Sound Zones as an Optimal Filtering Problem

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

Published in:
2018 52nd Asilomar Conference on Signals, Systems, and Computers

DOI (link to publication from Publisher):
10.1109/ACSSC.2018.8645268

Publication date:
2018

Document Version
Other version

SOUND ZONES AS AN OPTIMAL FILTERING PROBLEM

Jesper Kjær Nielsen, Taewoong Lee, Jesper Rindom Jensen, and Mads Græsbøll Christensen

Audio Analysis Lab, CREATE
Aalborg University, Denmark
{jkn,tlee,jrj,mgc}@create.aau.dk

ABSTRACT

The design of so-called sound zones is a fairly old idea which has recently gained some research attention. The idea is to control a loudspeaker array to reproduce a desired sound field in certain regions. In this paper, we show how the sound zone control problem can be solved using techniques from speech enhancement. Specifically, we first describe in detail the recently introduced variable span linear filtering (VSLF) framework which unifies the popular optimal filtering and subspace-based approaches to speech enhancement. We then show how the sound zone control problem can be solved using the VSLF framework and how a number of well-known sound zone control methods can be viewed as special cases of this solution. We also discuss in detail the differences between the speech enhancement problem and the sound zone control problem and argue that the VSLF framework is actually even better suited for controlling sound zones than for enhancing speech.

***Index Terms*—** Speech enhancement, variable span linear filtering, sound zones, acoustic contrast control, pressure matching

1. INTRODUCTION

Sometimes multiple persons either wish or have to share the same acoustic environment while simultaneously enjoying different audio programs. In, e.g., a car cabin [1, 2], the parents in the front seats might wish to listen to music while the children in the rear seats might wish to watch a cartoon. Another example is an outdoor festival [3] where multiple concerts might happen in parallel, often within hearing range of a residential area. In these and other applications [4–6], simply using headphones to obtain the desired effect either hampers social interactions or is infeasible. An alternative, yet immature technology, is the sound zone concept where the desired audio programs are reproduced in spatially confined regions, i.e., sound zones, by controlling a loudspeaker array. This idea was originatively conceived two decades ago [7] and has since attracted some research attention, mainly in the field of acoustics. A number of different methods have been proposed based on various design principles (see [8] for a recent overview), and they all seek to reproduce the desired audio program in a given zone with the least amount of distortion while minimising the leakage to the other zones.

The dominant sound zone control strategy is to pre-filter the loudspeaker signals with fixed FIR filters. Designing sound zones, therefore, essentially boils down to designing these filters so that they simultaneously optimise metrics related to signal distortion and zone leakage. In [9], the acoustic contrast control (ACC) method was introduced, and it sought to minimise the zone leakage by maximising the ratio of acoustic potential energies between a so-called bright and dark zone. The introduced notions of bright and dark zones were a convenient abstraction which allowed the control filters to be designed for each source and zone in isolation. That is, one zone was in turn considered to be the bright zone whereas the remaining zones were lumped together as a dark or quiet zone, and the multi-zone solution was obtained as a superposition of the individual solutions. The main disadvantage of the ACC approach is that the control filters are designed without penalising perturbations of the desired sound field in the bright zone. This was addressed in [10] with the introduction of the pressure matching (PM) method where the signal distortion in the bright zone was explicitly penalised in the control filter design. However, this improvement in the reconstruction quality of the desired sound field came at the expense of a much higher zone leakage, i.e., a smaller acoustic contrast between the bright and dark zone. Therefore, a number of methods [11–16] have sought to modify and/or combine ACC and PM so that a better trade-off between signal distortion and zone leakage is obtained.

Compared to sound zone control, speech enhancement is an older and much more mature field which is concerned with methods for retrieving a clean speech signal from a noisy mixture. Although this problem appears to be fundamentally different to that of designing sound zones, the main goal in speech enhancement is also to strike the balance between two conflicting requirements which are the speech distortion and the noise suppression [17]. More specifically, most speech enhancement methods involve the design of an FIR filter which filters out the noise from the noisy speech input signal while simultaneously preserving the clean speech signal in the output signal. This approach to doing speech enhancement is often referred to as optimal filtering [18] and, as the title of the present paper suggests, the optimal filtering principle can actually be adopted for the design of sound zones. We recently discovered this connection and used this insight to adapt the very flexible variable span linear filtering (VSLF) framework from speech enhancement [17,19] to sound zone control [20]. Interestingly, the resulting sound zone control framework (VAST) has many of the existing control methods such as ACC, PM, and their variations as special cases. Moreover, the framework can be used to show theoretically that ACC gives the maximum acoustic contrast and distortion while PM gives the minimum acoustic contrast and distortion, something which had only been supported by empirical evidence prior to the introduction of VAST. In this paper, we further explore the connections between those speech enhancement and sound zone control methods which are formulated as optimal filtering problems. We start by a detailed description of the VSLF framework in the context of single channel speech enhancement in Sec. 2 since VAST becomes much easier to derive and understand when rooted in a good understanding of the VSLF framework. The VAST framework is described in Sec. 3 where we also describe its special cases. Finally, we discuss the differences between the VSLF and VAST frameworks in Sec. 4.

