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, 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 on 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 which shapes 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 can be performed using variable span trade-off filters known from signal enhancement, and we show how these filters can also be used for trading of signal distortion in the target zone for acoustic contrast. Numerical validations under anechoic and reverberant environments were conducted, and the proposed method was evaluated via physical metrics including acoustic contrast and signal distortion as well as perceptual metrics such as the short-time objective intelligibility (STOI). The results confirm that, compared to existing nonadaptive sound zone control methods, a perceptual improvement can be obtained by the proposed signal-adaptive and perceptually optimized variable span trade-off (AP-VAST) control method.

Index Terms—Human auditory system, masking effect, personal sound, 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 loudspeaker array. Typically, two types of sound zones are considered: a bright zone and a dark zone. The bright zone is a confined region in which a desired (or target) 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 [2–4], pedestrian alert systems [5], mobile devices [6], personal computer [7], and other applications [8]. [2].

Many different methods for designing the loudspeaker control filters have been proposed over the last two decades since the concept was first introduced in [10]. 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, to create circular sound zones, typically using a circular or a linear loudspeaker array. Most mode matching methods are based on the ideas in [11], with some examples are [12]–[15]. All these methods design the control filters by minimizing physical metrics, except for [15] where the sound zones where optimized for preserving speech privacy.

The ACC methods are designed to maximize the ratio of the acoustic potential energies in the bright and dark zones, and this is achieved by solving a generalized eigenvalue problem [16]. 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 [2–4]. Various variations of the original ACC method have been proposed. These include the energy difference method [17], the planarity control method [19], and the broadband ACC (BACC) method [20]. 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 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 has been proposed in [21]–[23].

The PM methods produce control filters which minimize the difference between the reproduced and desired sound fields in the bright and dark zones. The original method was proposed in [24], [25]. 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, the method sometimes referred to as ACC-PM has been proposed [26], and it is

T. Lee, J. K. Nielsen, and M. G. Christensen were with the Department of Architecture, Design, and Media Technology, Aalborg University, Aalborg, 9000, Denmark (e-mail: tlee@create.aau.dk)
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 [27]. We note in passing that, despite its name, ACC-PM is actually not a combination of the ACC and PM methods and was originally not introduced as ACC-PM in neither [26] nor [27] but [28], [29]. The method has also been referred to as PM-ACC [1] or BACC-PM [29]. A true combination of the BACC and PM methods in the time-domain has recently been proposed in [30], [31].

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 [13]. Unfortunately, however, a much higher contrast than 15 dB is needed. In [32], [33], it was 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, and it measures the ratio of either the acoustic potential energy or loudness between the reproduced and interfering sound fields in a given zone (see [34] for more on this).

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 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. Consequently, it is not possible to trade-off the reproduction error in one frequency bin for the reproduction error in another frequency bin. In the time-domain, this reshaping of the reproduction errors is possible, but has, to the best of our knowledge, not been explored before.

In this paper, we propose a signal-adaptive sound zone control method which takes both the short-term input signal characteristics and the human auditory system into account. That is, the loudspeaker control filters are continuously calculated based on input signal statistics and masking curves updated on a segment-by-segment level. 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 1990’s [35], the so-called 13 dB miracle [36] 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) [37], [38]. In the sound zones application, we have reproduction errors instead of quantization errors. By using masking curves for designing weighting filters which 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 signal-adaptive sound zone control method will be based on the variable span linear filters (VSLF) from signal enhancement [39] since it allows us to trade-off the acoustic contrast for the reconstruction error and has many existing sound zone control methods such as PM, BACC, and ACC-PM as special cases [30], [31], [40].

The paper is organized as follows: in Sec. [I] the proposed 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 also explained how the input signals are segmented in block 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, the signal distortion, and the TIR, but also by perceptual metrics including the short-time objective intelligibility (STOI) [41], PEMU-Q metrics perceptual similarity measure (PSM) and instantaneous PSM (PSMt) [42]. 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 figures shows $L$ loudspeakers, with the $l$th loudspeaker having the finite impulse response (FIR) control filter with filter coefficients $q_l$ 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 problems for each zone. Throughout this paper, we, therefore, only focus on solving such a bright and dark zone problem, and we use subscripts $B$ and $D$ to represent the bright and dark zones, respectively. The reproduced sound
pressure $p_m[n]$ at microphone position or control point $m$ is represented by the convolution between input signal $x[n]$, the $L$ control filters $\{q_l\}_{l=1}^L$ of length $J$, and the $L$ room impulse responses (RIRs) $\{h_{ml}\}_{l=1}^L$ of length $K$, i.e.,

