\mu)}{\boldsymbol{a}_{\text{P-VAST}}^T(V,\mu)\boldsymbol{a}_{\text{P-VAST}}(V,\mu)}. \tag{34}$$

If we consider V and $V+1$, respectively, then we obtain

$$\gamma\left(\boldsymbol{q}_{\text{P-VAST}}(V,\mu)\right) = \frac{M_{\mathrm{D}}}{M_{\mathrm{B}}} \frac{\sum_{v=1}^{V} |a_v|^2 \lambda_v}{\sum_{v=1}^{V} |a_v|^2}, \tag{35}$$

$$\gamma\left(\boldsymbol{q}_{\text{P-VAST}}(V+1,\mu)\right) = \frac{M_{\mathrm{D}}}{M_{\mathrm{B}}} \frac{\sum_{v=1}^{V+1} |a_v|^2 \lambda_v}{\sum_{v=1}^{V+1} |a_v|^2}, \tag{36}$$

where a_v is the vth element in $\boldsymbol{a}_{\text{P-VAST}}(V,\mu)$. Subtracting (36) from (35) and reducing to common denominator lead us to have

$$\begin{aligned}
&\gamma\left(\boldsymbol{q}_{\text{P-VAST}}(V,\mu)\right) - \gamma\left(\boldsymbol{q}_{\text{P-VAST}}(V+1,\mu)\right) \\
&= \frac{M_{\mathrm{D}}}{M_{\mathrm{B}}} \frac{|a_{V+1}|^2 \Big(\sum_{v=1}^{V} |a_v|^2 \lambda_v - \lambda_{V+1} \sum_{v=1}^{V} |a_v|^2 \Big)}{\sum_{v=1}^{V} |a_v|^2 \sum_{v=1}^{V+1} |a_v|^2} \\
&= \frac{M_{\mathrm{D}}}{M_{\mathrm{B}}} \frac{|a_{V+1}|^2 \Big\{ \sum_{v=1}^{V} |a_v|^2 \big(\lambda_v - \lambda_{V+1}\big) \Big\}}{\sum_{v=1}^{V} |a_v|^2 \sum_{v=1}^{V+1} |a_v|^2}.
\end{aligned} \tag{37}$$

Since $|a_v|^2$ and $|a_{V+1}|^2$ are nonnegative and $\lambda_v \geq \lambda_{V+1}$ (the equality holds when $\lambda_v = 0$), the acoustic contrast monotonically decreases for increasing V, i.e., $\gamma\left(\boldsymbol{q}_{\text{P-VAST}}(V,\mu)\right) \geq \gamma\left(\boldsymbol{q}_{\text{P-VAST}}(V+1,\mu)\right)$.

B. Signal distortion

If we plug $\boldsymbol{q}_{\text{P-VAST}}(V,\mu)$ from (27) into (12) and (13), we obtain

$$\begin{aligned}
\tilde{\mathcal{S}}_{\mathrm{B}}(\boldsymbol{q}_{\text{P-VAST}}(V,\mu)) &= \tilde{\sigma}_d^2 - 2\tilde{\boldsymbol{r}}_{\mathrm{B}}^T \boldsymbol{U}_V \boldsymbol{G}^{-1} \boldsymbol{U}_V^T \tilde{\boldsymbol{r}}_{\mathrm{B}} \\
&\quad + \tilde{\boldsymbol{r}}_{\mathrm{B}}^T \boldsymbol{U}_V \boldsymbol{G}^{-1} \boldsymbol{\Lambda}_V \boldsymbol{G}^{-1} \boldsymbol{U}_V^T \tilde{\boldsymbol{r}}_{\mathrm{B}} \\
&= \tilde{\sigma}_d^2 - \sum_{v=1}^{V} \frac{\lambda_v + 2\mu}{(\lambda_v + \mu)^2} |\boldsymbol{u}_v^T \tilde{\boldsymbol{r}}_{\mathrm{B}}|^2,
\end{aligned} \tag{38}$$

$$\begin{aligned}
\tilde{\mathcal{S}}_{\mathrm{D}}(\boldsymbol{q}_{\text{P-VAST}}(V,\mu)) &= \tilde{\boldsymbol{r}}_{\mathrm{B}}^T \boldsymbol{U}_V \boldsymbol{G}^{-2} \boldsymbol{U}_V^T \tilde{\boldsymbol{r}}_{\mathrm{B}} \\
&= \sum_{v=1}^{V} \frac{|\boldsymbol{u}_v^T \tilde{\boldsymbol{r}}_{\mathrm{B}}|^2}{(\lambda_v + \mu)^2},
\end{aligned} \tag{39}$$

