IMPROVING SPEECH PRIVACY IN PERSONAL SOUND ZONES

Jacob Donley[⋆]*, Christian Ritz*[⋆] *and W. Bastiaan Kleijn*[†]

[⋆] School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Australia
[†] School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

ABSTRACT

This paper proposes two methods for providing speech privacy between spatial zones in anechoic and reverberant environments. The methods are based on masking the content leaked between regions. The masking is optimised to maximise the speech intelligibility contrast (SIC) between the zones. The first method uses a uniform masker signal that is combined with desired multizone loudspeaker signals and requires acoustic contrast between zones. The second method computes a space-time domain masker signal in parallel with the loudspeaker signals so that the combination of the two emphasises the spectral masking in the targeted quiet zone. Simulations show that it is possible to achieve a significant SIC in anechoic environments whilst maintaining speech quality in the bright zone.

Index Terms— multizone soundfield reproduction, personal sound zones, speech privacy, speech intelligibility

1. INTRODUCTION

Using an array of loudspeakers, multizone soundfield reproduction [1] aims to provide listeners in a target zone with their own individual soundfield that does not interfere with other zones within the reproduction region. In some cases, it is desirable to create zones of quiet, where audio from neighbouring zones is suppressed or cancelled [1, 2, 3]. The multizone approach can be used for applications such as the creation of personal sound zones [4] in multi-participant teleconferencing, restaurants/cafés, entertainment/cinema, vehicle cabins and public announcement locations where the reproduction can be optimised to provide private quiet zones.

In order to keep the sounds zones personal it is necessary to minimise the interzone audio interference (leakage) to maximise the individual experience. The existence of leakage means that the reproduction of speech in a particular zone may be intelligible in other zones, deviating from the desired personal sound zones. Some of the earlier methods treat the leakage with hard constraints and attempt to completely remove it [1, 2]. This results in zones that are mostly free of the interference but this is difficult to achieve in situations where a desired soundfield in the bright zone is obscured by or directed to another zone, as the system requires reproduction signals many times the amplitude of what is reproduced within any zone. This is known as the *multizone occlusion problem* [1, 4, 5] and has been dealt with in various ways such as the control of planarity [6], orthogonal basis planewaves [3] and alleviated zone constraints [3, 7]. Reproduction in reverberant rooms has also been accomplished with enhanced acoustic contrast using sparse methods [8].

More recent work has focused on alleviating the constraint so that the amount of leakage is controlled by a weighting function [3, 7]. Allowing the sound to leak into other zones can improve the practicality of the system but decreases the individuality of zones.

Existing methods focus on single frequency soundfields, although there has been work attempting to create multizone soundfields for wideband speech [9]. More recently, work has been done [10] to extend a method [3] to the reproduction of weighted wideband speech soundfields by using the spatial weighting function. This is shown in [11] to allow each zone's acoustic content to be controlled by dynamic space-time-frequency weighting.

To maintain speech privacy amongst the zones it is necessary to keep the leaked speech unintelligible [12]. If the leaked speech is at a level below the threshold of hearing then it may be expected to start becoming inaudible and/or masked. To reproduce clear speech in a weighted multizone soundfield at a level of $60\,\mathrm{dBA}$ in a zone, known as the 'bright' zone, the level of leaked speech in the quiet zone could be reduced to around $30\,\mathrm{dBA}$ to $35\,\mathrm{dBA}$ [8, 11] which is still well above the threshold of hearing ($\approx 0\,\mathrm{dBA}$).

In this work it is shown for the first time, as far as the authors are aware, a difference, or contrast, in intelligibility across the personal sound zones which corresponds to private sound zones. Contributions are made by evaluating the objective intelligibility of reproduced speech and providing methods of control for increased privacy between zones as a baseline study. A method is provided and evaluated for increasing privacy in multizone speech soundfields in anechoic and reverberant environments by using noise to mask the leaked spectrum into the target quiet zone so that it becomes unintelligible. A third contribution is the description and analysis of an enhanced method for increasing privacy and at the same time improving perceived quality in reproductions, analysed using objective (instrumental) measures. This is achieved by performing a weighted multizone reproduction on the noise masker so that it has more influence in the target quiet zone and less in the target bright zone.

