Audio Engineering Society

Conference Paper 22

Presented at the AES 5th International Conference on Audio for Virtual and Augmented Reality
2024 August 19–21, Redmond, WA, USA

This conference paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This conference paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Effects of binaural rendering personalisation and reverberation on speech-on-speech masking

Rapolas Daugintis[1], Benoit Alary[2], Michele Geronazzo[1,3], and Lorenzo Picinali[1]

[1]*Audio Experience Design, Dyson School of Design Engineering, Imperial College London, UK*
[2]*STMS, IRCAM, Sorbonne Université, CNRS, Ministère de la Culture Paris, France*
[3]*Department of Engineering and Management, University of Padova, Italy*

Correspondence should be addressed to Rapolas Daugintis (r.daugintis21@imperial.ac.uk)

ABSTRACT

This study investigates the effect of head-related transfer function (HRTF) personalisation on understanding binaurally rendered target speech masked by interfering speakers in reverberant conditions. During a listening test, participants had to identify a correct colour-number combination from a virtual talker rendered in front of them while ignoring two interfering talkers positioned either in front or at the back. The sound was rendered with either an individual HRTF or one of two non-individual ones. These were selected for each participant as the best or the worst–matching from the same HRTF dataset, based on predictions of a computational auditory model for sound localisation. Two types of reverb from measured spatial room impulse responses (SRIRs) were applied to the speech: realistic dichotic reverberation decoded from 4^{th}-order Ambisonic SRIRs or diotic reverb based on the omnidirectional Ambisonic channel IR as a baseline. Preliminary results show that realistic dichotic reverb improves speech perception when interfering speech is co-located with the target. No significant differences were observed across HRTF conditions on a group level, but individual HRTF-related performance differences exist, requiring further intra-subject analyses and data collection to characterise the individual results.

1 Introduction

Virtual and augmented reality (VR/AR) applications often use non-individual head-related transfer functions (HRTFs) for binaural audio reproduction, which may impact the rendering quality. A significant, albeit small, HRTF-dependent effect has been observed when understanding anechoic speech masked by an interfering talker, rendered binaurally [1, 2, 3]. However, when recreating a realistic multi-talker conversation, especially in AR, reverberation mimicking a physical environment should be applied to the virtual sources. Perception of attended masked speech is altered by the reverberant sound field both because the interfering speech is less correlated between the two ears — making it harder to reduce its impact via the binaural unmasking mechanism — and because the target speech is blurred [4]. In a physical space, reverberation is spa-

tially distributed based on the room geometry and its surfaces [5, 6]. We hypothesise that, due to binaural unmasking effects, realistic reverb can affect the masked speech perception in a way that depends on the spatial characteristics of the reverb itself. Consequently, the choice of the HRTF may influence the perceived spatial distribution of the reverb around the listener and, in turn, the perception of speech.

To study the effect of HRTF and realistic binaural reverb reproduction on speech-in-noise understanding, a speech-on-speech masking task was designed. By asking the participants to identify binaurally rendered reverberant target speech while ignoring interfering speakers, the listening test examined the differences in speech identification performance across individual and non-individual HRTFs, as well as across different reverberation rendering and interfering speech positioning scenarios.

2 Methods

2.1 Reverberation measurements

Measured spatial room impulse responses (SRIRs) were used for this study. SRIRs were measured in a large seminar room with a high ceiling and three alcoves on each side of the longer axis, separated by columns and bookcases (see Fig. 1). A 32-capsule Eigenmike (mh acoustics) spherical microphone array was used to obtain 4^{th}-order Ambisonic SRIRs. Two positions at 1.5 m distance in front (Fig. 1) and back of the microphone array along the longer room axis were measured using Genelec 8030A loudspeaker and exponential sine sweep technique [7]. The loudspeaker distance was chosen based on the direct-to-reverberant ratio (DRR) to ensure sufficient direct sound energy and balance the test difficulty. The SRIRs were denoised and encoded in the spherical harmonics domain,



Fig. 1: SRIR measurement setup for the front position.

by replacing the noisy tail of the measurements with a synthesised decaying noise sequence extending the estimated reverberation slope, as described in [8]. Reverberation time (RT) and DRRs (calculated as the energy ratio between the first 30 samples from the direct impulse peak and the rest of the IR), calculated from the omnidirectional channel of the encoded 4^{th}-order Ambisonic SRIRs of both positions are presented in Table 1. Although RT is similar in both positions, the DRR at the back position is slightly higher than in the front. This might have been caused by the microphone stand alignment or asymmetry of absorptive surfaces in the room.

