Highlights

Combining Background Noise and Artificial Masking to Achieve Privacy in Sound Zones

Daniel Wallace, Jordan Cheer

- Communication privacy can be improved by using loudspeaker arrays to focus speech
- Artificial masking noise can be focused towards potential eavesdroppers to improve privacy
- The constant background noise in a space can be leveraged to further improve privacy
- Using both artificial and natural maskers reduces acoustic contrast requirements

Combining Background Noise and Artificial Masking to Achieve Privacy in Sound Zones

Daniel Wallace[1,*], Jordan Cheer

Institute of Sound and Vibration Research, University of Southampton

Abstract

A private sound zone can be created by focusing a spoken message towards a target listener using a loudspeaker array. In practice, however, the reproduced speech cannot be completely contained within the target zone due to practical limits on the directivity of the array. Despite these limitations, the privacy of the message can be maintained if the leaked speech is sufficiently masked by noise. Two possible sources of this masking noise are considered in this article: the ambient noise in the reproduction environment, and an additional masking signal radiated by the loudspeaker array. The present article demonstrates that the process of designing a private audio system is significantly affected by the presence of ambient noise. A key complication is that temporal fluctuations and spatial non-uniformity in the ambient noise can reduce its effectiveness as a masker. These features also make it more difficult to estimate the corresponding reduction in the intelligibility of speech in each listening zone. To mitigate this spatial and temporal variance, it is proposed that systems should be designed to rely only on the masking provided by the diffuse, quasi-stationary background noise component of the environmental noise. It is shown that when systems utilise a combination of the background noise and an additional, artificial masker, a lower level of acoustic contrast is required from the system, compared to the case where the masking is supplied by the background noise exclusively.

*Corresponding Author

Email addresses: D.J.Wallace@soton.ac.uk (Daniel Wallace), J.Cheer@soton.ac.uk (Jordan Cheer)

[1]Daniel Wallace was supported by the EPSRC Centre for Doctoral Training in Next Generation Computational Modelling Grant No. EP/L015382/1

Keywords:
Speech Privacy, Signal Processing, Auditory Masking, Sound Zones

1. Introduction

Loudspeaker arrays can be used to form spatially separated listening zones within a shared space. Such systems have found utility in open plan offices and museum exhibits [1], as entertainment systems for the home [2, 3, 4], and for personalised telecommunications using mobile devices and in vehicles [5, 6, 7]. When such systems are used to transmit speech, such as through a security partition at a bank counter or in the aforementioned telecommunications examples, it is important to preserve the privacy of the target listener. This can be achieved by using one sound zoning process to focus the target speech towards the target listener and using a second process to radiate additional masking noise into areas where other listeners are situated [8, 9], or by combining appropriately filtered noise signals [10]. Sound zoning methods can be broadly categorised into those that control the energy of the reproduced signals in each zone, such as Acoustic Contrast Control (ACC) [11], and those that seek to accurately reproduce a target signal, such as Pressure Matching [12]. Recent hybrids and generalisations of these methods have also been proposed [13, 14, 15], yielding perceptual improvements to the reproduced sound fields and the ability to tailor the capabilities of the system to the objectives of the system or the properties of the input signals.

The present article describes a method for specifying the technical requirements of a sound zoning system based on the predicted speech intelligibility in each listening zone and the masking effect of the environmental background noise present in the listening space, which has not been considered in previous work. The analysis presented in this article significantly expands on previous work by the authors [16] through the use of ambient noise recordings from a range of typical environments, rather than assuming a single, speech-shaped background noise spectrum and thus expands the practical relevance of this study. The proposed approach is based on the recognition that in spaces where privacy is a concern, there is also likely to be additional sources of noise, for example, due to the voices and activity of other people sharing the space. This additional ambient noise will decrease the intelligibility of speech reproduced in the bright zone, impairing the performance of the system from the perspective of the target listener. However,

noise is beneficial for the provision of privacy, as this will contribute to the masking of any speech that escapes the target region.

Figure 1 illustrates the proposed method for providing private listening zones in a noisy environment, using a combination of an artificial masking signal and the background noise present in the space. The speech and masking signals are independently filtered so that they are focused into the bright and dark zones respectively, and the zonal signals are analysed in terms of the Speech Intelligibility Index (SII) [17] and the A-weighted masker level. This analysis provides information pertaining to the goals of the system, namely, that the target message be delivered clearly to the target listener in the bright zone, that this message is unintelligible in the dark zone, and that the negative perceptual effects of the required additional masking are minimised [9]. It is possible to carry out the bulk of this evaluation indirectly, i.e. without requiring measurement microphones within each listening zone. Instead, an internal representation of the zonal signals can be synthesised by convolving the input signals with the electroacoustical transfer responses between the loudspeaker array and the zones. These responses are necessary for the production of the sound zoning filters, and can therefore be re-used. However, the SII evaluation also requires a measurement of the ambient noise in each zone, which must be captured in real-time and combined with the synthesised zonal signals. It is impractical to directly measure the ambient noise experienced at each listener position as microphones would need to be placed coincident with the listeners' ears. Remote microphone techniques for estimating these signals are also limited in utility as the locations of the ambient noise sources are not known a-priori [18]. However, situating a microphone at a convenient nearby location can provide an approximation to the ambient noise within each zone.

The processing steps required to obtain a reliable masking prediction from a remote ambient noise measurement forms a key part of the discussion in this article, and this is described in Section 2 alongside a brief review of how speech intelligibility is affected by noise. A range of example sound zoning systems are introduced in Section 3 and their performance is evaluated in terms of the level of acoustic contrast that is required to provide certain target levels of speech intelligibility in each listening zone. The results presented in this section show that when systems include an artificial masking signal, a continuous trade-off exists between requiring high levels of acoustic contrast and radiating high levels of additional noise into the reproduction environment. The main conclusions and suggestions for further work are

summarised in Section 4.

2. The Effects of Noise on Private Sound Zoning Systems

This section will present a review of the different characteristics of environmental noise and a discussion of how they may influence the performance of a private sound zoning system. Throughout this article, a distinction is made between the *ambient noise* and the *background noise* in a given environment, using the convention described in British Standard BS 4142:2014+A1:2019 [19]. This standard concerns the measurement and rating of industrial and commercial sound and usually requires measurements of a specific sound source to be corrected based on the level of other sound sources in the environment. The ambient sound level, $L_{Aeq,T}$, is described as the "equivalent continuous A-weighted sound pressure level of the totally encompassing sound in a given situation at a given time, usually from many sources near and far, at the assessment location over a given time interval, T" [19]. The background sound level, $L_{A90,T}$, is the "A-weighted sound pressure level that is exceeded by the residual [ambient] sound at the assessment location for 90% of a given time interval, T" [19]. The background noise level measured in a space therefore excludes contributions to the ambient sound that are intermittent, and are thus less effective at masking continuous speech [20].

In the following subsections, the spectral, temporal and spatial properties of the ambient and background noise will be explored using recordings from public spaces contained within the Ambisonic Recordings of Typical Environments (ARTE) database [21]. These ambient noise recordings were made with an Ambisonic measurement system, facilitating analysis of the full 3D sound field. The background noise in each of the 9 recorded environments is isolated by dividing each of the ambient noise recordings into 125 ms segments, corresponding to the "fast" time constant recommended in BS 4142 [19], then computing the equivalent A-weighted sound pressure level readings from each of these segments. As described above, the background noise level is at the tenth percentile of these data, and the corresponding samples from the original recording can be concatenated into a single file, one tenth the length of the original ambient recording. This can then be processed to determine the spectral and spatial properties of the background noise.

2.1. Spectral Effects

In order to rely on the background noise as an energetic masker in a private audio system, the spectrum of the background noise must be appropriate for this purpose. In most of the environments recorded in the ARTE database, speech is audible within the recordings, either as a single intelligible voice or as a babble of multiple talkers. However, speech is not the only contributor to the overall noise in these environments - other examples of noise include the break-in of traffic noise from outside, industrial noise from mechanical ventilation or air conditioning systems, and the noise associated with the movement of people.