$$p_m[n] = \sum_{l=1}^L \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} x[n-k-j] h_{ml}[k] q_l[j]$$

$$= \sum_{l=1}^L y_{ml}^T[n] q_l = y_m^T[n] q,$$

where

$$y_{ml}[n] = X[n] h_{ml},$$

$$h_{ml} = [h_{ml}[0] \cdots h_{ml}[K-1]]^T,$$

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

$$y_m[n] = [y_{m1}^T[n] \cdots y_{mL}^T[n]]^T,$$

$$q = [q_1^T \cdots q_L^T]^T.$$  

The known signal vector $y_{ml}[n]$ 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 $q$ so that the reproduced sound field $p_m[n]$ in the control points matches a desired sound field $d_m[n]$ in the same 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 emitting $x[n]$ from a virtual source $z$. Thus, the desired sound pressure at control point $m$ is defined as

$$d_m[n] = \begin{cases} (h_{mz} \ast x)[n] & m \in \mathcal{M}_B \\ 0 & m \in \mathcal{M}_D \end{cases},$$

where $\mathcal{M}_B$ and $\mathcal{M}_D$ are the set of microphone 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 dereverberate the sound fields in order to match the desired and reproduced sound fields if the desired sound field is defined under an anechoic environment.

In the sound zone control, two zones labeled $\alpha$ and $\beta$ are generally 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 $x_\alpha[n]$ are zone $\alpha$ and zone $\beta$, respectively, and those for audio input $x_\beta[n]$ are zone $\beta$ and zone $\alpha$, 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 reproduction error $\varepsilon[n]$, defined as the difference between the desired and reproduced sound pressures at all control points in a given zone such that

$$\varepsilon_m[n] = d_m[n] - p_m[n].$$

More generally, we can filter these reproduction errors by weighting filters $w_m[n]$ so that

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

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), then we obtain the weighted and reproduced sound pressure at control point $m$ as

$$\tilde{p}_m[n] = (w_m \ast p_m)[n] = \sum_{l=1}^L y_{ml}^T[n] q_l = \tilde{y}_m^T[n] q$$

where $\tilde{y}_{ml}^T[n]$ 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 $d_m[n]$ for each of the zones. This can describe how much distortion is present in the bright zone and how much power is remained in the dark zone. This allows us to define the weighted signal distortion power (SDP) $\tilde{S}_B(q)$ and the weighted residual error power (REP) $\tilde{S}_D(q)$, respectively, as

$$\tilde{S}_B(q) = \frac{1}{|\mathcal{M}_B| N} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_B} |\tilde{\varepsilon}_m[n]|^2$$

$$= \tilde{\sigma}_d^2 - 2q^T \tilde{R}_B + q^T \tilde{R}_B q,$$

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

where $N$ is the number of observations and

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

$$\tilde{R}_B = \frac{1}{|\mathcal{M}_B| N} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_B} \tilde{y}_m[n] \tilde{d}_m[n],$$

$$\tilde{R}_C = \frac{1}{|\mathcal{M}_C| N} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_C} \tilde{u}_m[n] \tilde{y}_m^T[n],$$

with $\tilde{R}_C$ for $C \in \{B, D\}$ being the spatial correlation matrix for the corresponding zone, $\tilde{R}_B$ 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 a convex optimization problem

$$\text{minimize } \tilde{S}_B(q) \text{ subject to } \tilde{S}_D(q) \leq \epsilon,$$

where $\epsilon$ is a nonnegative scalar which represents the power allowed in the dark zone. The Lagrangian function corresponding to the problem in (16) is

$$\mathcal{L}(q) = \tilde{S}_B(q) + \mu \tilde{S}_D(q),$$

where $\mu \geq 0$ is the Lagrange multiplier. As also assumed in the signal enhancement literature [39], this Lagrange multiplier
treated as a user-defined parameter which controls the trade-off between minimizing $\tilde{S}_B(q)$ and suppressing $\tilde{S}_D(q)$. We note in passing that minimizing (17) for $\mu = 1$ and $\mu$ equal to a constant produces the broadband PM solution and 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 $\tilde{R}_B$ and $\tilde{R}_D$ are real, symmetric, and at least semi-positive definite matrices. Provided that $\tilde{R}_D$ is positive definite, these properties allow us to compute a joint diagonalization for those two matrices \[31, 43\]. As exploited for signal enhancement \[39\], we can use this diagonalization to obtain more control of the trade-off between the SDP and the acoustic contrast. Specifically, we obtain this control by solving (17) with a low-rank approximation to the control filters in $q$. The two matrices $\tilde{R}_B$ and $\tilde{R}_D$ can be jointly diagonalized as