Two reverb conditions were tested: realistic dichotic, based on binaurally decoded SRIRs, and diotic, based on the convolution with the omnidirectional channel of the SRIRs only and applied equally to both ears. The latter was employed as a baseline (the direct path was always rendered dichotically by directly convolving the speech signals with an appropriate HRTF in both reverb conditions, as discussed in Sec. 2.3).

Table 1: RT (T_{30}) and DRR of the SRIRs, measured at two positions used in the test.

Frequency (Hz)	125	250	500	1000	2000	4000	8000
T_{30} (s), front	1.1	1.1	1.2	1.1	0.9	0.7	0.5
T_{30} (s), back	1.1	1.1	1.1	1.1	0.9	0.8	0.6
Broadband DRR (dB), front				1.9			
Broadband DRR (dB), back				2.8			

2.2 Speech-on-speech masking task

The test was based on a speech identification task, using the coordinate response measure (CRM) speech corpus [9]. By superposing competing audible semantic streams that have to be disentangled (i.e. simultaneous talkers), an informational masking scenario was devised. The choice was motivated by previous findings that such tasks benefit more from the perceived spatial separation of sources, as compared to purely energetic masking cases, when masking occurs only due to overlap of spectral energy (e.g. speech in background noise) [10].

During each trial, three simultaneous phrases of a form 'Ready [*call sign*] go to [*colour*] [*number*] now' were presented by different talkers. Participants had to listen to the target call sign 'Baron' and identify the heard colour-number combination while ignoring the other two interfering speakers. The test followed a procedure similar to the one described in [3]. The original recordings of the CRM corpus, low-pass filtered at 18 kHz (instead of the publicly available dataset, which is filtered at 8 kHz) were used in this study because of the influence of high-frequency content on vertical and front/back discrimination via the spectral pinnae cues [11]. Although the intelligibility of CRM corpus across different talkers, colours, and numbers is not fully balanced [12], the easy structure of the test, which allows for quick unsupervised response procedure, meant that multiple repetitions of each condition could be used to smooth the result variability.

2.3 Binaural rendering & HRTF selection

The sounds for this study were rendered binaurally. Two spatial positions were used: in front of the listener and directly opposite at the back. The target speech (corresponding to the call sign 'Baron' that participants had to attend to) was placed in front while both interferers were positioned either in front (co-located condition) or at the back (separated condition). These directions were chosen to limit the effect of better ear listening, and because they are known to be particularly difficult to externalise, so the quality of rendering due to the choice of an HRTF or reverberation condition could have a stronger effect on the task performance.

For the binaural reproduction, the direct sound part and the reverberation were rendered separately and summed up at the end, similarly to [6]. The SRIR peaks plus 30 samples were removed from the reverberation part. Instead, delayed and level-adjusted mono speech samples were rendered directly through the appropriate HRTF using SPARTA binauraliser[1].

For the reverberant part, the rendering pipeline depended on the reverberation condition. In the dichotic case, the same dry speech samples were convolved with the 4^{th}-order Ambisonic SRIRs using spat5[2] multichannel convolver and then decoded using SPARTA AmbiBIN[3] ambisonic-to-binaural decoder. The magnitude least-squares method was used for the decoding to preserve the relevant interaural time difference (ITD) cues in the low frequencies [13]. For the diotic reverb version, the dry speech samples were convolved with only the first channel of the Ambisonic SRIRs (which corresponds to the 0^{th}-order omnidirectional part of the SRIR) and the same reverberant signal was used for both headphone channels. The direct and the reverberant part of the output signal were summed up at the end and the resulting signal convolved with an individual headphone equalisation filter from the SONICOM HRTF database [14].