To demonstrate the effect of the different noise sources present in the considered environments on the spectra of both the ambient and background noise, Figure 2 shows the power spectral density (PSD) estimates of the zeroth order (omnidirectional) Ambisonic component of the ambient and background noise from four environments in the ARTE database. The ⅓-octave band normal speech spectrum level from the SII standard [17] is included for comparison, demonstrating that although a range of sound sources are present in each scene, the spectrum of both the ambient and background noise is similar to that of speech. The difference between the solid and dashed lines of each colour indicates the spectral content that is removed when the background noise is isolated from the overall ambient noise. This difference is maximal in areas covered by the standard speech spectrum, suggesting that the process of isolating the background noise removes some of the effect of voices. Furthermore, in each instance, there is little difference between the ambient and background noise PSDs below 125 Hz, indicating that the ambient noise is dominated by steady noise sources in this frequency range. This is most apparent in the *Library* scene, where the PSD is dominated by energy below 125 Hz, associated with the noise from air handling units. The upward spread of masking from low to high frequencies allows this low-frequency background noise to mask portions of speech, despite the differences in their respective spectra.

2.2. Temporal Effects

In systems described by the block diagram shown in Figure 1, the levels of the speech programme and artificial masking signals are adjusted to meet speech intelligibility constraints in each zone. This evaluation is affected by the ambient noise level, which can vary significantly with time, particularly in environments with a large number of discrete and independent noise sources.

When the ambient noise is relied upon by a system to provide a proportion of the required masking, it is desirable for this noise to be temporally stable as "glimpses" [20] of the target speech may become audible or intelligible if the noise level fluctuates.

In principle, it is possible to account for a rapid variation in the ambient noise level by simultaneously adjusting the level of artificial masking output by the system. This approach would provide a constant level of masking within the dark zone, but would lead to equivalently rapid fluctuations in the composition of the dark zone sound field. Fluctuation strength, defined as the depth of modulation at a rate between 0.5 and 20 Hz, has been identified as a contributor to the sensation of psychoacoustic annoyance [22]. It is proposed here, therefore, that a more perceptually appropriate scheme is to adapt the masking signal to the level and spectrum of the background noise, which typically varies much more slowly than that of the ambient noise.

To consider the effects of temporal variation in the ambient noise, Figure 3 shows the running A-weighted sound pressure level of the 9 extracts from the ARTE database, evaluated at 125 ms intervals. The samples that contribute to the evaluation of the background noise are shown in red, and it can be seen that, in the majority of the examples, these samples are distributed throughout the duration of each recording, and the dynamic range of these samples is low compared to that of the remaining ambient noise. This indicates that throughout the 1.5 - 2.5 minute recordings, the background noise level remains fairly constant. Over longer timescales, the background noise level can change, as evidenced by the *Church 1* and *Church 2* scenes, which were excerpted from a single longer recording. This variation poses challenges in the design of private audio systems as they must be capable of operating across a wide range of environmental noise conditions, adapting quickly enough to account for the masking effects of the changing background noise, but not so rapidly that the resulting fluctuation in speech and masker levels becomes a nuisance. The effects of variations in the background noise level on a range of private sound zoning systems are discussed quantitatively in Section 3.3.

2.3. Spatial Effects

A further comparison between the ambient and background noise in typical public environments concerns the spatial distribution and diffuseness of the noise field. These factors affect the degree of masking provided by the noise, and hence the specifications of a private sound zoning system. When a

target speech signal and an interferer originate from different azimuths, the intelligibility of the target speech is improved compared to the case where the target and interferer are co-located [23, 24], due to a phenomenon termed the Spatial Release from Masking (SRM). For the systems discussed in this article, SRM has the potential to reduce the effectiveness of the masking provided by the ambient noise in the environment. In order to assess the feasibility of incorporating the ambient noise into the masking predictions made by a speech privacy control system, it is necessary to predict the degree of SRM in situations with multiple sources of masking, and understand the spatial properties of typical ambient noise fields. These two aspects will be discussed in the remainder of this section.

SRM is attributed to two effects in the auditory system, both related to the signals received at each ear [23]. The first is the better-ear effect, which relates to the shadowing effect of the head causing the signal to noise ratio at each ear to be different when sources of speech and noise are spatially separated. However, in experiments where maskers of equal power are placed symmetrically with respect to the head, negating the better-ear effect, SRM is still observed [25, 26]. This indicates that the human auditory system also performs binaural processing on the individual ear signals, amalgamating them into a single percept [27]. This processing can resolve the interaural time and level differences that arise when the signals entering each ear are correlated, for example if they originate from a single point source. When the interaural signals are uncorrelated, such as is the case in a diffuse sound field, the effect of SRM is reduced. Binaural speech intelligibility metrics use models of these two techniques to estimate the degree of SRM associated with a given set of binaural speech and noise signals. However, using this type of evaluation in a private audio system would require a binaural measurement of the background noise at each listener position, which is impractical to obtain in a deployable system.

Several investigations have been carried out to quantify the degree of SRM when multiple sources of masking are arranged on the horizontal plane around the listener, and a review of research into this area is provided by Bronkhorst [23]. This paradigm is of particular relevance to the present problem of quantifying the relative effects of ambient and artificial noise on personal audio system performance, because as the number of discrete masking sources surrounding the listener increases, the masking environment becomes increasingly diffuse. To illustrate the effect of source positioning on SRM, Figure 4 presents data from a study by Bronkhorst and Plomp [28], in

which the Speech Reception Threshold (SRT) of meaningful sentences was measured with sources of modulated speech-shaped noise either co-located with the frontal talker, shown by a blue point in Figure 4, and spatially distributed around the listener. The difference in SRT between each of these conditions quantifies the SRM, and the results show that as the number of maskers increases and their spatial distribution becomes more homogeneous, the SRM decreases. This trend has been observed in several other studies, using a range of speech tests, masker locations and masking signals [25, 26, 29]. Based on these results, it has been proposed that the binaural processing centre, which would otherwise be able to provide SRM, can become "overloaded" [30] in complex acoustical scenes where several sources of masking operate simultaneously. The limit to this capacity has been estimated at between three and six individual sources of masking [30, 25]. The study by Yost [25] showed that when six continuous noise maskers were distributed in front of the listener, the measured SRM was 0 dB. In the context of private personal audio system development, this suggests that SRM could decrease the level of masking provided by foreground sources of noise such as nearby conversations, noise from footfall or equipment by up to 8 dB. Conversely, the level of SRM associated with distributed sources such as distant traffic noise or noise from a ventilation system is likely to be low, as this condition represents the mathematical limit of adding many discrete sources of masking to an acoustical scene. This indicates that the background noise is less likely to be affected by SRM than the overall ambient noise, thereby allowing the level of masking to be estimated using simple, monaural intelligibility metrics. Furthermore, the diffuse field assumption allows this background noise level to be estimated from a remote point, away from the listening zones, further increasing the practicality of the proposed system.

The spatial distribution of both the ambient and background noise can be tested using the ARTE database by decoding each Ambisonic recording to a circular loudspeaker array and comparing the output level of each loudspeaker. Figure 5 shows the L_{Aeq} of 16 loudspeakers arranged around the measurement position as they reproduce the ambient and background noise from each of the 9 public spaces in the ARTE database. Equivalent length samples of the ambient and background noise are used to produce each directivity plot, and each is normalised independently so that the maximum source output is set to 0 dB. The results in Figure 5 show that in each scene the background noise has a broader directivity profile than the ambient noise. This is quantified in Figure 5 using the directivity index (DI) [31] of the back-

ground and ambient noise, i.e. the ratio of the maximum loudspeaker energy to the mean energy, reported in dB. On average across the tested configurations, the background noise has a directivity index 2.4 dB lower than the corresponding ambient noise. The similarity between the background and ambient DIs in the *Cafe 2* environment is due to the microphone being situated close to a wall during the recording [21]. For comparison purposes, the ARTE database also includes an artificially generated diffuse scene, generated by recording uncorrelated speech-shaped noise samples from each of the loudspeakers in a 41 channel spherical array. The horizontal DI of this scene, measured using the same process as described above, is 0.8 dB. Measurement of a perfectly diffuse field, with DI = 0 dB, would require perfect matching between the microphones in the measurement array and between the loudspeakers in the source array.

