THE DESIGN OF PERSONAL AUDIO SYSTEMS FOR SPEECH TRANSMISSION USING ANALYTICAL AND MEASURED RESPONSES.

D. Wallace, J. Cheer

Institute of Sound and Vibration Research
University of Southampton

ABSTRACT

Personal Audio systems can be used to provide information and entertainment content in public spaces. Limitations in array directivity mean that speech information intended for a target region may remain intelligible elsewhere. This compromises privacy for target listeners and could prove distracting or annoying to passive listeners nearby. A system has previously been proposed whereby the intelligibility of this leaked speech is reduced by radiating an artificial masking signal into the dark zone; this masking signal has been optimised to minimise the potential for annoyance whilst achieving a predefined level of intelligibility in each zone, but only free-field responses were considered. In practice, systems located in public spaces will be adversely affected by noise and reverberation. This detriment to system performance can be quantified using engineering measures such as acoustic contrast, although the perceived performance as evaluated by users does not necessarily correspond. The present paper explores the effect of using analytical and measured transfer responses on speech intelligibility and system optimisation using a practical example of a personal audio system in a room.

***Index Terms*—** Personal Audio, Speech Intelligibility

1. INTRODUCTION

Many applications of zonal audio technology that have seen commercial success were designed for use in public spaces, such as museums, shops and exhibitions [1, 2]. The design requirements for a personal audio system [3] installed in a public space relate to the ability of the device to convey information to the target listener, whilst minimising the potential for annoyance and distraction [4] of others sharing the space. This information may be contained within prerecorded speech, such as the automated voice at a self-service checkout, or the voice of a member of staff behind a security screen at a bank, for instance. In the latter case, it is also desirable for the information content delivered to the intended listener to remain private. In this work and previous work by the authors, this objective is formalised by maximising the difference in speech intelligibility between the sound fields in the bright and dark zones, as evaluated by the Extended Short Time Objective Intelligibility (ESTOI) metric [5]. This speech intelligibility contrast is increased by using the array to radiate a secondary masking signal into the dark zone of the system [6]. The level and spectrum of this signal may be adjusted based on psychoacoustic metrics to reduce the potential for annoyance in the dark zone [7].

Implementations and analyses of sound zoning systems have been conducted in anechoic environments e.g. [8, 9, 10], with the inclusion of reflections from the head [11], individual room reflections [12] and general reverberation e.g. [3, 13, 14, 15]. Of these, the performance study carried out in [13] is of particular relevance to this work as it considers the leakage between programmes in adjacent bright zones in terms of perceived distraction, as well as acoustic contrast. Overall, correlation was found between the perceptual and physical metrics, with higher levels of acoustic contrast resulting in lower levels of distraction, though the strength of this relation varied with different combinations of programme material; interfering speech was found to be the most distracting. Olivieri et. al. [15] present results from informal listening tests which suggest that zoning filters created using free-field responses, as opposed to measurements in anechoic or reverberant conditions, provide subjectively higher audio quality. Although filters generated using the measured responses produced greater directivity than when using a free-field assumption, the perceived channel separation was similar.

The present paper investigates how the choice of transfer responses used to generate the zoning filters affects the speech intelligibility, for a system situated in a well-damped room. The effect of regularisation on the speech intelligibility is also explored in each case. Section 2 describes the system under test, the results are presented in Section 3 and concluding remarks are made in Section 4.

2. SYSTEM DESIGN

The personal audio system used in this work comprises of a 27 channel linear loudspeaker array with alternate drivers offset ± 15.2 mm vertically from a horizontal centreline and spaced

This work was supported by an EPSRC Doctoral Training Centre grant (EP/L015382/1)

horizontally at intervals of 35.1 mm; the array was previously described in [16]. The array was positioned in an acoustically treated room with dimensions 4.4 x 3.7 x 2.3 metres, and a mid-frequency reverberation time of $T_{mf} = 0.11$s, defined as the arithmetic average of 500 Hz, 1 kHz and 2 kHz octave band reverberation times. Impulse responses were captured from the loudspeaker array to a 20-channel array of microphones spaced on a 72 mm grid using a ten-second logarithmic sine sweep ($20\text{Hz} - 20\text{kHz}, f_s = 48\text{kHz}$) from each driver in turn [17]. When used to produce transfer responses, measured impulse responses were truncated after the reverberation time to reduce the effect of low level noise on the zoning filters.

Fig. 1 shows the configuration of the loudspeaker and microphone array during one bright zone measurement. The microphones in the grid are arranged in two groups of ten which both span the overall dimensions of the grid. This allows one set of impulse responses to be used in the optimisation of the zoning filters, whilst the other set can be used to evaluate the performance of the system, thus avoiding bias in the prediction of acoustic contrast and other metrics [13, 18]. Fig. 2 shows a plan view of the loudspeaker and microphone arrays.