This paper begins with an explanation of the weighted multizone speech soundfield method used in this work in Section 2. Noise masking and its relation to speech intelligibility and speech privacy are explained in Section 3. Results of the noise masking methods and conclusions are given in Section 4 and Section 5, respectively.

2. WEIGHTED MULTIZONE SPEECH SOUNDFIELDS

The following section provides an overview of the weighted orthogonal basis expansion synthesis [3] and the cylindrical harmonic expansion reproduction [2] used in this work to reproduce speech in one zone and suppress it in another. This initial step creates a wideband controllable contrast in the level between zones which is then used to reduce leakage between zones.

A multizone soundfield reproduction is depicted in Fig. 1. The circular reproduction region, $\mathbb{D}$, of radius R, contains three sub regions called the bright, quiet and unattended zone, denoted by $\mathbb{D}_b$, $\mathbb{D}_q$ and $\mathbb{D} \cap (\mathbb{D}_b \cup \mathbb{D}_q)'$, respectively. The radius of $\mathbb{D}_b$ and $\mathbb{D}_q$ is

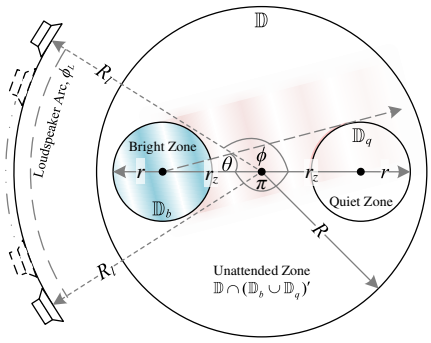

Fig. 1. A weighted multizone soundfield reproduction layout is shown. The shading depicts the desired bright zone soundfield partially directed towards the quiet zone causing the occlusion problem.

r and their centres are located on a circle of radius r_z concentric with $\mathbb{D}$. The angle of the desired planewave in $\mathbb{D}_b$ is θ and is reproduced by loudspeakers positioned on an arc of angle ϕ_L, radius R_l, concentric with $\mathbb{D}$ and with the first loudspeaker at angle ϕ.

Any arbitrary soundfield, including the reproduction of planewave speech, can be described by an infinite set of planewaves arriving from all angles [13]. In the orthogonal basis expansion approach to multizone soundfield reproduction [3] it is shown that a soundfield function, $S(\mathbf{x}, k)$, that fulfils the wave equation, where $\mathbf{x} \in \mathbb{D}$ is an arbitrary spatial sampling point and k is the wavenumber of the soundfield, can be described with an additional weighting function, $w(\mathbf{x})$. This weighting function provides relative importance to the reproduction in different zones and the weighted soundfield function used throughout this work can be written as

$$S(\mathbf{x}, k) = \sum_j P_j(k) F_j(\mathbf{x}, k), \tag{1}$$

where the coefficients for the orthogonal wavefields, $F_j(\mathbf{x}, k)$, for a given weighting function are $P_j(k)$ and $j \in \{1, \ldots, N\}$ where N is the number of basis planewaves [3].

The complex loudspeaker weights used to reproduce the soundfield in the time-frequency domain are defined as [14]

$$\tilde{Q}_l(k) = \sum_{m=-M}^{M} \frac{2e^{im\phi_l} \Delta\phi_s \sum_j \left(P_j(k) i^m e^{-im\phi_p}\right)}{i\pi H_m^{(1)}(kR_l)}, \tag{2}$$

where $M = \lceil kR \rceil$ is the truncation length [3], $i = \sqrt{-1}$, R and R_l are from Fig. 1, $\phi_p = (j-1)\Delta\phi$ are the wavefield angles, $\Delta\phi = 2\pi/N$, ϕ_l is the angle of the l^{th} loudspeaker from the horizontal axis and $\Delta\phi_s$ is the angular spacing of the loudspeakers. Here, P_j is chosen to minimise the difference between the desired soundfield and the actual soundfield [3]. In this work frequency, $f = kc/2\pi$ [13] and $c = 343\,\mathrm{m\,s^{-1}}$ is the speed of sound.