2.4. Summary

This section has presented an investigation into the characteristics of ambient noise in typical environments and considered how this may affect the design of a private sound zoning system. The ambient noise in an environment can be beneficial for private speech reproduction as it can, at least partially, mask any speech that leaks out of the target zone. However, the inherent spatial, spectral and temporal variability of the ambient noise in many typical environments can reduce the effectiveness of this masking. Additionally, in order to optimally set the input levels of the speech signal and the artificial masker in a private sound zoning system, the degree of masking produced by the ambient noise in each zone must be predicted accurately. The aforementioned variability in the ambient noise makes this process unreliable, potentially leading to a loss of privacy if the degree of masking provided by the ambient noise is overestimated. While the background noise contains less energy than the ambient noise, using this component poses less risk of over-predicting the masking, thereby increasing the overall reliability of the system whilst still gaining the advantages of using all available sources of masking.

On the other hand, by only considering the background noise in the design process, there is a risk that the masking effect of the overall ambient noise within the bright zone is underestimated, thereby degrading the intelligibility of speech intended for the target listener. However, although the ambient noise in the tested environments is more intense than the background noise, the level of additional masking provided by the ambient noise is limited by

SRM and listeners' ability to glimpse information in time-varying masking conditions. In summary, for all the typical environments studied in this section, the background noise:

- has a speech-shaped or low-pass frequency response, leading to the potential for the upward spread of masking into the speech range;
- is temporally stable, therefore limiting the likelihood of glimpses and reducing the rate at which the zonal signal levels must be updated;
- is more spatially diffuse than the overall ambient noise. This means that the background noise can be measured outside of the listening zones and listeners in the dark zone cannot rely on SRM to improve their ability to overhear the target message.

In unusual environments where the presented assumptions about the composition of the ambient and background noise are not met, such as where the background noise is caused by a single nearby source, the diffuse-field assumption for the background noise would not be appropriate. In this case, the background noise should be measured separately within each zone and a correction to the masking level based on the expected SRM should be applied.

3. Sound Zoning System Performance Evaluation

In addition to considering the impact of the environment on the performance of a private personal audio system, a key task for the designer is to specify the capabilities of the loudspeaker array in terms of the level of acoustic contrast that it must deliver. When speech and masker signals are radiated into the bright and dark zones respectively, the acoustic contrast ultimately controls the signal-to-noise ratio in each zone, which is well-correlated with the intelligibility of speech. Low levels of acoustic contrast lead to increased leakage between the zones, risking compromised privacy and poor speech clarity for the target listener. In order to achieve higher levels of acoustic contrast, it may be necessary to use a loudspeaker array with more elements or to provide additional room acoustic treatment, both of which have cost implications.

This section presents an evaluation of a range of systems with different acoustic contrast profiles, that have each been designed according the block

diagram in Figure 1. Each system is laid out symmetrically, with bright and dark zones situated to the right and left of the loudspeaker array respectively, as shown in Figure 6. The elements of two source arrays are marked on this figure, and the narrowband acoustic contrast for these two array geometries has been calculated based on transfer response measurements of the arrays in a well-damped listening room ($T_{60,mf}$ = 110 ms). The full 27 channel loudspeaker array, labelled in Figure 6 as Source Array (max), has been previously described by House et al. [32]. The four-channel array, labelled as Source Array (min), was constructed using a subset of the elements from the larger array, and a higher level of regularisation was used in the filter design process to further limit the acoustic contrast. The regularised ACC [11] method was used to design the sound zoning filters for the presented examples, but similar filters could also be designed using the other methods referenced in the Introduction. ACC was selected in this case as by definition of the optimisation process, the method maximises the Acoustic Contrast between adjacent zones, which translates into a maximal signal-to-noise ratio difference between the listening zones. Furthermore, ACC was chosen for reasons of simplicity, as the sound field control filters produced using this method only depend on the system geometry and the regularisation parameter; alternative methods also require the specification of a target sound field and the selection (or optimisation) of additional variables. From the narrowband acoustic contrast trace for each of the two source arrays, a ⅓-octave band acoustic contrast profile was constructed, and then several intermediate profiles were generated by interpolating between the two measured profiles. These profiles are shown in Figure 7, and characterise a set of systems with a wide range of frequency-dependent acoustic contrast levels, thereby encapsulating several practical methods for improving the acoustic contrast performance of a sound zoning system. In practice, for a given zonal geometry, low-frequency contrast may be improved by increasing the overall aperture of the array, and high frequency contrast can be increased by reducing the inter-element spacing [33]. The regularisation parameter can also be adjusted to control the contrast on a frequency-by-frequency basis, although this will also influence the robustness of the system to environmental changes and variations in the loudspeaker sensitivity and position [34].

For each of the systems described by the contrast profiles shown in Figure 7, the speech and masker levels are adjusted based on estimates of the spectral level of the background noise and the speech intelligibility requirements in each zone, as shown in Figure 1. A minimum acceptable level of intelli-

gibility, s_b, is specified in the bright zone, and a maximum acceptable level of intelligibility, s_d, is specified in the dark zone. To minimise the potential for annoyance and distraction to result from the introduction of the artificial masker, the A-weighted level of the masker is minimised in the optimisation, subject to the satisfaction of the intelligibility constraints.

Formally, the problem can be formulated as a constrained optimisation over the signal levels:

$$\begin{gathered} \text{Minimise} : L_A(\text{Masker in Dark Zone}) \\ \text{Subject to} : \text{SII}_b > s_b \;\&\; \text{SII}_d < s_d \end{gathered} \tag{1}$$

where L_A is the A-weighted sound pressure level, and the subscripts $\{\}_b$ and $\{\}_d$ refer to quantities evaluated in the bright and dark zones respectively.

The algorithm chosen for this optimisation process is Pattern Search [35], as this algorithm does not rely on the calculation of gradients in the cost function, which are shallow in regions where the dark zone signal is dominated by the background noise, potentially leading to gradient-based solvers halting prematurely. The algorithm searches a 2D parameter space formed by the programme and masker signal levels. Each level is allowed to vary $\pm$30 dB from the supplied background noise level. As the Pattern Search algorithm can make function evaluations at any coordinate point within this range, this optimisation can potentially require a large number of trials, so in order to provide flexibility and increase computational efficiency in simulating systems with a range of acoustic contrast profiles, a surrogate model is introduced, which takes advantage of the limited frequency resolution of the SII metric. At the front-end of the SII calculation process, signals are converted to a series of ⅓-octave band spectral levels, from 160 Hz to 8 kHz. Accordingly, in the proposed surrogate model, the signals received in the bright and dark zones are simulated by calculating the spectral levels of the original speech and masker signals and adjusting these based on the frequency response of the loudspeaker array and the pre-computed ⅓-octave band acoustic contrast profile, as shown in Figure 7. The same process is applied to the measured background noise and this contribution is added to each zonal signal equally, commensurate with the diffuse field assumption. This process is significantly less computationally demanding than the conventional direct method, which involves convolving the input signals with the appropriate sound zoning filterbank, then convolving these signals with the electroacoustical transfer responses between the loudspeaker array and each zone. Zonal signals generated in this way are simply converted to ⅓-octave

band spectra internally within the SII calculation, so the direct method can be considered an inefficient use of computational resources when the key optimisation constraints are based on the SII, rather than on the fine structure of the signals. More complex evaluation functions that take into account the masking effect of fluctuating noise, such as STOI [36], would necessitate the direct, signal-based approach. However, for the method utilised here, once the optimisation has concluded, the resulting optimal signal levels can be input into a single convolution-based array simulation to test the validity of the surrogate model.

3.1. Acoustic Contrast Requirements

The level of acoustic contrast provided by a loudspeaker array has a significant impact on the ability of a system to satisfy a certain pair of speech intelligibility constraints, s_b and s_d from Equation 1. For the present investigation, these constraints are set at $s_b = 0.60$ and $s_d = 0.05$. At an SII value of 0.60, connected speech is clearly intelligible despite the presence of background noise and the leakage of the masker into the bright zone, and at SII = 0.05, speech is essentially unintelligible [9].