\[U_{L,J}^T \tilde{R}_B U_{L,J} = \Lambda_{L,J}, \quad U_{L,J}^T \tilde{R}_D U_{L,J} = I_{L,J}, \quad (18)\]

where $\Lambda_{L,J} = \text{diag}(\lambda_1, \ldots, \lambda_{L,J})$ being a diagonal matrix containing the generalized eigenvalues in descending order, i.e., $\lambda_1 \geq \ldots \geq \lambda_{L,J} \geq 0$, $I_{L,J}$ being the $LJ \times LJ$ identity matrix, and $U_{L,J}$ being a nonsingular matrix containing the generalized eigenvectors sorted according to the eigenvalues. Note that $\tilde{R}_D$ is typically positive definite when $M_D \min(N, K + J - 1) \geq L_J$.

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

\[q = U_{L,J} a_{L,J}, \quad (19)\]

where $a_{L,J}$ is an $LJ$ coefficient vector. If we plug (19) into (11) and (12), then we obtain

\[\tilde{S}_B(U_{L,J} a_{L,J}) = \bar{\sigma}_a^2 - 2 a_{L,J}^T U_{L,J}^T \tilde{r}_B + a_{L,J}^T \Lambda_{L,J} a_{L,J}, \quad (20)\]

\[\tilde{S}_D(U_{L,J} a_{L,J}) = a_{L,J}^T \Lambda_{L,J} a_{L,J} = \|a_{L,J}\|^2. \quad (21)\]

Interestingly, we can observe from (21) that $\tilde{S}_D$ can be represented by $a_{L,J}$. Hence, this joint diagonalization leads us to analyze how $\tilde{S}_B$ and $\tilde{S}_D$ behave in terms of the eigen information. Furthermore, we benefit from introducing a $V$-rank approximation by forcing the $LJ-V$ smallest eigenvalues to 0 which directly reduces $\tilde{S}_D$. How this affects $\tilde{S}_B$ is explained later. Now we can approximate $q$ by using the first $V$ eigenvectors such that

\[q \approx U_V a_V, \quad (22)\]

where $1 \leq V \leq L_J$ and optimize $L$ over $a_V$ instead of $q$ directly. The cost function (17) is then

\[L(U_V a_V) = \bar{\sigma}_a^2 - 2 a_V^T U_V^T \tilde{r}_B + a_V^T \Lambda_V a_V + \mu a_V^T a_V. \quad (23)\]

The solution to this is analytically derived and given by

\[a_{P-VAST}(V, \mu) = \arg \min_{a_V} L(U_V a_V) = [\Lambda_V + \mu I_V]^{-1} U_V^T \tilde{r}_B. \quad (24)\]

Finally, we plug (24) into (22) and obtain the control filter as

\[q_{P-VAST}(V, \mu) = U_V a_{P-VAST}(V, \mu) = \sum_{v=0}^{V} u_v^T \tilde{r}_B \lambda_v + u_v, \quad (25)\]

where $\lambda_v$ and $u_v$ are the $v$th generalized eigenvalue and eigenvector, respectively.

Interestingly, we can obtain different solutions by varying $V$ and $\mu$, including also 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 PM (or Wiener) solution is obtained when $V = L_J$ and $\mu = 0$, the ACC-PM solution is obtained when $V = L_J$, and the MVDR solution is obtained $V = L_J$ and $\mu = 0$. As we have shown in the Appendix, the maximum acoustic contrast, but also the maximum SDP are obtained for $V = 1$. Increasing $V$ and keeping $\mu$ fixed decreases both, and we obtain the minimum SDP, but also the minimum acoustic contrast for the maximum number of eigenvectors, i.e., $V = L_J$. Thus, $V$ is a user parameter which can be used for controlling the trade-off between the acoustic contrast and SDP. Clearly, $\mu$ 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. ADAPTIVE PERCEPTUAL VAST (AP-VAST)

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 \[44\]. 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 SDP and the REP to exactly zero provided that the multiple-input/multiple-output inverse theorem (MINT) conditions are satisfied \[45, 46\], i.e., that we have more loudspeakers than control points and that the RIRs do not have any common zeros. As these conditions are seldom fulfilled in the sound zones application, we thus cannot avoid to make a reconstruction error, but we should 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., \cite{47}. This masking curve, which is an amplitude spectrum describing the sound pressure level below which any sound is inaudible, is used as the weighting filter $w_m[n]$ after being inverted. In other words, we apply a small weight to those parts of the spectrum where the masker has a high power, whereas we penalize reproduction errors more by applying a larger weight in those parts of the spectrum where the masker has a low or no power.

