Personal sound zones: Delivering interface free audio to multiple listeners

September 19, 2014

Terence Betlehem, Wen Zhang*, Mark A. Poletti, and Thushara D. Abhayapala

Abstract

Sound rendering is increasingly being required to extend over only certain regions of space for multiple listeners, known as personal sound zones, with minimum interference to listeners in other regions. In this article, we present a systematic overview of the major challenges that have to be dealt with for multi-zone sound control in a room. Sound control over multiple zones is formulated as an optimisation problem and a unified framework is presented to compare two state-of-the-art sound control techniques. While conventional techniques have been focusing on point-to-point audio processing, we introduce wave-domain sound field representations and active room compensation for sound pressure control over a region of space. The design of directional loudspeakers is presented and the advantages of using arrays of directional sources are illustrated for sound reproduction, such as greater control of sound fields over wide areas and reduced total number of loudspeaker units, thus making it particularly suitable for establishing personal sound zones.

I. INTRODUCTION

Sound recording and sound reproduction are becoming increasingly ubiquitous in our daily lives. The ultimate goal of sound reproduction is to recreate the full richness of sound fields including not only the sound content but also the spatial properties to give the listener full knowledge about both the sound source and acoustic environment. Spatial sound reproduction technologies so far have made tremendous

T. Betlehem and M. A. Poletti are with Intelligent Machines and Devices, Callaghan Innovation, Lower Hutt 5040, New Zealand. E-mail: {Terence.Betlehem, Mark.Poletti}@callaghaninnovation.govt.nz.

W. Zhang and T. D. Abhayapala are with the Research School of Engineering, College of Engineering and Computer Science, The Australian National University, Canberra ACT 0200, Australia. E-mail:{wen.zhang, thushara.abhayapala}@anu.edu.au.
progress in reproducing sound fields over fairly large regions of space using an array of loudspeakers. This introduces the idea of establishing **personal sound zones**, whereby interface free audio is delivered to multiple listeners in the same environment without physical isolation or use of headphones (Fig. 1). This concept has recently drawn attention due to a whole range of audio applications, from controlling sound radiation from a personal audio device, to creating individual sound zones in all kinds of enclosures (such as shared offices, passenger cars and exhibition centres) and generating quiet zones in noisy environments.

The first known demonstration of reproducing a sound field within a given region of space was conducted by Camras at IIT research institute in 1967, where an array of loudspeakers was distributed on the surface enclosing the selected region to control sound radiation and the listeners can move freely within the recreated environment [1]. The well-known ambisonics [2], wave field synthesis [3] and higher-order spherical harmonics based technique [4] were developed separately for more advanced spatial sound field reproduction over a large region of space. Druyvesteyn and Garas [5] firstly proposed the concept of a personal sound zone, i.e., reproducing sound within a desired region of space with reduced sound level elsewhere. Microsoft researchers later demonstrated their “Personal Audio Space” project at Microsoft Research TechFesta 2007, where a linear loudspeaker array consisting of 16 drivers was used to enhance the sound in one area while canceling sound waves in another area within the same physical space. Users reported that by stepping even a few paces outside the target region they could not hear the reproduced music. Researchers further extended this concept to develop personal audio for personal computers and televisions [6], as well as for mobile devices [7] and automobile cabins [8]. Such developments will make impact both in the workplace and for the general public.

The idea behind personal sound zones is to formulate a multi-zone sound control problem within
the same physical space as illustrated in Fig. 1. Here, multiple microphones and loudspeakers are used to control the reproduced sound fields. A preference is to use a single array of loudspeakers rather than separate arrays for each zone. This improves freedom and flexibility, allowing sound zones to be positioned dynamically and listeners to freely move between zones. When the system is implemented in reverberant enclosures, loudspeaker designs and audio processing are two key aspects to control sound radiation and to deal with the complexity and uncertainty associated with sound field reproduction. This article aims at reviewing these techniques to support the goal of establishing personal sound zones.

II. MULTI-ZONE SOUND CONTROL

In a general formulation, sound fields are produced over \( Q \) sound zones. Here \( M \) pressure controlling microphones are placed within each zone so that the zone sound fields are controlled by a total of \( QM \) matching points. The sound pressures measured at the microphone positions in each zone \( q \) are represented as a vector \( \mathbf{p}_q = [p(x_{q,1}, \omega), \ldots, p(x_{q,M}, \omega)]^T \) and given by

\[
\mathbf{p}_q = \mathbf{H}_q \mathbf{g},
\]

where \( \mathbf{g} = [g(y_1, \omega), \ldots, g(y_L, \omega)]^T \) denotes the vector of loudspeaker driving signals at a given frequency \( \omega \) to create personal audio sound scenes and \( \mathbf{H}_q \) represents a matrix of acoustic transfer functions (or acoustic impedances) between the loudspeaker drivers and the microphones in zone \( q \).

Sound control techniques can broadly be classified into two categories, acoustic contrast control (ACC) and pressure matching (PM), and we consider each in turn.

