2018

Multizone Soundfield Reproduction With Privacy- and Quality-Based Speech Masking Filters

Jacob Donley
University of Wollongong, jrd089@uowmail.edu.au

Christian H. Ritz
University of Wollongong, critz@uow.edu.au

W Kleijn
Victoria University of Wellington

Multizone Soundfield Reproduction With Privacy- and Quality-Based Speech Masking Filters

Abstract

Reproducing zones of personal sound is a challenging signal processing problem which has garnered considerable research interest in recent years. We introduce in this work an extended method to multizone soundfield reproduction which overcomes issues with speech privacy and quality. Measures of Speech Intelligibility Contrast (SIC) and speech quality are used as cost functions in an optimisation of speech privacy and quality. Novel spatial and (temporal) frequency domain speech masker filter designs are proposed to accompany the optimisation process. Spatial masking filters are designed using multizone soundfield algorithms which are dependent on the target speech multizone reproduction. Combinations of estimates of acoustic contrast and long term average speech spectra are proposed to provide equal masking influence on speech privacy and quality. Spatial aliasing specific to multizone soundfield reproduction geometry is further considered in analytically derived low-pass filters. Simulated and real-world experiments are conducted to verify the performance of the proposed method using semi-circular and linear loudspeaker arrays. Simulated implementations of the proposed method show that significant speech intelligibility contrast and speech quality is achievable between zones. A range of Perceptual Evaluation of Speech Quality (PESQ) Mean Opinion Scores (MOS) that indicate good quality are obtained while at the same time providing confidential privacy as indicated by SIC. The simulations also show that the method is robust to variations in the speech, virtual source location, array geometry and number of loudspeakers. Real-world experiments confirm the practicality of the proposed methods by showing that good quality and confidential privacy are achievable.

Keywords

quality-based, privacy-, reproduction, speech, soundfield, filters, multizone, masking

Disciplines

Engineering | Science and Technology Studies

Publication Details

J. Donley, C. Ritz & W. Bastiaan. Kleijn, "Multizone Soundfield Reproduction With Privacy- and Quality-Based Speech Masking Filters," IEEE ACM Transactions on Audio, Speech, and Language Processing, vol. 26, (6) pp. 1041-1055, 2018.

Multizone Soundfield Reproduction With Privacy and Quality Based Speech Masking Filters

Jacob Donley, *Student Member, IEEE,* Christian Ritz, *Senior Member, IEEE,* and W. Bastiaan Kleijn, *Fellow, IEEE*

Abstract

Reproducing zones of personal sound is a challenging signal processing problem which has garnered considerable research interest in recent years. We introduce in this work an extended method to multizone soundfield reproduction which overcomes issues with speech privacy and quality. Measures of Speech Intelligibility Contrast (SIC) and speech quality are used as cost functions in an optimisation of speech privacy and quality. Novel spatial and (temporal) frequency domain speech masker filter designs are proposed to accompany the optimisation process. Spatial masking filters are designed using multizone soundfield algorithms which are dependent on the target speech multizone reproduction. Combinations of estimates of acoustic contrast and long term average speech spectra are proposed to provide equal masking influence on speech privacy and quality. Spatial aliasing specific to multizone soundfield reproduction geometry is further considered in analytically derived low-pass filters. Simulated and real-world experiments are conducted to verify the performance of the proposed method using semi-circular and linear loudspeaker arrays. Simulated implementations of the proposed method show that significant speech intelligibility contrast and speech quality is achievable between zones. A range of Perceptual

Manuscript received August 9, 2017; revised November 28, 2017; accepted January 1, 2018. The work of J. Donley was supported in part by the Australian Government Research Training Program Scholarship and the University of Wollongong Global Challenges PhD Scholarship. The associate editor coordinating the review of this manuscript and approving it for publication was Assoc. Prof. Federico Fontana. (Corresponding author: Jacob Donley.) This paper was presented in part at the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, March 2016.

J. Donley and C. Ritz are with the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW, 2522 Australia (e-mail: jrd089@uowmail.edu.au; critz@uow.edu.au)

W. B. Kleijn is with the School of Engineering and Computer Science, Victoria University of Wellington, Wellington, 6140 New Zealand (e-mail: bastiaan.kleijn@ecs.vuw.ac.nz)

Evaluation of Speech Quality (PESQ) Mean Opinion Scores (MOS) that indicate good quality are obtained while at the same time providing confidential privacy as indicated by SIC. The simulations also show that the method is robust to variations in the speech, virtual source location, array geometry and number of loudspeakers. Real-world experiments confirm the practicality of the proposed methods by showing that good quality and confidential privacy are achievable.

Keywords

multizone soundfield reproduction, speech, privacy, intelligibility, quality.

I. INTRODUCTION

Personal sound zones, such as the individual sound environments provided to listeners by means of spatial multizone soundfield reproduction, without the need for physical barriers or headphones, have gained significant interest of researchers in recent years [1]–[3]. Some applications of personal sound zoning systems include vehicle cabin entertainment/communication systems, multi-participant teleconferencing, cinema surround sound systems and personal audio in restaurants/cafés [1], [4], [5]. In some cases, it is desirable to maintain quiet areas by cancelling or suppressing audio from adjacent zones. Quiet areas may be desired so that, for example, vehicle satellite navigation instructions may be heard by drivers without disturbing passengers or so that someone may read/work in silence while someone else listens to a talk show or news in the same room [6]. Limitations exist in the majority of work with multizone soundfield reproduction systems where sound is audible (and likely intelligible) for listeners in designated quiet zones and/or the perceived quality in target reproduction zones is degraded from interference caused by other zones [2], [3], [7], [8].

Multizone soundfield systems attempt to eliminate audio spatially leaked between zones [1], [9]–[12]. Multizone soundfield reproductions constraining quiet zones to zero energy may result in uncontrolled regions containing sounds many times the amplitude of the target bright zone. Techniques that improve performance in these situations optimise over spatial regions with planarity [13], basis plane-waves [14] and reduced constraints [14], [15]. Recent work [14]–[16] has shown that spatial weighting of importance for each zone can be used to control the amount of leakage and improve the performance of the multizone reproduction system.

Multizone soundfield reproductions designed for mono-frequent soundfields have been extended to wideband soundfields including speech [8], [16]–[18]. Recent research has investigated

the perceptual quality of multizone soundfields [2] and methods have been proposed to improve the quality using psychoacoustic models [16]. In this paper, we address open questions on the perception of leakage and what this means for speech privacy amongst zones.

Reproducing personal sound in public spaces, such as open-offices, brings concerns regarding privacy between zones. Existing methods do not specifically address the problem of information leaking between zones and may lead to the ability of users to deduce what content is being reproduced in other zones, e.g. in private teleconference meetings. Good speech privacy requires that the leaked speech signal is not intelligible [19], [20]. Although research has shown how to synthesise and reproduce wideband speech soundfields in multiple zones, state-of-the-art methods still lack the acoustic contrast between zones to provide speech privacy [2], [3], [7], [8]. For reproduction of speech at a level of $60\,\mathrm{dBA}$ in a target bright zone, state-of-the-art methods can provide a quiet zone level down to $\approx 35\,\mathrm{dBA}$ for zones large enough to fit a human listener and for sound arriving from any direction. However, in order to provide speech privacy in a quiet room, a consistent acoustic contrast of $\approx 60\,\mathrm{dBA}$ may be required, which would maintain a quiet zone level below the threshold of hearing ($\approx 0\,\mathrm{dBA}$). The level of acceptable interference while in different listening scenarios has also been studied and has shown that, in some scenarios, experienced listeners have an acceptability threshold of less than $-40\,\mathrm{dB}$ [21]. Most measurements of speech intelligibility, and thus privacy, are based on the mutual information conveyed between a speaker and listener [22]. The mutual information between different zones in a multizone reproduction scenario has been shown to be theoretically controllable, for the goal of maintaining speech privacy, by using spatially controlled masking [8]. However, it has only been shown that spatially controlled masking can improve speech privacy for theoretically ideal cases (using an impractical number of secondary sources). In practice, a reduced number of loudspeakers introduces deleterious phenomena such as spatial aliasing.

The fundamental problem of spatial aliasing in discretised soundfield reproductions has been investigated in [23] and shows that, in a multizone scenario, spatial aliasing can be considered another contributor to zone leakage. Analytical definitions have been formulated for the occurrence of aliasing in zoned soundfields [14], [23], [24] which can be used to account for its particular contribution to leakage. Another contributor to leakage is that caused by current multizone soundfield methods, where constraints on power and spatial error reduce acoustic contrast. It has been shown that acoustic contrast, and hence leakage, is frequency dependent [3], [13], [14], [16],

[25], [26] with most multizone soundfield synthesis and reproduction techniques, however, [16] does show that the leakage can be partly controlled per frequency. Frequency dependent leakage leads to an unknown spectral distortion of the audio content across different spatial regions.

In this work, a novel method consisting of several stages for improving speech privacy in personal sound zones is proposed and extends the authors' previous work in [8]. The proposed measure, Speech Intelligibility Contrast (SIC), which is based on mutual information between spatial regions, is used to maximise speech privacy in multizone soundfield reproductions. Optimisations are formulated to maximise SIC and instrumental measures of subjective quality after extending previous work from two dimensional (2-D) to three dimensional (3-D) wave equations.

The authors' previous work is further extended to incorporate novel multizone soundfield dependent spatial and spectral masker filters. The spatial masker filter is designed as a multizone soundfield filter which is dependent on the multizone soundfield reproduction scenario of the speech in the target bright zone. The spectral masker filters are designed as a combination of *a priori* estimates of the acoustic contrasts of both the masker signal and target speech signal multizone soundfield reproductions. Further, spectral shaping filters are designed to reduce the effects of aliasing, caused by discretised loudspeaker spacings, specifically on multizone soundfield reproductions. A combination of the proposed filters is used in masking the leaked speech in the quiet zone whilst leaving the target bright zone speech unimpaired.

The extended methods are analysed and evaluated to ensure a practical, systematic and robust procedure to improving speech privacy in personal sound zones. Experimental results are presented for both simulations and a real-world implementation using practical numbers of loudspeakers.

II. Weighted Multizone Speech Soundfields

This section overviews the soundfield synthesis and reproduction from the weighted orthogonal basis expansion [3], [14] and spherical harmonic expansion [27]–[30] methods, respectively. The methods described later in this paper rely on general properties (and combinations of properties) of multizone soundfield reproductions, such as acoustic contrast, loudspeaker layout, zone geometry and target zone soundfield wave fronts. The multizone techniques that can be used with the proposed methods are not limited to those described in this section, however,

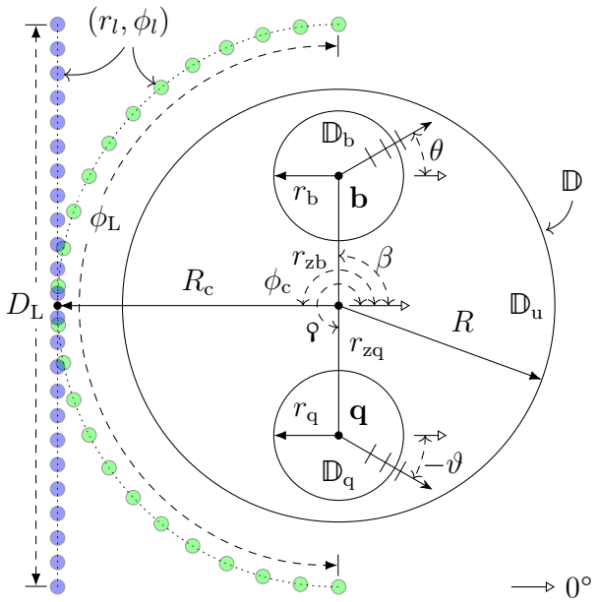


Fig. 1: A multizone soundfield reproduction layout is shown for a semi-circular (green) and linear (blue) loudspeaker array.

the descriptions in this section are given to facilitate the reader in understanding the proposed methods.

A. Notation, Definitions and Multizone Setup

Throughout this paper, the following notations are used: time-domain functions and their frequency-domain function transformation are represented in lowercase and uppercase italics, respectively. Vectors and matrices are represented by lowercase and uppercase bold face, respectively. The set of all real numbers is $\mathbb{R}$, $\mathbb{R}^+ \triangleq \{x : x \in \mathbb{R}, x \geq 0\}$, the set of all natural numbers starting at zero is $\mathbb{N}_0$, sets of indices are given by $[\![X]\!] \triangleq \{x : x \in \mathbb{N}_0, x < X\}$ and the unit imaginary number is $i = \sqrt{-1}$.

A personal sound zone system is depicted in Fig. 1 where the reproduction region, $\mathbb{D}$, of radius R contains three sub-regions denoted by $\mathbb{D}_\mathrm{b}$, $\mathbb{D}_\mathrm{q}$ and $\mathbb{D}_\mathrm{u} = \mathbb{D} \setminus (\mathbb{D}_\mathrm{b} \cup \mathbb{D}_\mathrm{q})$ called the bright, quiet and unattended zone, respectively. The radius of $\mathbb{D}_\mathrm{b}$ and $\mathbb{D}_\mathrm{q}$ are r_b and r_q, respectively and have centre points $\mathbf{b} \equiv (r_\mathrm{zb}, \beta)$ and $\mathbf{q} \equiv (r_\mathrm{zq}, \wp)$, respectively. Two separate loudspeaker geometries are shown for L loudspeakers with array centres at an angle of ϕ_c and distance R_c with the

lth loudspeaker position $\mathbf{l}_l \equiv (r_l, \phi_l), l \in [\![L]\!]$. The semi-circular array is concentric with $\mathbb{D}$, has a radius R_c and subtends an angle ϕ_L. The linear array is of length D_L. The loudspeakers are assumed to behave like omnidirectional point sources for simplicity. The angle of a desired point-source or plane-wave in $\mathbb{D}_\mathrm{b}$ is θ and in $\mathbb{D}_\mathrm{q}$ is ϑ. The wavenumber is given by $k = 2\pi f/c$, where f is frequency and c is the speed of sound propagation through a medium. In this work, c is assumed to be constant and therefore, f and k are interchangeable within a multiplicative constant, $2\pi/c$.