To compute the masking curve pertaining to control point $m$, we must first figure out which zone the control point is in. If it is in zone $\alpha$, say, the masking curve is computed from the desired signal $d^{(\alpha)}_m[n]$ at this control point when zone $\alpha$ 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 to be bright or dark zones. Thus, when zone $\alpha$ is considered to be the bright zone, the masking curve for the control points in the dark zone, i.e., zone $\beta$, are calculated from $d^{(\beta)}_m[n]$ and not $d^m[n] = 0$. The above discussion is summarized in Table I. If a zone is desired to be a quiet 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-by-segment level 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 $q_i$ is calculated at each of these time segments. For the segment-wise approach, the observation index $n_i$ considered so far is the local time-index, and it is related to the global time-index $n_i$ as

$$n_i = (N - \eta)(i - 1) + n, \quad i \in I,$$

where $I$ denotes the set of the segment indices, $\eta \in \{0, 1, \ldots, N - 1\}$ is the number of overlapping samples between segments, and $n = 0, 1, \ldots, N - 1$. This indexing is used in Fig. 3 which shows the implementation of the proposed sound zone control method which we refer to as adaptive and perceptually-optimized VAST (AP-VAST). Fig. 3 shows that the weighting as well as the filtering with the control filters are implemented in the short-time Fourier-transform (STFT) domain with 50% overlap and with identical analysis and synthesis windows given \cite{48}

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

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 $h_{ml}$ indicates that the RIRs have been measured or modelled in advance.

### TABLE I

| Zone  | $\alpha$ ($m \in \mathcal{M}_\alpha$) | $\beta$ ($m \in \mathcal{M}_\beta$) |
|-------|----------------------------------|----------------------------------|
| Desired signal | $d^{(\alpha)}_m[n_i]$ | $d^{(\beta)}_m[n_i]$ |
| Masker | $d^{(\alpha)}_m[n_i]$ | $d^{(\beta)}_m[n_i]$ |

Fig. 3. The block diagram to get the reproduced sound field at the given time segment $i$. Note that $\ast$ denotes the convolution in the time domain, $W_m[\omega, i]$ denotes the weighting filter in the frequency domain at frequency $\omega$ and time segment $i$ at control point $m$, FFT and IFFT denote the fast Fourier transform and its inverse, respectively, $\mathcal{M}_B, \mathcal{M}_D$ denotes a concatenated index set of $\mathcal{M}_B$ and $\mathcal{M}_D$, and $Q_i[\omega, i]$ denotes the control filters in the frequency domain at frequency $\omega$ and time segment $i$ for loudspeaker $l$.

IV. EXPERIMENTAL VALIDATION AND DISCUSSION

This section presents an evaluation of the proposed method under both anechoic and reverberant environments. Since the control and reproduction performances highly depend on various parameters such as the geometry of the loudspeaker array, the spatial distribution of the $M$ control points, the sampling frequency, the filter length $J$, the segment length $N$, and the input signal $x[n]$, it is very challenging to investigate all different possibilities. Thus, we here focus on how the performance of the proposed sound zone algorithm depends on the segment length $N$ and the control filter length $J$ using the system setup given in Fig. 4.
TABLE II
THE PARAMETER DETAILS FOR THE SIMULATIONS

| $L$ | 8 |
| $M_{BS}$, $M_{F}$ (control) | 25 |
| $M_{BS}$, $M_{F}$ (monitor) | 16 |
| $r_{c}$ | 2 m |
| $l_{g}$ | 0.2 m |
| $d_{w}$ | 0.05 m |
| $l_{c}$ | 2 m |
| sampling frequency | 8000 Hz |
| speed of sound, $c$ | 343 m/s |
| $K$ | 1000 |
| overlap | 50% |

TABLE III
THE NUMBER OF EIGENVECTORS $V$ AND THE USER DEFINED PARAMETER $\mu$ FOR PM, ACC-PM (ACM), P-VAST (Pq), AND AP-VAST (APa, APH, AND APq)

| $V$ | PM | ACM | Pq | APa | APH | APq |
|-----|-----|-----|-----|-----|-----|-----|
| $\mu$ | NA | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |

NA: not applicable

A. Set-up

As depicted in Fig. 4, a circular array with radius of $r_{c} = 2$ m with 8 omnidirectional loudspeakers evenly placed on the circumference was considered. The virtual sources were located outside of the array at the same location where was 0.5 m away from the 7th loudspeaker and 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 $l_{w} = 5$ cm spacing between the control points to cover the size of a human head. The control points are shown as green points in Fig. 4. The 16 gray points in each zone are monitor points which are used to assess how much the performance degrades when evaluated in between the control points. The centers of the two zones were $l_{c} = 2$ m apart from each other. For the spacing used here, the cut-off frequency for avoiding the spatial aliasing was approximately 3.4 kHz. We assumed that all loudspeakers and microphones as well as the virtual source were located in the same plane. All the parameters that are common in all experiments are summarized in Table II.

We evaluated the performance of AP-VAST and existing methods using both physical and perceptual metrics. As existing methods, we used ACC-PM [27], broadband PM (i.e., ACC-PM in [27] with $\xi = \mu/\left(1 + \mu\right) = 0.5$), frequency-domain ACC [16], and P-VAST [40] with averaged masking curves computed from 30 ms time segments, $V = L/4$ eigenvectors, and $\mu = 0.8$. As a baseline, we also evaluated the performance when no control (NC) was applied, i.e., that the control filters were all set equal to the Kronecker delta functions. For AP-VAST, the time-varying weighting filters were obtained from masking curves computed from 30 ms time segments. Moreover, $\mu = 0.8$ and $V = \{L/4, L/2, L/4\}$ were used. For the sake of conciseness, these three different values for $V$ in AP-VAST are referred as APs (all eigenvectors), APH (half of the eigenvectors), and APq (quarter of the eigenvectors) in the figures. We have also denoted ACC-PM by ACM and P-VAST by Pq (quarter of the eigenvectors) in the figures. As we are here only interested in comparing different ways of computing the control filters (i.e., the training in Fig. 3), we kept the implementation of testing part of Fig. 3 fixed for all methods. Specifically, we used the STFT-based method with 50% overlap between successive frames and the analysis and synthesis window in (27) in the testing part. We have summarized the values used for $V$ and $\mu$ in Table III.

For the input signals, three different sets were considered. The first set of input signals was 6 seconds of dialogues excerpt from the Disney movie “Zootopia” in two different languages: English and Danish. The second set was 4.5 seconds of track 49 (English female speech) and track 50 (English male speech) excerpt from the EBU SQAM database [49]. The third set was 4 seconds of track 69 (ABBA - Head over Heels) and track 70 (Eddie Rabbitt - Early in The Morning) excerpt from the EBU SQAM. Each of them was set to be $d_{m}(\alpha)[n]$ for $m \in M_{\alpha}$ and $d_{m}(\beta)[n]$ for $m \in M_{\beta}$, respectively. The energy of the two signals in a set was set to be identical, and the all the inputs signals were downsampled from 44.1 kHz to 8 kHz since we cannot really control the reproduction in between the control points in a zone for higher frequencies than the spatial aliasing frequency of approximately 3.4 kHz.

For the performance comparison, the typical physical metrics – AC, SDP, and TIR – as well as the perceptual metrics – STOI [41], PSM and PSMt [42] – were used. Note that

1Note that a regularization based on the truncated singular value decomposition (TSVD) was used [26] and that the magnitude normalization factor of the control filter was calculated as described in [26] 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.

2For the case of multiple bright zones, a family of perceptual source separation metrics in [60] (i.e., overall perceptual score (OPS), target-related perceptual score (TPS), interference-related perceptual score (IPS), and artifact-related perceptual score (APS)) were also reviewed. Unlike STOI, PSM, and PSMt, however, these metrics did not correlate very well with informal listening so we did not include them in this paper.
STOI was only used for the two data sets where speech was reproduced whereas PSMt and PSM were used for all three data sets. AC, SDP, and TIR are here defined as

$$\text{AC} = 10 \log_{10} \left( \frac{\sum_{n=0}^{N-1} \sum_{m \in M_B} |p_m[n]|^2}{\sum_{n=0}^{N-1} \sum_{m \in M_D} |p_m[n]|^2} \right),$$  

(28)

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

(29)

$$\text{TIR}^{(\alpha\beta)}_m = 10 \log_{10} \left( \frac{\sum_{n=0}^{N-1} |(\alpha_m[n])|^2}{\sum_{n=0}^{N-1} |(\beta_m[n])|^2} \right).$$  

(30)