A. Acoustic Contrast Control

Choi and Kim [9] firstly formulated the personal audio problem by creating two kinds of sound zones, the bright zone within which we want to reproduce certain sounds with high acoustic energy and the dark zone (or the quiet zone) within which the acoustic energy is kept at a low level. The principle of ACC is to maximise the contrast between the acoustic energy in the bright zone and in the dark zone. Among the \( Q \) sound zones, we specify the first zone as the bright zone and the remaining \( Q - 1 \) zones as the dark zones. The acoustic energy in the bright zone is defined from the sound pressures measured at the \( M \) matching points, that is

\[
E_b = \| \mathbf{H}_b \mathbf{g} \|^2 = \| \mathbf{H}_b \mathbf{g} \|^2 \quad \text{with} \quad \mathbf{H}_b = \mathbf{H}_1 \quad \text{and} \quad \| \cdot \| \text{ denoting the } \ell_2 \text{ norm.}
\]

Similarly, the acoustic energy in the dark zones is represented as

\[
E_d = \| \mathbf{H}_d \mathbf{g} \|^2 = \| \mathbf{H}_d \mathbf{g} \|^2 \quad \text{with} \quad \mathbf{H}_d = [\mathbf{H}_2^H, \ldots, \mathbf{H}_Q^H]^H \quad \text{and} \quad (\cdot)^H \text{ representing the Hermitian transpose.}
\]

In [9] the acoustic contrast, defined as a ratio between the average acoustic potential energy density produced in the bright zone to that in the dark zones, is maximised. The acoustic contrast maximising
method may perform well over the dark zones but may be unrobust to providing the desired maximum energy in the bright zone. To ensure the sound energy within different zones are optimised simultaneously, the problem can be reformulated as maximising the acoustic energy in the bright zone with the constraint that the energy in the dark zone is limited to a very small value $D_0$. In addition, a limit is imposed on the loudspeaker power consumption, i.e., $\|g\|^2 \leq E_0$, also known as the array effort. These constraints ensure that sound leakage outside the $Q$ zones not excessive and also that realised loudspeaker weights are chosen to ensure the implementation is robust to driver positioning errors and changes in the acoustic environment. The ACC problem can then be posed as

$$\max_g \|H_b g\|^2$$

subject to $\|H_d g\|^2 \leq D_0$ \hspace{1cm} (2a)

$\|g\|^2 \leq E_0$. \hspace{1cm} (2c)

The objective and the constraints are summarised into a single objective function represented using the Lagrangian [10],

$$\max L_c(g) = \|H_b g\|^2 - \lambda_1(\|H_d g\|^2 - D_0) - \lambda_2(\|g\|^2 - E_0), \quad \lambda_1, \lambda_2 \geq 0, \hspace{1cm} (3)$$

where $\lambda_1$ and $\lambda_2$ are Lagrange multipliers to adjust the relative importance of each condition (2b) and (2c). The solution that maximises the Lagrangian is obtained by taking the derivative of $L_c$ with respect to $g$ and equating it to zero, and is written as

$$\lambda_1 [H_d^H H_d + \frac{\lambda_2}{\lambda_1} I] g = [H_b^H H_b] g,$$ 

which is recognised as a generalised eigenvector problem. The optimum source strength vector $g_c$ is set as the eigenvector corresponding to the maximum eigenvalue of the matrix $[H_d^H H_d + \frac{\lambda_2}{\lambda_1} I]^{-1} [H_b^H H_b]$. The ratio of Lagrange multipliers $\lambda = \lambda_2/\lambda_1$ determines the trade-off between the performance and array effort and must be chosen iteratively for the constraint on the control effort to be satisfied. The formulation in (4) yields essentially the same answer as that in [8], or the so called indirect formulation in [10], which diagonally loads the matrix $H_d^H H_d$ before inverting it to improve the matrix condition number.

The formulation adopted here leads to a straightforward way for demonstrating the connection between the ACC method and the PM method, which will be explained next.
B. Pressure matching

The pressure matching (PM) method aims to reproduce a desired sound field in the bright zone at full strength, while producing silences in other zones. The idea comes from the traditional crosstalk-cancelation problem, where small regions of personal audio are created by controlling the pressures at discrete spatial points (microphone or listener positions). Multi-zone sound control is an extension of the traditional approach with a sufficiently dense distribution of matching points within all the zones. Given a target sound field \( p_{\text{des}} \) to be reproduced in the bright zone, the robust PM formulation can be written using an \( \ell_2 \) pressure matching objective along with the constraints on the sound energy in the dark zones and the array effort constraint,

\[
\min_g \| H_b g - p_{\text{des}} \|^2
\]

subject to

\[
\| H_d g \|^2 \leq D_0
\]

\[
\| g \|^2 \leq E_0.
\]

The problem can then be written as a Lagrangian cost function,

\[
\min_g L_p(g) = \| H_b g - p_{\text{des}} \|^2 + \lambda_1 (\| H_d g \|^2 - D_0) + \lambda_2 (\| g \|^2 - E_0), \quad \lambda_1, \lambda_2 \geq 0,
\]

where again \( \lambda_1 \) and \( \lambda_2 \) are Lagrange multipliers. The solution that minimises \( L_p \) is obtained by setting the derivative of \( L_p \) with respect to \( g \) to zero and is written as

\[
[H_b^H H_b + \lambda_1 H_d^H H_d + \lambda_2 I]^{-1} H_b^H p_{\text{des}}.
\]