Figure 8 shows the results of optimising speech and masker levels to meet these constraints, for each of the systems characterised by the acoustic contrast profiles shown in Figure 7. The broadband acoustic contrast, averaged over the speech frequency range, is used to identify each system. The background noise is sampled from the *Church 2* scene from the ARTE database, measured at $L_{A90} = 54.5$ dBA, and the masker is random noise with the same long-term average spectrum as speech, derived from the normal speech spectrum in the ANSI SII standard [17]. At the leftmost edge of the figure, the red shaded region indicates the levels of acoustic contrast where no valid solution to the optimisation problem can be found, as the acoustic contrast is too low to provide the required speech intelligibility contrast. Any increase in the programme level would unacceptably raise the dark zone intelligibility, and any increase in the masking signal level would result in excessive degradation to the programme signal in the bright zone. At a broadband acoustic contrast level of 9.2 dB, a feasible pair of signals is found; this combination results in both intelligibility constraints being met simultaneously. With this system configuration the required energy of the programme and masker signals are respectively 14 and 18 dB greater than the background noise level, potentially raising the likelihood that the dark zone sound field will cause

noise annoyance, compared to designs with more acoustic contrast and lower required signal levels.

At higher levels of acoustic contrast, the optimal programme level plateaus at 4 dB above the ambient noise level and the optimal masking signal level decreases at a rate of 2 dB for each increase of 1 dB in the broadband acoustic contrast. This gradient is observed because increasing the acoustic contrast affects both of the sound zoning processes; less speech is leaked from the bright zone into the dark zone, and less masker is leaked from the dark zone into the bright zone. The optimal programme level must remain constant in order to satisfy the bright zone intelligibility constraint in the presence of the constant background noise, so this allows the level of the masking signal to be doubly reduced. The gradient continues until the optimal masking signal level falls below the background noise level, at a broadband acoustic contrast level of 13.2 dB for this example.

After this transition point, the masking signal level falls sharply whilst the programme level remains constant, in order to overcome the background noise. The green shaded region denotes the range of systems that provide sufficient separation between zones for the masking signal to be omitted entirely. In all systems with lower acoustic contrast levels than this threshold value, the optimisation process yields pairs of signals that just meet the intelligibility constraints, i.e. $\mathrm{SII}_b = s_b$ and $\mathrm{SII}_d = s_d$. However, systems within the green region earn an additional degree of freedom with regard to the intelligibility constraints in each zone. Maintaining the programme level will result in a further improvement to privacy as the broadband acoustic contrast is increased, or alternatively, the programme level can be allowed to increase in order to improve bright zone intelligibility, whilst maintaining the previously set intelligibility limit in the dark zone. The latter approach is taken throughout this article.

State-of-the-art speech privacy control systems that emit both speech and masking signals, but do not account for the masking effect of the background noise, such as those described by Donley et al. [8] could also be represented in Figure 8. Such systems implicitly assume that the background noise is negligible and therefore must emit louder masking signals to achieve the same intelligibility constraints, compared to systems designed using the proposed approach. Conventional systems therefore have increased potential for noise annoyance, and additionally, in environments with significant background noise, risk poor speech intelligibility in the bright zone.

The difference in broadband acoustic contrast between the edges of the

red and green boundaries in Figure 8 quantifies the benefit of incorporating additional masking into a private personal audio system, in terms of the level of acoustic contrast that must be provided. In order to achieve the intelligibility constraints of $s_d = 0.05$ and $s_b = 0.60$ without using any additional masking, i.e. relying on the background noise alone, the system must provide a broadband acoustic contrast of 16.6 dB, potentially requiring a loudspeaker array with a significant number of transducers. When artificial masking is included, the minimum acoustic contrast requirement is 9.2 dB, reducing the technical requirements of the loudspeaker array system, but it must be noted that at this extreme, the necessary programme and masker levels significantly exceed the background noise level, potentially resulting in a perceptually unacceptable solution. Nevertheless, Figure 8 shows that there is a continuous trade-off between acoustic contrast requirements and signal levels, meaning that other target points on the curves could be selected based on operational requirements and/or listener preferences. Two examples would be to set the target acoustic contrast value to the point where the required masking signal level equals that of the programme, at 11.3 dB of broadband acoustic contrast, or when the masking signal level matches the background noise level, which for this example is at a broadband acoustic contrast value of 13.2 dB.

As a frequency-averaged acoustic contrast level is used to identify each system in the analysis above, the exact location of these transition points is also dependent on the frequency distribution of the acoustic contrast. For example, a system with a very high level of acoustic contrast across a narrow frequency band may exhibit worse performance with regard to privacy provision compared to a system with the same broadband average contrast, spread over the speech frequency range. The role of the system designer is to curate a set of acoustic contrast profiles that are achievable within the bounds of the design problem, akin to the set of curves shown in Figure 7, and then, with information from a background noise survey, choose a target profile from this selection based on the predicted speech and masker levels. The impact of the background noise statistics could be automatically included in this design process using the method proposed by Lee et. al. [15], which allows the desired shape of the frequency-dependent acoustic contrast to be informed by the masking properties of the masking signals in a given scene. A further constraint on this design problem is provided by the desired levels of speech intelligibility in each zone, and this effect is discussed in the following section.

3.2. Varying Speech Intelligibility Constraints

When alternative intelligibility constraints are placed on the zonal sound fields, different threshold values for the minimum required acoustic contrast levels are found. Figure 9 displays the optimal programme and masker signal levels for two different combinations of intelligibility constraints. Figure 9a shows how the acoustic contrast requirements change when the dark zone constraint, s_d, is relaxed from an SII value of 0.05 to 0.1, i.e. an increase in the permissible level of intelligibility in the dark zone. Figure 9b shows the effect of lowering the bright zone constraint, s_b, from 0.60 to 0.50, thereby accepting a lower level of bright zone intelligibility, compared to the case shown in Figure 8. Relaxing the intelligibility constraints in either of these two ways reduces the minimum acoustic contrast levels that are required, whether or not additional masking is used. These thresholds are indicated by the edges of the red and green regions. Additionally, the range of allowable acoustic contrast levels between these two boundaries decreases in size as the constraints are relaxed. This indicates that the more onerous the intelligibility constraints, the greater the advantage of providing additional, artificial masking, in terms of the reduced technical requirements for providing acoustic contrast.

When the dark zone intelligibility constraint is relaxed compared to the reference scenario presented in Figure 8, the result is a slight reduction in the programme and masker levels required at low acoustic contrast levels, e.g. at a broadband acoustic contrast level of 10 dB in Figure 9a, the signal levels are approximately 1 dB lower than the corresponding points in Figure 8. At this low level of contrast, the majority of the privacy provision is due to the additional masking noise, rather than the background noise in the space. As a higher level of speech intelligibility can be tolerated in the dark zone, both the programme and masker signals can be decreased in level compared to the reference scenario with $s_d = 0.05$. On the other hand, when the bright zone constraint is relaxed from $s_b = 0.60$ to $s_b = 0.50$, as shown in Figure 9b, the main effect is a reduction in the required programme level in systems with higher levels of acoustic contrast. Such systems have lower required masker levels and less secondary leakage of the masker into the bright zone, such that the predominant source of masking in the bright zone is from the background noise. Correspondingly, the programme level can be reduced to reflect the lower minimum intelligibility level stipulated by the bright zone constraint, s_b.

3.3. Variation in Background Noise Level

Although the background noise level in a space is more temporally stable than the overall ambient noise in typical environments, the background noise is likely to vary over the course of a day, due to changes in occupancy levels, ventilation settings and nearby traffic flow, for example. BS 4142 [19] reinforces that the background noise level is a fluctuating parameter and that obtaining a representative background noise level for the purposes of a noise survey may require statistical analysis of several measurement periods across a day, each usually no shorter than 15 minutes in length. In the context of private audio system design, basing the background noise estimate on these long-term $L_{A90,15\text{min}}$ readings risks underestimating the masking effect of the background noise across shorter periods of time, such as the duration of a spoken sentence or conversation. Selecting appropriate measurement times is therefore dependent on the individual installation environment - once this information has been gathered, a required level of acoustic contrast can be specified based on the maximum and minimum expected levels of background noise. This target acoustic contrast value can then be used during the process of designing a loudspeaker array and/or specifying the locations of the listening zones, as both of these factors have an impact on the acoustic contrast.

A slow fluctuation in the background noise level has been observed in studies of open plan office sound masking systems [37, 38], and in response, systems to schedule the level of masking, or slowly adapt it based on measured background levels, are available commercially [39]. Similar techniques could be employed in speech privacy control systems to reduce the level of additional masking input into the space, and to control the programme level to ensure good intelligibility for the target listener. In locations where significant fluctuation in the background noise level is common, system integrators should opt for designs where the majority of the masking signal level is directly controlled by the system, as opposed to relying on the masking provided by the background noise, as this provides maximum reliability for the central claim of speech privacy [40].