B. Multizone Soundfield Reproduction Method

Any arbitrary soundfield can be described by a set of plane-waves arriving from every angle [27], including speech soundfields. A soundfield function, $S(\mathbf{x}; k)$, that fulfils the wave equation, where $\mathbf{x} \in \mathbb{D}$ is an arbitrary spatial sampling point, can be defined with an additional spatial weighting function, $w(\mathbf{x})$, as shown in the orthogonal basis expansion approach [3], [14] to multizone soundfield reproduction. This weighting function allows for relative importance between zones to be specified for reproduction. The weighted soundfield function used in this work can be written as

$$S(\mathbf{x}; k) = \sum_{h \in [\![J]\!]} \mathcal{W}_h \, P_h(\mathbf{x}; k), \tag{1}$$

where, for a given weighting function, the coefficients, $\mathcal{W}_h$, are for a set of plane-wave soundfields, $P_h(\mathbf{x}; k)$, and J is the number of basis plane-waves [14].

The frequency domain complex loudspeaker weights used to reproduce the soundfield over the plane[1] are [14], [30], [31]

$$W_l(k) = \frac{\Delta\phi_\mathrm{s}}{2\pi i k} \sum_{\bar{m}=-\overline{M}}^{\overline{M}} \sum_{h \in [\![J]\!]} \frac{i^{\bar{m}} \exp(i\bar{m}(\phi_l - \rho_h))}{\hbar^{(1)}_{\bar{m}}(r_l k)} \mathcal{W}_h \tag{2}$$

where $\overline{M} = \lceil kR \rceil$ is the truncation length [14], $\hbar^{(1)}_\nu(\cdot)$ is a νth-order spherical Hankel function of the first kind, $\rho_h = 2\pi(h-1)/J$ are the plane-wave angles, ϕ_l is the angle of the lth loudspeaker from the horizontal axis and $\Delta\phi_\mathrm{s}$ is the angular spacing of the loudspeakers. Here, $\mathcal{W}_h$ is chosen to minimise the difference between the desired soundfield and the actual soundfield [14].

[1]Since the loudspeakers lie on a plane, an integration over elevation is carried out on the orthonormal spherical harmonics to simplify (2) and remove the dependence on elevation [27, Ch. 8].

To reproduce plane-wave speech soundfields, the set of loudspeaker signals can be found by applying $W_l(k)$ to the speech in the frequency domain and inverse transforming the signal back to the time domain. The set of framed loudspeaker signals in the time-frequency domain are given by[2]

$$Q_l(a,k) = W_l(k)\, Y(a,k) \tag{3}$$

where $Y(a,k)$ is the discrete Fourier transform of the ath overlapping windowed frame, from a total of A frames, of the input speech signal, $y(n)$. Each loudspeaker signal, $q_l(n)$, is reconstructed by performing overlap-add reconstruction with inverse transformed $Q_l(a,k)$ and the synthesis window. This results in the loudspeaker signals, which will reproduce the multiple zones.

Filtering each of the loudspeaker signals with their respective 3-D acoustic transfer function (ATF)[3] [27],

$$T(\mathbf{x},\mathbf{l};k) = \frac{\exp\big(ik\big\|\mathbf{x}-\mathbf{l}\big\|\big)}{4\pi\big\|\mathbf{x}-\mathbf{l}\big\|}, \tag{4}$$

and summing to give the superposition will result in the actual speech soundfield,

$$P^{(\mathrm{sp})}(\mathbf{x};a,k) = \sum_{l\in[\![L]\!]} Q_l(a,k)\, T(\mathbf{x},\mathbf{l}_l;k), \tag{5}$$

where $Q_l(a,k)$ is the time-frequency domain transform of $q_l(n)$ and $Q_l(k)$ is the frequency domain transform of $q_l(n)$.

The framed pressure can be observed as

$$p(\mathbf{x};a,n) = \mathrm{Re}\left\{ K^{-1} \sum_{m\in[\![K]\!]} P^{(\cdot)}(\mathbf{x};a,k_m) \exp\left(\frac{icnk_m}{2\hat{f}}\right)\right\}, \tag{6}$$

where $\mathrm{Re}\{\cdot\}$ returns the real part of its argument, $P^{(\cdot)}$ is any given soundfield function, $k_m \triangleq 2\pi\hat{f}m/cK$ and $\hat{f}$ is the maximum frequency. Performing overlap add on $p(\mathbf{x};a,n)$ then results in the pressure signal $p(\mathbf{x};n)$.

The soundfield can now be evaluated at any given point in the reproduction region for different input signals and $p(\mathbf{x};n)$ can be observed in the bright zone and quiet zone in order to estimate

---

[2]Note that recomputing $W_l(k)$ for each frame, a, is required for moving virtual sources and/or zones.

[3]2-D system models have also been shown to provide reasonable acoustic contrast in real-world environments [3]. 2-D ATFs are given by $T_{2\mathrm{D}}(\mathbf{x},\mathbf{l};k) = \frac{i}{4}\mathscr{H}_0^{(1)}\big(k\big\|\mathbf{x}-\mathbf{l}\big\|\big)$ [27].

the behaviour of the system. From (6) it is possible to analyse the speech intelligibility and quality in different zones in order to control the soundfield reproduction as described in the next section.

III. Speech Privacy and Intelligibility Contrast

This section discusses the relationship between speech privacy and intelligibility, and proposes the Speech Intelligibility Contrast (SIC) measure for improving privacy in personal sound zones. Two optimisations are provided as methods to control multizone soundfield reproductions to improve speech privacy where the latter of the described methods also yields quality control in reproductions.

A. The Speech Intelligibility Contrast (SIC)

It is noted that the relation between speech privacy and intelligibility is highly correlated. Two different privacy measures, the Speech Privacy Class (SPC) for closed spaces and the Articulation Index (AI) for open plan spaces, are published as standards ASTM E2638 [32] and ASTM E1130 [33], respectively. The SPC has been shown to be a good measure for high privacy scenarios [20] and with the two standard measures (SPC and AI) highly correlated to speech intelligibility, it is reasonable to maximise an intelligibility contrast measure to obtain speech privacy. A measure of intelligibility contrast has the benefit, over SPC and AI, of providing accurate estimations of speech privacy in different scenarios, such as reverberant rooms [34] and with time-frequency weighted noisy speech [35].

The basis of many objective intelligibility measures is an analysis of spectral band powers which have been shown to be highly correlated with subjective measures. A clean speech (talker) signal, $y_T(n)$, and a degraded speech (listener) signal, $y_L(n)$, with a high signal to noise ratio (SNR) will also attain high mutual information [22]. In this work, $\mathcal{I}_{\mathcal{M}}(y_L; y_T)$ is used to denote the intelligibility for two signals, $y_T(n)$ and $y_L(n)$. A proxy of the mutual information, such as that provided by the Short-Time Objective Intelligibility (STOI) [35] or Speech Transmission Index (STI) [34], is denoted by the measure, $\mathcal{M}$. The soundfield intelligibility, $\mathcal{I}_{\mathcal{M}}(p(\mathbf{x};\cdot); y) \in \{0, \ldots, 1\} \subsetneq \mathbb{R}$, of a signal, $y(n)$, at some spatial point, $\mathbf{x} \in \mathbb{D}$, is measured using the pressure signal, $p(\mathbf{x}; n)$.

The SIC is defined as [8]

$$\mathrm{SIC}_{\mathcal{M}} = \mathfrak{d}_{\mathrm{b}}^{-1} \int_{\mathbb{D}_{\mathrm{b}}} \mathcal{I}_{\mathcal{M}}(p(\mathbf{x};\cdot);y)\, d\mathbf{x} - \mathfrak{d}_{\mathrm{q}}^{-1} \int_{\mathbb{D}_{\mathrm{q}}} \mathcal{I}_{\mathcal{M}}(p(\mathbf{x};\cdot);y)\, d\mathbf{x}, \tag{7}$$

where $\mathfrak{d}_{\mathrm{b}} \triangleq \int_{\mathbb{D}_{\mathrm{b}}} 1\, d\mathbf{x}$ and $\mathfrak{d}_{\mathrm{q}} \triangleq \int_{\mathbb{D}_{\mathrm{q}}} 1\, d\mathbf{x}$ are the areas (sizes) of $\mathbb{D}_{\mathrm{b}}$ and $\mathbb{D}_{\mathrm{q}}$, respectively, and $\mathrm{SIC}_{\mathcal{M}}$ has a restricted domain such that $\mathcal{I}_{\mathcal{M}}, \forall \mathbf{x} \in \mathbb{D}_{\mathrm{b}}$ is greater than or equal to $\mathcal{I}_{\mathcal{M}}, \forall \mathbf{x} \in \mathbb{D}_{\mathrm{q}}$. The following subsection provides two methods to maximise $\mathrm{SIC}_{\mathcal{M}}$.

B. Privacy and Quality Control

To maximise the SIC, $\mathcal{I}_{\mathcal{M}}$ must be maximised at all points in $\mathbb{D}_{\mathrm{b}}$ whilst maintaining a minimum valued $\mathcal{I}_{\mathcal{M}}, \forall \mathbf{x} \in \mathbb{D}_{\mathrm{q}}$. In general, the higher mean SNR of $p(\mathbf{x};n)$ over $\mathbb{D}_{\mathrm{b}}$ the better, so reducing the mean SNR of $p(\mathbf{x};n)$ over $\mathbb{D}_{\mathrm{q}}$ naturally becomes the criteria to increase $\mathrm{SIC}_{\mathcal{M}}$. To maximise the SIC, noise is added to the arbitrary loudspeaker signals, $q_l(n)$, that are used to reproduce $p(\mathbf{x};n)$. It is assumed that $q_l(n)$ are designed to reproduce a mean amplitude of $p(\mathbf{x};n)$ over $\mathbb{D}_{\mathrm{b}}$ greater than that of $p(\mathbf{x};n)$ over $\mathbb{D}_{\mathrm{q}}$. A constrained optimisation is then formulated which is dependent on the reproduced signals in the quiet and bright zones as

$$\arg\max_{G} \mathrm{SIC}_{\mathcal{M}}, \text{ subject to: } G \in \mathbb{R}^{+}, \tag{8}$$

where the optimal noise levels, G, of $q_l(n)$ are found.

A private personal sound zone system would ideally support high perceptual quality in the bright zone whilst preserving maximum SIC. A trade-off between privacy and target quality is apparent when adjusting G of $q_l(n)$, as doing so reduces the quality of $p(\mathbf{x};n), \forall \mathbf{x} \in \mathbb{D}_{\mathrm{b}}$ due to the addition of error to $p(\mathbf{x};n), \forall \mathbf{x} \in \mathbb{D}$. Using a similar notation to $\mathcal{I}_{\mathcal{M}}$, the quality of $p(\mathbf{x};n), \forall \mathbf{x} \in \mathbb{D}_{\mathrm{b}}$ (the reproduction of $y(n)$) is any speech quality assessment model, $\mathcal{B}_{\acute{\mathcal{M}}}(p(\mathbf{x};\cdot);y) \in \{0,\ldots,1\} \subsetneq \mathbb{R}$, for a given measure, $\acute{\mathcal{M}}$, which is scaled to match that of $\mathcal{I}_{\mathcal{M}}$.

Now a new optimisation can be defined as

$$\begin{aligned} \arg\max_{G} \quad & \left(\mathrm{SIC}_{\mathcal{M}} + \frac{\lambda}{\mathfrak{d}_{\mathrm{b}}} \int_{\mathbb{D}_{\mathrm{b}}} \mathcal{B}_{\acute{\mathcal{M}}}\, d\mathbf{x}\right), \\ \text{subject to:} \quad & G \in \mathbb{R}^{+}, \\ & \mathcal{I}_{\mathcal{M}} \geq \mathcal{B}_{\acute{\mathcal{M}}}, \forall \mathbf{x} \in \mathbb{D}_{\mathrm{b}}, \end{aligned} \tag{9}$$

where the optimal noise levels, G, are defined in section IV and the importance of quality in the optimisation is controlled with the weighting parameter, $\lambda \in \mathbb{R}^{+}$.

The multi-stage process proposed in this paper aims to optimally choose the value of G to satisfy (9) whilst also constraining the amount of energy leaked between zones and meeting constraints due to spatial aliasing resulting from the use of a limited number of loudspeakers. The next section describes the spatial and spectral sound masker design approaches proposed in this work.

IV. Spatial and Spectral Sound Masking

In this section, a method for improving speech privacy between spatial zones in multizone soundfield reproduction scenarios is described. The intelligibility between $y_T(n)$ and $y_L(n)$ can be reduced by reducing the SNR as described in section III-A. The optimisations formulated in section III-B are realised by using spatially and spectrally weighted noise maskers. Spatial filters are defined using the multizone soundfield reproduction approach and vary depending on the target multizone speech soundfield, loudspeaker layout and zone geometry. Spectral shaping is described in the form of weighted predicted acoustic contrast ratios which are also dependent on the multizone reproduction of the target speech.

A. Spatial Sound Masking

To optimise the criteria in (9) a maximum mean SNR of $p(\mathbf{x};n)$ over $\mathbb{D}_\mathrm{b}$ and minimum mean SNR of $p(\mathbf{x};n)$ over $\mathbb{D}_\mathrm{q}$, is required. To achieve this, a time-domain Gaussian noise mask, $u(n)$, is projected into the spatial domain over $\mathbb{D}$ such that its reproduction becomes a multizone soundfield reproduction scenario. In this work, constraints are applied to the multizone reproduction of $u(n)$, which is quiet in $\mathbb{D}_\mathrm{b}$ and a plane-wave field in $\mathbb{D}_\mathrm{q}$, in order to simplify the optimisation of (8) and (9). The constraints are

$$\vartheta = \cos^{-1}\left(\frac{\overrightarrow{\mathbf{pq}}\cdot\hat{\mathbf{u}}_\mathrm{o}}{\left\|\overrightarrow{\mathbf{pq}}\right\|}\right), \tag{10}$$

so that the masker source is collocated with the leakage of the target bright zone soundfield reproduction (see section V-B and Appendix A for definitions of $\overrightarrow{\mathbf{pq}}$ and $\hat{\mathbf{u}}_\mathrm{o}$), and a new weighting function, $\widehat{w}(\mathbf{x})$, is constrained to an importance of 0.05, 1 and 100 in $\mathbb{D}_\mathrm{u}$, $\mathbb{D}_\mathrm{q}$ and $\mathbb{D}_\mathrm{b}$, respectively [14], [18]. The remainder of the multizone reproduction is the same as used to generate $Q_l(a,k)$ for the speech signal.