The direct and the reverberant (in the dichotic reverb scenario) parts of the sound were rendered binaurally using one of three different HRTFs: individual and two non-individual HRTFs from the SONICOM HRTF dataset [14]. These two were selected for each subject using a recently proposed non-individual HRTF matching procedure [15], based on the best and the worst predicted static sound localisation performance according to auditory model calculations [16]. For each participant, the two non-individual HRTFs were selected from the remaining 199 HRTFs, available from the SONICOM HRTF database.

2.4 Listening test setup & procedure

The test was administered via a laptop running Max 8 (Cycling '74) in an acoustically treated room (the same room was used to measure HRTFs; its acoustic parameters are reported in [14]). The sound was played via the Motu M2 audio interface and Sennheiser HD

[1] https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/#binauraliser

[2] https://forum.ircam.fr/projects/detail/spat/

[3] https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/#ambibin

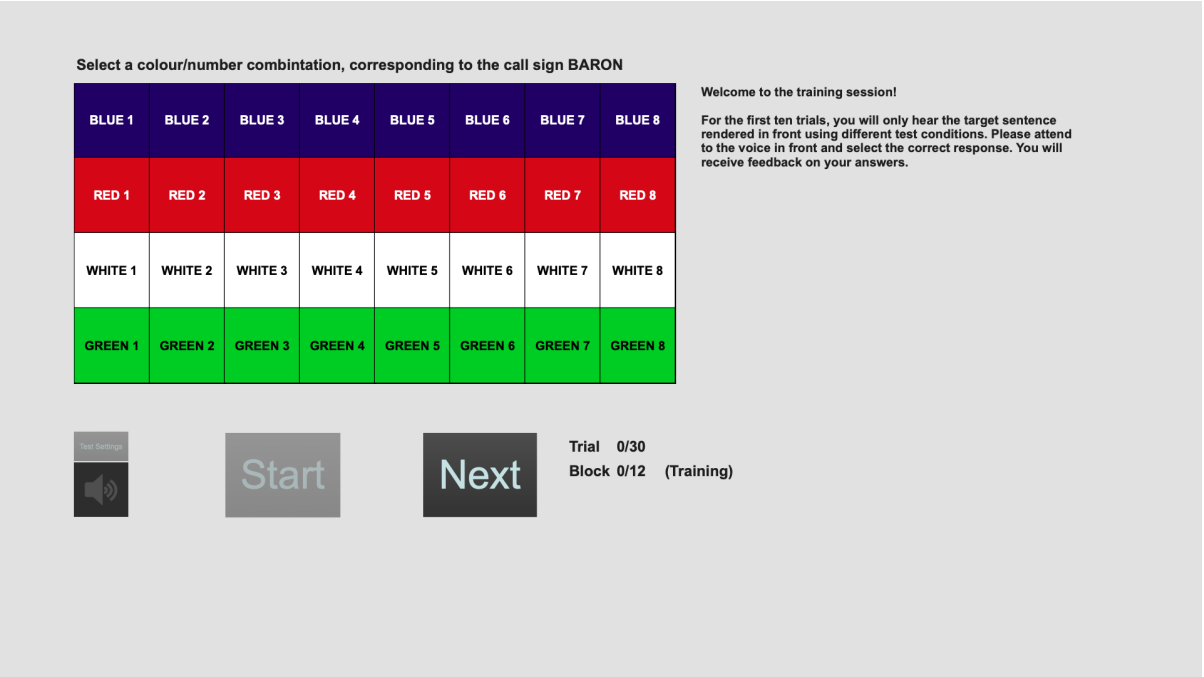


Fig. 2: Graphical user interface of the listening test.

650 headphones. The audio gain was fixed to the same level across all the participants (The sound pressure level at the ear canal was measured to be approximately 60-65 dBA during the playback).