2. SPEECH ENHANCEMENT USING VSLF

Speech enhancement is an important, but also difficult problem which has attracted significant research attention in the speech processing community for many decades (see [21, 22] for fairly recent overviews). Although many variations of the problem exist, we here restrict our attention to its most basic form since this is sufficient to convey our main points. Specifically, we consider the single channel problem where we wish to extract the clean speech sample $s(n) \in \mathbb{R}$ from a noisy mixture $y(n) \in \mathbb{R}$, i.e.,

$$y(n) = s(n) + e(n) \tag{1}$$

where $e(n) \in \mathbb{R}$ is additive noise. There is a number of reasons for why estimating $s(n)$ from $y(n)$ is difficult. First, the problem is under-determined in the absence of any prior information since we have two latent variables for every observation. Second, specific prior information is often not available since the speech and noise characteristics can vary significantly from person to person and from noise environment to noise environments. Some noise types such as babble noise are even very speech-like since it is mostly composed of multiple talking persons. Third, reverberation results in that the the noise component and the direct-path speech component become correlated which is difficult to take into account in general and is, therefore, often ignored. Fourth, and finally, speech and some noise types are highly non-stationary which means that segment-by-segment processing under a local stationarity assumption should be performed for short segments.

Although a number of different approaches to speech enhancement exist based on, e.g., spectral substraction [23], statistical models [24], binary masking [25], subspace techniques [26], and non-negative matrix factorisations [27], the majority of enhancement methods can be classified as an optimal filtering approach [18] or has such an interpretation. The basic idea in the optimal filtering approach is to design an FIR filter $\boldsymbol{h} \in \mathbb{R}^M$ which filters out the noise and preserves the clean speech from the noisy speech signal, i.e.,

$$\hat{s}(n) = \boldsymbol{h}^T \boldsymbol{y}(n) = \boldsymbol{h}^T \boldsymbol{s}(n) + \boldsymbol{h}^T \boldsymbol{e}(n) \tag{2}$$

where $\hat{s}(n)$ is an estimate of the nth clean speech sample and $\boldsymbol{y}(n) \in \mathbb{R}^M$ is defined as

$$\boldsymbol{y}(n) = \begin{bmatrix} y(n) & y(n-1) & \cdots & y(n-M+1) \end{bmatrix}^T . \tag{3}$$

Note that $\boldsymbol{s}(n)$ and $\boldsymbol{e}(n)$ are defined analogously to $\boldsymbol{y}(n)$. From (2), we see that the filter $\boldsymbol{h}$ should be defined so that the clean speech component is passed with as little distortion as possible, i.e., $\boldsymbol{h}^T \boldsymbol{s}(n) \approx s(n)$ and so that as much noise as possible is filtered out, i.e., $\boldsymbol{h}^T \boldsymbol{e}(n) \approx 0$. Unfortunately, these two design criteria are in general conflicting since requiring no speech distortion leads to no noise reduction and, conversely, requiring complete noise reduction leads to that $\hat{s}(n) = 0$. Therefore, a compromise must be made and the various optimal filtering methods are different ways of trading-off speech distortion for noise suppression. Unfortunately, it is far from obvious how these errors should be measured in terms of perceptual meaningful and practical metrics [28] so they are usually simply measured using the mean squared error criterion for mathematical tractability. Specifically, if we denote the speech distortion by $J_{\text{SD}}(\boldsymbol{h})$ and the noise suppression by $J_{\text{NS}}(\boldsymbol{h})$, their definitions are

$$\begin{aligned} J_{\text{SD}}(\boldsymbol{h}) &= \mathbb{E}\left[\Big(s(n) - \boldsymbol{h}^T \boldsymbol{s}(n)\Big)^2\right] \\ &= \boldsymbol{h}^T \boldsymbol{R}_{\text{s}} \boldsymbol{h} - 2\boldsymbol{h}^T \boldsymbol{r}_{\text{s}} + \boldsymbol{\iota}_1^T \boldsymbol{r}_{\text{s}} \end{aligned} \tag{4}$$