The goal is to solve (8) and (9) or, equivalently, to control the mean SNR of $p(\mathbf{x};n)$ over $\mathbb{D}_\mathrm{q}$ by finding another set of loudspeaker signals that would reproduce $u(n)$ in $\mathbb{D}_\mathrm{q}$ only. To do this, $u(n)$ is transformed to the frequency domain, framed as $U(a,k)$ and used in replacement of the input signal, $Y(a,k)$, in (3) to give

$$\widehat{Q}_l(a,k) = \widehat{W}_l(k)\, U(a,k), \tag{11}$$

where the masker loudspeaker signals, $\widehat{Q}_l(a,k)$, are found after new loudspeaker weights are derived from (2) as $\widehat{W}_l(k)$. Superposition gives the resulting masker soundfield as

$$P^{(\mathrm{m})}(\mathbf{x};a,k) = \sum_{l\in[\![L]\!]} \widehat{Q}_l(a,k)\, T(\mathbf{x},\mathbf{l}_l;k). \tag{12}$$

The masker soundfield is then added to the speech soundfield

$$P^{(\mathrm{sp,m})}(\mathbf{x};a,k) = P^{(\mathrm{sp})}(\mathbf{x};a,k) + \bar{G}P^{(\mathrm{m})}(\mathbf{x};a,k) \tag{13}$$

$$= \sum_{l\in[\![L]\!]} Q'_l(a,k)\, T(\mathbf{x},\mathbf{l}_l;k), \tag{14}$$

where $Q'_l(a,k)$ are the new loudspeaker signals and $\bar{G}$ is the relative gain adjustment given by the root mean square (RMS) value from all L loudspeaker signals,

$$\bar{G} \triangleq \frac{G}{K}\left(\frac{1}{L}\sum_{l\in[\![L]\!],m\in[\![K]\!]} \left|Q_l(k_m)\right|^2\right)^{1/2}. \tag{15}$$

Then, $\mathrm{SIC}_\mathcal{M}$ is obtained from (7) after $p(\mathbf{x};n)$ is found from (6) using $P^{(\mathrm{sp,m})}(\mathbf{x};a,k)$. Now $\mathrm{SIC}_\mathcal{M}$ can be used to optimise G from (15) through (13) with (8). Alternatively, though, similarly, $\mathrm{SIC}_\mathcal{M}$ and $\mathcal{B}_{\acute{\mathcal{M}}}$ can be used to optimise G with (9). The optimisation problem can now be analysed by measuring $\mathcal{I}_\mathcal{M}$ for $\mathbf{x}\in\mathbb{D}_\mathrm{b}\cap\mathbb{D}_\mathrm{q}$, $\mathcal{B}_{\acute{\mathcal{M}}}$ for $\mathbf{x}\in\mathbb{D}_\mathrm{b}$, $\mathrm{SIC}_\mathcal{M}$ and for various $G\in\mathbb{R}^+$.

B. Long Term Average Speech Spectrum

The average magnitude spectrum of speech has been well documented and is known as the Long-Term Average Speech Spectrum (LTASS) [36], [37]. In order to accurately mask the speech that is leaked into the quiet zone, the spectrum of the masker should closely match the spectrum of the leakage. At any measurement point in a speech soundfield the spectral shape will, on average, consist of the speech magnitude spectrum and spectral shaping caused by the system response. Speech Shaped Noise (SSN) is an appropriate masking signal for the speech component

of leaked content. To obtain SSN, framed Guassian noise is shaped to the LTASS as $U^{(\mathrm{sp})}(a,k)$ where $^{(\mathrm{sp})}$ denotes filtering for the speech spectrum. The magnitude response of the LTASS filter, $H^{(\mathrm{sp})}(k)$, can be approximated by either table 2 of [36], table 1 of [37] or by finding the mean sound pressure level (SPL) for a set of speech samples, e.g.,

$$\left|H^{(\mathrm{sp})}(k)\right|^2 = \frac{2}{BN^2}\sum_{b\in[\![B]\!]}\left|\sum_{n\in[\![N]\!]} h_b^{(\mathrm{sp})}(n)\exp\left(\frac{-icnk}{2\hat{f}}\right)\right|^2 \tag{16}$$

where $h_b^{(\mathrm{sp})}(n)\in\mathbb{R}$ is the bth non-overlapping frame from the sequence of $\widehat{N}$ speech samples, B is the number of frames and $B=\lceil\widehat{N}/N\rceil$. The SSN, $U^{(\mathrm{sp})}(a,k)$, can then be used in (11) to obtain $Q_l'(a,k)$ from (14) via (13) and (12).

C. A Priori Reproduction Spectrum Estimation

Even though the multizone reproduction system aims to match the desired input signal spectrum in the bright zone it does not guarantee that the quiet zone spectrum that is leaked remains the same shape. In fact, the spectrum of the quiet zone will vary significantly depending on many factors, such as the geometrical positioning of zones, virtual sources and secondary sources, and the type of reproduction technique used.

It is possible, however, to form an *a priori* estimate of the leaked spectrum by either knowing or estimating the inverse of the underlying acoustic contrast in the system. The inverted acoustic contrast can be found by either the ratio of energies between zones or by assuming a uniform (temporal) frequency spectrum in the bright zone. The system magnitude response in $\mathbb{D}_{\mathrm{q}}$ can be estimated using the soundfield $P^{(\mathrm{sp})}(\mathbf{x};a,k)$ reproduced from $Q_l(a,k)$, as

$$\left|H^{(\mathrm{q})}(k)\right| = \frac{1}{A}\sum_{a\in[\![A]\!]}\left(\frac{\mathfrak{d}_{\mathrm{b}}\int_{\mathbb{D}_{\mathrm{q}}}\left|P^{(\mathrm{sp})}(\mathbf{x};a,k)\right|d\mathbf{x}}{\mathfrak{d}_{\mathrm{q}}\int_{\mathbb{D}_{\mathrm{b}}}\left|P^{(\mathrm{sp})}(\mathbf{x};a,k)\right|d\mathbf{x}}\right)^{1/2}, \tag{17}$$

where $^{(\mathrm{q})}$ denotes a filter for the leaked quiet zone spectrum.

In practical reproductions it may be unnecessary to shape the noise spectrum above some aliasing frequency, k_{u}, as the leakage would boost high frequencies which can be seen later in Fig. 6. A more practical filter can be approximated as,

$$\left|H^{(\mathrm{q}')}(k)\right| = \begin{cases}\left|H^{(\mathrm{q})}(k)\right|, & k<k_{\mathrm{u}}\\ \left|H^{(\mathrm{q})}(k_{\mathrm{u}})\right|, & k\geq k_{\mathrm{u}}\end{cases}, \tag{18}$$

which ensures no shaping above k_{u}.

The leakage spectrum filter, $H^{(\mathrm{q}')}(k)$, can be used alongside the LTASS filter from (16) to obtain a good approximation of the leaked speech spectrum. The Gaussian noise, $U^{(\mathrm{sp},\mathrm{q}')}(a,k)$, shaped to $H^{(\mathrm{sp})}(k)$ and $H^{(\mathrm{q}')}(k)$, then matches accurately the leaked speech in the quiet zone up to the aliasing frequency and can then be used in (11) to obtain $Q_l'(a,k)$ from (14) via (13) and (12).

D. Secondary Leakage

Leakage between zones is a feature of multizone reproductions regardless of the target reproduction signal. When reproducing a multizone masking soundfield which matches the leaked speech in the target quiet zone there will also be leakage of the masker back into the target bright zone, we term this the *secondary leakage*. The shape of the secondary leakage may detrimentally influence both $\mathrm{SIC}_{\mathcal{M}}$ and $\mathcal{B}_{\acute{\mathcal{M}}}$ which shows the importance of the masker spectrum in the optimisation of (9). Ideally, a spectrum which influences both $\mathrm{SIC}_{\mathcal{M}}$ and $\mathcal{B}_{\acute{\mathcal{M}}}$ to equal extent, or to satisfy (9), is needed.

In this work we propose the use of a secondary leakage filter, $H^{(\mathrm{b})}(k)$, to determine a masker spectrum which has equal influence on $\mathrm{SIC}_{\mathcal{M}}$ and $\mathcal{B}_{\acute{\mathcal{M}}}$. As seen from the target quiet zone, the leaked spectrum back into the target bright zone is estimated using the soundfield $P^{(\mathrm{m})}(\mathbf{x};a,k)$ reproduced from $\widehat{Q}_l(a,k)$, and the secondary leakage spectrum is found as

$$\left|H^{(\mathrm{b})}(k)\right| = \frac{1}{A}\sum_{a\in[\![A]\!]}\left(\frac{\mathfrak{d}_{\mathrm{q}}\int_{\mathbb{D}_{\mathrm{b}}}\left|P^{(\mathrm{m})}(\mathbf{x};a,k)\right|d\mathbf{x}}{\mathfrak{d}_{\mathrm{b}}\int_{\mathbb{D}_{\mathrm{q}}}\left|P^{(\mathrm{m})}(\mathbf{x};a,k)\right|d\mathbf{x}}\right)^{1/2}, \tag{19}$$

where $^{(\mathrm{b})}$ denotes a filter for the secondary leakage spectrum.

Following the same reasoning for (18), the secondary leakage filter that ensures no shaping above k_{u} is

$$\left|H^{(\mathrm{b}')}(k)\right| = \begin{cases}\left|H^{(\mathrm{b})}(k)\right|, & k < k_{\mathrm{u}}\\ \left|H^{(\mathrm{b})}(k_{\mathrm{u}})\right|, & k \geq k_{\mathrm{u}}\end{cases}, \tag{20}$$

which is used to obtain the masker spectrum which has controllable influence on intelligibility and quality as

$$\left|H^{(\mathcal{IB})}(k)\right| = \left|H^{(\mathrm{sp})}(k)\right|\frac{\left|H^{(\mathrm{q}')}(k)\right|^{1-\lambda}}{\left|H^{(\mathrm{b}')}(k)\right|^{\lambda}}. \tag{21}$$

The influence of the spectrum on intelligibility over quality can be controlled with the parameter $\lambda \in \{0, \ldots, 1\} \subsetneq \mathbb{R}$, unlike λ, which does not control the shape of the spectrum.

The spectral maskers in this section have been derived for a single target speech signal. The methods are also applicable for cases where separate speech signals in each zone are desired, however, because the leaked speech between zones is not controlled, further reductions in quality may occur. Methods for controlling the leaked spectrum between zones, which may then improve quality, have been proposed in [16], [18].

V. Reducing Loudspeakers and Aliasing

A fundamental issue with wideband soundfield synthesis is the high number of secondary sources required for alias free reproduction of speech or music. In this section the consequent effect of aliasing on multizone soundfields is described and an analytical approach to reduce the effect is presented.

A. Grating Lobe Motivated Masker Filtering

For a sound zoning system to remain practical it should be possible for a small number of loudspeakers to provide high SIC and quality. A fundamental problem with the reduction in the number of loudspeakers is spatial aliasing which gives rise to grating lobes capable of impeding the different zones and cannot be spatially controlled with soundfield synthesis.

Since filtering the target bright zone signal will knowingly alter the quality of the reproduced content it is sensible to shape only the portion of the (temporal) frequency spectrum of the masker signal without spatial aliasing artefacts. If the masker signal is dominant at frequencies where its grating lobes directly impede the target bright zone then the quality will be significantly reduced. Band-limiting the masker signal, $u(n)$, by applying a low-pass, denoted by $^{(\mathrm{lp})}$, filter, $H^{(\mathrm{lp})}(k)$, with a cutoff frequency of k_{u} (some aliasing frequency) will eliminate this effect, however, the masker signal will then not be able to mask speech in the stopband. Any low-pass filter can be used, for instance, a Chebyshev Type I [38] is

$$\left|H^{(\mathrm{lp})}(k)\right| = \left(1 + \left(\varepsilon\, \mathcal{T}_{\breve{n}}(k/k_{\mathrm{u}})\right)^2\right)^{-1/2}, \tag{22}$$

where $\mathcal{T}_{\breve{n}}(\cdot)$ is a Chebyshev polynomial [38] of the first kind with order $\breve{n}$ and ε is the maximum allowable passband ripple. The noise signal, $u(n)$, is filtered with $H^{(\mathrm{lp})}(k)$ to obtain $U^{(\mathrm{lp})}(a, k)$.

Fortunately, the frequency spectrum of speech is dominant at lower frequencies [36], [37] and so the majority of information leaked can still be masked effectively from the low-pass filtered masker signal. To perform the spatially weighted masking, (11) is used with noise signal $U^{(\mathrm{lp})}(a,k)$ and $Q_l'(a,k)$ is found from (14) via (13) and (12).

B. Grating Lobe Prediction

The grating lobes can be accurately predicted if the loudspeaker array and zone geometry is known. The next two sub-subsections provide an analytical approach to finding the frequency where grating lobes touch the quiet zone for both circular and linear loudspeaker arrays.

1) Circular Array Grating Lobes: For a truncation length of $\overline{M}' = \lceil kR' \rceil$ [14], [39], where R' is the radius of the smallest circle (concentric with $\mathbb{D}$) encompassing all zones, and by using the part circle method [14], [31], it is possible to formulate an approximation for the upper frequency limit, k_{u}, at which aliasing will begin to occur. The minimum number of required loudspeakers is given by [14], [24]

$$L \geq \left\lceil \frac{\phi_{\mathrm{L}}\big(2\overline{M}+1\big)}{2\pi} \right\rceil + 1, \tag{23}$$

substituting the truncation length, $\overline{M}'$, and rearranging gives [40]

$$\hat{k}_{\mathrm{u}} = \frac{2\pi(L-1) - \phi_{\mathrm{L}}}{2R'\phi_{\mathrm{L}}}, \tag{24}$$

however, this provides the frequency where the centre of the grating lobe is at least R' from the centre of the reproduction, not accounting for zone positions. In many cases it is possible to use a frequency higher than $\hat{k}_{\mathrm{u}}$ where the grating lobes do not travel through the quiet zone. That is to say, aliasing artifacts can be tolerated in $\mathbb{D}_{\mathrm{u}}$ depending on relative locations of the zones thus redefining the aliasing to that occurring in $\mathbb{D}_{\mathrm{q}}$, not $\mathbb{D}$. The aim is to find a new $\hat{k}_{\mathrm{u}}$ by deriving a replacement for $2R'$. To aid the derivations, Fig. 2 shows a circular array with auxiliary values.