The graphical user interface (GUI) of the test is shown in Fig. 2. The test was divided into 12 blocks, 30 trials each. The blocks corresponded to the combinations of 3 HRTF and 2 reverberation conditions, each repeated twice. Within each block, the co-located and separated maskers conditions were interleaved. Therefore, there were 30 trials for each HRTF, reverberation, and maskers condition in the test. Participants were unaware of the different conditions but were asked to concentrate on the target speaker, which was always rendered in front of the listener.

Before the test, participants were presented with a familiarisation block. The first 10 trials contained only the target speech, while the next 20 both the target and one masker. During the training, HRTF and reverberation conditions, as well as masker positions, were randomised without replacement. After each trial of the training session, the participants were informed whether their answers were correct.

A total of 13 participants performed the test. All of them were either native English speakers or had resided in an English-speaking country for a sustained period.

3 Results

The main results from the listening test are presented in Fig. 3. The boxplot shows the percentage of correct responses (colour-number combinations) across participants when the masker speakers were separated from the target and placed at the back (left panel) and when they were co-located with the target at the front (right panel). Results in each panel are also subdivided by the reverberation and HRTF conditions.

To understand the effect of maskers, reverb, and HRTF conditions on the percentage of correct responses, a three-way repeated measures analysis of variance (RM-ANOVA) was performed. The normality of the distributions was confirmed by the Shapiro–Wilk test and the sphericity by Mauchly's test. RM-ANOVA revealed a statistically significant effect of the maskers' position ($F(1,12) = 179.294$, $p < 0.0001$), indicating, as expected, that it was more challenging to understand the target speech when the interfering speech

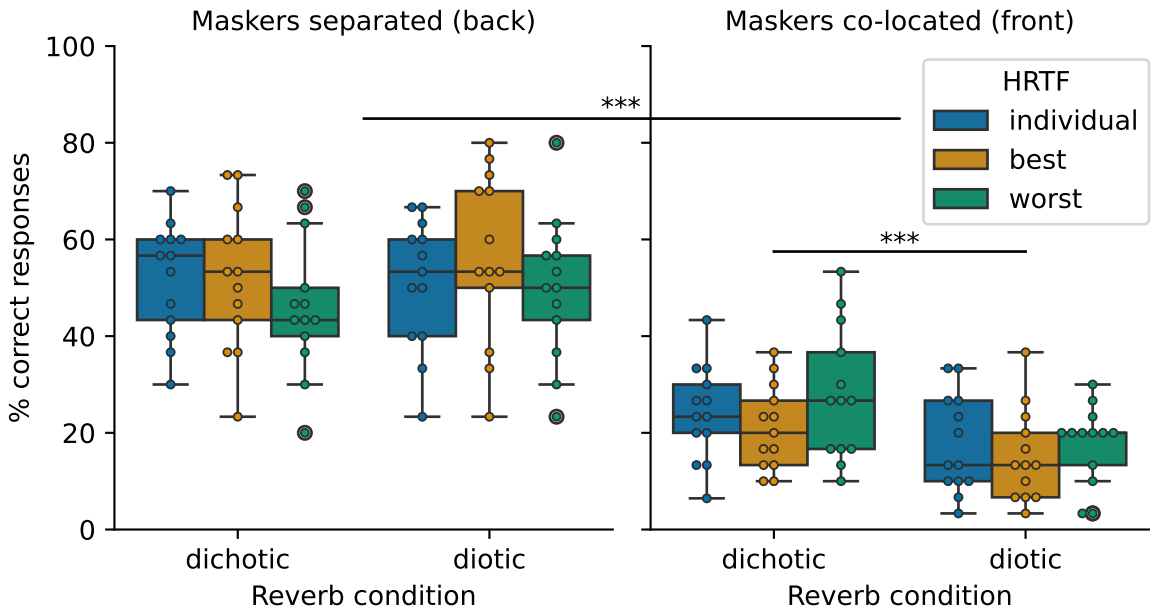


Fig. 3: Percentage correct responses for different maskers positions and different reverb and HRTF conditions. Points represent individual performances and boxplots show medians and the inter-quartile range of the performance distributions. '***' denotes significant differences ($p < 0.001$) in distributions between maskers' positions and between reverb conditions in the case of the co-located maskers.