In order to reproduce planewave speech soundfields $\tilde{Q}_l(k)$ must be applied to the speech in the time-frequency domain and inverse transformed back to the time-domain to obtain the set of loudspeaker signals. This can be done by means of a Gabor transform or any unitary time-frequency transformation as

$$\tilde{q}_{al}(n) = \frac{1}{2K} \sum_{m=0}^{K-1} \tilde{Q}_l(m\Delta k) \tilde{Y}_a(m\Delta k) e^{i\pi mn/K}, \tag{3}$$

where $\tilde{Y}_a(k)$ is the discrete Fourier transform of the a^{th} overlapping windowed frame of the input speech signal, $y(n)$. Each loudspeaker signal, $q_l(n)$, is reconstructed by performing overlap-add reconstruction with the synthesis window. This results in the loudspeaker signals, which will reproduce the multiple zones.

The observed signals, $p(\mathbf{x}, n)$, can be found at any arbitrary point in the soundfield by convolving each of the loudspeaker signals with the transfer function, $H(\mathbf{x}, \mathbf{x}_l, k)$, and summing, as

$$p(\mathbf{x}, n) = \frac{1}{2K} \sum_l \sum_{m=0}^{K-1} Q_l(m\Delta k) H(\mathbf{x}, \mathbf{x}_l, m\Delta k) e^{i\pi mn/K}, \tag{4}$$

where $\mathbf{x}_l$ is the position of the l^{th} loudspeaker and $Q_l(k)$ is the time-frequency transform of $q_l(n)$. The soundfield can now be evaluated at any given point in the reproduction region for different input signals and the resulting pressure, $p(\mathbf{x}, n)$, can be observed in the bright zone and quiet zone. From this it is possible to analyse the speech intelligibility in each zone as presented in the following section.

3. PRIVATE SOUND ZONES

This section discusses the relationship between speech privacy and intelligibility and how they are affected in a multizone soundfield reproduction scenario. The use of the Speech Intelligibility Contrast (SIC) is proposed for improving the privacy in personal sound zones.

3.1. Speech Privacy and Intelligibility Contrast

A measure is required to optimally design and evaluate the performance of a method to control privacy in the multizone soundfield reproduction. The relationship between speech intelligibility and privacy is highly correlated. Two measurement standards currently published for assessing speech privacy in closed and open plan spaces are ASTM E2638 [15] and ASTM E1130 [16], respectively. These standards are based on two different measures, which are the Speech Privacy Class (SPC) and the Articulation Index (AI). Both are highly correlated to speech intelligibility and the SPC has been shown to be a better measure for higher privacy situations [12] making it reasonable to maximise a measure of intelligibility contrast between zones to obtain privacy.

It has been shown that objective intelligibility measures are highly correlated with subjective measures and are based on analysing spectral band powers. High mutual information between the clean speech (talker), $y(n)$, and the degraded speech (listener), $p(\mathbf{x}, n)$ from (4), is attained at high signal to noise ratio (SNR) [17], hence indicating that reducing the SNR, for example by adding noise, reduces intelligibility. In this work the intelligibility for two signals $x_1(n)$ and $x_2(n)$ is denoted as $\mathcal{I}(x_1; x_2)$. The particular measure $\mathcal{M}$ can be the mutual information, such as that provided by the Short-Time Objective Intelligibility (STOI) [18] or Speech Transmission Index (STI) [19]. The intelligibility of the pressure signal at a spatial point $\mathbf{x}$ and the signal $y(n)$ is then $\mathcal{I}_{\mathcal{M}}(p(\mathbf{x}, \cdot); y)$.