Here, the work in [23] is extended to the multizone reproduction scenario to define a zone based limit for the grating lobe. Similar to the work in [23], a point, $\mathbf{p}$, is positioned on the loudspeaker arc at distance R_{c} and with angle

$$\alpha = \theta - \sin^{-1}\left(\frac{d^{\perp}_{\overrightarrow{\mathbf{pb}}}}{R_{\mathrm{c}}}\right) + \pi, \tag{25}$$

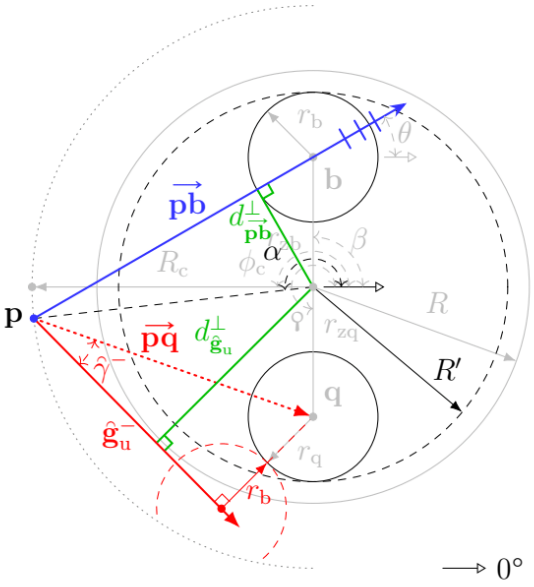


Fig. 2: Auxiliary entities of a circular array multizone soundfield reproduction layout. The plane-wave vector ($\overrightarrow{\mathbf{pb}}$) is blue, the grating lobe limit ($\hat{\mathbf{g}}^{-}_\mathrm{u}$ is shown) found using (28) is red and the frequency limit ($\hat{k}'_\mathrm{u}$) is computed with (31) using the perpendicular distances ($d^{\perp}_{\hat{\mathbf{g}}_\mathrm{u}}$ and $d^{\perp}_{\overrightarrow{\mathbf{pb}}}$) that are shown in green.

where $\mathbf{p}$ is the origin for grating lobes as shown in Fig. 2 and

$$d^{\perp}_{\overrightarrow{\mathbf{pb}}} = \left| r_\mathrm{zb} \sin(\beta - \theta) \right|. \tag{26}$$

The first spectral repetition of grating lobes have a width equal to the bright zone diameter [23]. The outer-most tangent from the origin of a circle of radius $r_\mathrm{b} + r_\mathrm{q}$ at $\mathbf{q}$ which intersects $\mathbf{p}$, corresponds to the centre of a grating lobe whose edge touches $\mathbb{D}_\mathrm{q}$. Vector notation is used when finding the tangent and hence the newly defined aliasing frequency.

The vector $\overrightarrow{\mathbf{pq}}$ (derived in Appendix A) can be rotated[4] about $\mathbf{p}$ to equate the centre of the grating lobe. The maximum allowable angle of the grating lobe before impeding $\mathbb{D}_\mathrm{q}$ is one of the two angles:

$$\hat{\gamma}^{\pm} = \pm \sin^{-1}\left(\frac{r_\mathrm{b} + r_\mathrm{q}}{\left\| \overrightarrow{\mathbf{pq}} \right\|} \right), \tag{27}$$

[4]The rotational matrix, denoted by $\mathring{\mathbf{R}}(\cdot)$, is defined in Appendix A.

where $\|\cdot\|$ denotes the Euclidean norm. Therefore the grating lobe of the upper frequency limit due to aliasing is one of the two tangents[5]:

$$\hat{\mathbf{g}}_{\mathrm{u}}^{\pm} = \mathring{\mathbf{R}}\big(\hat{\gamma}^{\pm}\big) \cdot \overrightarrow{\mathbf{pq}}. \tag{28}$$

The perpendicular distance from $\hat{\mathbf{g}}_{\mathrm{u}}^{\pm}$ to the origin,

$$d_{\hat{\mathbf{g}}_{\mathrm{u}}^{\pm}}^{\perp} = \frac{\left|\hat{\mathbf{p}}^{\mathsf{T}} \cdot \left(\mathring{\mathbf{R}}\big(\frac{\pi}{2}\big) \cdot \hat{\mathbf{g}}_{\mathrm{u}}^{\pm}\right)\right|}{\left\|\hat{\mathbf{g}}_{\mathrm{u}}^{\pm}\right\|}, \tag{29}$$

where $^{\mathsf{T}}$ is a transposition of the vector, can be used to determine the correct tangent as

$$d_{\hat{\mathbf{g}}_{\mathrm{u}}}^{\perp} = \max\Big(d_{\hat{\mathbf{g}}_{\mathrm{u}}^{+}}^{\perp}, d_{\hat{\mathbf{g}}_{\mathrm{u}}^{-}}^{\perp}\Big). \tag{30}$$

The corresponding circular array aliasing frequency, $\hat{k}'_{\mathrm{u}}$, can then be found by replacing $2R'$ in (24) with $d_{\hat{\mathbf{g}}_{\mathrm{u}}}^{\perp} + d_{\overrightarrow{\mathbf{pb}}}^{\perp}$, as

$$\hat{k}'_{\mathrm{u}} = \max\left(\frac{2\pi(L-1) - \phi_{\mathrm{L}}}{\Big(d_{\hat{\mathbf{g}}_{\mathrm{u}}}^{\perp} + d_{\overrightarrow{\mathbf{pb}}}^{\perp}\Big)\phi_{\mathrm{L}}}, \hat{k}_{\mathrm{u}}\right). \tag{31}$$

2) Linear Array Grating Lobes: Similar to the derivation for a circular array, the linear array solution uses the tangents from the origin of the grating lobe to a circle of radius $r_{\mathrm{b}} + r_{\mathrm{q}}$ at point $\mathbf{q}$. Fig. 3 shows a linear array with auxiliary values. For a linear array the point of origin of the grating lobe, $\mathbf{p}$, is found from the intersection of the unit plane-wave vector and the loudspeaker array unit vector.

Using $\overrightarrow{\mathbf{pq}}$ for a linear array (see Appendix B) in (27) and (28) yields $\bar{\mathbf{g}}_{\mathrm{u}}^{\pm}$ for a linear array. The maximum of the two angles between $\bar{\mathbf{g}}_{\mathrm{u}}^{\pm}$ and $\overrightarrow{\mathbf{pb}}$ (see Appendix B) gives the maximum allowable grating lobe angle for a linear array by

$$\psi^{\pm} = \cos^{-1}\left(\frac{\bar{\mathbf{g}}_{\mathrm{u}}^{\pm} \cdot \overrightarrow{\mathbf{pb}}}{\left\|\bar{\mathbf{g}}_{\mathrm{u}}^{\pm}\right\| \cdot \left\|\overrightarrow{\mathbf{pb}}\right\|}\right), \tag{32}$$

$$\bar{\gamma} = \max\big(\psi^{+}, \psi^{-}\big). \tag{33}$$

The linear array aliasing frequency [41, eq. (5.61)] is then

$$\bar{k}_{\mathrm{u}} = \frac{2\pi(L-1)}{D_{\mathrm{L}}(\sin(\bar{\gamma} - \Theta) + \sin(\Theta))}, \tag{34}$$

[5]The two tangents stem from the sign of (27) and are denoted by $\pm$.

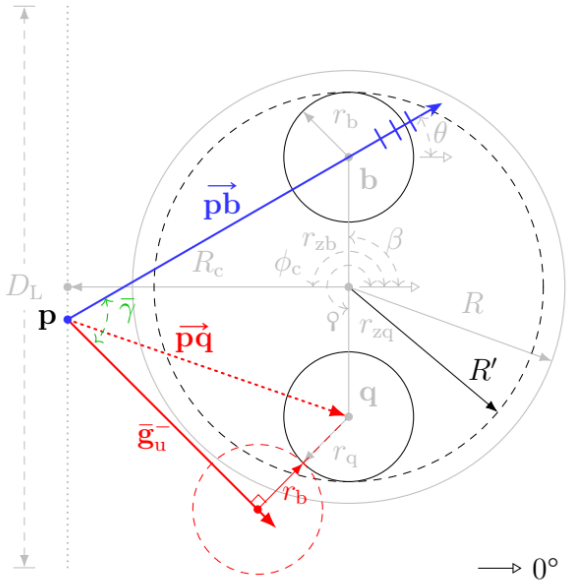


Fig. 3: Auxiliary entities of a linear array multizone soundfield reproduction layout. The plane-wave vector ($\overrightarrow{\mathbf{pb}}$) is blue, the grating lobe limit ($\bar{\mathbf{g}}_\mathrm{u}^-$ is shown) found following section V-B2 is red and the frequency limit ($\bar{k}_\mathrm{u}$) is computed with (34) using the maximum allowable grating lobe angle ($\bar{\gamma}$) that is shown in green.

where $\Theta = |\pi - |\theta| - |\phi_\mathrm{c}||$, $\theta \in \{-\pi, \ldots, \pi\}$ and $\phi_\mathrm{c} \in \{-\pi, \ldots, \pi\}$.

The upper cut-off frequency from aliasing is k_u and equates to either $\hat{k}'_\mathrm{u}$ or $\bar{k}_\mathrm{u}$ depending on the loudspeaker array geometry. k_u can be used in a low-pass filter which can then be applied to any masker signal, $U(a, k)$, in (3).

C. Example Aliasing Artefacts

The impinging effect of the grating lobe into $\mathbb{D}_\mathrm{b}$ as ϑ is varied is shown in Fig. 4. Other studies have investigated the effect of aliasing with differing numbers of loudspeakers [42]. For the cases in Fig. 4 where the optimal k_u is found (i.e. in the left column) the mean energy in $\mathbb{D}_\mathrm{b}$ remains low. When the angle of the desired wave front in $\mathbb{D}_\mathrm{q}$ is moved, the energy in $\mathbb{D}_\mathrm{b}$ increases as the grating lobe traverses the zone which is undesirable and shows the importance of accurately computing k_u. Re-evaluating k_u for changes in ϑ will ensure the energy of the grating lobe in $\mathbb{D}_\mathrm{b}$ is kept low.

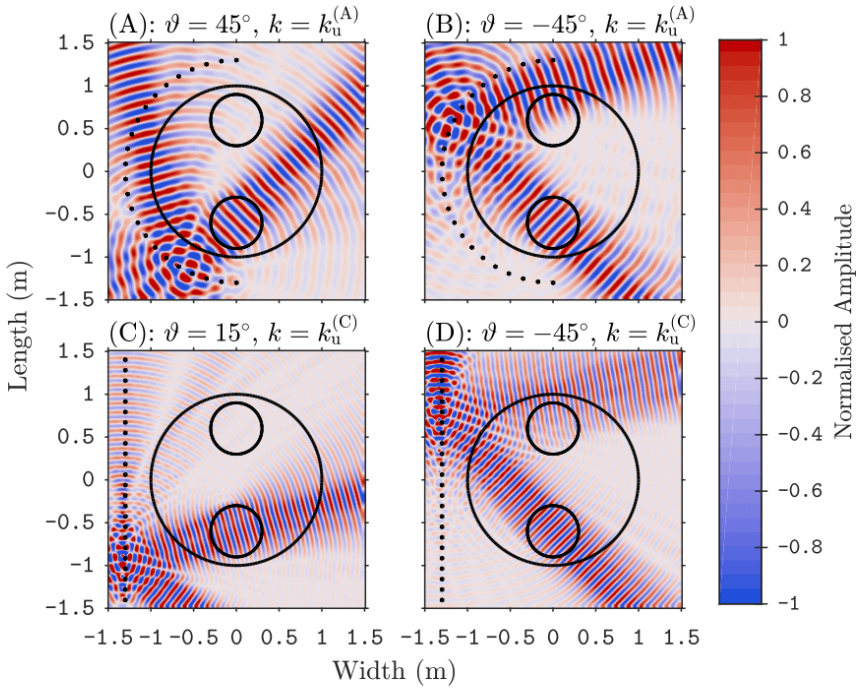


Fig. 4: The real part of the masker soundfields at a particular aliasing frequency are shown to illustrate the impinging effect of grating lobes on $\mathbb{D}_{\mathrm{b}}$. The left column shows the optimal k_{u} whereas the right column shows the masker grating lobe entering $\mathbb{D}_{\mathrm{b}}$ after ϑ has been changed from that in the left column. Soundfields are shown for a semi-circular array and linear array in the top and bottom rows, respectively. The soundfield parameters are the same as those given in section VII-A with $L = 24$. The angle of the wave front, ϑ, in $\mathbb{D}_{\mathrm{q}}$ is labelled for each plot and is chosen to best illustrate the interference in $\mathbb{D}_{\mathrm{b}}$. The superscript of k_{u} indicates which multizone setup in the figure was used to calculate k_{u}. The upper frequency limits are $k_{\mathrm{u}}^{(\mathrm{A})} = 35.6\,\mathrm{m}^{-1}$ and $k_{\mathrm{u}}^{(\mathrm{C})} = 59.6\,\mathrm{m}^{-1}$ corresponding to temporal frequencies of $1.94\,\mathrm{kHz}$ and $3.25\,\mathrm{kHz}$, respectively.

VI. Reproduction Filtering

For a set of arbitrary magnitude responses, $\{|H^{(g)}(k)|\}_{g\in\mathbb{G}}$, where g denotes a particular filter, the complex symmetric linear-phase FIR filter can be found using a frequency-sampling method as

$$H^{(g)}(k) = \left|H^{(g)}(k)\right| \exp\left(\frac{-i\pi k}{2\Delta k}\right), \tag{35}$$

where $\Delta k \triangleq k_m/m$ is the wavenumber spacing. For the more general case where multiple magnitude responses are designed, their product will result in the complex cascaded filter bank

$$H^{(\mathbb{G})}(k) = \prod_{g\in\mathbb{G}} H^{(g)}(k), \tag{36}$$

where a window can be applied to the time transformed filter impulse response.

The arbitrary magnitude linear-phase FIR cascaded filter bank, $H^{(\mathbb{G})}(k)$, can now be applied to any system input signal, such as the speech, $Y(a,k)$, or the masker, $U(a,k)$, by convolution, e.g.,

$$U^{(\mathbb{G})}(a,k) = U(a,k)H^{(\mathbb{G})}(k), \tag{37}$$

which can then be used in (3) or (11) instead of $Y(a,k)$ or $U(a,k)$, respectively, to synthesise $Q'_l(a,k)$ from (14) via (13) and (12).

VII. Results and Discussion

This section presents objective intelligibility results for the bright and quiet zones in anechoic reproduction environments and discusses the SIC and quality trade-off.