was co-located. There was also a significant interaction between the maskers' position and the reverb type ($F(1,12) = 27.131$, $p < 0.001$). Additional two-way RM-ANOVAs were performed on the two separate groups split by maskers' position. It showed a significant effect of reverberation type in the case of co-located maskers ($F(1,12) = 28.998$, $p < 0.001$). The figure reveals that when the maskers were co-located the task was on average more difficult to perform in the diotic reverb scenario. No significant effect of reverberation was found in the separated maskers case, as well as no significant differences between HRTF conditions.

3.1 Spatial release from masking

The improvement of speech perception when the target and masking speech sources are spatially separated, known as spatial release from masking (SRM), was analysed similarly to [3] by calculating the difference between individual performances in separated vs co-located scenarios. Fig. 4 shows SRM in terms of percentage point differences in performances across participants for different reverberation and HRTF conditions. Two-way RM-ANOVA was performed to understand the effect of reverb and HRTF conditions on SRM. It revealed significant differences between the reverb conditions ($F(1,12) = 27.133$, $p < 0.001$). On average, the SRM was higher in the diotic reverberation case: while the performance was similar between the reverberation conditions when the maskers were separated from the target, it was impinged by the diotic reverb more than the dichotic one in the co-located maskers scenario.

On the other hand, the differences between HRTFs didn't meet the significance criteria ($F(2,24) = 2.889$, $p = 0.075$). A slight trend of lower median SRM with the worst HRTF condition (especially in the dichotic case) can be observed in the figure but data from more participants might be required to resolve its significance.

3.2 Comparison with objective metrics

The presented results indicate that, at a group level, the speech perception performance does not depend on

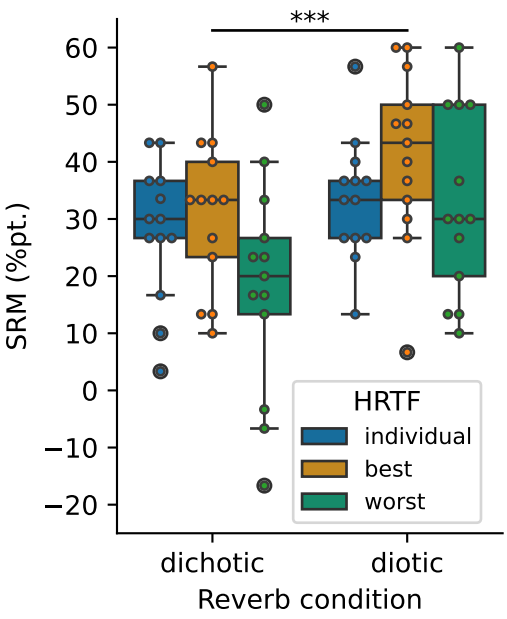


Fig. 4: Spatial release from masking as percentage point differences between test performance with separated vs co-located maskers for different reverb and HRTF conditions. Points represent individual results and boxplots show medians and the inter-quartile range of the distributions. '***' shows significant differences ($p < 0.001$) in distributions between different reverb conditions.