Traditionally, these metrics are defined for input signals with flat spectra, but since we here consider a signal-adaptive approach we calculated the metrics for the input signals being used. Since AC is defined as a zone-wise metric, AC was calculated zone-wise. On the other hand, SDP, TIR, and the perceptual metrics were all calculated point-wise. In total, 50 control points and 32 monitor points were used, and we present the results as an average across all applicable data sets. More specifically, the results for AC were calculated from 6 samples (two zones and three data sets), the results for SDP, TIR, PSM, and PSMt were calculated from 150 control points or 96 monitor points, and the results for STOI were calculated from 100 control points or 64 monitor points. In all figures, the mean values are illustrated by crosses and the error bars illustrate the 95% confidence intervals. The default values were used to calculate STOI, PSM, and PSMt, and we used the freely available MATLAB implementations of these metrics for obtaining the results [41], [42]. Finally, the RIRs were just delayed and attenuated impulses in the anechoic case and calculated using the RIR generator [51], which is a MATLAB implementation of the image source methods [52], in the reverberant case. Specifically, the 48 kHz RIRs were generated and then resampled to 8 kHz.

B. Experiments

In this section, we investigated three different scenarios. In the first scenario, we compared AP-VAST to the other non-adaptive approaches in terms of the metrics described in the section above. The performance on both the control and monitor points were evaluated using a control filter length of \( J = 120 \) samples (i.e., 15 ms at a sampling frequency of 8 kHz which corresponds to a frequency grid of 66.67 Hz in the ACC method) for all methods and a segment length of \( N = 480 \) samples (i.e., 60 ms at a sampling frequency of 8 kHz) for AP-VAST. For the other methods, all samples were used for calculating the control filters. In the second scenario, we used the same setup as described above, except for using a reverberation time (\( T_{60} \)) of 200 ms. In the third and last scenario, the performance of AP-VAST was evaluated as a function of the segment length \( N \) in the same reverberant environment as in scenario two. In the following three section, we describe the results obtained in each scenario.

1) Anechoic environment: The results from the first scenario are shown in Fig. 5. Panel (a) shows the AC obtained by all methods. For the existing methods, we see that the ACC method significantly outperformed PM and ACC-PM (ACM). The performance of AP-VAST (APa, AP, and APq) and P-VAST (Pq) had all approximately the same performance as ACC. A general trend, however, seems to be that the performance increased with an increasing \( V \) which is consistent.
with the fact that $V$ is a user-parameter trading-off AC for SDP. This can also be observed in panel (b) where the SDP is shown. There, all methods had a similar performance, except for the ACC method which was significantly worse than the rest. The poor performance of the ACC method stems from that it did not preserve the spatial direction of the sound field which is penalized by the SDP metric. This lack of phase control is a well-known limitation of the ACC method [13]. Since the TIR metric does not penalize the lack of phase control, the ACC method performed much better in panel (c) where it outperformed the PM and ACC-PM methods. The highest TIRs were obtained by AP-V AST and P-V AST with only a quarter of the eigenvectors. Compared the AC results in panel (a), we see nearly the same trend for the TIR in panel (c), illustrating the similarity of the AC and TIR metrics.

The results for the perceptual metrics are shown panels (d)–(f). For STOI in panel (d), we see that all method, except for the case of no control produced intelligible signals. For the PSM and PSMt metrics, the AP-V AST, P-V AST, and ACC methods all outperformed PM and ACC-PM, with AP-V AST being slightly better than the all other methods. This suggests that a perceptually better reproduction was achieved by using AP-V AST even though it did not obtain the lowest SDP. Finally, we see that the grid of control points was fine enough for the frequency range under consideration.

2) Reverberant environment: For a reverberant environment, a room with $T_{60} = 200$ ms and volume $140 \, \text{m}^3$ was considered which can be regarded as a normal room. Compared to the anechoic environment, the measured performances in Fig. 6 all got worse across all metrics, in particular for the ACC method. The AC and TIR metrics in panel (a) and panel (c), respectively, were approximately 5 dB lower compared to the anechoic case, whereas the SDP in panel (b) was approximately 2 dB higher. For the PSM and PSMt metrics in panel (e) and panel (f), respectively, we see that AP-V AST and P-V AST clearly outperformed all other methods, with AP-V AST again being slightly better than P-V AST. Finally, we again see that there was no major difference between the performances measured in the control point and the monitor points.