Assuming linearity and time-invariance, playback through the array is simulated by first convolving the programme material with the zoning filters and then convolving these filtered signals with the measured impulse responses, before summing the contributions from each loudspeaker at each microphone. For example, in the frequency domain, each microphone signal $\mathbf{y}_j$ in the bright zone can be represented as

$$\mathbf{y}_j = \sum_{i=1}^{L} \mathbf{G}_{Bij} \mathbf{q}_i x, \tag{1}$$

where x is the input signal, $\mathbf{q}_i$ is the zoning filter applied to the i^{th} loudspeaker and $\mathbf{G}_{Bij}$ is the transfer response from the i^{th} loudspeaker to the j^{th} microphone in the bright zone.

The programme signal is a 30 second sample of sentences from the CSTR VCTK Corpus [19], spoken by a range of male and female British English speakers and the masking signal is speech shaped noise generated from the programme sample, low-pass filtered at 3 kHz. This masker provides a trade-off between effective masking of speech and low perceived annoyance according to Zwicker and Fastl's Psychoacoustic Annoyance metric [7, 20].

Fig. 3 shows an example of a measured impulse response from the centre of the loudspeaker array to a microphone near to the centre of the bright zone and the corresponding transfer response magnitude. The corresponding simulated free-field response is also shown, which decreases at high frequencies as the ideal impulse lies between samples in the digital representation.

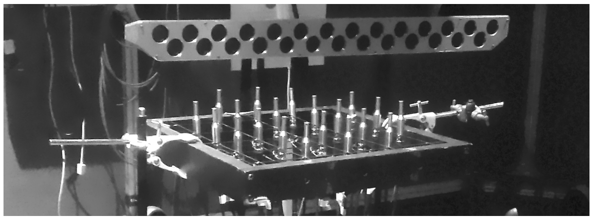

Fig. 1. Foreground: 20-channel microphone array; Background: 27-channel loudspeaker array.

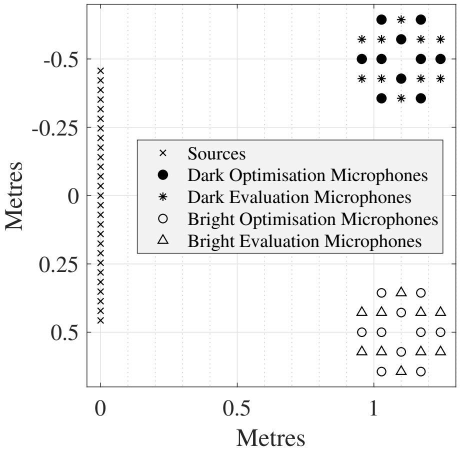
-0.5
-0.25
0
0.25
0.5
Metres
0
0.5
1
Metres
× Sources
● Dark Optimisation Microphones
✳ Dark Evaluation Microphones
○ Bright Optimisation Microphones
△ Bright Evaluation Microphones

Fig. 2. Plan view of bright and dark zones relative to loudspeaker array. Microphone and loudspeaker array centres are 1.22 metres above floor level.

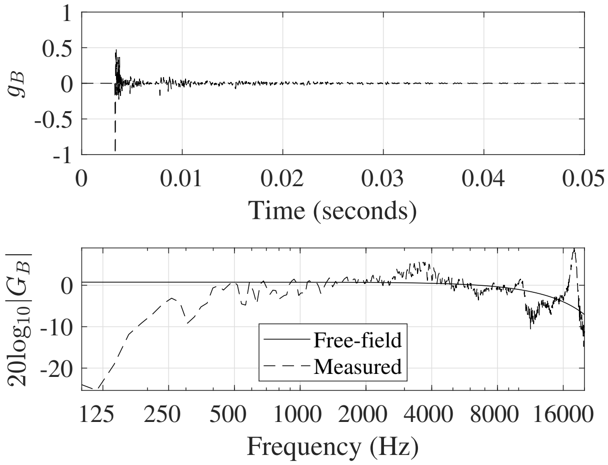

Fig. 3. Upper panel: Example measured impulse response from the centre loudspeaker to the centre of the bright zone. Lower panel: Corresponding measured transfer response magnitude (dashed line) with free-field transfer response for comparison.

3. RESULTS

Results are reported for two different approaches to calculating the zoning filters. In the first case, the geometric positions of the microphones and loudspeakers are used to simulate the free-field transfer responses. A baffled piston approximation is used for the loudspeakers, and microphones are assumed to be compact omnidirectional receivers. In the second instance, measured transfer responses are used to design the filters.