$$J_{\text{NS}}(\boldsymbol{h}) = \mathbb{E}\left[\Big(0 - \boldsymbol{h}^T \boldsymbol{e}(n)\Big)^2\right] = \boldsymbol{h}^T \boldsymbol{R}_{\text{e}} \boldsymbol{h} \tag{5}$$

where $\mathbb{E}[\cdot]$, $\boldsymbol{R}_{\text{s}} = \mathbb{E}[\boldsymbol{s}(n)\boldsymbol{s}^T(n)]$, and $\boldsymbol{R}_{\text{e}} = \mathbb{E}[\boldsymbol{e}(n)\boldsymbol{e}^T(n)]$ are the expectation operator, the clean speech covariance matrix, and the noise covariance matrix, respectively. Moreover, $\boldsymbol{r}_{\text{s}} = \boldsymbol{R}_{\text{s}}\boldsymbol{\iota}_1$ is a correlation vector with $\boldsymbol{\iota}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T$. Note that we have here assumed that the clean speech and noise components are zero mean and wide sense stationary (WSS) stochastic processes. Whereas the former assumption is reasonable for audio signals, the latter only approximately holds for a short speech segments of length $N \geq M$, say. Consequently, the optimal filtering methods are typically implemented on a segment-by-segment basis, and we here focus on the processing for just one such segment. We also limit ourselves to the case where the statistics is assumed known prior knowledge. For practical speech enhancement algorithms, this is, of course, unrealistic, but it turns out to be fulfilled for the sound zone control problem.

Based on the simple objective functions for speech distortion and noise suppression defined in (4) and (5), respectively, we can formulate a combined optimal filtering objective as

$$J_{\text{OF}}(\boldsymbol{h}) = J_{\text{SD}}(\boldsymbol{h}) + \mu J_{\text{NS}}(\boldsymbol{h}) \tag{6}$$

where μ is a non-negative scalar. An interpretation of μ and the combined objective $J_{\text{OF}}(\boldsymbol{h})$ is that they are the Lagrange multiplier and Lagrangian, respectively, associated with the convex problem

$$\begin{aligned} \underset{\boldsymbol{h} \in \mathbb{R}^M}{\text{minimise}} \quad & J_{\text{SD}}(\boldsymbol{h}) \\ \text{subject to} \quad & J_{\text{NS}}(\boldsymbol{h}) \leq \beta \boldsymbol{\iota}_1^T \boldsymbol{r}_{\text{s}} \end{aligned} \tag{7}$$

where $\beta \in [0, 1]$ is a user parameter which controls how important noise suppression is relative to speech distortion. Unfortunately, the relationship between μ and β is not simple, except for in a few extreme special cases[1]. Instead of β, the Lagrange multiplier μ is, therefore, often treated as a user parameter since this also turns the joint and iterative optimisation over both $\boldsymbol{h}$ and μ into a simple quadratic optimisation problem having a simple analytical solution.

Although μ in (6) provides us with a handle for controlling the trade-off between speech distortion and noise suppression, we can obtain an even better control over this trade-off by essentially enforcing a low rank approximation to the clean speech covariance matrix $\boldsymbol{R}_{\text{s}}$. This is the main rationale behind the variable span linear filtering (VSLF) framework [17, 19] and also the principle which unifies the optimal filtering and subspace approaches to speech enhancement. To derive this, first consider the joint diagonalisation (also sometimes referred to as the generalised eigenvalue decomposition or matrix pencil) of the positive semi-definite clean speech covariance matrix $\boldsymbol{R}_{\text{s}}$ and positive definite noise covariance matrix $\boldsymbol{R}_{\text{e}}$ [29, p. 106]

$$\boldsymbol{B}^T \boldsymbol{R}_{\text{s}} \boldsymbol{B} = \boldsymbol{\Lambda} \tag{8}$$

$$\boldsymbol{B}^T \boldsymbol{R}_{\text{e}} \boldsymbol{B} = \boldsymbol{I}_M \tag{9}$$

where $\boldsymbol{I}_M$ is the $M \times M$ identity matrix, $\boldsymbol{\Lambda}$ is a diagonal matrix containing the M non-negative and real-valued eigenvalues of $\boldsymbol{R}_{\text{e}}^{-1}\boldsymbol{R}_{\text{s}}$ in descending order, and $\boldsymbol{B} \in \mathbb{R}^{M \times M}$ is a non-singular matrix containing the eigenvectors of $\boldsymbol{R}_{\text{e}}^{-1}\boldsymbol{R}_{\text{s}}$ sorted according to the eigenvalues in $\boldsymbol{\Lambda}$. Since $\boldsymbol{B}$ is a non-singular matrix, any filter vector $\boldsymbol{h}$ can be written as a linear combination of the eigenvectors in $\boldsymbol{B}$ in (4), i.e.,

$$\boldsymbol{h} = \boldsymbol{B}\boldsymbol{a} \tag{10}$$

[1] For example, setting $\mu = 0$ corresponds to setting $\beta = 1$, and letting $\mu \to \infty$ corresponds to letting $\beta \to 0^+$. However, these special cases are uninteresting since they just produce the trivial solutions $\boldsymbol{h} = \boldsymbol{0}$ and $\boldsymbol{h} = \boldsymbol{\iota}_1$, respectively.