Equation (7) may be solved using an interior point algorithm to choose appropriate values of \( \lambda_1 \) and \( \lambda_2 \) to satisfy the constraints [11]. A simpler formulation is to set the parameter \( \lambda_1 = 1 \), which implies applying equal effort to matching the pressure in the bright zone and minimising the energy in the dark zone. This gives the original formulation of mutli-zone sound control as in [12] but has an added robustness constraint on the array effort as presented in [13], that is \( g_p = [H_b^H H_b + H_d^H H_d + \lambda_2 I]^{-1} H_b^H p_{\text{des}} \). This solution is also identical to that of the ACC method given (i) the choice of target pressures in the bright zone is an ACC solution, \( p_{\text{des}} = H_b g_c \) and (ii) identical constraints in \( E_0 \) and \( D_0 \) are met. This demonstrates that the formulation in the PM approach to sound field reproduction problem subsumes the ACC problem. Chang and Jacobsen [14] investigated a combined solution of these two techniques, \( g_{cb} = [(1 - \kappa) H_b^H H_b + \kappa H_d^H H_d]^{-1} (1 - \kappa) H_b^H p_{\text{des}} \), which is actually same as the one presented in (7) with the regularisation term omitted. The tuning parameter \( \kappa \) is equivalent to the tuning parameter \( \lambda_1 \). The design has been shown effective for reproducing plane wave sound fields at frequencies even above
Fig. 2. A plane wave of 500 Hz from 45° is reproduced in the bright zone (red circle) using pressure matching whilst deadening the sound in the dark zone (blue circle) using 30 loudspeakers placed on a circle of radius $R = 3\ m$ and each zone is of radius $r = 0.6\ m$ as shown in (a). Plot (b) shows the acoustic contrast versus the array effort and the mean-square reproduction error in the bright zone using the ACC method (blue line) and the PM method (red line).

The Nyquist frequency with good contrast control, thus providing the potential to reduced the number of loudspeakers required and increase the zone sizes and upper operating frequencies using the PM method.

The PM approach gives an explicit solution to obtain the loudspeaker driving signals and does not require solving an eigenvector problem, as is required in the case of acoustic contrast optimisation. PM is especially suitable for the situation that different constraints are imposed on each sound zone when the listeners require different quality of listening experience. However a series of Lagrange multipliers need to be determined and a generalised eigenvalue solution is no longer possible. Instead convex-optimisation methods like the interior-point method should be used [11]. The PM approach also imposes an objective on the phase of reproduced sound fields within the bright zone, and thus provides a better holographic image compared to the contrast control method. Figure 2(b) demonstrates that the ACC method always maintains a high level of contrast between the bright and dark zone using a small array effort, but a high reproduction error also indicates that the reproduced sound field may swirl around the listener in different directions. On the other hand, the pressure-matching approach achieves small reproduction error whilst higher contrast may be obtained by choosing an appropriate desired sound field. Preliminary perceptual tests were found to generally agree with the simulation results however the source signal itself significantly affects the system performance [15].
While the least-squares solutions in the frequency domain seems to provide a great deal of simplicity and flexibility, the positions of the loudspeakers and the matching points within sound zones must be chosen judiciously for good reproduction performance. Representing sound fields in the wave domain or mode domain as in (8) can provide physical insights into these critical issues [16]. Dimensionality analysis tells us that for PM over $Q$ sound zones, the number of loudspeakers required is determined by the upper frequency or wave number $k$ of operation, the number of sound zones and the size of each sound zone [16]. Here we assume that each sound zone is a circle or sphere of radius $r_0$ located at the origin $O_q$ as shown in Fig. 1, although without loss of generality each sound zone could be of arbitrary shape. The minimum number $L$ is about $Q(2kr_0 + 1)$ for two-dimensional (2D) reproduction and $Q(kr_0 + 1)^2$ for three-dimensional (3D) reproduction, respectively [4].

C. Discussion

Practical implementation: When a small number of loudspeakers are used, for example three speakers used in a mobile device, current personal audio systems can only achieve limited performance, i.e., $\sim 10$ dB contrast level between bright and dark zones [7]. An array of nine sources has been implemented for personal audio systems in televisions and personal computers, in an anechoic chamber achieving over 19 dB contrast under ideal conditions [6]. However, in terms of practical implementation in a car cabin, Cheer et al. [8] demonstrated that the optimised level of acoustic contrast obtained from the ACC method may not be able to be achieved because of errors and uncertainties and the least-squares based PM approach may provide a more robust solution. In addition, multi-zone reproduction is fundamentally constrained whenever attempting to reproduce a sound field in the bright zone that is directed to or obscured by another zone. This is known as the “occlusion problem” [11], [12].

Loudspeaker positions: Using the compressive sensing idea, the formulation of multi-zone sound field reproduction can be regularised with the $\ell_1$ norm of the loudspeaker weights and solved using the Least-absolute shrinkage and selection operator (Lasso) [17]. The assumption here is that the desired sound field can be reproduced by a few loudspeakers, which are placed close to the direction of the virtual source and are sparsely distributed in space. This can produce a low sound levels outside the bright zones and hence can reduce the interference to the dark zone.
**Wave Domain Sound Field Representation**

The Helmholtz wave equation can be solved to express any sound field as a weighted sum of basis functions,

\[ p(x, \omega) = \sum_{n=1}^{\infty} \alpha_n(\omega) \beta_n(x, \omega), \quad (8) \]

where \( \alpha_n(\omega) \) are sound field coefficients corresponding to mode index \( n \), \( \beta_n(x, \omega) \) are basis functions with the orthogonality property

\[ \langle \beta_n, \beta_m \rangle \triangleq \int_C \beta_n^*(x, \omega) \beta_m(x, \omega) \, dx = \xi_n(\omega) \delta_{nm}. \]

The sound field within a control region \( C \) can be represented using a finite number of basis functions, i.e., \( n \in [1, N] \) and \( \xi_n(\omega) = \langle \beta_n, \beta_n \rangle \) is the strength of each mode over the control zone.