Acoustic contrast control is used to design the zoning filters, following the 'indirect method' from Section II.b. of [9]. According to this method, filters $\mathbf{q}$ are optimised by selecting the eigenvector corresponding to the largest eigenvalue of $[\mathbf{G}_D^H\mathbf{G}_D + \beta\mathbf{I}]^{-1}[\mathbf{G}_B^H\mathbf{G}_B]$, where $\mathbf{G}_B$ and $\mathbf{G}_D$ are the electroacoustical response matrices from the source array to the microphones in the bright and dark zones respectively, β is a regularisation parameter and $\{\}^H$ denotes the Hermitian transpose. At each frequency $\beta = \beta_0\kappa$, where κ is the condition number of $[\mathbf{G}_D^H\mathbf{G}_D]$. The proportionality constant β_0 is varied to provide a range of levels of regularisation.

In the acoustic contrast control filter design process, the matrix $\mathbf{G}_D^H\mathbf{G}_D$ must be inverted. The close proximity of the loudspeakers and microphones within the source and receiver arrays can cause $\mathbf{G}_D^H\mathbf{G}_D$ to be ill conditioned. Regularisation reduces the condition number of this matrix to improve numerical stability, but has additional audible effects, such as flattening the frequency response in the bright zone and reducing the achievable level of acoustic contrast.

In order to maintain consistency between bright zone signals under different regularisation conditions, a 1/3 octave band equaliser is applied to the programme signal to equalise the programme in the bright zone to approximate the transfer response magnitude of a single loudspeaker (Fig. 3); the use of a 1/3 octave band equaliser is a simplistic means to avoid over-equalisation of narrow bands which may lead to poor robustness. Further to this, the input signal level is set such that the sound pressure level of the programme in the bright zone is 60 dB SPL. The level of the masking signal is adjusted iteratively to give an average ESTOI level of 0.05 averaged across microphones in the dark zone. Informal listening tests confirm that at this level of degradation, speech may be considered essentially unintelligible.

3.1. Acoustic Contrast

Acoustic contrast is defined as the ratio of the mean squared pressures in the bright and dark zones, and is calculated at each frequency as

$$AC = 10\log_{10}\left(\frac{\mathbf{q}^H\mathbf{G}_B^H\mathbf{G}_B\mathbf{q}}{\mathbf{q}^H\mathbf{G}_D^H\mathbf{G}_D\mathbf{q}}\right). \qquad (2)$$

Fig. 4 shows the variation in measured acoustic contrast when free-field (upper panel) and measured (lower panel) transfer responses are used in the bright zone filter design process for different levels of regularisation.

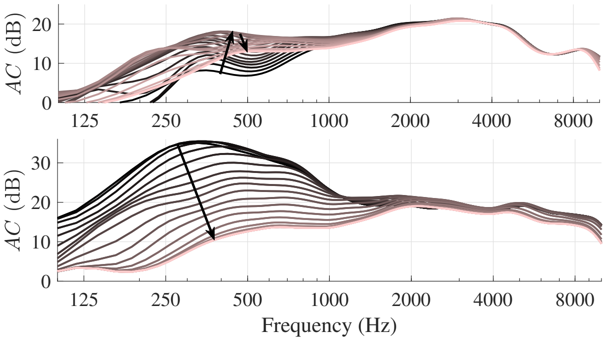

Fig. 4. Acoustic contrast with an increasing β_0 from 10^{-28} to 10^{-16} indicated by colour change from dark to light. Arrows indicate increasing regularisation. Upper plot: Free-field filters; Lower plot: Measured filters.

The lower panel in Fig. 4 shows the variation in acoustic contrast with different levels of regularisation when the filters are calculated using measured transfer responses. At regularisation levels below $\beta_0 = 10^{-28}$, the transfer responses for which the filters are optimised match the environment closely, so a high level of acoustic contrast is achieved. Exact matching is prevented through the use of separate optimisation and evaluation microphones (Fig. 2). As regularisation increases, a mismatch between the responses assumed in the filter calculation and the actual response in the room is introduced, so acoustic contrast decreases, particularly at frequencies below 2 kHz. Above $\beta_0 = 10^{-16}$, this regularisation term dominates $[\mathbf{G}_D^H\mathbf{G}_D]$, reducing the acoustic contrast control to the simpler brightness control method [21], that is, maximising the level in the bright zone alone. This is confirmed by noting the similarity between the lightest lines in the upper and lower panels of Fig. 4.