To illustrate how changes in the background noise level affect the specification of acoustic contrast levels and the optimisation of signal levels, two examples are shown in Figure 10, where the optimal signal levels required to meet intelligibility constraints of $s_b = 0.6$ and $s_d = 0.05$ are predicted using two further environments from the ARTE database, corresponding to a cafe and a train station, which have $L_{A90,2.5\text{min}}$ values of 59.4 dBA and 67.2

dBA respectively. Across the range of tested systems, the predicted optimal signal levels increase with the background noise level, but the characteristic shape of each pair of curves remains similar to those predicted for the quieter *Church 2* scene, shown in Figure 8. The *Train Station* scene has an L_{A90} value 13 dB greater than that measured in the *Church 2* scene, but the two boundaries, at which the system becomes feasible, and where the system may be configured to omit additional masking noise, are found to be at levels of broadband acoustic contrast less than 1 dB higher than in the quieter *Church 2* environment. Comparison of the results presented in Figures 9 and 10 suggests that the level of acoustic contrast that a system must provide depends more strongly on the chosen speech intelligibility constraints than the background noise level. The zonal filtering process is a linear filtering operation, and as such, changes to the background noise level do not affect the relative signal to noise ratios within each zone. Instead, the slight increase in required acoustic contrast levels in areas with more background noise can be attributed to nonlinearities in the SII algorithm. Even without the degrading effect of ambient noise, speech reproduced at an unnaturally high level is judged to be less intelligible than speech produced at regular conversation levels [17]. To overcome this effect, slightly higher levels of acoustic contrast are required from systems installed in noisier environments.

4. Conclusions

The privacy of a communications system can be improved by using an array of loudspeakers to focus speech towards a target listener, and by ensuring that any leakage of this speech into other areas is sufficiently masked. The present article has discussed how the ambient noise in a reproduction environment can be leveraged by such a system to provide a proportion of this masking, thereby reducing the level of artificial masking that must be emitted into the environment. The proposed system optimisation process demands that certain speech intelligibility targets are met in each listening zone, based on estimates of the signals emitted by the array, and measurements of the ambient noise. The degree of energetic masking provided by the ambient noise in each listening zone can be complex to predict, as public spaces typically contain many distinct noise sources, leading to temporal and spatial variation in the composition of the dark zone sound field. Analysis of the background noise component of the ambient noise in a range of recorded environments has revealed that this component is typically more temporally

stable and spatially diffuse, rendering it a more reliable and effective source of masking than the unprocessed ambient noise.

The results presented in this paper show that a mixture of artificial masking and background noise can be used to provide a private sound zoning system. As well as showing the minimum levels of acoustic contrast that are required to obtain a feasible system and to omit the artificial masker altogether, the reliance of each system on the background noise can be evaluated by comparing the relative levels of the artificial masker and background noise. This highlights a practical trade-off in the design of private audio systems for use in noisy environments. The optimisation process described in Section 3 minimises the level of the artificial masker in order to improve the perceived acceptability of the dark zone sound field. By consequence, systems with higher levels of acoustic contrast are implicitly more reliant on accurate assessment of the masking provided by the background noise. Errors in this process could compromise the privacy of the target listener. Less capable systems (in terms of their acoustic contrast performance) can achieve identical speech intelligibility levels within each zone, compared to a system with a higher acoustic contrast performance, by increasing the level of artificial masking, a practice that risks the system being regarded as a source of noise pollution. Two suggestions for compromises between these two extremes have been proposed. In one example, the level of acoustic contrast can be specified to yield equal levels of speech and masker within their respective zones. With a higher level of acoustic contrast, the masking signal can be attenuated to match the background noise level.

Long-term changes in the background noise level, for example across the course of a day, have a significant impact on the level at which the speech and masking signals must be reproduced. However, this variation has been shown to have relatively little impact on the level of acoustic contrast that must be provided for a given level of privacy performance. Much more significant in this regard is the specification of speech intelligibility targets in each zone. As the desired speech intelligibility contrast increases, the broadband acoustic contrast requirements also increase, particularly for the threshold at which the additional masking signal can be omitted entirely.

Further experimentation is necessary to precisely determine the detrimental effect of additional, artificial noise in environments that are already noisy. This may include the application of more complex speech intelligibility metrics than the SII to evaluate the effect of time-varying and directional noise on speech intelligibility, and the conduction of listening tests to assess the degree

of privacy achieved by the proposed systems in different environments. Previous experiments by the authors on similar systems have shown that when the masker is the dominant noise source in an environment, its acceptability is strongly correlated with its loudness [9]. When artificial masking is used in combination with ambient noise, the acceptability is likely to also depend on the spatial, spectral and temporal content of the masker, evaluated in the context of the surrounding environment.

References

[1] W. F. Druyvesteyn, J. Garas, Personal Sound, Journal of the Audio Engineering Society 45 (9) (1997) 685–701.
URL http://www.aes.org/e-lib/inst/browse.cfm?elib=7843

[2] M. F. Simon Galvez, S. J. Elliott, J. Cheer, Personal Audio Loudspeaker Array as a Complementary TV Sound System for the Hard of Hearing, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences E97.A (9) (2014) 1824–1831. doi:10.1587/transfun.E97.A.1824.

[3] J.-H. Chang, C.-H. Lee, J.-Y. Park, Y.-H. Kim, A realization of sound focused personal audio system using acoustic contrast control., The Journal of the Acoustical Society of America 125 (4) (2009) 2091–2097. doi:10.1121/1.3082114.

[4] K. R. Baykaner, C. Hummersone, R. Mason, S. Bech, The acceptability of speech with interfering radio program material, in: Proc. 136th Audio Engineering Society Convention, Berlin, 2014, pp. 1–9.
URL http://www.aes.org/e-lib/browse.cfm?elib=17167

[5] J. Cheer, S. J. Elliott, M. F. S. Gálvez, Design and implementation of a car cabin personal audio system, AES: Journal of the Audio Engineering Society 61 (6) (2013) 412–424.

[6] S. J. Elliott, M. Jones, An active headrest for personal audio, The Journal of the Acoustical Society of America 119 (5) (2006) 2702. doi:10.1121/1.2188814.

[7] S. J. Elliott, J. Cheer, H. Murfet, K. R. Holland, Minimally radiating sources for personal audio, The Journal of the Acoustical Society of America 128 (4) (2010) 1721–1728. doi:10.1121/1.3479758.

[8] J. Donley, C. H. Ritz, W. B. Kleijn, Multizone Soundfield Reproduction With Privacy and Quality Based Speech Masking Filters, IEEE/ACM Transactions on Audio Speech and Language Processing 26 (6) (2018) 1037–1051. doi:10.1109/TASLP.2018.2798804.

[9] D. Wallace, J. Cheer, Design and evaluation of personal audio systems based on speech privacy constraints, Journal of the Acoustical Society of America 147 (4) (2020) 2271–2282. doi:10.1121/10.0001065.

[10] A. Chaman, Y. Liu, J. Casebeer, I. Dokmanić, Multipath-enabled Private Audio with Noise, 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, (2019), pp. 685-689. doi: 10.1109/ICASSP.2019.8683045.

[11] J.-W. Choi, Y.-h. Kim, Generation of an acoustically bright zone with an illuminated region using multiple sources, The Journal of the Acoustical Society of America 111 (4) (2002) 1695–1700. doi:10.1121/1.1456926.

[12] O. Kirkeby, P. A. Nelson, Reproduction of plane wave sound fields, The Journal of the Acoustical Society of America 94 (5) (1993) 2992–3000. doi:10.1121/1.407330.

[13] F. Olivieri, F. M. Fazi, S. Fontana, D. Menzies, P. A. Nelson, Generation of Private Sound with a Circular Loudspeaker Array and the Weighted Pressure Matching Method, IEEE/ACM Transactions on Audio Speech and Language Processing 25 (8) (2017) 1579–1591. doi:10.1109/TASLP.2017.2700945.

[14] P. J. Jackson, F. Jacobsen, P. Coleman, J. Abildgaard Pedersen, Sound field planarity characterized by superdirective beamforming, in: Proceedings of Meetings on Acoustics, Vol. 19, Acoustical Society of America, Montreal, 2013. doi:10.1121/1.4800877.