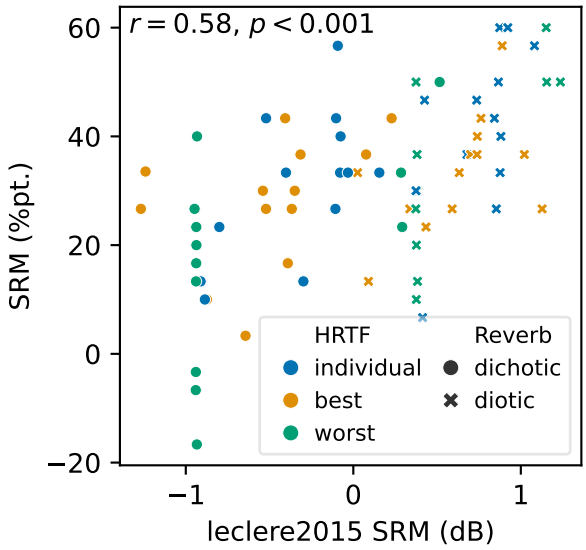


Fig. 5: Spatial release from masking (SRM) as a percentage-point difference of participants' performance between the co-located and separated maskers position, plotted against the SRM (in dB), calculated from BRIRs using *leclere2015* auditory model. Each point represents an individual result with a specific HRTF and reverb condition.

the specific HRTF selection used in this study. However, performance differences between HRTF conditions exist on an individual level. To assess if these differences could be explained by some other objective metrics, an auditory speech intelligibility model was used. The model used — *leclere2015* from the Auditory Modelling Toolbox [17] — estimates SRM as an effective ratio (in dB) between a reverberant speech target and a spatially separated static broadband noise masker [18]. The model was supplied with the binaural room impulse responses (BRIRs) from the target and the separated masker positions, which were recorded from the Max patcher output to account for the full audio rendering pipeline. The model relies on a few assumptions about the cut-off between useful early reflections and detrimental late reverberation, which are room-dependent, but the default room-independent parameters were used in this exercise. For each participant, the SRMs with all three HRTFs and two reverb conditions were estimated.

The model estimates in dB are plotted against the measured SRM in percentage points and shown in Fig. 5. The linear relationship between the two metrics was assessed by calculating Pearson's correlation coefficient. It revealed a moderate positive correlation between the model output and the listening test results ($r = 0.58$, $p < 0.001$). Generally, the model predicted mostly negative SRM values for the dichotic reverb scenarios and positive SRM for the diotic ones, but the absolute values might not be easy to interpret without the model calibration. Furthermore, the worst HRTFs are clustered around a few *leclere2015* SRM values. This bias is a result of the HRTF selection procedure, which often favours the same HRTF to be chosen as the 'worst' for multiple subjects [15]. Besides this clustering of the 'worst' HRTF, there are no measurable differences between the distributions associated with different HRTF conditions.

4 Discussion

The study results suggest that when reproducing a multi-speaker virtual environment, there is some benefit in using fully directional dichotic reverberation as compared to diotic binaurally-correlated reverb of the same strength. These results are somewhat in line with previous findings of speech intelligibility improvement when artificial reverberation is decorrelated between the two ears [19] or increased speech reception threshold in noise and voice interferers when binaural reverberation is used over the diotic one [4]. However, the effect in this study is only present in the more difficult scenario of the co-located target and maskers. One possible explanation is that the separated condition is easier to perform, resulting in participants not having to rely on subtle spatial reverberation cues. In contrast, the participants found the co-located condition very difficult. In general, for the majority of the incorrect responses (88% of the cases) the participants selected the colour and/or the number, uttered by one of the two interfering speakers, so 11% (1/9th) correct response rate could be considered a chance level. Median responses for the co-located maskers and diotic reverberation condition balances just above this threshold, indicating that many participants were guessing the answers during this condition. Therefore, adjusting the difficulty level by manipulating the DRR and number of interfering speakers could lead to a more sensitive test and accentuated differences between conditions.

Nonetheless, the diotic reverberation condition used in this test can only be seen as a worst-case baseline since it does not represent a realistic reverberation scenario. Highly correlated binaural reverberant signals that are not filtered through HRTFs offer little externalisation (besides the direct part). It would be interesting to compare these results with a more realistic reverberation that is uncorrelated between the ears, but acoustically isotropic. Furthermore, the test performance may be dependent on the exact acoustics of the measured space, so contrasting the isotropic reverberation with a highly anisotropic one is subject to future research.