From the upper panel in Fig. 4, which shows the performance achieved when using free-field responses to calculate the filters, it can be seen that the maximum level of contrast is lower than that achieved by the filters designed using the measured responses, regardless of the regularisation level. This difference in the upper performance limit is because while the free-field responses only contain the time of arrival of the direct sound field the measured transfer responses contain information about the time of arrival of both direct sound and early reflections in the impulse response, as shown in Fig. 3, as well as a component due to diffuse reverberation.

The upper panel of Fig. 4 also shows that when β_0 is low, the acoustic contrast is low compared to the corresponding results presented in the lower panel calculated using the measured responses, particularly in low- to mid-frequencies. This is due to the more significant mismatch between the free-field responses used in the filter design process and the physical room responses. As the system is regularised, the effect is equivalent to adding a random component to the free-field transfer responses [9, 14], which approximates the diffuse reverberation component in the measured responses. The acoustic contrast achieved by the filters designed using the free-field responses is maximised when the level of regularisation is sufficient to ensure robustness to the difference between the free-field and physical responses. Continuing to increase regularisation beyond this point results in a transition to brightness control, as in the case using measured filters.

3.2. Effect of Regularisation on Speech Intelligibility

It has been shown in the previous section that the system optimised using the measured responses is capable of achieving a higher level of acoustic contrast than when the free-field responses are used. However, of primary importance to the design of personal audio systems for the reproduction of information content is the intelligibility of the speech in the bright zone. Fig. 5 shows ESTOI, averaged across bright zone evaluation microphones with different levels of regularisation used in the control process which creates the bright zone. To maximise the bright zone intelligibility, the zonal filters must effectively reduce the cross-talk between zones, but must not introduce distortions to the temporal fine structure of the target speech signal. The regularisation level for the zoning process which focusses the masker into the dark zone is fixed at the level which gives maximum acoustic contrast with each type of transfer response, since preserving the temporal structure of the masking noise is of no benefit.

Firstly, the performance of the system is considered while operating in quiet, with no masking signal. These ideal conditions give the upper limit for the bright zone intelligibility when each type of response is used in the filter design process. The black traces in Fig. 5 show that when using both free-field responses (solid line) and measured responses (dashed line), the general trend shows a gradual increase in ESTOI with β_0. With no additional noise in the environment to degrade intelligibility, the distortion to the speech signal must be attributed to the filters themselves. As regularisation increases, the level of this distortion decreases and ESTOI improves, but the value attained with measured responses never exceeds that when free-field responses are used. This is the first indication that although higher levels of acoustic contrast are predicted when using measured responses, the maximum attainable intelligibility using this method is not necessarily higher than when using free-field data in the design of the zoning filters.

The grey traces in Fig. 5 show how ESTOI in the bright zone depends on β_0 when a masking signal is directed into the dark zone, and adjusted to provide ESTOI = 0.05 in that region. Leakage of the masker into the bright zone decreases the intelligibility level compared to the case where no masker is present. This difference is most significant for the filters designed using free-field responses.

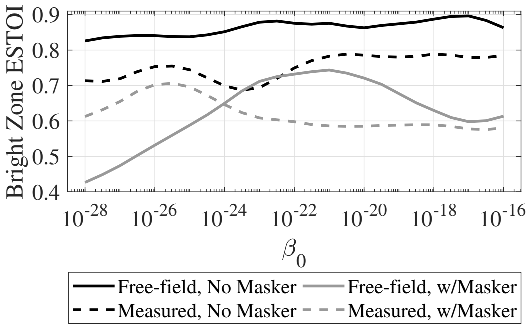

Fig. 5. ESTOI of programme material in the bright zone with and without masking, using filters constructed using free-field and measured responses

The dashed traces in Fig. 5 represent the bright zone ESTOI with masking (grey) and without masking (black), using the filters derived from the measured responses. At low regularisation, the intelligibility is similar to the case with no masker. This further supports the indication that the audible distortion caused by the filters themselves limits bright zone intelligibility and outweighs the benefit of additional acoustic contrast at these low levels of β_0. As regularisation increases, acoustic contrast decreases (as shown in Fig. 4) and consequently the required masking signal level increases, resulting in reduced intelligibility in the bright zone caused by leakage of the masker from the dark zone into the bright zone.

The maximum values of the grey traces in Fig. 5 show that slightly higher intelligibility is predicted when using free-field responses in the filter design process, compared to measurements. With free-field responses, ESTOI values greater than 0.7 (the maximum value when measured responses are used) can be achieved over a range of β_0 that spans four orders of magnitude, indicating relative insensitivity to the choice of regularisation parameter.