where $\boldsymbol{a} \in \mathbb{R}^M$ contains the weights. By inserting this expression for $\boldsymbol{h}$ in the objective functions for the speech distortion and noise suppression in (4) and (5), respectively, we readily obtain

$$J_{\text{SD}}(\boldsymbol{B}\boldsymbol{a}) = \boldsymbol{a}^T\boldsymbol{\Lambda}\boldsymbol{a} - 2\boldsymbol{a}^T\boldsymbol{\Lambda}\boldsymbol{B}^{-1}\boldsymbol{\iota}_1 + \boldsymbol{\iota}_1^T\boldsymbol{r}_{\text{s}} \tag{11}$$

$$J_{\text{NS}}(\boldsymbol{B}\boldsymbol{a}) = \boldsymbol{a}^T\boldsymbol{a}\ . \tag{12}$$

To understand how the joint diagonalisation allows us to trade off speech distortion for noise suppression, we now make a $V(\leq M)$-rank approximation to $\boldsymbol{R}_{\text{s}}$ by forcing the $M-V$ smallest eigenvalues to 0. Inserting this V-rank approximation in the speech distortion objective in (4) gives

$$J_{\text{SD}}(\boldsymbol{B}\boldsymbol{a}) \approx \boldsymbol{a}_V^T\boldsymbol{\Lambda}_V\boldsymbol{a}_V - 2\boldsymbol{a}_V^T\boldsymbol{\Lambda}_V\boldsymbol{U}_V^T\boldsymbol{\iota}_1 + \text{const.} \tag{13}$$

where $\boldsymbol{\Lambda}_V$ and $\boldsymbol{U}_V$ are the upper left $V \times V$ submatrix of $\boldsymbol{\Lambda}$ and the first V columns of $\boldsymbol{U} = \boldsymbol{B}^{-T}$, respectively. Note that the approximation is exact if $\text{rank}(\boldsymbol{R}_{\text{s}}) = V$. From (13), we see that the speech distortion objective only depends on the first V elements, denoted as $\boldsymbol{a}_V$, of the vector $\boldsymbol{a}$. As seen from (12), the remaining $M - V$ elements only increase the noise suppression objective and should, therefore, be set to zero. Thus, if we define the filter vector $\boldsymbol{h}$ as a linear combination of the first V eigenvectors, i.e.,

$$\boldsymbol{h} = \boldsymbol{B}_V\boldsymbol{a}_V\ , \tag{14}$$

we make the V-rank approximation to $\boldsymbol{R}_{\text{s}}$ and obtain the solution for $\boldsymbol{h}$ resulting in the largest noise suppression among all filter vectors satisfying the under-determined set of linear equations $\boldsymbol{a}_V = \boldsymbol{U}_V^T\boldsymbol{h}$.

We are now finally able to derive the VSLF optimal filter. To do that, we insert the expression for the filter vector in (14) into the combined objective function in (6) and obtain

$$\boldsymbol{a}_{\text{VSLF}}(V,\mu) = \underset{\boldsymbol{a}_V \in \mathbb{R}^M}{\text{argmin}}\, J_{\text{OF}}(\boldsymbol{B}_V\boldsymbol{a}_V) = \left(\boldsymbol{\Lambda}_V + \mu\boldsymbol{I}_V\right)^{-1}\boldsymbol{B}_V^T\boldsymbol{r}_{\text{s}}$$

from which it readily follows that

$$\boldsymbol{h}_{\text{VSLF}}(V,\mu) = \boldsymbol{B}_V\boldsymbol{a}_{\text{VSLF}}(V,\mu) = \sum_{v=1}^{V}\frac{\boldsymbol{b}_v^T\boldsymbol{r}_{\text{s}}}{\lambda_v + \mu}\boldsymbol{b}_v \tag{15}$$