where $\boldsymbol{G} = \boldsymbol{\Lambda}_V + \mu \boldsymbol{I}_V$. Interestingly, we can observe that $\tilde{\mathcal{S}}_{\mathrm{B}}(\boldsymbol{q}_{\text{P-VAST}}(V,\mu))$ decreases and $\tilde{\mathcal{S}}_{\mathrm{D}}(\boldsymbol{q}_{\text{P-VAST}}(V,\mu))$ increases monotonically for increasing V, respectively, because all variables in (38) and (39) are nonnegative. Finally, we plug (38) and (39) into (18), then we obtain

$$\mathcal{L}(\boldsymbol{q}_{\text{P-VAST}}(V,\mu)) = \tilde{\sigma}_d^2 - \sum_{v=1}^{V} \frac{|\boldsymbol{u}_v^T \tilde{\boldsymbol{r}}_{\mathrm{B}}|^2}{\lambda_v + \mu}, \tag{40}$$

which decreases for increasing V. Therefore, we can obtain the minimum reproduction error when all eigenvectors are used and μ is a positive value, which means that the residual power in the dark zone is still being controlled. Note that we obtain the minimum signal distortion if $\mu = 0$.

ACKNOWLEDGMENT

The authors would like to thank Dr.-Ing. M. Schneider and Assoc. Prof. E. A. P. Habets for providing the RIRs.

REFERENCES

[1] J. Brunskog, F. M. Heuchel, D. C. Nozal, M. Song, F. T. Agerkvist, E. F. Grande, and E. Gallo, "Full-scale outdoor concert adaptive sound field control," in *23rd Int. Congr. Acoust.*, Aachen, Germany, Sep. 2019, pp. 1170–1177.

[2] J. Cheer, S. J. Elliott, and M. F. Simón Gálvez, "Design and implementation of a car cabin personal audio system," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 412–424, Jul. 2013.

[3] J.-W. Choi, "Subband optimization for acoustic contrast control," in *Proc. 22nd Int. Congr. Sound Vib.*, Florence, Italy, 2015, pp. 849–856.

[4] P. N. Samarasinghe, W. Zhang, and T. D. Abhayapala, "Recent advances in active noise control inside automobile cabins: Toward quieter cars," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 61–73, Nov. 2016.

[5] H. So and J.-W. Choi, "Subband optimization and filtering technique for practical personal audio systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brighton, UK, May 2019, pp. 8494–8498.

[6] S. Widmark, "Causal MSE-optimal filters for personal audio subject to constrained contrast," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 5, pp. 972–987, May 2019.

[7] D. Quinn, J. Mitchell, and P. Clark, "Development of a next-generation audible pedestrian alert system for EVs having minimal impact on environmental noise levels: Project eVADER," in *Inter-Noise*, Melbourne, Australia, Nov. 2014.

[8] J. Cheer, S. J. Elliott, Y. Kim, and J.-W. Choi, "Practical implementation of personal audio in a mobile device," *J. Audio Eng. Soc.*, vol. 61, no. 5, pp. 290–300, 2013.

[9] J.-H. Chang, C.-H. Lee, J.-Y. Park, and Y.-H. Kim, "A realization of sound focused personal audio system using acoustic contrast control," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2091–2097, Apr. 2009.

[10] J.-M. Lee, T. Lee, J.-Y. Park, and Y.-H. Kim, "Generation of a private listening zone; Acoustic parasol," in *20th Int. Congr. Acoust.*, Sydney, Australia, Aug. 2010.

[11] M. F. Simón Gálvez, S. J. Elliott, and J. Cheer, "Personal audio loudspeaker array as a complementary TV sound system for the hard of hearing," *IEICE Trans. Fundamentals*, vol. E97-A, no. 9, pp. 1824–1831, 2014.

[12] W. F. Druyvesteyn and J. Garas, "Personal sound," *J. Audio Eng. Soc.*, vol. 45, no. 9, pp. 685–701, 1997.

[13] Y. J. Wu and T. D. Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 6, pp. 1711–1720, 2011.

[14] W. Zhang, T. D. Abhayapala, T. Betlehem, and F. M. Fazi, "Analysis and control of multi-zone sound field reproduction using modal-domain approach," *J. Acoust. Soc. Am.*, vol. 140, no. 3, pp. 2134–2144, Sep. 2016.

[15] W. Zhang, P. Samarasinghe, H. Chen, and T. D. Abhayapala, "Surround by sound: A review of spatial audio recording and reproduction," *Appl. Sci.*, vol. 7, no. 6, May 2017, Art. no. 532.

[16] W. Jin, W. B. Kleijn, and D. Virette, "Multizone soundfield reproduction using orthogonal basis expansion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, Canada, May 2013, pp. 311–315.