In this work the SIC is defined as

$$\mathrm{SIC}_{\mathcal{M}} = \frac{1}{\|\mathbb{D}_b\|} \int_{\mathbb{D}_b} \mathcal{I}_{\mathcal{M}}\, d\mathbf{x} - \frac{1}{\|\mathbb{D}_q\|} \int_{\mathbb{D}_q} \mathcal{I}_{\mathcal{M}}\, d\mathbf{x}, \tag{5}$$

where $\|\mathbb{D}_b\|$ and $\|\mathbb{D}_q\|$ are the sizes of $\mathbb{D}_b$ and $\mathbb{D}_q$, respectively, and the domain is restricted such that $\mathcal{I}_{\mathcal{M}}$ for any $\mathbf{x} \in \mathbb{D}_b$ is greater than or equal to $\mathcal{I}_{\mathcal{M}}$ for any $\mathbf{x} \in \mathbb{D}_q$. The following two subsections provide two methods to maximise $\mathrm{SIC}_{\mathcal{M}}$.

3.2. Improving Multizone Privacy

To maximise the SIC, $\mathcal{I}_{\mathcal{M}}$ must be zero at all points in $\mathbb{D}_q$ whilst maintaining maximum $\mathcal{I}_{\mathcal{M}}$ at all points in $\mathbb{D}_b$. Ideally, the mean SNR of $p(\mathbf{x}, n)$ over $\mathbb{D}_b$ should be maintained as high as possible, so to increase $\mathrm{SIC}_{\mathcal{M}}$ the mean SNR of $p(\mathbf{x}, n)$ over $\mathbb{D}_q$ should be reduced. To maximise the SIC noise is added to $q_l(n)$ under the constraint that the mean amplitude of $p(\mathbf{x}, n)$ over $\mathbb{D}_q$ is less than that of $p(\mathbf{x}, n)$ over $\mathbb{D}_b$. This then becomes a constrained optimisation dependent on the reproduced signals in the bright and quiet zones as

$$\max_{G \in \mathbb{R}} \mathrm{SIC}_{\mathcal{M}}, \tag{6}$$

where the noise levels, G_{dB}, of $q_l(n)$ are optimised.

To increase the SIC a time-domain noise mask, $u(n)$, is added to each loudspeaker signal, $q_l(n)$, which is derived from its time-frequency domain representation from (3). Noise is added at different gain, G_{dB}, relative to the maximum amplitude among L loudspeaker signals, $A = \max(\{q_l(n) : l = 1, ..., L\})$. The noise mask is added as

$$q_l'(n) = q_l(n) + u(n) A 10^{\frac{G_{\mathrm{dB}}}{20\,\mathrm{dB}}}, \tag{7}$$

where the new loudspeaker signals are $q_l'(n)$. In this work $u(n)$ is chosen to be uniform white noise with no directivity and this method is referred to as the 'Flat Mask' due to its spatial and spectral uniformity.

Then by transforming $q_l'(n)$ for use in (4), $\mathrm{SIC}_{\mathcal{M}}$ is obtained from (4) and (5). Now $\mathrm{SIC}_{\mathcal{M}}$ can be optimised with (6) using G_{dB} in (7). However, this method does not control $u(n)$ in the spatial domain and so the mean $\mathcal{I}_{\mathcal{M}}$ over $\mathbb{D}_b$ is also reduced even though the SIC is maintained.