The modal basis functions for source distributions and sound fields arranged in cylindrical coordinates and spherical coordinates can be written as [18]

\[ p_{2D}(x, \omega) = \sum_{\nu=-N}^{N} \alpha_\nu(\omega) J_{\nu}^{(2D)}(kr) \exp(i\nu\phi) \quad (9a) \]

\[ p_{3D}(x, \omega) = \sum_{\nu=0}^{N} \sum_{\mu=-\nu}^{\nu} \alpha_{\nu\mu}(\omega) J_{\nu}^{(3D)}(kr) Y_{\mu}^\nu(\theta, \phi), \quad (9b) \]

where \( \exp(\cdot) \) and \( Y_{\mu}^\nu(\cdot) \) are complex exponentials and spherical harmonics, respectively; \( J_{\nu}^{(2D)}(kr) \) and \( J_{\nu}^{(3D)}(kr) \) are functions representing the 2D and 3D mode amplitudes at radius \( r \), respectively. Given the radius of the control region \( r_0 \) and wave number \( k \), the truncation number \( N \approx kr_0 \) [4] and we have the following dimensionality results \( N_{2D} = 2kr_0 + 1 \) and \( N_{3D} = (kr_0 + 1)^2 \). This gives the Nyquist sampling condition for a uniform circular array \( (M \geq N_{2D}) \) and a spherical array \( (M \geq N_{3D}) \), respectively.

**Further remarks:** While the reproduction error has been widely used to quantify the performance of sound field rendering methods, a planar wavefront may be reproduced whose direction of propagation does not coincide with the desired direction, which may give a high reproduction error. In [19], the cost function of the ACC method is refined to optimise the extent to which a sound field resembles a plane wave. A constraint is imposed on the plane-wave energy within the bright zone over a range of incoming directions, thus optimising the spatial aspects of the sound field for ACC. Simulation results demonstrate that a circular array of 48 equally-spaced loudspeakers produces consistently-high contrast and a planar target sound zone of radius 0.15 m for frequencies up to 7 kHz.
III. ACTIVE ROOM COMPENSATION

One challenge in the personal audio problem is room reverberation. A strong wall reflection may ruin the personal audio listening experience [15]. Room reverberation can be corrected for by using active room compensation, provided the acoustic transfer function (ATF) matrices are determined. For static room environments these matrices may be pre-measured but for time-varying environments they must be determined adaptively. In this section, methods for determining and correcting for these matrices to compensate room responses over space are described. The room compensation approaches described here are more robust at low frequencies. At high frequencies, a reverberant sound field is diffuse. Compensation is extremely sensitive to small changes within the room and cannot be practically compensated for without very fast filter adaptation. Personal sound systems may not be able to compensate for these variations. Instead diffuse components may be treated as noise and the system made robust to them.

We summarise the advances made for the case of a single zone with the ATF matrix, $H \equiv H_1$, using wave-domain or modal-space processing. These approaches demonstrate the challenges inherent in applying room compensation to the multi-zone problem. We also review a crosstalk-cancelation approach to the multizone case which utilises impulse response reshaping.

A. Modal-Space Processing

Based on the wave-domain sound field representation (8), the sound field $p(x, \omega)$ can be expressed as in (9). The ATF $H_\ell(x, \omega)$ from each loudspeaker $\ell$ to a point $x$ inside the sound control zone can also be parameterised, as

\[
H_\ell^{(2D)}(x, \omega) = \sum_{\nu=-N}^{N} \gamma_{\nu\ell}(\omega) J_{\nu}^{(2D)}(kr) \exp(i\nu\phi),
\]

\[
H_\ell^{(3D)}(x, \omega) = \sum_{\nu=0}^{N} \sum_{\mu=-\nu}^{\nu} \gamma_{\nu\mu\ell}(\omega) J_{\nu}^{(3D)}(kr) Y_{\mu}^{\nu}(\theta, \phi),
\]

where $\gamma_{\nu\ell}(w)$ and $\gamma_{\nu\mu\ell}(w)$ are ATF coefficients. The sound pressure vector $p$ and ATF matrix $H$ can then be written in matrix form,

\[
p = B\alpha,
\]

\[
H = BT,
\]

where $B$ is the $M \times N$ matrix of basis functions evaluated at each of the $M$ microphone positions defined $[B]_{mn} = \beta_{n}(x_m, \omega)$, $\alpha$ is an $M$-long vector of sound field coefficients and $T$ is the $N \times L$ matrix of the ATF coefficients defined $[T]_{n\ell} = \gamma_{n\ell}$ and $N$ is either $N_{2D}$ or $N_{3D}$. The pressure matching problem
Fig. 3. Listening room compensation using wave-domain adaptive filtering. The free-field transformed loudspeaker excitation signals $\tilde{g}$ are used in a reverberant room with the filter matrix $\tilde{C}$ to compensate for the ATFs in matrix $H$.

of (5a), in the mode domain becomes $\Gamma g = \alpha_{\text{des}}$, where $\alpha_{\text{des}}$ is the $N$-long vector of coefficients for the desired sound field. The compensation problem can then be solved in off-line manner by determining the least-squares solution [20].