On the other hand, the HRTF classification based on modelled sound localisation does not appear to relate to the speech reception performance in a statistically significant way. Nevertheless, a minor pattern, favouring individual and 'best' HRTFs over the 'worst' one starts to emerge in the SRM results. Thus, data from more subjects might be needed to draw a more robust conclusion on the role of HRTF personalisation. Although the correct perception of the source location has been previously shown to play a role in unmasking the target speech [10], other factors such as binaural and spectral signal-level differences between the two BRIR positions might have a stronger influence on the performance, irrespective of the HRTF individualisation. This is somewhat illustrated by the moderate result correlation with the speech intelligibility model, which is only based on signal-level BRIR differences. Although the interpretation of the modelled results in this scenario is limited because the model only considers static energetic masking and doesn't account for the informational masking part, better control of the test conditions, correcting for the signal-level BRIR differences, could provide more meaningful results. Finally, alternative HRTF selection methodologies and a bigger participant sample size might help better separate the HRTF personalisation effect from purely signal-level differences.

Further improvements to the study could be made by administering the test in a more immersive virtual environment, which visually matches the rendered acoustic space. Better visual-acoustic congruence may aid the sense of externalisation of the sound source and may in turn impact the observed effect, providing a more realistic assessment of the speech unmasking phenomena.

5 Summary

This paper presented a headphone-based speech-on-speech masking test, which examined the effect of individual vs non-individual HRTF selection and realistic dichotic vs diotic reverberation rendering on understanding reverberant masked speech. The target speech was rendered virtually in front of the listener while listeners had to ignore the interfering speech maskers, which were either co-located or separated and positioned at the back. In general, the results show that the target speech is easier to understand when interfering speech is separated and rendered at the back. Furthermore, a statistically significant improvement in speech perception is observed when dichotic reverb is used over the diotic one, but only when the maskers are co-located with the target speaker. Finally, no statistically significant effect of HRTF selection was observed on a group level. It is worthwhile to note that individual HRTF-related differences in performance exist across the participants, requiring further intra-subject analysis of different HRTF conditions. The findings of this

study might be valuable for designing efficient VR/AR applications with a focus on social group interactions and telepresence.

6 Acknowledgements

This research is part of the SONICOM project (EU Horizon 2020 RIA grant agreement ID: 101017743). RD also received the Global Fellows Fund from Imperial College London (as part of the UK's Turing Scheme) for a research visit to IRCAM, where the study was conceived. We thank Dr Olivier Warusfel, Dr Mathieu Lavandier and Dr Virginia Best for their input in the ideation stages of the study. We also thank Dr Virginia Best for providing the original recordings of the CRM corpus.

References

[1] Cuevas-Rodriguez, M., Gonzalez-Toledo, D., Reyes-Lecuona, A., and Picinali, L., "Impact of Non-Individualised Head Related Transfer Functions on Speech-in-Noise Performances within a Synthesised Virtual Environment," *J. Acoust. Soc. Am.*, 149(4), pp. 2573–2586, 2021, doi: 10.1121/10.0004220.

[2] Zenke, K. and Rosen, S., "Spatial Release of Masking in Children and Adults in Non-Individualized Virtual Environments," *J. Acoust. Soc. Am.*, 152(6), pp. 3384–3395, 2022, doi: 10.1121/10.0016360.

[3] González-Toledo, D., Cuevas-Rodríguez, M., Vicente, T., Picinali, L., Molina-Tanco, L., and Reyes-Lecuona, A., "Spatial Release from Masking in the Median Plane with Non-Native Speakers Using Individual and Mannequin Head Related Transfer Functions," *J. Acoust. Soc. Am.*, 155(1), pp. 284–293, 2024, doi:10.1121/10.0024239.

[4] Lavandier, M. and Culling, J. F., "Speech Segregation in Rooms: Monaural, Binaural, and Interacting Effects of Reverberation on Target and Interferer," *J. Acoust. Soc. Am.*, 123(4), pp. 2237–2248, 2008, doi:10.1121/1.2871943.

[5] Alary, B., Massé, P., Schlecht, S. J., Noisternig, M., and Välimäki, V., "Perceptual Analysis of Directional Late Reverberation," *J. Acoust. Soc. Am.*, 149(5), pp. 3189–3199, 2021, doi:10.1121/10.0004770.