3.3. Improving Multizone Privacy and Quality

Ideally a private personal sound zone system would have a maximum SIC whilst maintaining high perceptual quality in the bright zone. Adding $u(n)$ to $q_l(n)$ adds error to $p(\mathbf{x}, n)$ for all $\mathbf{x}$ which as a side-effect reduces the quality of $p(\mathbf{x}, n)$ for any $\mathbf{x} \in \mathbb{D}_b$ and a trade-off between target quality and privacy becomes necessary. Following a similar notation to $\mathcal{I}_{\mathcal{M}}$, the quality of $p(\mathbf{x}, n), \mathbf{x} \in \mathbb{D}_b$ degraded from $y(n)$ is any speech quality assessment model of measure, $\acute{\mathcal{M}}$, denoted by $\mathcal{B}_{\acute{\mathcal{M}}}(p(\mathbf{x}, \cdot); y) \in \{0, ..., 1\}$, scaled to match that of $\mathcal{I}_{\mathcal{M}}$. Now a new optimisation can be defined as

$$\max_{G_{\mathrm{dB}} \in \mathbb{R}} \mathrm{SIC}_{\mathcal{M}} + \frac{\lambda}{\|\mathbb{D}_b\|} \int_{\mathbb{D}_b} \mathcal{B}_{\acute{\mathcal{M}}}\, d\mathbf{x}, \tag{8}$$

where the noise levels, G_{dB}, are defined below, λ is a weighting parameter for the importance of quality in the optimisation and $\mathcal{I}_{\mathcal{M}} \geq \mathcal{B}_{\acute{\mathcal{M}}}$ for $\mathbf{x} \in \mathbb{D}_b$. This optimisation also requires minimum mean SNR of $p(\mathbf{x}, n)$ over $\mathbb{D}_q$ and maximum mean SNR of $p(\mathbf{x}, n)$ over $\mathbb{D}_b$ achieved here by applying zone weighting to $u(n)$.

To simplify the optimisation of (8) in this work, constraints are applied to the multizone reproduction of $u(n)$, which is a planewave field in $\mathbb{D}_q$ and quiet in $\mathbb{D}_b$. The constraints are $\theta = 0^\circ$, so that the masker source is collocated with the leakage, and a new weighting function, $\bar{w}(\mathbf{x})$, is constrained to an importance in $\mathbb{D}_q$ of unity, 10^4 in $\mathbb{D}_b$ and 0.05 in the unattended zone. The remainder of the multizone reproduction is the same as used to generate $q_l(n)$ for $y(n)$.

The goal is to find another set of loudspeaker signals that would reproduce $u(n)$ in $\mathbb{D}_q$ to control the mean SNR of $p(\mathbf{x}, n)$ over $\mathbb{D}_q$, therefore solving (8). To do this, $u(n)$ is transformed to the time-frequency domain as $\tilde{U}_a(k)$ and used as the input signal in (3). New loudspeaker weights, $\tilde{Q}_{'l}(k)$, are derived from (2). Then, from (3), the loudspeaker signals, $\hat{q}_l(n)$, are reconstructed and these become the new noise mask signals as

$$q_l''(n) = q_l(n) + \hat{q}_l(n) A 10^{\frac{G_{\mathrm{dB}}}{20\,\mathrm{dB}}}, \tag{9}$$

where the new loudspeaker signals are $q_l''(n)$ with noise levels, G_{dB}. In this work this method is referred to as the 'Zone Weighted Mask' due to the masker signal being dependent on the multizone scenario.

Then by transforming $q_l''(n)$ for use in (4), $\mathrm{SIC}_{\mathcal{M}}$ is obtained from (4) and (5). Now $\mathrm{SIC}_{\mathcal{M}}$ can be optimised with (8) using G_{dB} in (9). The optimisation problem can now be analysed by measuring $\mathcal{I}_{\mathcal{M}}$ for $\mathbf{x} \in \mathbb{D}_b \cap \mathbb{D}_q$, $\mathcal{B}_{\acute{\mathcal{M}}}$ for $\mathbf{x} \in \mathbb{D}_b$ and for various G_{dB}.