3) The influence of $N$ with a fixed $J$: In this section, we evaluated the performances of AP-V AST for different segment length $N$ and compared the performance to the non-adaptive P-V AST and the no control methods. We used the same reverberant environment as considered above. The control filter length $J$ was again equal to 120 filter coefficients. We considered the segments length of durations 60 ms, 120 ms, 500 ms, 1000 ms, and the input signal length (i.e., 6.0, 4.5, and 4.0 seconds, respectively) referred to as all in Fig. 7. The input signal length is also what was used in P-V AST, and this means that the only difference between AP-V AST-all and P-V AST was if on masking curve computed from 30 ms segments or on masking curve for the entire signal was used. From Fig. 7 we see that AP-V AST-all seemed to be slightly better than P-V AST across all metrics, although the performance differences were quite small. For the AC and TIR metrics in panel (a) and panel (c), respectively, we see that a longer segment length led to a better performance. A general observation from the figure is that AP-V AST outperformed P-V AST in almost all scenarios.

\[^{5}\text{According to [11], a STOI score of 0.80 or more maps to approximately 100\% speech intelligibility.}\]
Fig. 7. The mean values (crosses) and the 95% confidence intervals (solid lines) of the performance evaluation metrics in reverberant environment ($T_{60} = 200$ ms) with respect to the length of the time segment $N$: physical metrics of (a) AC, (b) SDP, (c) TIR and perceptual metrics of (d) STOI, (e) PSM and (f) PSMt are shown. AC was calculated zone-wise, whereas the rest of the metrics were calculated point-wise. The candidates of the segments length of durations $N$ are 60 ms, 120 ms, 500 ms, 1000 ms, and the input signal length which is plotted as all. Note that three different values for the number of eigenvectors in AP-VAST are denoted as APa (all eigenvectors), APh (half of the eigenvectors), APq (quarter of the eigenvectors). Besides, P-VAST using quarter of the eigenvectors is denoted as Pq, ACC-PM is denoted as ACM, and NC for without any control, respectively.

C. Computational complexity

The experiments showed that AP-VAST outperformed the existing, nonadaptive methods in terms of both physical and perceptual metrics. Thus, there is a clear benefit to making a sound zone control algorithm signal-adaptive. 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 $O$ from [43] to denote the computational complexity of an algorithm. By using this notation, the joint diagonalization in AP-VAST and P-VAST has a computational complexity of order $O(L^3 J^3)$ 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 $O(L^3 J^3 I)$ where $I$ is the number of segments. The joint diagonalization is performed from the spatial statistics. If these are computed naively, the computational complexity is $O(M N L^2 J^2)$ for every segment where $M$ and $N$ are the number of control points and the segment length, respectively. PM and ACC-PM, on the other hand, require that a large least-squares problem is solved which is of the same order of complexity as the joint diagonalization, i.e., $O(L^3 J^3)$. Since the ACC method is solved in the frequency domain, we get many smaller problems, one for every frequency bin, instead of one big problem. This results in a complexity of $O(L^3)$ for computing the joint diagonalization in one of $J$ frequency bins and a complexity of $O(L^2)$ for forming the spatial statistics.

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 was achieved by taking the characteristics of input signals as well as of 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. Using this model, masking thresholds were calculated from the given input signal and used as perceptual weighting filters applied to the input signals. This 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 which trades-off the acoustic contrast and the signal distortion. Through validations in both anechoic and reverberant environments, the performances in terms of the typical physical metrics – AC, TIR, and SDP – as well as the perceptual metrics – STOI, PSM, and PSMt – were measured. The performances across all metrics, zones, and input signals were fairly consistent, all indicating that the proposed method provide a perceptually better reproduction of the desired sound field, even though the physical metrics are not necessarily better.

APPENDIX A

A. Acoustic contrast

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

$$\gamma(q) = \frac{M_D q^T R_B q}{M_B q^T R_D q},$$

(31)
and if we plug \( q_{\text{P-VAST}}(V, \mu) \) from (35) into \( \gamma(q) \), then it yields
\[
\gamma(q_{\text{P-VAST}}(V, \mu)) = \frac{M_D}{M_B} \alpha_{\text{P-VAST}}^T(V, \mu) \Lambda V \alpha_{\text{P-VAST}}(V, \mu).
\]
(32)