where λ_v and $\boldsymbol{b}_v$ are the vth eigenvalue and eigenvector, respectively. Note that V and μ are user-defined parameters which control the trade-off between the speech distortion and noise suppression in different ways. Moreover, note that $\boldsymbol{b}_v^T\boldsymbol{r}_{\text{s}}$ is proportional to λ_v. Consequently, the elements in $\boldsymbol{h}_{\text{VSLF}}(V,\mu)$ are automatically set to 0 if $\lambda_v = 0$, unless $\mu = 0$, which means that the VSLF solution is the same for all $V \geq \text{rank}(\boldsymbol{R}_{\text{s}})$. Interestingly, we can also use the VSLF solution to derive simple expressions for the speech distortion and the noise suppression. By inserting the VSLF solution in (15) into the objectives in (4) and (5), we obtain

$$J_{\text{SD}}(\boldsymbol{h}_{\text{VSLF}}(V,\mu)) = \boldsymbol{\iota}_1^T\boldsymbol{r}_{\text{s}} - \sum_{v=1}^{V}\frac{\lambda_v + 2\mu}{(\lambda_v+\mu)^2}|\boldsymbol{b}_v^T\boldsymbol{r}_{\text{s}}|^2 \tag{16}$$

$$J_{\text{NS}}(\boldsymbol{h}_{\text{VSLF}}(V,\mu)) = \sum_{v=1}^{V}\frac{1}{(\lambda_v+\mu)^2}|\boldsymbol{b}_v^T\boldsymbol{r}_{\text{s}}|^2\ . \tag{17}$$

Thus, the speech distortion and noise suppression are non-increasing or non-decreasing, respectively, for an increasing V, thus confirming the trade-off between speech distortion and noise suppression theoretically. Moreover, the combined objective reduces to

$$J_{\text{OF}}(\boldsymbol{h}_{\text{VSLF}}(V,\mu)) = \boldsymbol{\iota}_1^T\boldsymbol{r}_{\text{s}} - \sum_{v=1}^{V}\frac{1}{\lambda_v+\mu}|\boldsymbol{b}_v^T\boldsymbol{r}_{\text{s}}|^2 \tag{18}$$

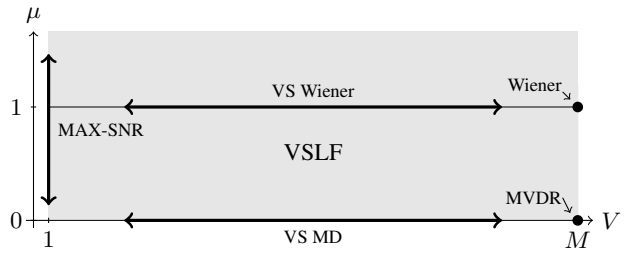

Fig. 1. An illustration of how some of the traditional optimal filtering enhancement algorithms are special cases of the VSLF framework.

which is non-increasing with increasing V. A consequence of this is that we should always use all M eigenvectors if we want to minimise $J_{\text{OF}}(\boldsymbol{h}_{\text{VSLF}}(V,\mu))$ over V, regardless of how μ is selected. At first, this observation might seem to render the joint diagonalisation a mere academic exercise since the VSLF solution for $V = M$ can be obtained directly as the minimiser of the objective in (6). However, it is important to remember that the above analysis is based on the unrealistic case of known statistics. Moreover, the objective function is not necessarily a good measure of the perceptual quality of the enhanced speech signal. Consequently, the joint diagonalisation is still a very useful tool for trading-off signal distortion for noise suppression when tuning a speech enhancement algorithm.

In summary, the key strength of the VSLF solution is the explicit control over the trade-off between the speech distortion and noise suppression through the user parameters V and μ. Interestingly, many of the well-known optimal filtering methods such as the maximum SNR ($V = 1$), Wiener ($V = M$ and $\mu = 1$), and MVDR ($V = M$ and $\mu = 0$) filters are also special cases of the VSLF solution, which is illustrated in Fig. 1. For a much more thorough discussion on the special cases of the VSLF solution, we refer the interested reader to [17, 19].