A. Experimental Setup

The geometrical layout of Fig. 1 is evaluated, where $r_{\mathrm{zb}} = r_{\mathrm{zq}} = 0.6\,\mathrm{m}$, $r_{\mathrm{b}} = r_{\mathrm{q}} = 0.3\,\mathrm{m}$, $R = 1.0\,\mathrm{m}$ and $\beta = \wp/3 = 90\,^\circ$. The loudspeaker arrays have $R_{\mathrm{c}} = 1.3\,\mathrm{m}$ and $\phi_{\mathrm{c}} = 180\,^\circ$. The part circle loudspeaker array is an arc which subtends an angle of $\phi_{\mathrm{L}} = 180\,^\circ$. The linear loudspeaker array has a length of $D_{\mathrm{L}} = (L-1)\Delta D_{\mathrm{L}}$ where $\Delta D_{\mathrm{L}} = 12.2\,\mathrm{cm}$ is the spacing between adjacent loudspeakers (designed to match Genelec 8010A loudspeakers). The values of $\theta = \{0\,^\circ, 24.8\,^\circ, 46.1\,^\circ\}$ are used for the angle of the desired plane-wave virtual source in the bright zone for the part circle array and $\theta = \{0\,^\circ, 24.8\,^\circ, 42.7\,^\circ\}$ are for the line array. Using (10), values of θ correspond to $\vartheta = \{-46.1\,^\circ, -24.8\,^\circ, 0\,^\circ\}$ and $\vartheta = \{-42.7\,^\circ, -24.8\,^\circ, 0\,^\circ\}$ for the part circle and line array, respectively. These angles are chosen such that the grating lobe for speech impedes $\mathbb{D}_{\mathrm{q}}$ at the same angle that the maskers grating lobe impedes $\mathbb{D}_{\mathrm{b}}$. The relationship is symmetrical about $\theta = 24.8^\circ$ and three values are chosen.

A pseudo-random selection, constrained to have a male to female speaker ratio of $50:50$, was used to determine Twenty files from the TIMIT corpus [43] for the evaluation. Input speech signals with a sampling frequency of $16\,\mathrm{kHz}$ are framed with $50\,\%$ overlapping $64\,\mathrm{ms}$ windows and transformed using an FFT to the time-frequency domain. The loudspeaker signals, $Q'_l(a,k)$, are synthesised using the methods described in section II and section IV. The number of loudspeakers used for the simulated reproductions are $L = \{16, 24, 32, 114\}$ where, for the cases

in this work, aliasing problems below $8\,\mathrm{kHz}$ are avoided in the reproduction using $L = 114$ for the semi-circular array [14], [24]. For the case when $L = 114$ for a linear array, $\Delta D_{\mathrm{L}} = 3.63\,\mathrm{cm}$ to prevent aliasing below $8\,\mathrm{kHz}$ and the speed of sound is $c = 343\,\mathrm{m\,s^{-1}}$. The noise masker gain levels, G, are varied ranging from $-40\,\mathrm{dB}$ to $20\,\mathrm{dB}$ in (15) for use in (13).

The anechoic reproductions are analysed with $\mathrm{SIC}_{\mathrm{STOI}}$ and $\mathcal{B}_{\mathrm{PESQ}}$ which evaluate the performance using the STOI [35] and Perceptual Evaluation of Speech Quality (PESQ) [44] measures, respectively. Thirty-two receivers are positioned randomly in each zone for recordings which are then analysed. Time-frequency weighted noisy speech, like the simulated recordings in this work, is well suited to the STOI measure. The PESQ measure is a good instrumental measure for quality of speech. The STOI and PESQ are measured in this work for each file and receiver combination using the clean, $y(n)$, and degraded, $p(\mathbf{x}; n)$, speech signals. A spatial average of the quality and intelligibility results over each zone is then performed following (7) and (9).

B. Soundfield Error and Planarity

The accuracy of the reproduced soundfield in $\mathbb{D}_{\mathrm{b}}$ is evaluated using the mean squared error (MSE) as defined in [3] and the planarity measure as defined in [13], [42]. Results for the MSE and planarity in the frequency domain are provided in Fig. 5, where the target angle for the soundfield in the bright zone, θ, is varied from $-30°$ to $55°$. As the target bright zone angle is varied, the masker angle, ϑ, is computed using (10). The results show that the MSE in $\mathbb{D}_{\mathrm{b}}$ is consistently low below the aliasing frequency with an average error of $-30.3\,\mathrm{dB}$ for the semi-circular array and $-30.2\,\mathrm{dB}$ for the linear array. While the MSE increases above the aliasing frequency, it is still significantly low with an average of $-20.9\,\mathrm{dB}$ for the semi-circular array and $-24.0\,\mathrm{dB}$ for the linear array. It is also apparent that the planarity remains consistently high above the aliasing frequency, indicating that the shape of the wave front remains planar as the grating lobes impede $\mathbb{D}_{\mathrm{b}}$. The average planarity in $\mathbb{D}_{\mathrm{b}}$ above the aliasing frequency is 84.3% for the semi-circular array and 88.1% for the linear array. These results indicate that the spatial error is significantly low in the bright zone for a wide range of target bright zone angles when using the proposed methods.

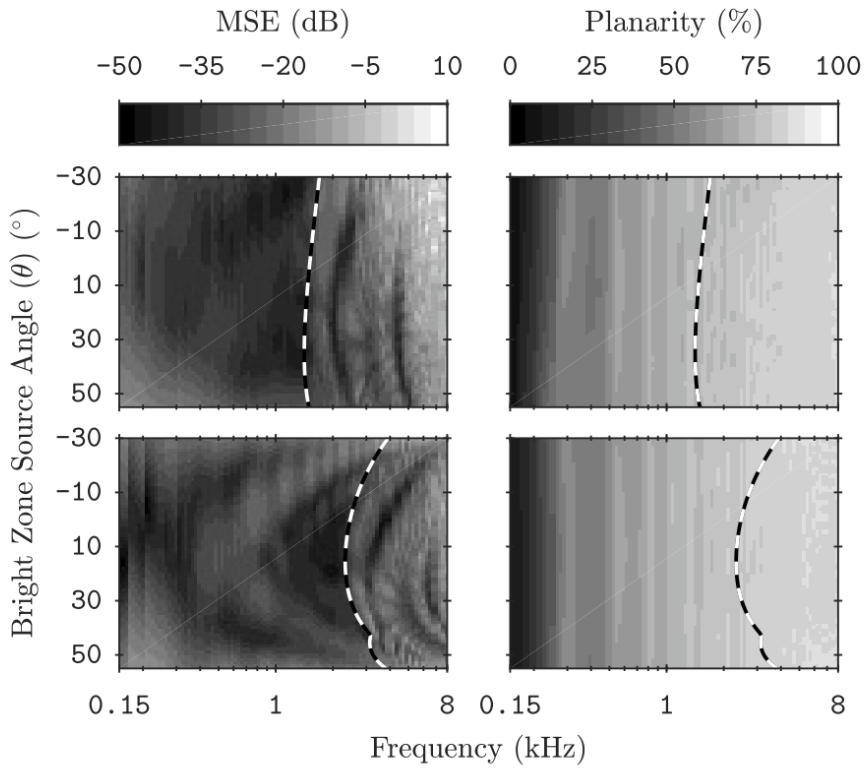


Fig. 5: The MSE and planarity of $\mathbb{D}_{\mathrm{b}}$ in the frequency domain are shown as θ is varied. Results for the semi-circular and linear loudspeaker array are given in the top and bottom rows, respectively, and the MSE and planarity are shown in the left and right column, respectively. As the target bright zone angle, θ, is varied the corresponding masker angle, ϑ, is found with (10). The number of loudspeakers is $L = 24$ and the masker gain is $G = -10\,\mathrm{dB}$. The remainder of the setup is as described in section VII-A. The black and white dashed lines show the aliasing frequency as computed using the methods described in section V.

C. Masker Filtering: Design and Comparison

The filters from (16), (18), (20) and (22) are $H^{(\mathrm{sp})}(k)$, $H^{(\mathrm{q}')}(k)$, $H^{(\mathrm{b}')}(k)$ and $H^{(\mathrm{lp})}(k)$, respectively, which are shown in Fig. 6 (A) along with the intermediate filters, $H^{(\mathrm{q})}(k)$ and $H^{(\mathrm{b})}(k)$, from (17) and (19), respectively. The LTASS is $H^{(\mathrm{sp})}(k)$, the leakage into $\mathbb{D}_{\mathrm{q}}$ is shaped by $H^{(\mathrm{q})}(k)$, the secondary leakage into $\mathbb{D}_{\mathrm{b}}$ is shaped by $H^{(\mathrm{b})}(k)$ and the low pass grating lobe filter is $H^{(\mathrm{lp})}(k)$. Using the experimental setup in section VII-A, a cascaded masker filter bank, $H^{(\mathbb{G})}(k)$, is obtained using (36) with $\mathbb{G} = \{\mathcal{IB}, \mathrm{lp}\}$ and for $\lambda \in \{0.0, 0.5, 1.0\}$. Also shown is the spectrum of the proposed filtered masker in both $\mathbb{D}_{\mathrm{q}}$ (Fig. 6 (B)) and in $\mathbb{D}_{\mathrm{b}}$ (Fig. 6 (C)) for

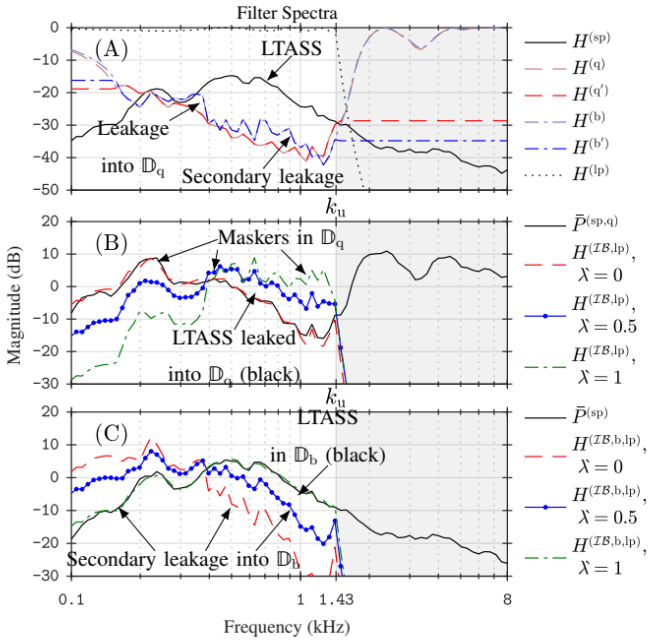


Fig. 6: Example filter spectra are shown. Individual filter responses are displayed in A with comparisons to the average leaked pressure magnitude in $\mathbb{D}_q$ and $\mathbb{D}_b$ shown in B and C, respectively. Descriptive labels are provided for various spectra. Responses are averaged over $1/12$th octave bands. The bandwidth of aliasing above k_u is shaded.

the various λ. The mean LTASS leaked over $\mathbb{D}_q$, denoted in this work as $\bar{P}^{(sp,q)}(k)$ and shown in Fig. 6 (B), is found using (6) and (16) with 32 virtual receivers and responses are averaged over the receiver positions. Similarly the mean LTASS over $\mathbb{D}_b$ is denoted as $\bar{P}^{(sp)}(k)$ and shown in Fig. 6 (C). It can be seen in Fig. 6 that $H^{(\mathbb{G})}(k)$ is a much closer match to the average leaked spectrum in $\mathbb{D}_q$ when $\lambda = 0.0$ and is closer to $\bar{P}^{(sp)}(k)$ when $\lambda = 1.0$. A trade-off between these two results is shown where $\lambda = 0.5$.

To measure the accuracy of the filters with respect to the leaked spectrum to be masked, a symmetrical variant of the Itakura-Saito (IS) [45] distance is used, the COSH spectral distance. The COSH distance used in this work is given by

$$E^{(g)}_{\mathrm{COSH}}(\mathbf{x}) = K^{-1} \sum_{m \in [\![K]\!]} \left(\cosh \left(\ln \frac{\left| H^{(g)}(k_m) \right|}{\breve{P}(\mathbf{x}; k_m)} \right) - 1 \right), \tag{38}$$

where $H^{(g)}(k)$ is the filter to be measured, $\breve{P}(\mathbf{x}; k)$ is the pressure spectrum at $\mathbf{x}$ and $E^{(g)}_{\mathrm{COSH}}(\mathbf{x})$ is the COSH distance for all K frequencies. To evaluate the leaked spectrum, the mean COSH

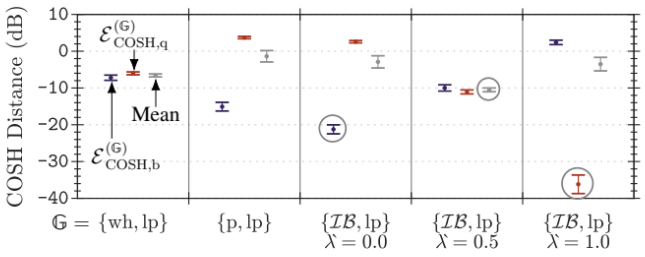


Fig. 7: Mean COSH distances, $\mathcal{E}^{(g)}_{\mathrm{COSH,z}}$, for different noise maskers and zones are shown. The columns indicate the different noise maskers and each column contains three values which are $\mathcal{E}^{(\mathbb{G})}_{\mathrm{COSH,b}}$ (left), $\mathcal{E}^{(\mathbb{G})}_{\mathrm{COSH,q}}$ (middle) and the mean of both zones (right). COSH distance values are given in decibels and the smallest distance for each set is circled. The 95% confidence intervals shown are calculated over the area of each zone.

distance over some zone, $\mathbb{D}_{\mathrm{z}}$, of size $\mathfrak{d}_{\mathrm{z}}$ for $\mathrm{z} \in \{\mathrm{b}, \mathrm{q}\}$, is found as

$$\mathcal{E}^{(g)}_{\mathrm{COSH,z}} = \mathfrak{d}_{\mathrm{z}}^{-1} \int_{\mathbb{D}_{\mathrm{z}}} E^{(g)}_{\mathrm{COSH}}(\mathbf{x})\, d\mathbf{x}. \tag{39}$$

The values in Fig. 7 given by (39) show that the proposed cascaded filter, $\{\mathcal{IB}, \mathrm{lp}\}$, provides a masker spectrum with the least mean distance to the spectrum of the speech in $\mathbb{D}_{\mathrm{b}}$ and leaked speech in $\mathbb{D}_{\mathrm{q}}$ with $\lambda = 0.5$ at $-10.5\,\mathrm{dB}$ when compared to white noise ($\{\mathrm{wh}, \mathrm{lp}\}$), pink noise ($\{\mathrm{p}, \mathrm{lp}\}$), $\lambda = 0.0$ and $\lambda = 1.0$.