4. RESULTS

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

4.1. Multizone Reproduction Evaluation

The layout of Fig. 1 is evaluated, where $r = 0.3\,\mathrm{m}$, $r_z = 0.6\,\mathrm{m}$, $R = 1\,\mathrm{m}$ and $R_l = 1.5\,\mathrm{m}$. The value of $\theta = \{0^\circ, 15^\circ, 90^\circ\}$ for the angle of the desired planewave virtual source in the bright zone. These angles are chosen to represent multizone occlusion scenarios. Input speech signals sampled at $16\,\mathrm{kHz}$ are transformed to the frequency domain using an FFT and $64\,\mathrm{ms}$ windows with 50% overlapping. The loudspeaker signals, $q_l'(n)$ and $q_l''(n)$, are generated using the methods outlined in section 2 and 3. The reproduction is performed for $L = 295$ and $\phi_L = 2\pi$ which, for the cases in this work, is free of aliasing problems below $8\,\mathrm{kHz}$ [2, 3].

The zone weights are constant and are chosen so that the bright zone weight is unity, the unattended zone weight is 0.05 the reproduction importance of the bright zone following [3, 10] and the weight of the quiet zone is set to 10^4. Frequency dependent zone weighting and signal filtering may give further improvements. The noise masking methods, 'Flat Mask' and 'Zone Weighted Mask', are applied with G_{dB} ranging from $-40\,\mathrm{dB}$ to $20\,\mathrm{dB}$ in (7) and (9).

Speech files for the evaluation were taken from the TIMIT corpus [20]. Twenty files were randomly selected such that the selection was constrained to have a male to female speaker ratio of $50:50$.

Three reverberant rooms and one anechoic are evaluated. The rooms walls have an absorption coefficient of 0.3 and are $4\,\mathrm{m} \times 9\,\mathrm{m} \times 3\,\mathrm{m}$, $8\,\mathrm{m} \times 10\,\mathrm{m} \times 3\,\mathrm{m}$ and $9\,\mathrm{m} \times 14\,\mathrm{m} \times 3\,\mathrm{m}$, sizes that were selected to match a small office, medium office and restaurant/café, respectively. The multizone setup is placed in the centre of the rooms and recordings are analysed from both zones where 32 receivers are positioned randomly in each zone. Room reflections are simulated using the image method [21] with approximately 446×10^3, 206×10^3 and 149×10^3 images for each of the respective rooms, at $0.5\,\mathrm{s}$ in length and sampling frequency of $16\,\mathrm{kHz}$.

The reproductions are analysed using the STOI, STI and Perceptual Evaluation of Speech Quality (PESQ) [22] measures to evaluate the performance with $\mathrm{SIC}_{\mathrm{STOI}}$ and $\mathrm{SIC}_{\mathrm{STI}}$ in anechoic and reverberant environments, respectively. The STOI measure is designed for the prediction of time-frequency weighted noisy speech like the simulated recordings in this work. The STI measure is currently

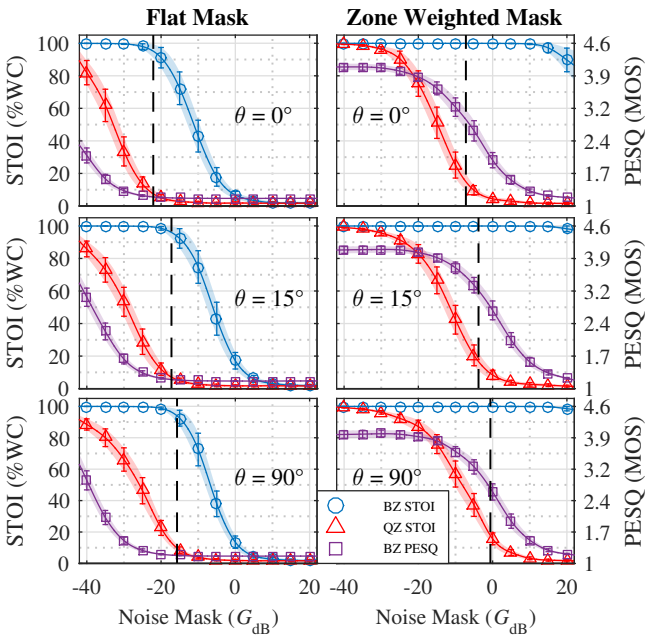

Fig. 2. Mean STOI and PESQ are shown for the anechoic environment and 95% confidence intervals are indicated. BZ and QZ are the bright and quiet zone, respectively. Black dashed lines indicate optimum G_{dB} and $\lambda = 1$.