3. SOUND ZONE CONTROL USING VAST

On first sight, the discussion in the previous section on optimal filtering based speech enhancement might seem unrelated to the sound zone control problem. As alluded to in the introduction, however, one immediate parallel between the two problems is the trade-off between two conflicting design criteria which for the sound zone control problem are the signal distortion in the bright zone and the leakage to the dark zone. To define these criteria mathematically, we use the sketch in Fig. 2. Here, 'B' and 'D' denote the bright and dark zones, respectively, which are each formed by grouping a collection of microphones or control points. We index each of these control points using m and define the collection of bright and dark zone indices as $\mathcal{M}_{\text{B}}$ and $\mathcal{M}_{\text{D}}$, respectively. In the bright zone, we wish to reproduce some desired signal $d_m(n) \in \mathbb{R}$ where $m \in \mathcal{M}_{\text{B}}$ whereas the control points in the dark zone, i.e., $m \in \mathcal{M}_{\text{D}}$, should ideally measure 0. For a control point in either the bright or dark zone, the reproduced sound pressure is a summation of L signals which, for each loudspeaker and control filter index $l \in \{1,\ldots,L\}$, are obtained as a convolution between the known input signal $x(n) \in \mathbb{R}$, the unknown control filter $\boldsymbol{q}_l \in \mathbb{R}^J$, and the known[2] room impulse

[2]We assume that these impulse response have been pre-measured or can be accurately simulated.

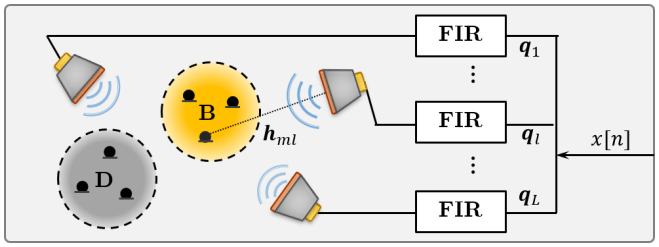

Fig. 2. A sketch of how a loudspeaker array is controlled via a set of FIR control filters to produce a bright and a dark zone.

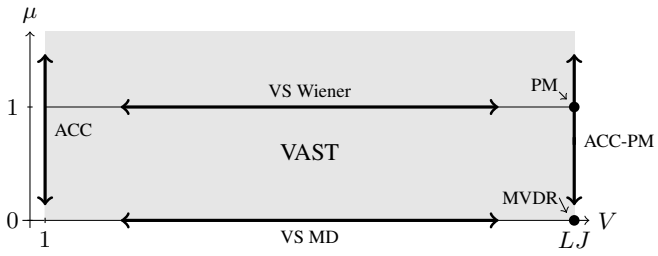

Fig. 3. An illustration of how some of the traditional sound zone methods are special cases of the VAST framework.

response $\boldsymbol{h}_{ml} \in \mathbb{R}^K$, i.e.,

$$p_m(n) = \sum_{l=1}^{L} \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} x(n-k-j) h_{ml}(k) q_l(j) \tag{19}$$

$$= \sum_{l=1}^{L} \boldsymbol{q}_l^T \boldsymbol{y}_{ml}(n) = \boldsymbol{q}^T \boldsymbol{y}_m(n) \tag{20}$$

where we have defined

$$\boldsymbol{y}_{ml}(n) = \boldsymbol{X}(n) \boldsymbol{h}_{ml} \tag{21}$$

$$\boldsymbol{X}(n) = \begin{bmatrix} x(n) & \cdots & x(n-K+1) \\ \vdots & \ddots & \vdots \\ x(n-J+1) & \cdots & x(n-K-J+2) \end{bmatrix} \tag{22}$$

$$\boldsymbol{y}_m(n) = \begin{bmatrix} \boldsymbol{y}_{m1}^T(n) & \cdots & \boldsymbol{y}_{mL}^T(n) \end{bmatrix}^T \tag{23}$$

$$\boldsymbol{q} = \begin{bmatrix} \boldsymbol{q}_1^T & \cdots & \boldsymbol{q}_L^T \end{bmatrix}^T . \tag{24}$$

Note that $\boldsymbol{y}_{ml}(n) \in \mathbb{R}^J$ is known and what we refer to as the uncontrolled pressures at control point m originating from loudspeaker l. The control filters in $\boldsymbol{q}$, however, are unknown and what we are trying to design. To do that, we define the signal distortion objective $J_{\mathrm{B}}(\boldsymbol{q})$ for the bright zone and the residual error objective $J_{\mathrm{D}}(\boldsymbol{q})$ as

$$J_{\mathrm{B}}(\boldsymbol{q}) = \frac{1}{N|\mathcal{M}_{\mathrm{B}}|} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_{\mathrm{B}}} \| d_m(n) - \boldsymbol{q}^T \boldsymbol{y}_m(n) \|_2^2 \tag{25}$$

$$J_{\mathrm{D}}(\boldsymbol{q}) = \frac{1}{N|\mathcal{M}_{\mathrm{D}}|} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_{\mathrm{D}}} \| 0 - \boldsymbol{q}^T \boldsymbol{y}_m(n) \|_2^2 . \tag{26}$$