[17] J. Donley, C. Ritz, and W. B. Kleijn, "Multizone soundfield reproduction with privacy- and quality-based speech masking filters," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1041–1055, Jun. 2018.

[18] J.-W. Choi and Y.-H. Kim, "Generation of an acoustically bright zone with an illuminated region using multiple sources," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1695–1700, Apr. 2002.

[19] M. Shin, S. Q. Lee, F. M. Fazi, P. A. Nelson, D. Kim, S. Wang, K. Park, and J. Seo, "Maximization of acoustic energy difference between two spaces," *J. Acoust. Soc. Am.*, vol. 128, no. 1, pp. 121–131, Jul. 2010.

[20] P. Coleman, P. J. B. Jackson, M. Olik, and J. A. Pedersen, "Personal audio with a planar bright zone," *J. Acoust. Soc. Am.*, vol. 136, no. 4, pp. 1725–1735, Oct. 2014.

[21] S. J. Elliott and J. Cheer, "Regularisation and robustness of personal audio systems," ISVR Technical Memorandum 995, Tech. Rep., 2011.

[22] 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.*, Vancouver, BC, Canada, May 2013, pp. 341–345.

[23] Y. Cai, M. Wu, L. Liu, and J. Yang, "Time-domain acoustic contrast control design with response differential constraint in personal audio systems," *J. Acoust. Soc. Am.*, vol. 135, no. 6, pp. EL252–EL257, Jun. 2014.

[24] D. H. M. Schellekens, M. B. Møller, and M. Olsen, "Time domain acoustic contrast control implementation of sound zones for low-frequency input signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 365–369.

[25] O. Kirkeby and P. A. Nelson, "Reproduction of plane wave sound fields," *J. Acoust. Soc. Am.*, vol. 94, no. 5, pp. 2992–3000, Nov. 1993.

[26] M. A. Poletti, "An investigation of 2D multizone surround sound systems," in *Proc. 125th Conv. Audio Eng. Soc.*, San Francisco, CA, USA, Oct. 2008.

[27] M. B. Møller, M. Olsen, and F. Jacobsen, "A hybrid method combining synthesis of a sound field and control of acoustic contrast," in *Proc. 132nd Conv. Audio Eng. Soc.*, Budapest, Hungary, Apr. 2012, P. 8627.

[28] J.-H. Chang and F. Jacobsen, "Sound field control with a circular double-layer array of loudspeakers," *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4518–4525, Jun. 2012.

[29] M. F. Simón 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, Language Process.*, vol. 23, no. 11, pp. 1869–1878, Nov. 2015.

[30] M. B. Møller and M. Olsen, "Sound zones: On performance prediction of contrast control methods," in *Proc. Audio Eng. Soc. Int. Conf. Sound Field Control*, Guildford, UK, Jul. 2016.

[31] F. M. Heuchel, D. Caviedes Nozal, F. T. Agerkvist, and J. Brunskog, "Sound field control for reduction of noise from outdoor concerts," in *Proc. 145th Conv. Audio Eng. Soc.*, New York, NY, USA, Oct. 2018, Paper 10107.

[32] T. Lee, J. K. Nielsen, J. R. Jensen, and M. G. Christensen, "A unified approach to generating sound zones using variable span linear filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, Canada, Apr. 2018, pp. 491–495.

[33] W. F. Druyvesteyn, R. M. Aarts, A. Asbury, P. Gelat, and A. Ruxton, "Personal sound," in *Proc. Inst. Acoust.*, vol. 16, no. 2, 1994, pp. 571–585.

[34] J. Rämö, S. Marsh, S. Bech, R. Mason, and S. H. Jensen, "Validation of a perceptual distraction model in a complex personal sound zone system," in *Proc. 141st Conv. Audio Eng. Soc.*, Los Angeles, CA, USA, Sep. 2016, Paper 9665.

[35] J. Rämö, L. Christensen, S. Bech, and S. Jensen, "Validating a perceptual distraction model using a personal two-zone sound system," in *Proc. Mtgs. Acoust.*, vol. 30, Boston, MA, USA, Jun. 2017, p. 050003.

[36] J. Francombe, P. Coleman, M. Olik, K. R. Baykaner, P. J. B. Jackson, R. Mason, S. Bech, M. Dewhirst, J. A. Pedersen, and M. Dewhirst, "Perceptually optimized loudspeaker selection for the creation of personal sound zones," in *Proc. 52nd Int. Conf. Audio Eng. Soc.: Sound Field Control*, Guildford, UK, Sep. 2013.