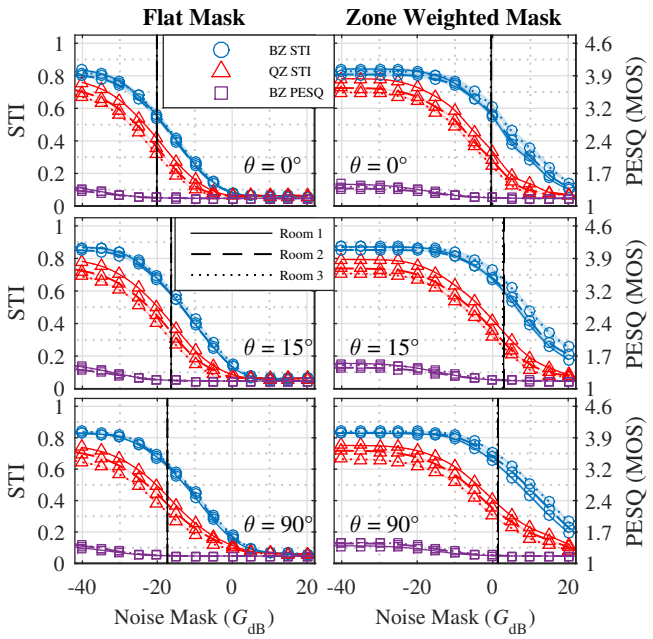

Fig. 3. Mean STI and PESQ are shown for the small office, medium office and restaurant/café labelled as Room 1, Room 2 and Room 3, respectively. BZ and QZ are the bright and quiet zones, respectively. Vertical black lines indicate optimum G_{dB} and $\lambda = 1$.

the only choice for a reverberant objective intelligibility measure. A good objective measure for speech quality is the PESQ measure.

The STOI and PESQ are measured in this work with the clean, $y(n)$, and degraded, $p(\mathbf{x}, n)$, speech for each file and receiver combination. The STI is measured for each receiver using the systems impulse response found with a logarithmic sine sweep. The intelligibility and quality results are then averaged over each zone like that of (5) and (8). This results in three object measures, two weighting methods, four rooms, 13 levels of added noise, 20 speech files and 64 receiver positions totalling $\approx 332,800$ data points.

4.2. Intelligibility Contrast from Noise-Based Sound Masking

Fig. 2 shows that by using the 'Flat Mask' method to obtain privacy between zones it is possible to obtain upwards of 85% $\mathrm{SIC}_{\mathrm{STOI}}$ but this is only possible within a small range of G_{dB} ($-25\,\mathrm{dB}$ to $-20\,\mathrm{dB}$). The range remains the same size as the angle is increased but the G_{dB} which is required to maintain $\mathrm{SIC}_{\mathrm{STOI}}$ is increased to approximately $-15\,\mathrm{dB}$. In each case of the 'Flat Mask' method the signal in the bright zone is of poor quality as shown by the corresponding PESQ curve (which is undesirable).

It can be seen that θ has a small impact on the range of increased $\mathrm{SIC}_{\mathrm{STOI}}$ and it is possible to maintain above 80% $\mathrm{SIC}_{\mathrm{STOI}}$ for different angles. The effect of θ is only minor due to the large zone weighting used in the reproduction process. Fig. 2 shows that with a small change in angle, $15°$, $\mathrm{SIC}_{\mathrm{STOI}}$ remains the same and the PESQ curve starts to rise.