[6] Engel, I., Henry, C., Amengual Garí, S. V., Robinson, P. W., and Picinali, L., "Perceptual Implications of Different Ambisonics-based Methods for Binaural Reverberation," *J. Acoust. Soc. Am.*, 149(2), pp. 895–910, 2021, doi:10.1121/10.0003437.

[7] Farina, A., "Advancements in Impulse Response Measurements by Sine Sweeps," in *Proc. Audio Eng. Soc. (AES) Conv.*, p. 7121, Vienna, Austria, 2007.

[8] Massé, P., Carpentier, T., Warusfel, O., and Noisternig, M., "Denoising Directional Room Impulse Responses with Spatially Anisotropic Late Reverberation Tails," *Appl. Sci.*, 10(3), p. 1033, 2020, doi:10.3390/app10031033.

[9] Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D., "A Speech Corpus for Multitalker Communications Research," *J. Acoust. Soc. Am.*, 107(2), pp. 1065–1066, 2000, doi: 10.1121/1.428288.

[10] Brungart, D. S., "Informational and Energetic Masking Effects in the Perception of Two Simultaneous Talkers," *J. Acoust. Soc. Am.*, 109(3), pp. 1101–1109, 2001, doi:10.1121/1.1345696.

[11] Martin, R. L., McAnally, K. I., Bolia, R. S., Eberle, G., and Brungart, D. S., "Spatial Release from Speech-on-Speech Masking in the Median Sagittal Plane," *J. Acoust. Soc. Am.*, 131(1), pp. 378–385, 2012, doi:10.1121/1.3669994.

[12] Brungart, D. S., "Evaluation of Speech Intelligibility with the Coordinate Response Measure," *J. Acoust. Soc. Am.*, 109(5), pp. 2276–2279, 2001, doi:10.1121/1.1357812.

[13] Schörkhuber, C., Zaunschirm, M., and Höldrich, R., "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proc. Annu. Ger. Conf. Acoust. (DAGA)*, Munich, Germany, 2018.

[14] Engel, I., Daugintis, R., Vicente, T., Hogg, A. O. T., Pauwels, J., Tournier, A. J., and Picinali, L., "The SONICOM HRTF Dataset," *J. Audio Eng. Soc. (AES)*, 71(5), pp. 241–253, 2023, doi: 10.17743/jaes.2022.0066.

[15] Daugintis, R., Barumerli, R., Picinali, L., and Geronazzo, M., "Classifying Non-Individual

Head-Related Transfer Functions with A Computational Auditory Model: Calibration And Metrics," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023, doi:10.1109/ICASSP49357.2023.10095152.

[16] Barumerli, R., Majdak, P., Geronazzo, M., Meijer, D., Avanzini, F., and Baumgartner, R., "A Bayesian Model for Human Directional Localization of Broadband Static Sound Sources," *Acta Acust.*, 7, p. 12, 2023, doi:10.1051/aacus/2023006.

[17] Majdak, P., Hollomey, C., and Baumgartner, R., "AMT 1.x: A Toolbox for Reproducible Research in Auditory Modeling," *Acta Acust.*, 6(19), 2022, doi:10.1051/aacus/2022011.

[18] Leclère, T., Lavandier, M., and Culling, J. F., "Speech Intelligibility Prediction in Reverberation: Towards an Integrated Model of Speech Transmission, Spatial Unmasking, and Binaural de-Reverberation," *J. Acoust. Soc. Am.*, 137(6), pp. 3335–3345, 2015, doi:10.1121/1.4921028.

[19] Orduña-Bustamante, F., Padilla-Ortiz, A., and Torres-Gallegos, E. A., "Binaural Speech Intelligibility through Personal and Non-Personal HRTF via Headphones, with Added Artificial Noise and Reverberation," *Speech Commun.*, 105, pp. 53–61, 2018, doi:10.1016/j.specom.2018.10.009.