[37] M. Buerger, C. Hofmann, C. Frankenbach, and W. Kellermann, "Multizone sound reproduction in reverberant environments using an iterative least-squares filter design method with a spatiotemporal weighting function," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, New Paltz, NY, USA, Oct. 2017.

[38] S. J. Elliott, J. Cheer, J.-W. Choi, and Y. Kim, "Robustness and regularization of personal audio systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 2123–2133, Sep. 2012.

[39] M. Schneider and E. A. P. Habets, "Iterative DFT-domain inverse filter optimization using a weighted least-squares criterion," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 1957–1969, Dec. 2019.

[40] K. Brandenburg and T. Sporer, "NMR and masking flag: Evaluation of quality using perceptual criteria," in *Proc. 11th Int. Conf. Audio Eng. Soc*, Portland, OR, USA, May 1992, pp. 169–179.

[41] M. Bosi and R. E. Goldberg, *Introduction to digital audio coding and standards*, 1st ed. Dordrecht, The Netherlands: Kluwer, 2003.

[42] K. Brandenburg and G. Stoll, "ISO/MPEG-1 audio: A generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.*, vol. 42, no. 10, pp. 780–792, Oct. 1994.

[43] K. Brandenburg, C. Faller, J. Herre, J. D. Johnston, and W. B. Kleijn, "Perceptual coding of high-quality digital audio," *Proc. IEEE*, vol. 101, no. 9, pp. 1905–1919, Sep. 2013.

[44] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[45] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, Nov. 1995.

[46] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[47] J. R. Jensen, J. Benesty, and M. G. Christensen, "Noise reduction with optimal variable span linear filters," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 631–644, Apr. 2016.

[48] J. Benesty, M. G. Christensen, and J. R. Jensen, *Signal enhancement with variable span linear filters*, 1st ed. Singapore: Springer, Feb. 2016.

[49] J. K. Nielsen, T. Lee, J. R. Jensen, and M. G. Christensen, "Sound zones as an optimal filtering problem," in *Proc. 52th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Oct. 2018, pp. 1075–1079.

[50] 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.*, Brighton, UK, May 2019, pp. 136–140.

[51] ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality levels of coding systems," International Telecommunication Union (ITU), Geneva, Switzerland, Oct. 2015.

[52] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[53] ITU-R BS.1387-1, "Method for objective measurements of perceived audio quality," International Telecommunication Union (ITU), Geneva, Switzerland, Nov. 2001.

[54] R. Huber and B. Kollmeier, "PEMO-Q — A new method for objective audio quality assessment using a model of auditory perception," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[55] G. H. Golub and C. F. Van Loan, *Matrix computations*, 4th ed. Baltimore, MD, USA: The Johns Hopkins University Press, Feb. 2013.

[56] B. C. J. Moore, *An introduction to the psychology of hearing*, 6th ed. Leiden, The Netherlands: Brill, Apr. 2013.

[57] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[58] P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Inverse filter design and equalization zones in multichannel sound reproduction," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 3, pp. 185–192, May 1995.

[59] 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. 1292–1304, Jun. 2005.

[60] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, Apr. 2000.

[61] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 6, pp. 969–978, Jun. 1990.

[62] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.

[63] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep., Sep. 2010, Ver. 2.1.20141124.

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

[65] EBU-Tech 3253, "Sound quality assessment material recordings for subjective tests," European Broadcasting Union (EBU), Genova, Switzerland, Sep. 2008.

[66] M. Schneider and E. A. P. Habets, "An iterative least-squares design method for filters with constrained magnitude response in sound reproduction," in *Proc. 43rd German Annu. Conf. Acoust. — DAGA*, Kiel, Germany, Mar. 2017, pp. 1347–1350.

[67] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA — A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, 2018, Art. 8.

[68] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, 1937.

[69] C. Mendonça and S. Delikaris-Manias, "Statistical tests with MUSHRA data," in *Proc. 144th Conv. Audio Eng. Soc.*, Milan, Italy, May 2018, Paper 10006.

[70] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, Dec. 1965.

[71] P. Royston, "Remark AS R94: A remark on algorithm AS 181: The W-test for normality," *J. Roy. Stat. Soc. Appl. Statist.*, vol. 44, no. 4, pp. 547–551, 1995.

[72] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*, 5th ed. Boca Raton, FL, USA: Chapman & Hall/CRC Press, Jul. 2011.

[73] D. Rey and M. Neuhäuser, "Wilcoxon-signed-rank test," in *International encyclopedia of statistical science*, 1st ed., M. Lovric, Ed. Berlin, Heidelberg: Springer, 2011, pp. 1658–1659.

[74] S. M. Ross, *Introduction to probability and statistics for engineers and scientists*, 3rd ed. Burlington, MA, USA: Elsevier, Jul. 2004.