[15] T. Lee, J. K. Nielsen, M. G. Christensen, Signal-Adaptive and Perceptually Optimized Sound Zones With Variable Span Trade-Off Filters, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020). doi:10.1109/TASLP.2020.3013397.

[16] D. Wallace, J. Cheer, Combining artificial and natural background noise in personal audio systems, in: Proceedings of the IEEE Sensor Array

and Multichannel Signal Processing Workshop, Vol. 2018-July, 2018. doi:10.1109/SAM.2018.8448847.

[17] ANSI, ANSI/ASA S3.5-1997 (R2017) Methods for Calculation of the Speech Intelligibility Index (1997).

[18] J. Cheer, S. J. Elliott, E. Oh, J. Jeong, Application of the remote microphone method to active noise control in a mobile phone, The Journal of the Acoustical Society of America 143 (4) (2018) 2142–2151. doi:10.1121/1.5031009.

[19] BSI, BS 4142:2014+A1:2019 Methods for rating and assessing industrial and commercial sound (2019).

[20] M. Cooke, A glimpsing model of speech perception in noise, The Journal of the Acoustical Society of America 119 (3) (2006) 1562–1573. doi:10.1121/1.2166600.

[21] A. Weisser, J. M. Buchholz, C. Oreinos, J. Badajoz-Davila, J. Galloway, T. Beechey, G. Keidser, The Ambisonic Recordings of Typical Environments (ARTE) database, Acta Acustica united with Acustica 105 (4) (2019) 695–713. doi:10.3813/AAA.919349.

[22] H. Fastl, E. Zwicker, Psychoacoustics: Facts and models, 3rd Edition, Springer, 2007. doi:10.1007/978-3-540-68888-4.

[23] A. W. Bronkhorst, The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple Talker Conditions, Acta Acustica united with Acustica 86 (1) (2000) 117–128.

[24] J. Swaminathan, C. R. Mason, T. M. Streeter, V. Best, E. Roverud, G. Kidd, Role of Binaural Temporal Fine Structure and Envelope Cues in Cocktail-Party Listening, Journal of Neuroscience 36 (31) (2016) 8250–8257. doi:10.1523/JNEUROSCI.4421-15.2016.

[25] W. A. Yost, Spatial release from masking based on binaural processing for up to six maskers, The Journal of the Acoustical Society of America 141 (3) (2017) 2093–2106. doi:10.1121/1.4978614.

[26] G. L. Jones, R. Y. Litovsky, A cocktail party model of spatial release from masking by both noise and speech interferers, The Journal of the Acoustical Society of America 130 (3) (2011) 1463–1474. doi:10.1121/1.3613928.

[27] N. I. Durlach, Equalization and Cancellation Theory of Binaural Masking-Level Differences, The Journal of the Acoustical Society of America 35 (8) (1963) 1206–1218. doi:10.1121/1.1918675.

[28] A. W. Bronkhorst, R. Plomp, Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing, The Journal of the Acoustical Society of America 92 (6) (1992) 3132–3139. doi:10.1121/1.404209.

[29] M. L. Hawley, R. Y. Litovsky, J. F. Culling, The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer, The Journal of the Acoustical Society of America 115 (2) (2004) 833–843. doi:10.1121/1.1639908.

[30] J. Peissig, B. Kollmeier, Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners, The Journal of the Acoustical Society of America 101 (3) (1997) 1660–1670. doi:10.1121/1.418150.

[31] B. N. Gover, J. G. Ryan, M. R. Stinson, Microphone array measurement system for analysis of directional and spatial variations of sound fields, The Journal of the Acoustical Society of America 112 (5) (2002) 1980–1991. doi:10.1121/1.1508782.

[32] C. House, S. Dennison, D. G. Morgan, N. Rushton, G. V. White, J. Cheer, S. Elliott, Personal Spatial Audio in Cars Development of a loudspeaker array for multi-listener transaural reproduction in a vehicle, in: Proceedings of the Institute of Acoustics, Vol. 39, 2017.

[33] F. Winter, F. Schultz, G. Firtha, S. Spors, A Geometric Model for Prediction of Spatial Aliasing in 2.5D Sound Field Synthesis, IEEE/ACM Transactions on Audio Speech and Language Processing 27 (6) (2019) 1031–1046. doi:10.1109/TASLP.2019.2892895.

[34] S. J. Elliott, J. Cheer, J.-w. Choi, Y. Kim, Robustness and Regularization of Personal Audio Systems, IEEE Transactions on Audio, Speech and Language Processing 20 (7) (2012) 2123–2133.

[35] M. A. Abramson, Pattern Search Filter Algorithms for Mixed Variable General Constrained Optimization Problems, Ph.D. thesis, Rice University (2002).

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech, IEEE Transactions on Audio, Speech and Language Processing 19 (7) (2011) 2125–2136. doi:10.1109/TASL.2011.2114881.

[37] V. Hongisto, J. Ker, Simple model for the acoustical design of open-plan offices, Acta Acustica united with Acustica 90 (2004) 481–485.

[38] L. Lenne, P. Chevret, J. Marchand, Long-term effects of the use of a sound masking system in open-plan offices: A field study, Applied Acoustics 158 (2020) 107049. doi:10.1016/j.apacoust.2019.107049.

[39] A. L'Espérance, A. Boudreau, L.-A. Boudreault, F. Gariépy, R. Mackenzie, Adaptive Volume Control for Sound Masking, Noise-Con 2017 (2017) 678–686.

[40] R. Pirn, Acoustical Variables in Open Planning, The Journal of the Acoustical Society of America 49 (5A) (1971) 1339–1345. doi:10.1121/1.1912506.

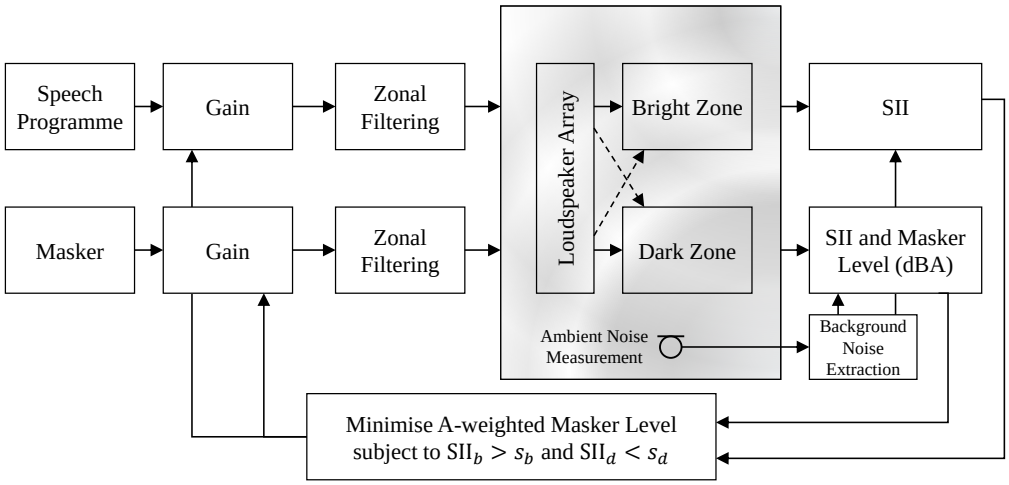


Figure 1: Block diagram of a private personal audio system operating in a noisy environment.

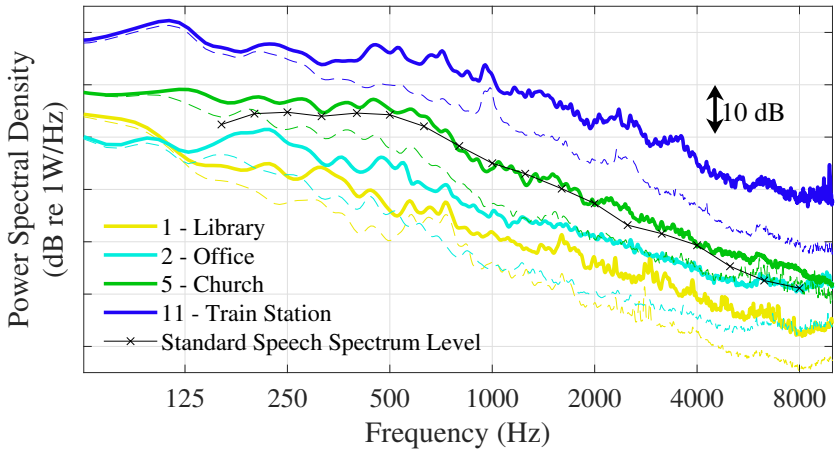


Figure 2: [Colour Online] Power spectral density estimates of four ambient noise signals from the ARTE database (solid lines), and their corresponding background noises (dashed lines). Numbered legend entries refer to their identification in the ARTE database [21].