An adaptive mode-domain approach was devised in [21]. The ATF matrix can be further parameterised

$$\begin{align*} H &= UJ\Gamma, \end{align*}$$

(12)

where $U$ is a tall Vandermonde matrix (2D) or spherical harmonic matrix (3D) with the property that $U^H U = I$ and $J$ is a diagonal matrix of the mode amplitudes at the microphone positions. The vector of microphones signals $p = Hg$ are hence transformed into mode-domain coefficients through $\alpha = J^{-1} U^H p$. For modest levels of room reverberation, $\Gamma$ can be expressed as the sum of an anechoic room component and a small reverberant component. By approximating the reverberation as small, a simple iterative procedure for choosing $g$ to drive $\alpha$ to $\alpha_{\text{des}}$ can be formulated. Reverberant compensation methods [20], [21] may have difficulties in practice with pre-ringing artefacts, but these artefacts may be reduced by using more advanced MIMO polynomial filter design [22].

B. Active Listening Room Compensation with Wave Domain Adaptive Filtering

Active listening room compensation can be used to make a reverberant room problem look like an anechoic room problem [23]. By applying a compensation filter matrix to the input loudspeaker signals, the uncompensated anechoic-room driving signals can then be used. The essence of the problem is to minimise the error energy $e^H e$, where

$$e = H_0 g - H C g$$

and $H_0$ is the anechoic-room ATF matrix and $C$ is an $L \times L$ compensation filter matrix. This effectively chooses the filter matrix $C$ to drive the net transfer function matrix $HC$ to the anechoic-room ATF
In massive multichannel problems for which the number of loudspeakers $L$ and microphones $M$ are large, the resultant matrices are large and may have issues with computational requirements (for filtered x-RLS) and convergence rates (for filtered x-LMS). The poor convergence can be solved using eigenspace adaptive filtering [23] by performing a generalised singular value decomposition (SVD) to diagonalise the system. Unfortunately the SVD still incurs a high computational cost.

The problem can fortunately be solved computationally effectively by using a wave-domain approach. If the microphones are arranged over a uniform circular array of radius $r$ and the sources are arranged over a concentric uniform circular array then the anechoic-room ATF matrix may be parameterised

$$H_0 = U J K^H V^H,$$

where $\Gamma_0$ is a matrix of ATF coefficients corresponding to the anechoic room, $K$ is a diagonal matrix of Hankel functions and $V$ is a tall Vandermonde matrix (2D) or a spherical harmonic matrix (3D). Matrix $V$ possesses the property $V^H V = I$, provided that at least one loudspeaker is present for each mode to be controlled and in total the number of loudspeakers, i.e., $L \geq N_{2D}$ or $L \geq N_{3D}$.

The wave-domain adaptive filtering (WDAF) approach is to transform the signals at the microphones and the loudspeaker signals into the wave domain through the transform $T_1$ and $T_3$, then adaptively calculate the mode-domain compensation signals $\tilde{C}(w)$, and transform the compensated loudspeaker signals back using the inverse transform $T_2$ as depicted in Fig. 3. If the compensation filter matrix $\tilde{C}(w)$ is forced to be diagonal, then each of its diagonal entries can be determined from decoupled adaptive filters. This would explicitly solve the problems of computational complexity that appeared in multi-point compensation techniques. While it is straightforward to chose $T_1$ and $T_3$ to do so, in reality $T_2$ cannot always be chosen without a-priori knowledge of the ATF matrix. However, [23], [24] show that the system can be partially diagonalised by choosing $T_1 = V^H$, $T_2 = V$, and $T_3 = U^H$.

C. System Identification of ATF Matrix

The ATFs change in a room as people move about and as temperature changes. Since active room compensation in particular is sensitive to this phenomenon, it is better if the ATFs are determined adaptively. Similar to active listening room compensation, this task can be performed efficiently in the wave domain while transforms can be used to part-diagonalise the reverberant room ATF matrix [24].

The advantages of WDAF and the mode-domain approaches are that, i) sound pressure is controlled over the entire control region and not just at specific points and ii) they represent the problem with a
Fig. 4. Crosstalk cancelation for delivering a time-domain signal $s$ to the top microphone whilst cancelling the signals at the remaining $Q - 1$ microphones.

Fig. 5. Shortening of impulse responses to 50 msec in a room of reverberation time 250 msec using relaxed multichannel least squares (left), the relaxed minimax approach in [25] (centre) and the ratio optimisation approach of [26] (right).

Reduced number $N_{2D} < M$ (or $N_{3D} < M$) of parameters, which reduces the complexity and reduces the correlation in the elements of the ATF matrix since the filters are part decoupled. This helps speed the convergence of adaptive filtering.

Since many more microphones and loudspeakers are required for a 3D control zone, active room compensation is more practically deployed in 2D scenarios. However, 2D compensation cannot satisfactorily correct for roof and floor reflections, so sound absorbers must be employed to eliminate these effects.

### D. Impulse Response Reshaping

Multiple listening zones may also be achieved by using crosstalk cancelation. Here, each of $Q$ signal is delivered to a listening position whilst cancelling the crosstalk paths to the remaining $Q - 1$ positions using $L$ loudspeakers and, for monaural signals, $M = 1$ microphone in each zone. As shown in Fig. 4, this problem is solved by implementing crosstalk-cancelation filters. The basic idea of the impulse response reshaping approach is that fully equalising the delivered paths is unnecessary. It is more robust and efficient to reshape these impulse responses.