If we now define the positive semi-definite spatial covariance matrices $\boldsymbol{R}_{\mathrm{B}}$ and $\boldsymbol{R}_{\mathrm{D}}$ as well as the spatial correlation vector $\boldsymbol{r}_{\mathrm{B}}$ as

$$\boldsymbol{R}_{\mathrm{C}} = \frac{1}{N|\mathcal{M}_{\mathrm{C}}|} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_{\mathrm{C}}} \boldsymbol{y}_m(n) \boldsymbol{y}_m^T(n) \quad \text{for C} \in \{\mathrm{B}, \mathrm{D}\} \tag{27}$$

$$\boldsymbol{r}_{\mathrm{B}} = \frac{1}{N|\mathcal{M}_{\mathrm{B}}|} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_{\mathrm{B}}} \boldsymbol{y}_m(n) d_m(n) , \tag{28}$$

we readily obtain that

$$J_{\mathrm{B}}(\boldsymbol{q}) = \boldsymbol{q}^T \boldsymbol{R}_{\mathrm{B}} \boldsymbol{q} - 2 \boldsymbol{q}^T \boldsymbol{r}_{\mathrm{B}} + \sigma_{\mathrm{d}}^2 \tag{29}$$

$$J_{\mathrm{D}}(\boldsymbol{q}) = \boldsymbol{q}^T \boldsymbol{R}_{\mathrm{D}} \boldsymbol{q} \tag{30}$$

where we have also defined the constant

$$\sigma_{\mathrm{d}}^2 = \frac{1}{N|\mathcal{M}_{\mathrm{B}}|} \sum_{n=0}^{N-1} \sum_{m \in \mathcal{M}_{\mathrm{B}}} d_m^2(n) . \tag{31}$$

By comparing the objectives in (29) and (30) to the objectives in (4) and (5), we see that they have exactly the same quadratic form. Thus, the sound zone control filter design can be viewed as an optimal filtering problem and be solved using the VSLF framework, provided that $\boldsymbol{R}_{\mathrm{D}}$ has full rank. From (27), we see that $N|\mathcal{M}_{\mathrm{D}}|$ must be at least greater than or equal to LJ to satisfy this.

To differentiate the VSLF solution for sound zones from the VSLF solution for speech enhancement, we introduced the VAST acronym in [20]. Analogously to (15), the VAST filter is given by

$$\boldsymbol{q}_{\mathrm{VAST}}(V, \mu) = \sum_{v=1}^{V} \frac{\boldsymbol{b}_v^T \boldsymbol{r}_{\mathrm{B}}}{\lambda_v + \mu} \boldsymbol{b}_v \tag{32}$$

which can also be used to show that $J_{\mathrm{B}}(\boldsymbol{q})$, $J_{\mathrm{D}}(\boldsymbol{q})$, and $J_{\mathrm{B}}(\boldsymbol{q}) + \mu J_{\mathrm{D}}(\boldsymbol{q})$ are non-increasing, non-decreasing, and non-increasing, respectively, for an increasing V. Moreover, we can also use the VAST framework to show that the acoustic contrast defined as

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

which is an important metric for sound zone control, is non-increasing for an increasing V when the VAST solution is used. This observation is extremely interesting since the ACC and PM solutions are two extreme special cases of the VAST solution for $V = 1$ and for $V = LJ$ and $\mu = 1$, respectively (see also Fig. 3 for some special cases of VAST). Thus, the VAST framework can be used to show theoretically that you not only trade-off signal distortion in the bright zone for the residual error in the dark zone, but also trade-off signal distortion in the bright zone for the acoustic contrast between the bright and dark zones.

4. DISCUSSION

So far, we have focused on the similarities between the speech enhancement and sound zone control problem. However, there are also some very important differences which have consequences for VAST. First, the main difficulties in speech enhancement, we initially listed in Sec. 2, are actually not (or much smaller) problems for the sound zone control. That is, we have direct access to the individual signals and, therefore know the exact statistics defined in (27) and (28). Consequently, non-stationary signals do not lead to major problems since we do not have to estimate their statistics from short segments. Second, most (if not all) existing sound zone control methods assume Oracle knowledge of the room impulse responses (RIRs) from the loudspeakers to the control points. In practice, these are typically pre-measured in a separate calibration step so that the zones do not have to contain physical microphones during play-back. However, having Oracle knowledge of the RIRs is typically an optimistic assumption in practice since the RIRs cannot be