D. Speech Privacy Results

A descriptive comparison of the effectiveness and robustness of the methods outlined throughout this paper is presented in this subsection. Results for instrumentally measured intelligibility and quality are given so the reader may intuitively interpret the relationships between noise masking, quality and privacy. The robustness of the methods is conveyed through consistent results when varying the target bright zone virtual source angle, the array geometry and the number of available loudspeakers. The varying effectiveness of the methods is shown via results for different masking spectra and spectrum weighting parameters.

Figure 8, Fig. 9 and Fig. 12 all show results for the semi-circular and linear array in the left and right column, respectively. The figures all include variation in θ, microphone positions and speech in the 95% confidence intervals. Variation in the spectrum weight and shaping as determined by λ is shown along the rows of Fig. 8 with white noise in the first row for comparison. Variation in

the loudspeaker count, L, is shown along the rows of Fig. 9. A discussion on the aforementioned variables is given in the following sub-subsections.

1) Angle: While consistently applying spatial weighting to all, or part, of the reproduction it is still natural for the acoustical brightness contrast performance to vary depending on θ. Figure 8 contains the variations due to the different θ in its confidence intervals which are still considerably small and show the method's robustness to variance in θ.

2) Spectrum Shape and Weighting: While Fig. 8 (A, B) show a good separation between the two $\mathcal{I}_{\mathrm{STOI}}$ results, the wideband white masker (without the grating lobe filter, $\{\mathrm{lp}\}$) that is used still keeps the $\mathcal{B}_{\mathrm{PESQ}}$ low in the region where $\mathrm{SIC}_{\mathrm{STOI}}$ is high. To allow for both high valued $\mathcal{B}_{\mathrm{PESQ}}$ and $\mathrm{SIC}_{\mathrm{STOI}}$, the spectrum is shaped and the results in Fig. 8 (C–H) show how $\mathcal{B}_{\mathrm{PESQ}}$ and $\mathrm{SIC}_{\mathrm{STOI}}$ can be tuned with the parameter λ. The hypothesis that low valued λ improves masking performance over $\mathbb{D}_{\mathrm{q}}$ to increase $\mathrm{SIC}_{\mathrm{STOI}}$ and high valued λ reduces masking effects over $\mathbb{D}_{\mathrm{b}}$ to increase $\mathcal{B}_{\mathrm{PESQ}}$ is confirmed in Fig. 8. The case where $\lambda = 0.5$ gives on average the best separation between the two $\mathcal{I}_{\mathrm{STOI}}$ results whilst maintaining a high valued $\mathcal{B}_{\mathrm{PESQ}}$. For cases where $\mathrm{SIC}_{\mathrm{STOI}}$ is required to be high and $\mathcal{B}_{\mathrm{PESQ}}$ is of less importance, $\lambda = 1.0$ may sometimes provide slightly better results than $\lambda = 0.5$, as can be seen in Fig. 8 (G).

3) Array Geometry: The two different array geometries evaluated are the semi-circular array and linear array where results are shown in the first and second column, respectively, in Fig. 8 and Fig. 9. The main observable difference is that the linear array provides slightly less contrast between the two $\mathcal{I}_{\mathrm{STOI}}$ and therefore a slightly smaller range of high valued $\mathrm{SIC}_{\mathrm{STOI}}$. This difference in contrast has more influence on the resulting $\mathcal{B}_{\mathrm{PESQ}}$ in Fig. 8, however, it is still possible to obtain high valued $\mathcal{B}_{\mathrm{PESQ}}$ and high valued $\mathrm{SIC}_{\mathrm{STOI}}$. It should be noted that the loudspeaker spacing, ΔD_{L}, is constant for all results in Fig. 8. To investigate the effect of differing L, the loudspeaker spacing is varied for results in Fig. 9 which show that better performance is acquired for smaller values of ΔD_{L} and for a larger number of loudspeakers, L. The semi-circular array performs better than the linear array with the same ΔD_{L} which is caused by the fact that the semi-circular array has a higher low-frequency acoustical brightness contrast between $\mathbb{D}_{\mathrm{b}}$ and $\mathbb{D}_{\mathrm{q}}$. The higher contrast here is a result of the apparent angular window of the array to the multiple zones. However, the linear array does have a slightly higher k_{u} compared to the semi-circular array when L and ΔD_{L} are the same between array geometries. The linear arrays higher k_{u} does not provide a better $\mathrm{SIC}_{\mathrm{STOI}}$ or $\mathcal{B}_{\mathrm{PESQ}}$, though, because the loss in low

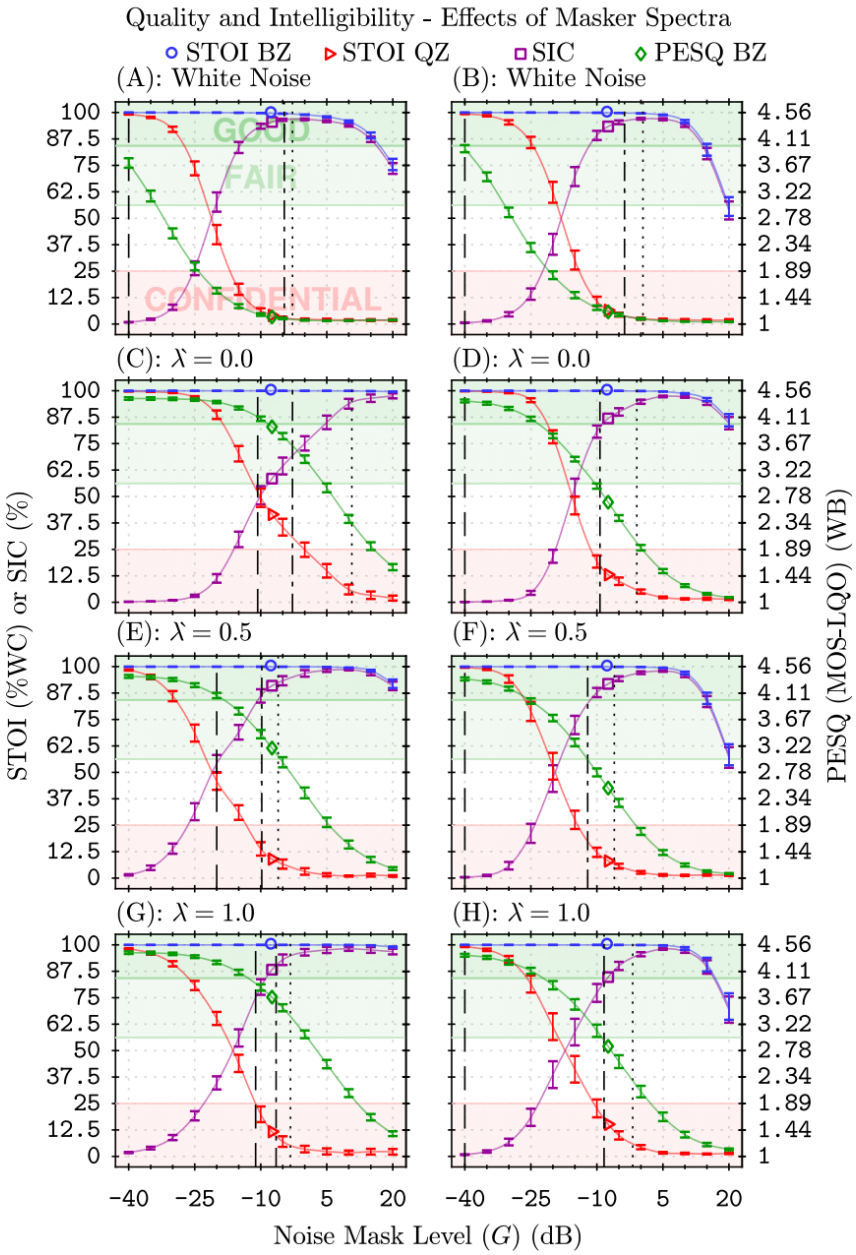


Fig. 8: Mean STOI and PESQ are shown for different masking spectra and different array types with $L = 24$. A and B are for a white noise, C and D are $\lambda = 0.0$, E and F are $\lambda = 0.5$ and G and H are $\lambda = 1.0$. The left column is for semi-circular array reproductions and the right column is for linear array reproductions. Optimum G (dB) is indicated by the vertical black dotted lines for $\lambda = 0.33$, dash-dot lines for $\lambda = 1.0$ and dashed lines for $\lambda = 3.0$. Good and fair PESQ MOS scores [44] are labelled and shaded in green and confidential speech privacy [33] is labelled and shaded in red. BZ and QZ are the bright and quiet zone, respectively. 95% confidence intervals over θ, microphone positions and speech variation are given.

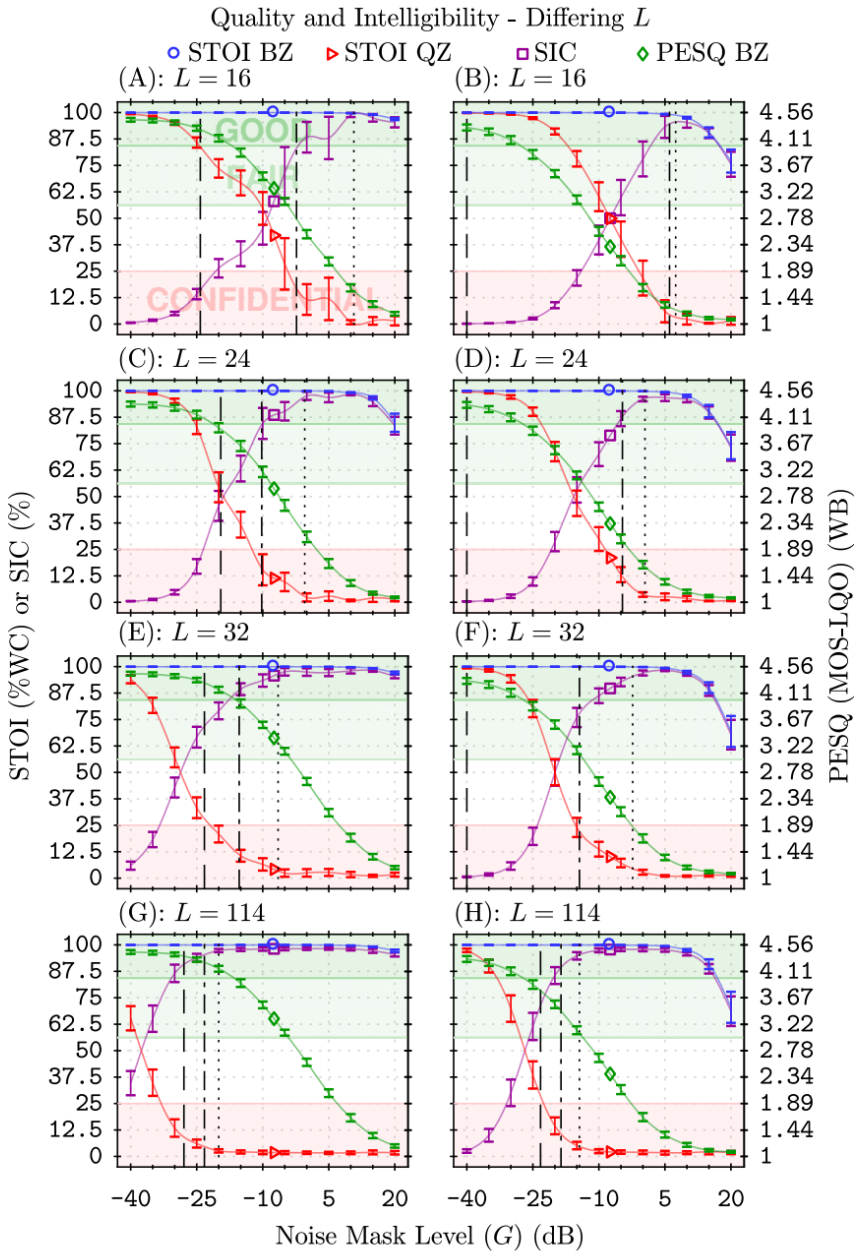


Fig. 9: Mean STOI and PESQ are shown for different L and different array types with $\lambda = 0.5$. Each loudspeaker count is presented in a row where A and B are $L = 16$, C and D are $L = 24$, E and F are $L = 32$ and G and H are $L = 114$. The left column is for semi-circular array reproductions and the right column is for linear array reproductions where $D_{\mathrm{L}} = \phi_{\mathrm{c}} R_{\mathrm{c}}$. Optimum G (dB) is indicated by the vertical black dotted lines for $\lambda = 0.33$, dash-dot lines for $\lambda = 1.0$ and dashed lines for $\lambda = 3.0$. Good and fair PESQ MOS scores [44] are labelled and shaded in green and confidential speech privacy [33] is labelled and shaded in red. BZ and QZ are the bright and quiet zone, respectively. 95% confidence intervals over θ, microphone positions and speech variation are given.

frequency contrast for the linear array reduces these values more so, primarily due to the high energy speech content at low frequencies and the only slightly larger k_{u}.

4) Loudspeaker Count: The loudspeaker count, L, and, more specifically, the loudspeaker spacing, ΔD_{L}, have a large influence on both the performance and the practicality of the system. The influence on performance is shown in Fig. 9 where as the loudspeaker count increases for a semi-circular array (and hence the speaker spacing decreases) the separation between the two $\mathcal{I}_{\mathrm{STOI}}$ results increases, as does $\mathcal{B}_{\mathrm{PESQ}}$. The minimum number of loudspeakers in the semi-circular array which still attains good $\mathcal{B}_{\mathrm{PESQ}}$ and $\mathrm{SIC}_{\mathrm{STOI}}$ is the case where $L = 24$, which is good motivation for the number of real-world loudspeakers to use. As the linear array may either use a differing number of loudspeakers with a fixed ΔD_{L} or with a fixed D_{L}, in this work Fig. 9 presents results for a fixed $D_{\mathrm{L}} = \phi_{\mathrm{c}} R_{\mathrm{c}}$ as this maintains a constant valued k_{u}, consistent with the semi-circular array for direct comparison. Results related to a potentially more practical scenario, where ΔD_{L} is fixed, and proportional to the dimensions of a smaller real-world loudspeaker, the reader is referred to Fig. 8. Simulations for varying L with the linear array follow the same trend as those for the semi-circular array where, as L increases, both $\mathcal{B}_{\mathrm{PESQ}}$ and $\mathrm{SIC}_{\mathrm{STOI}}$ also increase.