If we consider \( V \) and \( V + 1 \), respectively, then we obtain
\[
\gamma(q_{\text{P-VAST}}(V, \mu)) = \frac{M_D}{M_B} \sum_{v=1}^{V} |a_v|^2 \lambda_v,
\]
(33)
\[
\gamma(q_{\text{P-VAST}}(V + 1, \mu)) = \frac{M_D}{M_B} \sum_{v=1}^{V+1} |a_v|^2 \lambda_v,
\]
(34)
where \( a_v \) being the \( v \)th element in \( \alpha_{\text{P-VAST}}(V, \mu) \). Subtracting (34) from (33) and reducing to common denominator lead us to have
\[
\gamma(q_{\text{P-VAST}}(V, \mu)) - \gamma(q_{\text{P-VAST}}(V + 1, \mu)) = \frac{M_D}{M_B} \left( \sum_{v=1}^{V} |a_v|^2 \lambda_v - \sum_{v=1}^{V+1} |a_v|^2 \lambda_v \right)
= \frac{M_D}{M_B} |a_{V+1}|^2 \left( \sum_{v=1}^{V} |a_v|^2 \lambda_v - \sum_{v=1}^{V+1} |a_v|^2 \lambda_v \right).
\]
(35)

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 an increasing \( V \), i.e., \( \gamma(q_{\text{P-VAST}}(V, \mu)) \geq \gamma(q_{\text{P-VAST}}(V + 1, \mu)) \).

B. Signal distortion

If we plug \( q_{\text{P-VAST}}(V, \mu) \) from (25) into (11) and (12), we obtain
\[
\hat{S}_B(q_{\text{P-VAST}}(V, \mu)) = \hat{\sigma}_B^2 - 2 \bar{r}_B^T U_V G^{-1} U_V^T \bar{r}_B
+ \bar{r}_B^T U_V G^{-1} \Lambda_V G^{-1} U_V^T \bar{r}_B,
= \hat{\sigma}_B^2 - \sum_{v=1}^{V} |a_v|^2 \lambda_v + 2 \mu \frac{2}{(\lambda_v + \mu)^2} |\bar{r}_B|^2,
\]
(36)
\[
\hat{S}_D(q_{\text{P-VAST}}(V, \mu)) = \bar{r}_B^T U_V G^{-2} U_V^T \bar{r}_B,
= \sum_{v=1}^{V} |u_v^T \bar{r}_B|^2 \lambda_v + \mu.
\]
(37)

where \( G = \Lambda_V + \mu I_V \). Interestingly, we can observe that \( \hat{S}_B(q_{\text{P-VAST}}(V, \mu)) \) and \( \hat{S}_D(q_{\text{P-VAST}}(V, \mu)) \) decreases and increases monotonically for an increasing \( V \), respectively, because all variables in (36) and (37) are nonnegative. Finally, we plug (36) and (37) into (17) and obtain
\[
L(q_{\text{P-VAST}}(V, \mu)) = \hat{\sigma}_B^2 - \sum_{v=1}^{V} |u_v^T \bar{r}_B|^2 \lambda_v + \mu,
\]
(38)

which decreases for an increasing \( V \). Therefore, we can obtain the minimum reproduction error when all eigenvectors are used and \( \mu \) is a positive value which means that the power in the dark zone is still being controlled. Note that we obtain the minimum signal distortion if \( \mu = 0 \).
input signals,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Shanghai, China, Mar. 2016, pp. 365–369.

[24] 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.

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

[26] 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.

[27] 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.

[28] Y. Cai, M. Wu, and J. Yang, “Sound reproduction in personal audio systems using the least-squares approach with acoustic contrast control constraint,” J. Acoust. Soc. Am., vol. 135, no. 2, pp. 734–741, Feb. 2014.

[29] M. B. Møller and M. Olsen, “Sound zones: On performance prediction of contrast control methods,” in Proc. AES Int. Conf., Guildford, UK, 2016.

[30] 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, AB, Canada, Apr. 2018, pp. 491–495.

[31] 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.

[32] 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.

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

[34] J. Francois, 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., Guildford, UK, Sep. 2013, pp. 1–9.

[35] 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.

[36] M. Bosi and R. E. Goldberg, Introduction to digital audio coding and standards. Springer Science & Business Media, 2012, vol. 721.

[37] 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.

[38] K. Brandenburg, C. Fuller, J. Herec, J. D. Johnston, and W. B. Kleijn, “Perceptual coding of high-quality digital audio,” Proc. IEEE, vol. 101, no. 9, pp. 1918–1919, Sep. 2013.

[39] J. R. Jensen 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.

[40] 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.

[41] 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.

[42] 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.

[43] G. H. Golub and C. F. Van Loan, Matrix computations, 3rd ed. Baltimore: The Johns Hopkins University Press, 1996.

[44] B. C. J. Moore, An introduction to the psychology of hearing, 6th ed. Brill, 2013.

[45] 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.

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