measured without errors [30] and are normally time-varying due to, e.g., temperature changes [31]. Third, the correlation vectors $\boldsymbol{r}_{\mathrm{s}}$ and $\boldsymbol{r}_{\mathrm{B}}$ have slightly different interpretations, unless the desired signal $d_m(n)$ in the bright zone is defined to be the uncontrolled pressures $y_m(n)$. Usually, however, the desired signal is defined as the pressure originating from a virtual source, and this means a distortion of 0 cannot always be obtained, even for the MVDR solution which results in zero distortion for the speech enhancement problem. Fourth, whereas the clean speech covariance matrix is typically much more sparse than the noise covariance matrix in speech enhancement, a similar relationship between the spatial covariance matrices in the bright and dark zones seems not to exist. How this influences the choice of V is still an open question, but we anticipate that a larger V should generally be used for VAST than for VSLF. Fifth, the vector and matrix dimensions in the sound zone control problem are usually much larger than in the speech enhancement problem. Consequently, the computational cost of computing the control filters is much larger. The typical approach to circumvent this is to compute the control filters offline for generic signals such as white Gaussian noise sources or Dirac's delta functions. This means that all frequencies are given equal weight in the optimisation which is clearly suboptimal when highly coloured signals such as speech and audio are played back.

Despite that the VSLF framework was originally developed with the speech enhancement problem as the its main application, the above discussion actually reveals that it is even better suited for the sound zone control problem, primarily because we there have direct access to the individual signals and, therefore, also the statistics. The main disadvantage of applying the VSLF framework to the sound zone control problem is the computational cost and memory requirements. However, the problem is highly structured, and we believe that this can be exploited to develop fast algorithms which either solve the problem exactly or approximately.

5. REFERENCES

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

[2] X. Liao, J. Cheer, S. J. Elliott, and S. Zheng, "Design array of loudspeakers for personal audio system in a car cabin," in *Proc. Int. Congr. Sound Vib.*, 2016.

[3] F. Heuchel, D. Caviedes Nozal, and F. T. Agerkvist, "Sound field control for reduction of noise from outdoor concerts," in *Conv. Audio Eng. Soc.* Audio Engineering Society, 2018.

[4] 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, 2009.

[5] S. J. Elliott, J. Cheer, H. Murfet, and K. R. Holland, "Minimally radiating sources for personal audio," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 1721–1728, 2010.

[6] J.-M. Lee, T. Lee, J.-Y. Park, and Y.-H. Kim, "Generation of a private listening zone; Acoustic parasol," in *Proc. Int. Congr. Acoust.*, 2010.

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

[8] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 81–91, 2015.

[9] J. Choi and Y. 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.

[10] M. Poletti, "An investigation of 2-D multizone surround sound systems," in *Conv. Audio Eng. Soc.*, 2008, number 125.

[11] 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, 2012.

[12] M. B. Møller, M. Olsen, and F. Jacobsen, "A hybrid method combining synthesis of a sound field and control of acoustic contrast," in *Conv. Audio Eng. Soc.*, 2012.

[13] M. F. Simón-Gálvez, S. J. Elliott, and J. Cheer, "Time domain optimization of filters used in a loudspeaker array for personal audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 1869–1878, 2015.

[14] 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. 252–257, 2014.

[15] M. B. Møller and M. Olsen, "Sound zones: On performance prediction of contrast control methods," in *AES Int. Conf. Sound Field Control*, 2016.

[16] 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.*, 2016, pp. 365–369.

[17] J. Benesty, M. G. Christensen, and J. R. Jensen, *Signal enhancement with variable span linear filters*, vol. 7, Singapore, Singapore: Springer Science+Business Media Singapore Pte Ltd, 2016.

[18] J. Benesty, M. G. Christensen, J. R. Jensen, and J. Chen, "A brief overview of speech enhancement with linear filtering," *EURASIP J. on Advances in Signal Process.*, vol. 2014, no. 1, pp. 162, 2014.

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

[20] 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.*, 2018.

[21] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Boca Raton, FL, USA: CRC Press, 2. edition, 2013.

[22] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, 2015.

[23] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[25] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1486–1494, 2009.

[26] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech enhancement: A signal subspace perspective*, Elsevier, 2014.

[27] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[28] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 47–56, 2011.

[29] J. N. Franklin, *Matrix theory*, Mineola, New York, USA: Dover Publications Inc., 2000.

[30] M. B. Møller, J. K. Nielsen, E. Fernandez-Grande, and S. K. Olesen, "On the influence of transfer function noise on low frequency pressure matching for sound zones," in *Proc. IEEE Sensor Array and Multich. Signal Process. Workshop*, 2018, pp. 331–335.

[31] M. Olsen and M. B. Møller, "Sound zones: on the effect of ambient temperature variations in feed-forward systems," in *Conv. Audio Eng. Soc.*, 2017.