VIII. Real-World Implementation

To compliment simulations, a practical real-world implementation has been evaluated in anechoic conditions. This section provides details of the hardware, calibration and recorded results.

A. Hardware Setup

The multizone audio reproduction systems described in section VII-A were implemented in a flat-walled multilayered anechoic chamber measuring $4.8\,\mathrm{m} \times 3.3\,\mathrm{m} \times 2.4\,\mathrm{m}$. The systems consisted of 24 loudspeakers evenly spaced on a semi-circle of radius $1.3\,\mathrm{m}$ and a line of length $2.8\,\mathrm{m}$ as shown in Fig. 10. Recordings of the reproduced speech were received using $4 \times$ Behringer ECM8000 measurement microphones in each zone, positioned equidistant along a $0.3\,\mathrm{m}$ diameter circle (concentric with the zone). The loudspeaker models were all Genelec 8010A studio monitors with a free field frequency response of $74\,\mathrm{Hz}$ to $20\,\mathrm{kHz}$ ($\pm 2.5\,\mathrm{dB}$). The loudspeakers and microphones were driven by $3 \times$ Behringer ADA8200 8-channel input/output

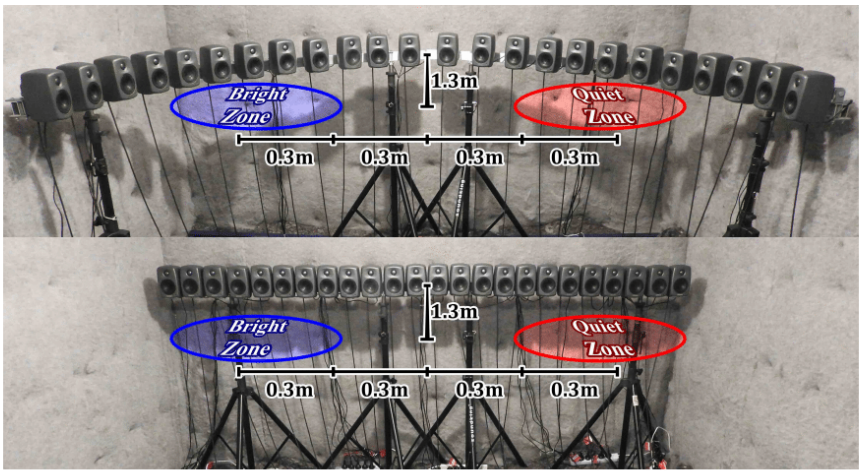


Fig. 10: The two real-world multizone implementations are pictured. The semi-circular and linear array are shown on the top and bottom, respectively. The bright zone (blue) on the left and the quiet zone (red) on the right are separated by $1.2\,\mathrm{m}$ and each have a radius of $0.3\,\mathrm{m}$. The centre of the reproduction region is midway between both zones and is $1.3\,\mathrm{m}$ from the centre of the loudspeaker array.

audio interfaces connected to a computer via an RME HDSPe RayDAT 36-channel input/output soundcard. The software used to generate, playback and record the multizone soundfield was Mathworks' MATLAB R2017a.

B. System Calibration and Response

In order to ensure a flat magnitude response and correct phase response for all loudspeakers, a calibration procedure is performed. The calibration is the application of system equalisation filters computed from inverse system transfer functions found by using an exponential sine sweep (ESS) method[6] [46]. Prior to applying the inverse filters, the loudspeaker signals, ($q_l'(n)$ from $Q_l'(a,k)$) are upsampled by interpolating with a factor of 3 from $16\,\mathrm{kHz}$ to $48\,\mathrm{kHz}$, to match that of the reproduction system due to the sampling frequency mismatch. The band-pass inverse filters are then convolved with the upsampled loudspeaker signals. Soundfield recordings are performed using the upsampled calibrated loudspeaker signals and in order to compare with simulated recordings, the $48\,\mathrm{kHz}$ sampled recordings are downsampled to $16\,\mathrm{kHz}$ by a factor of

3 with decimation.

C. Simulated and Real-World Comparison

To confirm that the calibration procedure allows for a flat magnitude response in the target bright zone, within the accuracy of the loudspeakers (i.e. $\pm 2.5\,\mathrm{dB}$), the response over $\mathbb{D}_\mathrm{b}$ and $\mathbb{D}_\mathrm{q}$ is measured by reproducing and recording a multizone weighted ESS. Afterwards, the $\mathrm{SIC}_\mathrm{STOI}$ and $\mathcal{B}_\mathrm{PESQ}$ are computed and compared with simulated results using speech samples and measured ATFs.

1) Sound Pressure Levels: The SPL is found for $\theta = 24.8°$ and results do not vary significantly for different values of θ (as explained in VII-D1). Figure 11 shows that the real-world multizone magnitude response over $\mathbb{D}_\mathrm{b}$ is flat and lies within $\pm 2.5\,\mathrm{dB}$, even after the signal has been processed and other system noises have been included. The real-world SPL over $\mathbb{D}_\mathrm{q}$ also agrees with simulated SPL over $\mathbb{D}_\mathrm{q}$ with only slight variations when using measured ATFs as shown in Fig. 11. The average SPL up to $\min(\hat{k}'_\mathrm{u}, \bar{k}_\mathrm{u})$ over $\mathbb{D}_\mathrm{q}$ for the real-world scenario is considerably low at $-25.5\,\mathrm{dB}$ for the semi-circular array and $-24.9\,\mathrm{dB}$ for the linear array. The equivalent acoustic brightness contrast, following [9], [10], between $\mathbb{D}_\mathrm{b}$ and $\mathbb{D}_\mathrm{q}$ for the real-world scenario is $25.6\,\mathrm{dB}$ for the semi-circular array and $25.0\,\mathrm{dB}$ for the linear array.

2) Speech Intelligibility Contrast and Quality: The $\mathcal{I}_\mathrm{STOI}$ and $\mathcal{B}_\mathrm{PESQ}$ in Fig. 12 are seen to be almost identical between the real-world and simulated results. Figure 11 suggests this would likely be the case.

For the real-world case using a semi-circular array, $\lambda = 0.5$ and $\lambda = 0.33$ gives optimal $G = -3.26\,\mathrm{dB}$, the results obtained are $\mathrm{SIC}_\mathrm{STOI} = 96.4\,\%$ and $\mathcal{B}_\mathrm{PESQ} = 2.52\,\mathrm{MOS}$ indicating confidential speech privacy and better than poor speech quality, respectively. $\lambda = 1.0$ gives $G = -9.77\,\mathrm{dB}$, $\mathrm{SIC}_\mathrm{STOI} = 85.9\,\%$ and $\mathcal{B}_\mathrm{PESQ} = 3.22\,\mathrm{MOS}$ indicating confidential privacy and better than fair quality, respectively, and $\lambda = 3.0$ gives $G = -19.1\,\mathrm{dB}$, $\mathrm{SIC}_\mathrm{STOI} = 50.0\,\%$ and $\mathcal{B}_\mathrm{PESQ} = 3.92\,\mathrm{MOS}$ indicating normal privacy and better than fair quality (close to good quality), respectively. The results show that λ successfully controls the trade-off between speech privacy

[6]The ESS is generated as a $10\,\mathrm{s}$ sweep from $100\,\mathrm{Hz}$ to $10\,\mathrm{kHz}$ with a $1\,\mathrm{s}$ buffer of silence before and after. The system is set to a sampling frequency of $48\,\mathrm{kHz}$ after which the ESS is reproduced one loudspeaker at a time and recorded from the centre of $\mathbb{D}$. The calibration filters are computed from the recordings with a length of $0.5\,\mathrm{s}$ and are regularised so that the maximum pass-band gain is $60\,\mathrm{dB}$ and stop-band gain is $-6\,\mathrm{dB}$.

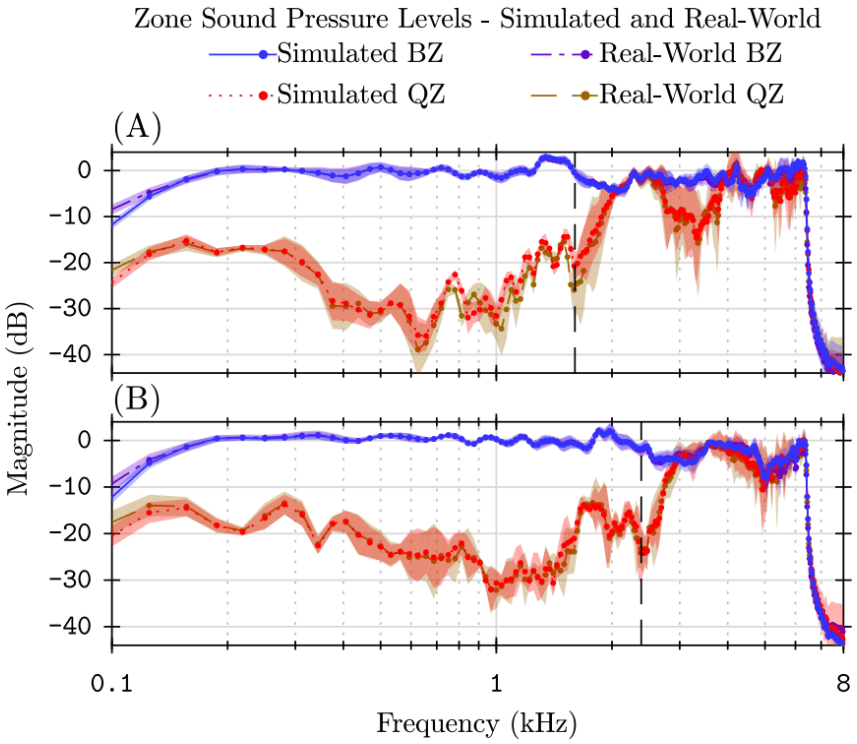


Fig. 11: Mean SPLs are shown for the simulated and real-world cases with $L = 24$ for a semi-circular array (A) and linear array (B) where $\theta = 24.8°$. 95 % confidence intervals over the microphone positions in each zone are shaded and the vertical black dashed line is k_{u}. BZ and QZ are the bright and quiet zone, respectively.

and speech quality where a lower value λ emphasises privacy and higher valued λ emphasises quality. The results obtained are either as good or better than those reported in [8], however, in the case of this work, the results are obtained with significantly fewer loudspeakers ($L = 24$ (8.14%) instead of $L = 295$) and with the use of noisy real-world equipment.

For the real-world case using a linear array, $\lambda = 0.5$ and $\lambda = 0.33$ gives optimal $G = -3.72\,\mathrm{dB}$, the results obtained are $\mathrm{SIC}_{\mathrm{STOI}} = 95.5\,\%$ and $\mathcal{B}_{\mathrm{PESQ}} = 2.17\,\mathrm{MOS}$ indicating confidential privacy and better than poor quality, respectively. $\lambda = 1.0$ gives $G = -13.5\,\mathrm{dB}$, $\mathrm{SIC}_{\mathrm{STOI}} = 79.7\,\%$ and $\mathcal{B}_{\mathrm{PESQ}} = 3.21\,\mathrm{MOS}$ indicating confidential privacy and better than fair quality, respectively, and when $\lambda = 3.0$ gives $G = -18.6\,\mathrm{dB}$, $\mathrm{SIC}_{\mathrm{STOI}} = 56.6\,\%$ and $\mathcal{B}_{\mathrm{PESQ}} = 3.64\,\mathrm{MOS}$ indicating normal privacy and better than fair quality, respectively. These results show that the real-world linear array performs just as well as the real-world semi-circular

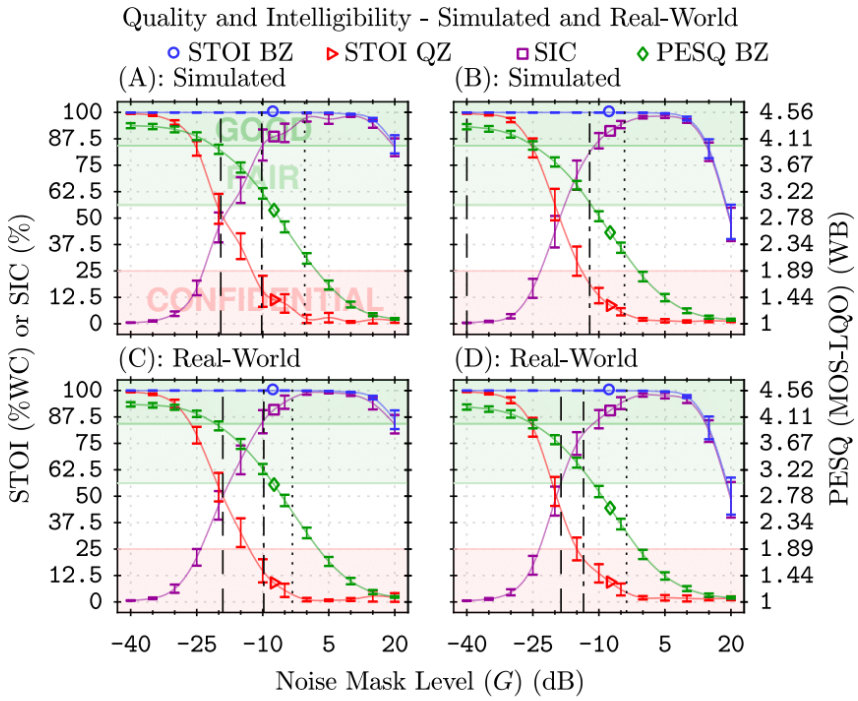


Fig. 12: Mean STOI and PESQ are shown for the simulated (A–B) and real-world (C–D) anechoic environment with $\theta = 24.8°$, $\lambda = 0.5$. The left column is for semi-circular array reproductions and the right column is for linear array reproductions where $D_{\mathrm{L}} = \phi_c R_c$. Optimum G (dB) is indicated by the vertical black dotted lines for $\lambda = 0.33$, dash-dot lines for $\lambda = 1.0$ and dashed lines for $\lambda = 3.0$. Good and fair PESQ MOS scores [44] are labelled and shaded in green and confidential speech privacy [33] is labelled and shaded in red. BZ and QZ are the bright and quiet zone, respectively. 95% confidence intervals over θ, microphone positions and speech variation are given.

array and that λ still successfully controls the trade-off between speech privacy and speech quality. This is fortuitous as a linear array is a more practical implementation for box-shaped rooms.