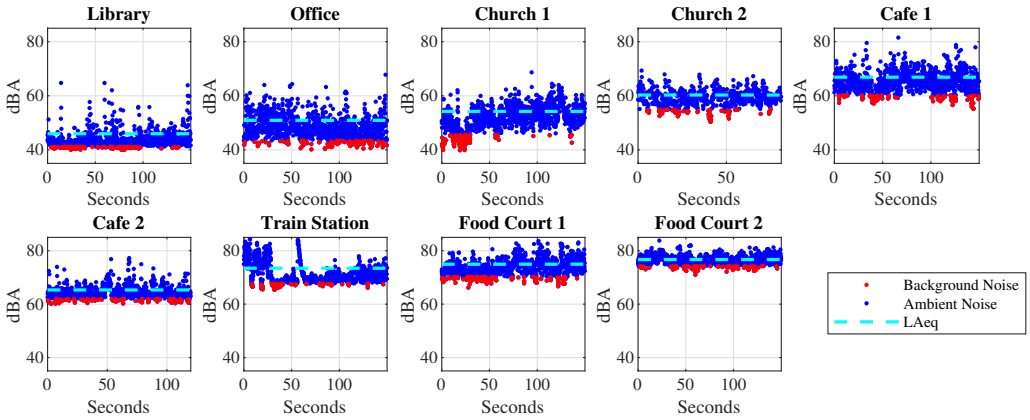


Figure 3: [Colour Online] A-weighted sound pressure level of 125 ms samples of ambient noise from the ARTE database [21]. Samples contributing to the lowest 10% of this dataset (red points) constitute the background noise.

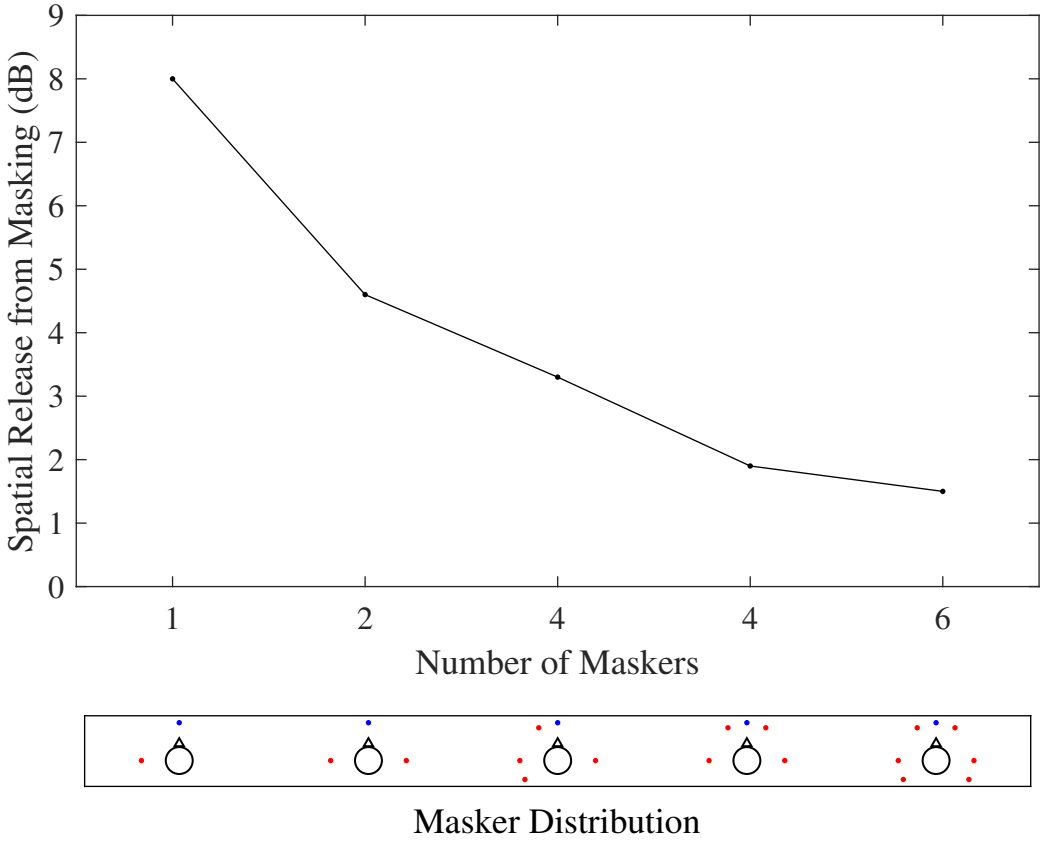


Figure 4: [Colour Online] Spatial Release from Masking (SRM) from multiple maskers (red points) distributed around the listener with respect to a frontal talker (blue points). The abscissa increases with increasing diffuseness of the masking conditions, due to an increasing number or a widened spatial distribution of maskers. Data from Ref. [28].

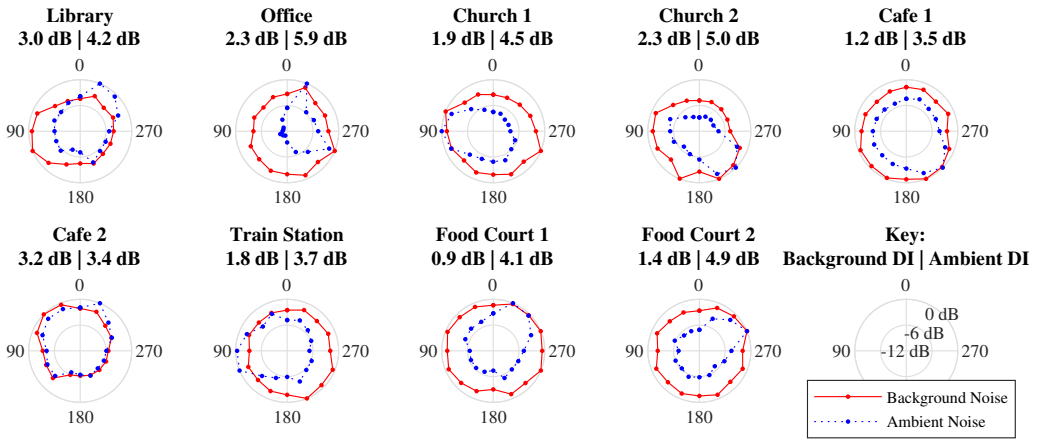


Figure 5: [Colour Online] Horizontal directivity of the background and ambient noise in the 9 public environments of the ARTE database [21]. Decibel values below each figure title refer to the Directivity Index (DI) of the background noise and ambient noise respectively.

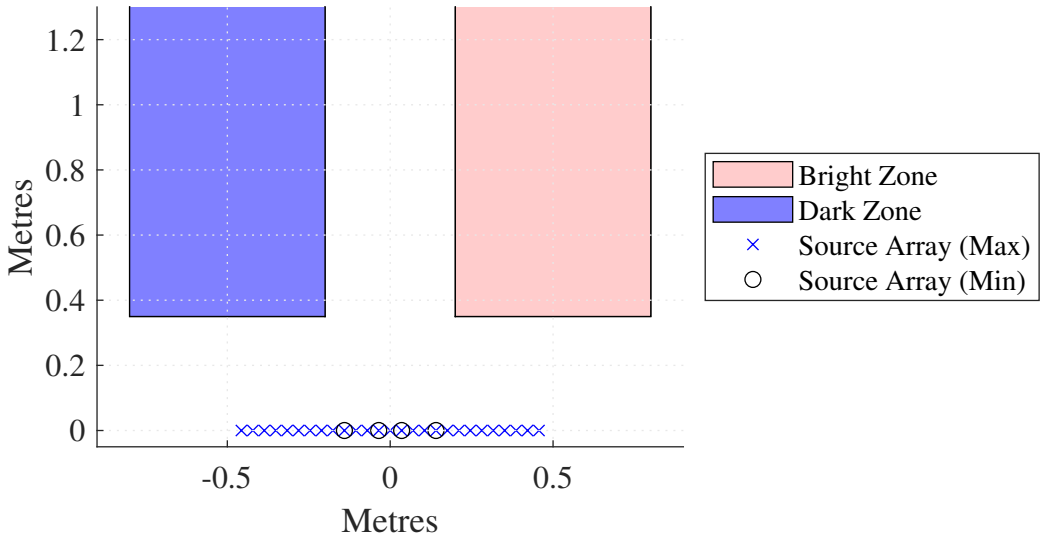


Figure 6: [Colour Online] Location of loudspeaker array and symmetrical bright and dark zones.