Using impulse response reshaping, the early reflections of the delivered paths are reinforced whilst late reverberation and crosstalk are minimised [26]. Here, by defining windows on these desired and
undesired ATF components $w_q^{(d)}$ and $w_q^{(u)}$ respectively in each zone $q$, the ratio of undesired-to-desired components is minimised

$$\min_{\hat{g}} \log \frac{\|W_u \hat{r}\|_{p_u}}{\|W_d \hat{r}\|_{p_d}},$$

where $\hat{r}$ represents the global impulse response given a concatenated vector of crosstalk cancelation filters $\hat{g} \triangleq [\hat{g}_1^T, \ldots, \hat{g}_L^T]^T$ and a block-Toeplitz matrix $\hat{H}$ representing the room impulse responses, i.e., $\hat{r} = \hat{H} \hat{g}$, $W_u \triangleq \text{Diag}(w_1^{(u)}, \ldots, w_Q^{(u)})$, and $W_d \triangleq \text{Diag}(w_1^{(d)}, \ldots, w_Q^{(d)})$. Different $p_d$- and $p_u$- norms may be chosen for the desired and undesired components but it has been shown perceptually favourable to choose norms which approximate the infinite norm. Equation 14 can be solved analytically for the $p_u = p_d = 2$ case where it reduces to a generalised Rayleigh quotient. In general, (14) is solved using steepest descent methods [26]. A relaxed multichannel approach using least squares and minimax metrics [25] may include regularisations to reduce the array effort below that of the ratio-based approach in [26]. These approaches are compared in Fig. 5 for simulation with $L = 3$ and $Q = 2$ in a room with a reverberation time of 250 ms using only short 75 msec-long filters.

Impulse response reshaping in principle can be applied to the pressure-matching and modal-space approaches of creating personal sound zones. More robust and efficient filters can be obtained than equalisation that cancel the undesirable late reverberation whilst leaving in some beneficial early reflections. Unfortunately this problem must be formulated in the time domain, which results in a computationally-intractable massive multichannel problem. The future development of lower-complexity convex optimization algorithms may permit practical solutions to these large problems.

IV. DIRECTIONAL SOURCES

The use of directional sources can provide advantages over conventional loudspeakers, whose directivity is omnidirectional at low frequencies and is not typically controllable. Directional sources that provide multiple modes of sound radiation can be used with active compensation to produce sound arriving from angles where there are no sources by reflecting sound from room surfaces and can also cancel unwanted reverberation (Fig. 6).

In a multi-listener situation a single directional loudspeaker can reduce unwanted radiation of sound to other listeners by maximising the direct sound to the intended recipient relative to the reverberant field. A loudspeaker with directivity $D$ and radiating acoustic power $W$ in an ideal Sabinian space produces a direct sound intensity $I_{\text{dir}} = WD/(4\pi r^2)$ and a reverberant sound intensity of $I_{\text{rev}} = 4W/R'$ where $R' = S\epsilon/(1 - \epsilon)$ is the room constant, $S$ the room surface area and $\epsilon$ the mean absorption coefficient of
the room surfaces. The direct to reverberant intensity ratio is thus

\[ \text{DRR} = \frac{DR^I}{4\pi r^2}. \]  

(15)

Increasing the directivity thus allows the direct sound at the listener to be increased relative to the reverberant sound. Equivalently, the reverberant field is reduced by \(1/\text{DRR}\).

Standard loudspeakers typically have insufficient directivity to provide a significant enhancement of direct sound in a reverberant space. High directivity can be achieved using traditional array techniques such as delay and sum beamforming, but the array size must be large at low frequencies to achieve significant directivity. For practical use superdirectional arrays are required, which achieve higher directivities than an array with uniform amplitude weightings [28]. Superdirectivity can be achieved using linear differential arrays, where the transducer weights have alternating signs, or by using circular and spherical arrays, where the weights are obtained from trigonometric or spherical harmonic functions, respectively. Such loudspeakers are termed higher order sources (HOSs), and can produce multiple radiation patterns which are described by cylindrical or spherical harmonics.

Because superdirectional arrays are compact relative to their directivity, they may be built into a single unit, and we therefore assume here that a directional source is a single unit, typically of similar dimension to a standard loudspeaker. This section considers the design of directional loudspeakers and their application to maximum directivity, then considers the advantages of using arrays of directional sources, which allow greater control of sound fields over wide areas and are particular suitable for establishing personal sound zones.
A. Spherical arrays

The sound field produced by an arbitrary source of maximum radius \(a\) positioned at the origin and radiating a complex frequency \(\exp(i\omega t)\) is represented in the wave domain as in (9b) \[18\]

\[
p(r, \theta, \phi, w) = \sum_{\nu=0}^{N} \sum_{\mu=-\nu}^{\nu} \alpha_{\nu}^{\mu}(w) h^{(2)}_{\nu}(kr) Y_{\mu}^{\nu}(\theta, \phi), r \geq a, \tag{16}
\]

where \(h^{(2)}_{\nu}(kr)\) is the spherical Hankel function of the second kind, i.e., the radial function to represent the mode amplitude at \(r\) and \(\alpha_{\nu}^{\mu}(w)\) are sound field coefficients. Same as the dimensionality analysis in the wave domain, we will assume that the directivity of the source can be described by a maximum order \(N\) so that \(\nu \in [0, N]\).