4. CONCLUSIONS

Personal Audio systems designed for conveying speech information may be designed using analytical responses based on the geometrical positions of the source array and listening zones, or measured transfer responses from the room in which playback will occur. When evaluated in terms of the maximum achievable acoustic contrast alone, filters derived from measured data perform better due to the close matching between the optimisation and the playback environments. However, when the ESTOI algorithm is used to assess the intelligibility of speech in the bright zone, zoning filters based on regularised free-field responses are preferred as they offer similar levels of intelligibility alongside the obvious advantages of simplicity in implementation, robustness to changes in room reverberation and cost effectiveness in production when compared to measured responses.

5. REFERENCES

[1] OgilvyNewZealand, “All Good Bananas - Listen to your conscience,” Internet: goo.gl/pjKUiW 02/03/2011 [03/10/2018].

[2] Holosonics Research Labs Inc., “Press Release,” Internet: goo.gl/Pvf49F 08/05/2014 [03/10/2018].

[3] W F Druyvesteyn and J Garas, “Personal Sound,” *Journal of the Audio Engineering Society*, vol. 45, no. 9, pp. 685–701, 1997.

[4] Jon Francombe, Russell Mason, Martin Dewhirst, and Søren Bech, “Elicitation of attributes for the evaluation of audio-on-audio interference,” *The Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2630–2641, 2014.

[5] Jesper Jensen and Cees H. Taal, “An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[6] Jacob Donley, Christian H. Ritz, and W. B. Kleijn, “Multizone Soundfield Reproduction With Privacy and Quality Based Speech Masking Filters,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 4, pp. 1–15, 2018.

[7] Daniel Wallace and Jordan Cheer, “Optimisation of Personal Audio Systems for Intelligibility Contrast,” in *Proc. 144th Audio Engineering Society Convention*, 2018.

[8] Ji-ho Chang, Chan-hui Lee, Jin-young Park, and Yang-hann Kim, “A realization of sound focused personal audio system using acoustic contrast control,” *Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2091–2097, 2014.

[9] Stephen J Elliott, Jordan Cheer, Jung-woo Choi, and Youngtae Kim, “Robustness and Regularization of Personal Audio Systems,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 7, pp. 2123–2133, 2012.

[10] Stephen J Elliott, Jordan Cheer, Harry Murfet, and Keith R Holland, “Minimally radiating sources for personal audio.,” *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1721–8, 2010.

[11] Ji-Ho Chang, Jin-Young Park, and Yang-Hann Kim, “Scattering effect on the sound focused personal audio system,” *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3060, 2009.

[12] Marek Olik, Philip J. Jackson, and Philip Coleman, “Influence of low-order room reflections on sound zone system performance,” in *Proceedings of Meetings on Acoustics*, 2013, vol. 19, pp. 015058–015058.

[13] Marek Olik, Jon Francombe, Philip Coleman, Philip J B Jackson, Martin Olsen, Martin Møller, Russell Mason, and Søren Bech, “A Comparative Performance Study of Sound Zoning Methods in a Reflective Environment,” *Audio Engineering Society Conference: 52nd International Conference: Sound Field Control - Engineering and Perception*, 2013.

[14] Marcos F. Simón-Gálvez, Stephen J. Elliott, and Jordan Cheer, “The effect of reverberation on personal audio devices,” *The Journal of the Acoustical Society of America*, vol. 135, no. 5, pp. 2654–2663, 2014.

[15] Ferdinando Olivieri, Mincheol Shin, Filippo M Fazi, Philip A Nelson, and Peter Otto, “Loudspeaker array processing for multi-zone audio reproduction based on analytical and measured electroacoustical transfer functions,” *Proceedings of the 52nd Audio Engineering Society International Conference*, 2013.

[16] C. House, S. Dennison, D. G. Morgan, N. Rushton, G. V. White, J. Cheer, and 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*, 2017, vol. 39. pt. 2.

[17] Angela Farina, “Simultaneous measurement of impulse response and distortion with a swept-sine technique,” in *Proc. AES 108th conv, Paris, France*, 2000.

[18] Michael A. Akeroyd, John Chambers, David Bullock, Alan R. Palmer, A. Quentin Summerfield, Philip A. Nelson, and Stuart Gatehouse, “The binaural performance of a cross-talk cancellation system with matched or mismatched setup and playback acoustics,” *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 1056–1069, 2007.

[19] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald, “CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound],” 2017.

[20] Hugo Fastl and Eberhard Zwicker, *Psychoacoustics: Facts and models*, pp. 328, Springer, 3rd edition, 2007.

[21] Jung-Woo Choi and Yang-Hann Kim, “Generation of an acoustically bright zone with an illuminated region using multiple sources,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1695–1700, 2002.