The effect of the spatially weighted noise maskers can be clearly seen in Fig. 2 where the use of a 'Zone Weighted Mask' improves $\mathrm{SIC}_{\mathrm{STOI}}$ across all scenarios. The maximum improvement occurs when G_{dB} is between $-5\,\mathrm{dB}$ and $20\,\mathrm{dB}$ and provides a $\mathrm{SIC}_{\mathrm{STOI}}$ of greater than 95% for every scenario. Even when the occlusion problem is present it is still possible to obtain privacy with greater than 95% $\mathrm{SIC}_{\mathrm{STOI}}$ when G_{dB} is between $0\,\mathrm{dB}$ and $15\,\mathrm{dB}$.

Another benefit of using the 'Zone Weight Mask' is that the quality of the bright zone reproduction is increased within the region where $\mathrm{SIC}_{\mathrm{STOI}}$ is significantly large. With a $\mathrm{SIC}_{\mathrm{STOI}}$ of greater than 70% it is also possible to obtain a PESQ of greater than 3.4 reducing to 3.2 and 2.8 for a $\mathrm{SIC}_{\mathrm{STOI}}$ of 80% and 90%, respectively. This shows the trade-off between reproduction quality and zone privacy which is controlled using λ and may depend on the application of the private multizone system.

With the multizone reproduction in different reverberant rooms it can be seen in Fig. 3 that a contrast in intelligibility is still possible without room equalisation. The quality is reduced most likely due to uncontrolled early reflections inhibiting the bright zone, however, the $\mathrm{SIC}_{\mathrm{STI}}$ still remains high at various G_{dB} albeit reduced from an ideal anechoic environment. The maximum $\mathrm{SIC}_{\mathrm{STI}}$ is 40% and occurs with the 'Zone Weighted Mask' for Room 3.

5. CONCLUSIONS

This paper has investigated speech privacy between bright and quiet zones in multizone reproduction scenarios. Methods have been proposed and evaluated for increasing the speech intelligibility contrast (SIC) in anechoic and reverberant environments showing that added noise can be used to mask the leaked spectrum to provide a significant SIC of higher than 95%. It has also been shown that it is possible to maintain quality in the bright zone with a PESQ MOS of 3.2 whilst providing a SIC above 80% by using space-time domain masker signals and that speech privacy can be achieved in reverberant rooms using the methods outlined in this paper. Future work will look into further improvement of the quality and privacy in reverberant environments as well as a reduction in the number of required loudspeakers.

6. REFERENCES

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

[2] Y. J. Wu and T. D Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 1711–1720, 2011.

[3] W. Jin, W. B. Kleijn, and D. Virette, "Multizone soundfield reproduction using orthogonal basis expansion," in *Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. 2013, pp. 311–315, IEEE.

[4] T. Betlehem, W. Zhang, M. Poletti, and T. D. Abhayapala, "Personal Sound Zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Process. Mag.*, vol. 32, pp. 81–91, 2015.

[5] T. Betlehem and P. D. Teal, "A constrained optimization approach for multi-zone surround sound," in *Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. 2011, pp. 437–440, IEEE.

[6] P. Coleman, P. Jackson, M. Olik, and J. A. Pedersen, "Personal audio with a planar bright zone," *J. Acoust. Soc. of Am.*, vol. 136, pp. 1725–1735, 2014.

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

[8] W. Jin and W. B. Kleijn, "Theory and design of multizone soundfield reproduction using sparse methods," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 2343–2355, 2015.

[9] 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, pp. 378–387, 2013.

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

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

[12] 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.

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

[14] 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, pp. 107–116, 2009.

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

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

[17] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. Petkov, B. Sauert, and P. Vary, "Optimizing speech intelligibility in a noisy environment: A unified view," *IEEE Signal Process. Mag.*, vol. 32, pp. 43–54, 2015.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, 2011.

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

[20] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. of Am.*, vol. 65, pp. 943–950, 1979.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. 2001, pp. 749–752, IEEE.