The most direct method for constructing a loudspeaker that can produce a controllable directivity is to mount a number of drivers in a spherical baffle of radius \(a\) \[29\]. The general behaviour of such a source is most simply explained by deriving the sound field due to a sphere with arbitrary surface velocity

\[
v (\theta_s, \phi_s, t, w) = e^{i\omega t} \sum_{\nu=0}^{N} \sum_{\mu=-\nu}^{\nu} \zeta_{\nu}^{\mu}(w) Y_{\nu}^{\mu}(\theta_s, \phi_s),
\]

where \((\theta_s, \phi_s)\) is the driver position on the sphere. The exterior field has the general form of (16). The expansion coefficients are found by calculating the radial velocity for the general case, and requiring that they equal (17), i.e.,

\[
\alpha_{\nu}^{\mu}(w) = -i\rho c \frac{\zeta_{\nu}^{\mu}(w)}{h^{(2)}_{\nu}(ka)}
\]

and the sound field, including the effect of mass-controlled drivers, is

\[
p(r, \theta, \phi, t, w) = -\frac{i\rho c e^{i\omega t}}{k} \sum_{\nu=0}^{N} \sum_{\mu=-\nu}^{\nu} \zeta_{\nu}^{\mu}(w) \frac{h^{(2)}_{\nu}(kr)}{h^{(2)}_{\nu}(ka)} Y_{\nu}^{\mu}(\theta, \phi), r \geq a.
\]

Hence, each coefficient of the surface velocity produces a corresponding mode of radiation whose polar response is governed by a spherical harmonic.

The normalised magnitude of the mode responses for orders 0 to 5 are shown in Fig. 7 (a). For all modes greater than order \(\nu = 0\) the response reduces at low frequencies. All modes of order \(\nu\) become active at a frequency approximately given by

\[
ka = \nu \text{ or } f = \frac{\nu c}{2\pi a}.
\tag{18}
\]

This means that it is not possible to create high order directivities at low frequencies. The spherical loudspeaker is omnidirectional at low frequencies, and can produce increasing directivities as more modes become active above frequencies given by (18).
In practice, the surface velocity in (17) must be approximated using a discrete array of $L_0$ drivers positioned on the sphere. Ideally the drivers are positioned so that they are spaced equally from each other which produces the most robust approximation to the integration over the sphere required to approximate each spherical harmonic. This is possible if the drivers are placed in the center of the faces of platonic solids, allowing up to 20 drivers (for the icosahedron). Higher numbers of drivers can be used using numerically optimised integration nodes for the sphere.

A simple way to model the discrete approximation is to assume each driver is a point source. The sound field due to a point source on a sphere then models a single driver, and the sound fields due to $L_0$ point sources allows the calculation of the total field. However this approach ignores the directivity of each driver, which becomes significant at high frequencies. A more accurate model of the drivers which is mathematically tractable is to model each one as a spherical cap vibrating radially [29].

The sampling of the sphere means that the spherical loudspeaker is unable to generate spherical harmonic terms above the spatial Nyquist frequency of the array. This may be derived by noting that there are a total of $\mathcal{N} = (N + 1)^2$ spherical harmonics up to order $N$. Controlling this number of modes using $L_0$ loudspeakers is possible for $L_0 \geq \mathcal{N}$. At a given frequency, the maximum mode order that can be radiated is $N = ka$. Hence, the spatial Nyquist frequency is

$$f_{N_{yq,3D}} = \frac{c(\sqrt{L_0} - 1)}{2\pi a}.$$  

(19)

The number of drivers required for a sphere of radius $a$ to produce $N$th order directional responses up
to a frequency $f$ is given by

$$L_{3D} = \left( \frac{2\pi a f}{c} + 1 \right)^2.$$ 

For example, a 3rd order speaker with radius $a = 0.1 m$ and a Nyquist frequency of 4 kHz would require 70 drivers. This is a large number of drivers, and motivates the investigation of simpler approaches such as cylindrical and line arrays.

### B. Cylindrical arrays

A simpler approach may be taken if the directional loudspeaker is only required to produce directivity in a 2D plane. This is commonly the case for sound reproduction in the home where stereo and 5.1 surround formats are ubiquitous. A circular array requires less drivers than a spherical array for the same spatial Nyquist frequency. To see this, consider a sphere where $L_0$ drivers are placed on the equator instead of equally spaced around the sphere. This arrangement allows the generation of sectorial spherical harmonics where $\nu = |\mu|$ which produce radiation with lobes only in the $(x, y)$ plane. The driver spacing is now $2\pi a/L_0$ and the spatial Nyquist frequency is

$$f_{Nyq,2D} = \frac{c(L_0 - 1)}{4\pi a}. \tag{20}$$

The number of drivers for a given 2D spatial frequency is

$$L_{2D} = \frac{4\pi a f}{c} + 1.$$ 

Comparing (20) with (19), the 2D Nyquist frequency can be much higher than the 3D Nyquist frequency for the same number of drivers. The limitation of the circular array is that the transducer layout does not provide sufficient vertical directivity at high frequencies and the source begins to produce unwanted radiation lobes in elevation. To reduce these lobes the transducers must either have greater aperture in elevation, or a line array must be used to control the vertical directivity. Since a line array is more effective mounted on a cylinder than on a sphere a practical alternative to the spherical array for the 2D case is a cylindrical baffle in which are mounted multiple circular arrays (Fig. 6). Such a geometry can still use less transducers than the spherical case, for the same spatial Nyquist frequency.