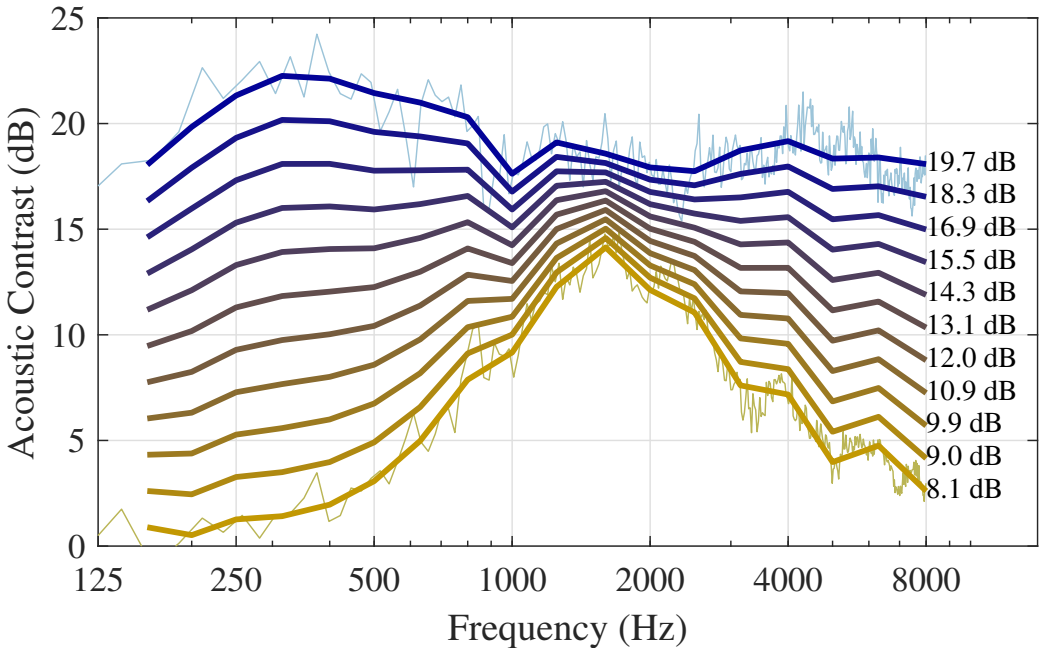


Figure 7: [Colour Online] Series of ⅓-octave band acoustic contrast profiles used to simulate several levels of system performance. The measured narrowband acoustic contrast used to generate the highest and lowest curves are also shown using light lines, and the labels refer to the broadband, frequency averaged acoustic contrast level.

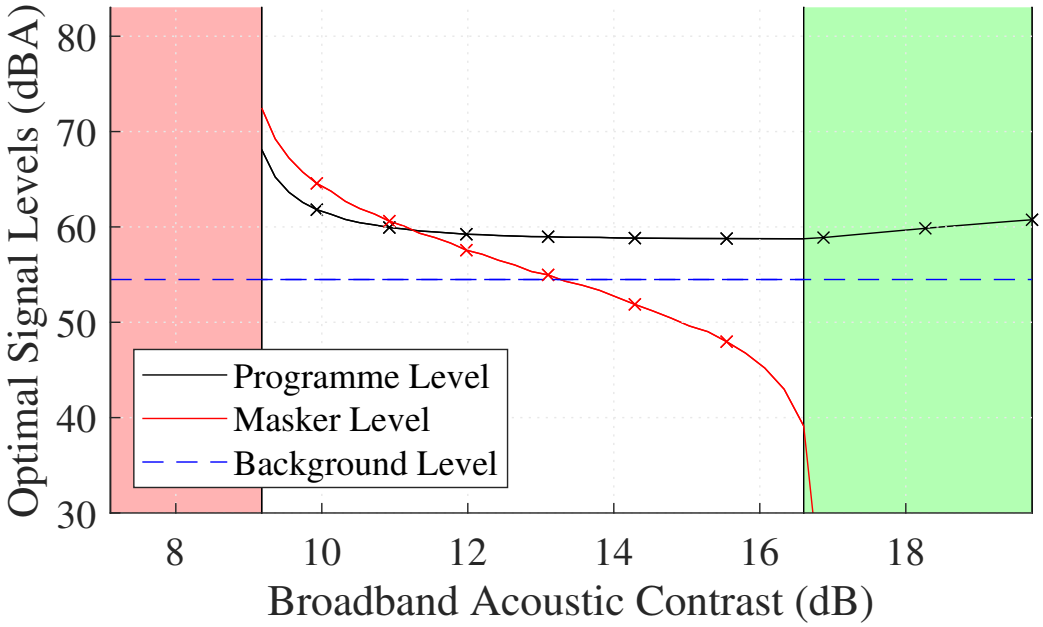
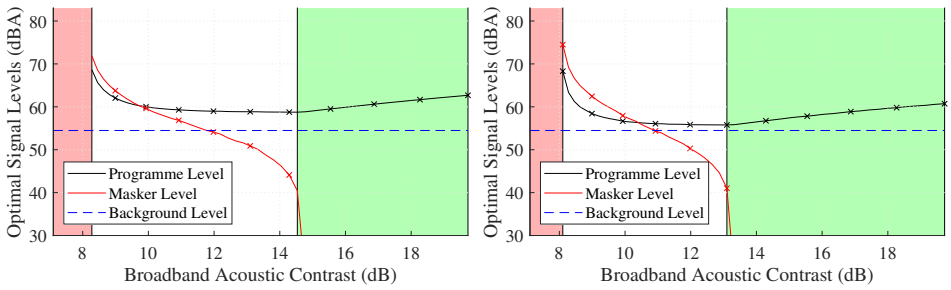


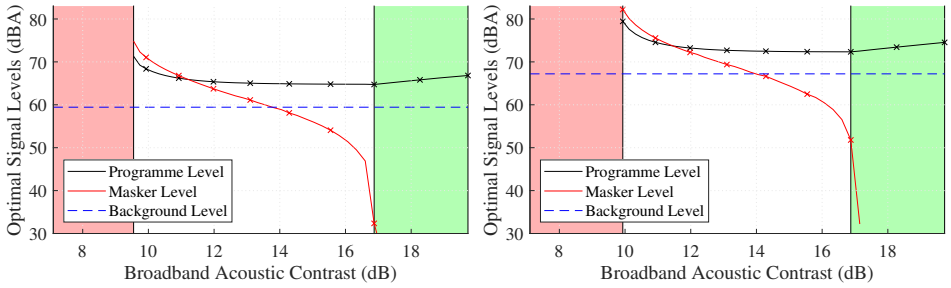
Figure 8: [Colour Online] Predictions of optimal programme and masker signals based on intelligibility constraints of $s_b = 0.60$ and $s_d = 0.05$, for each acoustic contrast profile shown in Figure 7. The background level indicated at 54.5 dBA is from the *Church 2* scene of the ARTE database [21].



(a) $s_b = 0.60$, $s_d = 0.10$.

(b) $s_b = 0.50$, $s_d = 0.05$.

Figure 9: [Colour Online] Predictions of optimal programme and masker signals with the speech intelligibility constraints indicated in each subcaption, for each acoustic contrast profile shown in Figure 7. The background level indicated at 54.5 dBA is from the *Church 2* scene of the ARTE database [21].



(a) *Cafe 1*, $L_{A90} = 59.4$ dBA. (b) *Train Station*, $L_{A90} = 67.2$ dBA.

Figure 10: [Colour Online] Predictions of optimal programme and masker signals in the ARTE environments indicated in each subcaption, for each acoustic contrast profile shown in Figure 7. $s_b = 0.60$ and $s_d = 0.05$ for both plots.