IX. Conclusion

We proposed a method for improving the speech privacy and quality in multizone soundfield reproductions by using robust spatial and temporal frequency domain filters on masking signals. Practical implementations are facilitated by the proposed methods; masking filters are analytically derived in order to avoid spatial aliasing artefacts and secondary leakage is accounted for using weighting parameters on *a priori* estimates of multizone spectral leakage. The practical benefits

include robustness to variations in the reproduced speech, virtual source location and array geometry, and a significantly reduced number of the required loudspeakers.

Results have shown that it is necessary to account for multizone leakage when performing masking or when high quality reproductions are required. It is also shown that estimating the aliasing frequency is of importance when the loudspeaker count and geometry can vary. A more robust estimation of the aliasing frequency has also been shown to provide more reliable results. System performance is also dependent on the acoustic contrast between the zones which may vary depending on the reproduction technique used and the real-world equipment setup and calibration. The results presented verify the benefits of the proposed method for practical implementations. The analytically derived filters and optimal gains are shown capable of providing good and fair MOS ratings for speech quality whilst providing normal and confidential privacy, respectively, via measured SIC values in simulated environments. The real-world implementation, and the results thereof, confirm the practicality of the proposed methods by also showing that good and fair speech quality, with respective normal and confidential speech privacy, can be reproduced amongst personal sound zones.

Future work could include investigations on the perceived annoyance of different sound maskers and their influence on cognitive performance. Evaluations of simultaneous reproductions of speech in multiple zones and the effect of joint optimisations using temporal and spatial filters are also potential topics for future work.

APPENDIX A
CIRCULAR ARRAY AUXILIARY VALUES

The rotated grating lobe vector, $\overrightarrow{\mathbf{pq}}$, points from the circular array grating lobe origin, $\hat{\mathbf{p}}$ (at angle α and radius R_c), to the quiet zone origin, $\mathbf{q}$ (at angle φ and radius r_zq), and is given by

$$\hat{\mathbf{p}} = R_\mathrm{c} \cdot \mathring{\mathbf{R}}(\alpha) \cdot \hat{\mathbf{u}}_\mathrm{o}, \tag{40}$$

$$\mathbf{q} = r_\mathrm{zq} \cdot \mathring{\mathbf{R}}(\varphi) \cdot \hat{\mathbf{u}}_\mathrm{o}, \tag{41}$$

$$\overrightarrow{\mathbf{pq}} = \mathbf{q} - \hat{\mathbf{p}}, \tag{42}$$

where $\hat{\mathbf{u}}_{\mathrm{o}}$ is a unit column vector at the origin and

$$\mathring{\mathbf{R}}(\delta) \triangleq \begin{bmatrix} \cos(\delta) & -\sin(\delta) & 0 \\ \sin(\delta) & \cos(\delta) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{43}$$

is a rotational matrix for a given angle δ.

APPENDIX B
LINEAR ARRAY AUXILIARY VALUES

The centre point of $\mathbb{D}_{\mathrm{b}}$ and the loudspeaker array are

$$\mathbf{b} = r_{\mathrm{zb}} \cdot \mathring{\mathbf{R}}(\beta) \cdot \hat{\mathbf{u}}_{\mathrm{o}} \quad \text{and} \tag{44}$$

$$\mathbf{c} = R_{\mathrm{c}} \cdot \mathring{\mathbf{R}}(\phi_{\mathrm{c}}) \cdot \hat{\mathbf{u}}_{\mathrm{o}}, \tag{45}$$

respectively, and the solution for the intersection is

$$\mathbf{a}_1 = \begin{bmatrix} \cos(\theta) & \sin(\theta) & 0 \end{bmatrix}^{\mathsf{T}}, \tag{46}$$

$$\mathbf{a}_2 = \begin{bmatrix} \cos\big(\phi_{\mathrm{c}} - \frac{\pi}{2}\big) & \sin\big(\phi_{\mathrm{c}} - \frac{\pi}{2}\big) & 0 \end{bmatrix}^{\mathsf{T}}, \tag{47}$$

$$\begin{bmatrix} s_1 & s_2 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} \mathbf{a}_1 & -\mathbf{a}_2 \end{bmatrix}^{\dagger} \cdot (\mathbf{c} - \mathbf{b}), \tag{48}$$

where † denotes a Moore-Penrose pseudoinverse. The intersecting point for the linear array is then

$$\bar{\mathbf{p}} = \mathbf{b} + s_1 \mathbf{a}_1 = \mathbf{c} + s_2 \mathbf{a}_2. \tag{49}$$

Inserting $\bar{\mathbf{p}}$ in replacement of $\hat{\mathbf{p}}$ in (42) yields a new $\overrightarrow{\mathbf{pq}}$ for a linear array and $\overrightarrow{\mathbf{pb}} = \mathbf{b} - \bar{\mathbf{p}}$.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers whose comments have been invaluable to the content and overall presentation of the paper.

REFERENCES

[1] T. Betlehem *et al.*, "Personal Sound Zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 81–91, Mar. 2015, doi: 10.1109/MSP.2014.2360707.

[2] K. Baykaner *et al.*, "The relationship between target quality and interference in sound zone," *J. Audio Eng. Soc.*, vol. 63, no. 1/2, pp. 78–89, Jan. 2015, doi: 10.17743/jaes.2015.0007.

[3] W. Jin and W. B. Kleijn, "Theory and design of multizone soundfield reproduction using sparse methods," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2343–2355, Dec. 2015, doi: 10.1109/TASLP.2015.2479037.

[4] H. Chen *et al.*, "In-car noise field analysis and multi-zone noise cancellation quality estimation," in *Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, IEEE, 2015, pp. 773–778.

[5] P. N. Samarasinghe *et al.*, "Recent advances in active noise control inside automobile cabins: Toward quieter cars," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 61–73, Nov. 2016, doi: 10.1109/MSP.2016.2601942.

[6] L. Ward *et al.*, "The effect of situation-specific non-speech acoustic cues on the intelligibility of speech in noise," in *Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017.

[7] W. Zhang *et al.*, "Analysis and control of multi-zone sound field reproduction using modal-domain approach," *J. Acoust. Soc. Am.*, vol. 140, no. 3, pp. 2134–2144, Sep. 2016, doi: 10.1121/1.4963084.

[8] J. Donley *et al.*, "Improving speech privacy in personal sound zones," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2016, pp. 311–315.

[9] J.-W. Choi and Y.-H. Kim, "Generation of an acoustically bright zone with an illuminated region using multiple sources," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1695–1700, Apr. 2002, doi: 10.1121/1.1456926.

[10] J.-H. Chang *et al.*, "A realization of sound focused personal audio system using acoustic contrast control," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2091–2097, Apr. 2009, doi: 10.1121/1.3082114.

[11] M. Shin *et al.*, "Maximization of acoustic energy difference between two spaces," *J. Acoust. Soc. of Am.*, vol. 128, no. 1, p. 121, Jul. 2010, doi: 10.1121/1.3438479.

[12] W. Zhang *et al.*, "Surround by Sound: A Review of Spatial Audio Recording and Reproduction," *Appl. Sciences*, vol. 7, no. 6, p. 532, May 2017, doi: 10.3390/app7050532.

[13] P. Coleman *et al.*, "Personal audio with a planar bright zone," *J. Acoust. Soc. Am.*, vol. 136, no. 4, pp. 1725–1735, Oct. 2014, doi: 10.1121/1.4893909.

[14] W. Jin *et al.*, "Multizone soundfield reproduction using orthogonal basis expansion," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2013, pp. 311–315.

[15] H. Chen *et al.*, "Enhanced sound field reproduction within prioritized control region," in *INTER-NOISE and NOISE-CON Congr. and Conf. Proc.*, vol. 249, Inst. of Noise Control Eng., 2014, pp. 4055–4064.

[16] J. Donley and C. Ritz, "Multizone reproduction of speech soundfields: A perceptually weighted approach," in *Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, IEEE, 2015, pp. 342–345.

[17] N. Radmanesh and I. S. Burnett, "Generation of isolated wideband sound fields using a combined two-stage lasso-LS algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 378–387, Feb. 2013, doi: 10.1109/TASL.2012.2227736.

[18] J. Donley and C. Ritz, "An efficient approach to dynamically weighted multizone wideband reproduction of speech soundfields," in *China Summit Int. Conf. Signal Inform. Process. (ChinaSIP)*, IEEE, Jul. 2015, pp. 60–64.

[19] J. S. Bradley and B. N. Gover, "A new system of speech privacy criteria in terms of Speech Privacy Class (SPC) values," pp. 1–5, 2010.

[20] B. N. Gover and J. S. Bradley, "ASTM metrics for rating speech privacy of closed rooms and open plan spaces," *Canadian Acoust.*, vol. 39, pp. 50–51, 2011.

[21] J. Francombe *et al.*, "Determining the threshold of acceptability for an interfering audio programme," in *Audio Eng. Soc. Conv. 132*, Audio Eng. Soc., 2012.

[22] W. B. Kleijn *et al.*, "Optimizing speech intelligibility in a noisy environment: A unified view," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 43–54, Mar. 2015, doi: 10.1109/MSP.2014.2365594.

[23] F. Winter *et al.*, "On Analytic Methods for 2.5-D Local Sound Field Synthesis Using Circular Distributions of Secondary Sources," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 914–926, May 2016, doi: 10.1109/TASLP.2016.2531902.

[24] Y. J. Wu and T. D. Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1711–1720, Aug. 2011, doi: 10.1109/TASL.2010.2097249.

[25] P. Coleman *et al.*, "Numerical optimization of loudspeaker configuration for sound zone reproduction," in *Int. Congr. Sound and Vibration*, IIAV, 2014, pp. 1–8.

[26] T. Okamoto, "Generation of multiple sound zones by spatial filtering in wavenumber domain using a linear array of loudspeakers," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2014, pp. 4733–4737.

[27] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.

[28] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, Nov. 2005.

[29] P. N. Samarasinghe *et al.*, "3D soundfield reproduction using higher order loudspeakers," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2013, pp. 306–310.

[30] M.-f. Zha *et al.*, "3D multizone soundfield reproduction using spherical harmonic analysis," in *China Summit Int. Conf. Signal Inform. Process. (ChinaSIP)*, IEEE, Jul. 2015, pp. 625–629.

[31] Y. J. Wu and T. D. Abhayapala, "Theory and design of soundfield reproduction using continuous loudspeaker concept," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 107–116, Jan. 2009, doi: 10.1109/TASL.2008.2005340.

[32] *Standard test method for objective measurement of the speech privacy provided by a closed room*. ASTM Int. E2638-10, 2010.

[33] *Standard test method for objective measurement of speech privacy in open plan spaces using articulation index*. ASTM Int. E1130-08, 2008.

[34] *Sound system equipment-Part 16: Objective rating of speech intelligibility by speech transmission index*. IEC 60268-16, 2003.

[35] C. H. Taal *et al.*, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011, doi: 10.1109/TASL.2011.2114881.

[36] D. Byrne *et al.*, "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2108–2120, Oct. 1994, doi: 10.1121/1.410152.

[37] *Artificial voices*. Int. Telecommun. Union (ITU), ITU-T Rec. P.50, 1999.

[38] T. W. Parks and C. S. Burrus, *Digital Filter Design*. New York, NY, USA: John Wiley & Sons, 1987.

[39] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 697–707, 2001, doi: 10.1109/89.943347.

[40] J. Donley *et al.*, "Reproducing personal sound zones using a hybrid synthesis of dynamic and parametric loudspeakers," in *Asia-Pacific Signal & Inform. Process. Assoc. Annu. Summit and Conf. (APSIPA ASC)*, IEEE, Dec. 2016, pp. 1–5.

[41] Y.-H. Kim and J.-W. Choi, *Sound Visualization and Manipulation*. Singapore: John Wiley & Sons Singapore Pte. Ltd., Sep. 2013.

[42] P. Coleman *et al.*, "Acoustic contrast, planarity and robustness of sound zone methods using a circular loudspeaker array," *J. Acoust. Soc. Am.*, vol. 135, no. 4, pp. 1929–1940, Apr. 2014, doi: 10.1121/1.4866442.

[43] J. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

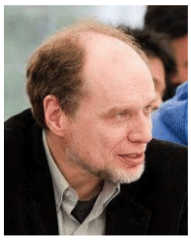
[44] A. W. Rix *et al.*, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2001, pp. 749–752.

[45] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Int. Congr. Acoust.*, Tokyo, Japan, 1968, pp. 17–20.

[46] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio Eng. Soc. Conv. 122*, Audio Eng. Soc., 2007.

Jacob Donley (S'13–GSM'14) received the B.Eng (Hons1) degree in computer engineering from the University of Wollongong, Australia, in 2014 and is currently pursuing the Ph.D. degree in signal processing at the University of Wollongong. He has received scholarships from Telstra Corporation Limited, the Department of Education and Training, and the University of Wollongong in 2010, 2014 and 2014, respectively. Jacob has worked as an image analyst and software engineer at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in 2013/14 and has been a lecturer at the University of Wollongong and Western Sydney University, Australia, in 2016 and 2017, respectively. His research interests are in the areas of multichannel audio (recording, rendering and reproduction), spatial audio, speech, psychoacoustics, active noise control, room acoustics and signal processing.

Christian Ritz (M'97–SM'08) received his B.E. degree in electrical engineering and B.Math degree both from the University of Wollongong, Wollongong, Australia, in 1998. He received his Ph.D. degree in 2003 from the University of Wollongong, Wollongong, Australia. He joined the University of Wollongong in 2003 and is currently an Associate Professor there. His current research interests include spatial audio signal processing, multichannel speech signal processing and multimedia signal processing.



W. Bastiaan Kleijn (M'88–SM'97–F'99) received the Ph.D. degree in electrical engineering from Delft University of Technology, The Netherlands (TU Delft); an M.S.E.E. degree from Stanford University; and a Ph.D. degree in soil science and an M.Sc. degree in physics from the University of California, Riverside. He is a Professor at Victoria University of Wellington, New Zealand, and Technical University of Delft, The Netherlands (part-time). He was a Professor and Head of the Sound and Image Processing Laboratory at The Royal Institute of Technology (KTH), Stockholm, Sweden, from 1996 until 2010 and a founder of Global IP Solutions, a company that provided the original audio technology to Skype and was later acquired by Google. Prior to 1996, he was with the Research Division of AT&T Bell Laboratories in Murray Hill, New Jersey. He is an IEEE Fellow.