The radiation of sound for the cylindrical case can be approximated by assuming that the cylinder is infinite and that each driver is represented as a surface velocity distribution in height $z$ and azimuth angle $\phi$ [27]. Its produced mode responses are shown in Fig. 7 (b). The responses are similar to those for the spherical source, and the activation frequencies are the same. The limitation of this analysis is that in practice a truncated cylinder must be used, leading to variations of the mode response magnitude around the infinite cylinder values due to diffraction from the ends of the cylinder.
C. Line arrays

The simplest array for providing high directivity is a line array, which produces an axisymmetric polar response. While this does not provide the full control of 3D or 2D radiation that the spherical and cylindrical arrays do, it is sufficient for maximising the direct to reverberant ratio. It has the same limitation as the circular and spherical arrays, that is difficult to create high order responses at low frequencies. However, the line array allows an order $N$ response to be produced using $L_0 = N + 1$ transducers as opposed to $(N + 1)^2$ using a spherical array or $2N + 1$ for a circular array (assuming no vertical directivity control). The maximum directivity produced in 3D is $D = (N + 1)^2$.

An order $N$ loudspeaker with this directivity will produce the maximum direct to reverberant ratio for an on-axis listener. The simplest case, $N = 1$, results in a polar response $p(\theta) = 0.25 + 0.75 \cos(\theta)$ which has a directivity of 4 [7]. The first order response can be implemented using $N = 2$ coupled or uncoupled drivers, or more simply, using a single driver and controlling the radiation from the rear of the driver, although the directivity results can be less accurate with frequency [7].

D. Arrays of directional sources

If multiple directional loudspeakers are available, then it becomes possible to create multiple zones of sound. Multizone reproduction requires a large number of monopole loudspeakers. The use of directional sources allows the production of multizone fields using significantly less loudspeaker units. In effect, a large number of drivers are grouped into a small number of physical devices to allow the creation of complex sound fields.

It has been shown that an array of $L \ Nth$ order sources operating in free-field conditions has a spatial Nyquist frequency of approximately $2N$ times that of the same geometry monopole array [31]. Results better than free-field can be achieved in a reverberant room by using the techniques discussed in [32]. In this case the directional sources are able to exploit room reflections to provide directions of arrival other than those directly from the sources. The use of $L$ higher order sources, each of which can produce up to order $N$ responses, can produce a similar accuracy of reconstructed field to $L(2N + 1)$ monopole loudspeakers in the 2D case, and $L(N+1)^2$ loudspeakers in the 3D case. For example, Fig. 8 shows the sound field reproduction error achieved using a circular array of five higher order loudspeakers in comparison with an array of forty five monopole sources. For a virtual source angle of 72° (the desired source position is equal to the first loudspeaker position), the error is similar to that produced
Fig. 8. Least squares error of reproduction as a function of frequency for an array of five fourth order sources at 36° exactly between a pair of loudspeakers (dashed) and 72° coinciding with a loudspeaker (dashed), and a circular array of 45 omnidirectional line sources (unbroken) in a 2D rectangular room of dimensions 6.4 by 5 m and with wall reflection coefficients of 0.7.

by the monopole sources. At the angle of 36° (the desired source halfway between two loudspeakers), the error is about 10 dB higher than the monopole case but still reasonably accurate, particularly at low frequencies. Reproduction has been achieved over a 1 m diameter using only 5 loudspeaker units, with room de-reverberation. The simulation is limited to 2 kHz bandwidth for computational complexity reasons. The worst-case reproduction error will be below −10 dB up to around 3 kHz. The bandwidth and reproduction radius of accurate reproduction can be extended by using more sources and higher orders, creating sufficient space for multiple listeners listening to independent sound fields.

The use of higher order sources can be viewed as an optimisation problem with a constraint on the total number of loudspeaker units in the room. The only way to improve reproduction in such a case is to add capability to the existing loudspeakers. Higher order sources offer a practical approach to providing the control of the high-spatial-dimension sound fields that are required for creating multiple personal sound zones. For example, the reproduction of sound in $Q$ zones of radius $r_0$, up to a spatial frequency $k_{\text{max}}$, using $L$ HOSs requires a maximum order per source of

$$N = \left\lceil \frac{Q(k_{\text{max}}r_0 + 0.5)}{L} - 0.5 \right\rceil. \quad (21)$$

For 8 kHz reproduction over regions of radius 0.2 m, the order is $N = 10$ for $L = 10$ sources and $N = 6$ for $L = 15$ sources. Such numbers are achievable in moderate to large-sized rooms.
V. Summary and Future Opportunities

In this article, we presented, according to our involvement and insights, the audio processing and loudspeaker design aspects that support the goal of establishing personal sound zones. The problems that have been explored include multi-zone sound control, wave-domain active room compensation and directional loudspeaker design, which allow for sound control over spatial regions. A high-performance personal audio system would likely address many of these aspects in its design. In sound field control, interference mitigation and room compensation robust to changes and uncertainty in the acoustic environment remain as challenging problems. Yet future opportunities exist in i) higher-order surround sound using an array of directional sources and wave-domain active room compensation to perform multi-zone sound control in reverberant enclosures and ii) personal audio devices using multiple sensors to establish personal sound zones by efficiently cancelling crosstalk and using distributed beamforming.

REFERENCES


