DOI: 10.61782/fa.2025.0813

FORUM ACUSTICUM
EURONOISE 2025

SPATIAL RELEASE FROM MASKING AND LISTENING EFFORT: DIFFERENCES BETWEEN BINAURAL AND FREE-FIELD PRESENTATION

Nicola La Magna[1*] **Katarina Poole** [1] **Lorenzo Picinali**[1]
[1] Dyson School of Design Engineering, Imperial College London, London, UK

ABSTRACT

Speech intelligibility decreases in noisy environments, especially for individuals with hearing impairment. Spatial separation of speech and competing masking sounds improves understanding through spatial release from masking (SRM), though its cognitive demand remains understudied. Virtual Reality (VR) has been successfully employed in the past for testing and training auditory perception. This can be achieved using loudspeaker systems or headphones, in the latter case requiring Head-Related Transfer Functions (HRTFs). Individual HRTFs are known to improve speech-in-noise (SIN) understanding in horizontal and median planes, although benefits may diminish under informational masking conditions. This study investigates SRM performance and listening effort (LE) using the Coordinate Response Measure (CRM) corpus in a speech-on-speech task. Participants experience free-field loudspeaker rendering and binaural presentations (individual and non-individual HRTFs) in target-masker spatial configurations on horizontal and median planes. Performance is measured via identification accuracy and pupillometry. We hypothesize similar performance for free-field and both HRTF conditions in the horizontal plane due to the reliance on interaural differences. On the median plane, free-field is expected to outperform individual HRTFs, with non-individual HRTFs performing worse due to poor spectral cues representation that will induce increased effort. Understanding SRM and LE differences in free-field and binaural settings can advance our understanding of spatial hearing mechanisms in complex environment, and potentially speed up the adoption of virtual acoustics techniques in hearing research.

Keywords: *Binaural Rendering, SRM, Listening Effort, Loudspeaker, Phantom Source*

**Corresponding author*: n.la-magna@imperial.ac.uk.

1. INTRODUCTION

Understanding speech in noisy environments is a crucial aspect of auditory perception, communication and turn taking in everyday life. It is therefore not surprising that neural and structural specializations have evolved to perform this complex task. For instance, the cognitive phenomenon known as the ”cocktail party effect” [1] describes our brain’s ability to focus on a specific auditory stimulus while ignoring other background sounds cognitive. Moreover, speech intelligibility in such conditions benefits under Spatial Release from Masking (SRM). It refers to the improvement in understanding speech when the target and masking sounds are spatially separated. In this scenario, in the horizontal plane humans relies on interaural cues (i.e., time differences and level differences), while in the median plane the pinnae and upper body result in spectral cues that aid in the inference of sound source elevation and front-back discrimination [2]. As a result, different auditory experiments has been used to test SRM benefits, including sounds rendered through loudspeaker systems [3] or binaurally via Head-Related Transfer Functions (HRTFs) [4,5]. While SRM has been widely studied, its cognitive demands and benefits under different rendering conditions remain less explored.

1.1 Related Works

The perception of speech in complex environments is influenced by several factors, including the adverse effects of environmental noises and interfering voices. Speech intelligibility often decreases in sounds fields that present these conditions, especially for those individuals that are affected by hearing impairments [6]. The mechanism which allow our brain to focus on a specific target while ignoring others are described as the ”cocktail party effect” [7]. This is mainly possible thanks to spatial acoustic cues such as interaural time difference (difference in time of arrival between the two ears, ITD) and interaural level difference (difference in level between the two ears, ILD) which, together with monaural spectral cues, help distinguish conversations based on their position in space. Together, these phenomena

contribute to spatial release from masking (SRM) [8], which is the improvement in detection of a target (e.g., speech) when maskers are spatially separated compared with when they are co-located with the target itself. However, the magnitude of this benefit depends on many factors, including stimulus types, location, task, and hearing abilities of the participants [9, 10]. One of these factors may be represented by each individual cognitive processing load. Also known as listening effort, processing load depends on the interaction between cognitive capacity and task demands. As proposed by the The Framework for Understanding Effortful Listening (FUEL) model, other aspects can influence it such as working memory, attention, resource allocation and motivation [11]. While listening effort has been studied in different SRM spatial configurations [10, 12], the cognitive demand across different spatial cues and rendering presentations remains understudied. In fact, SRM can be assessed through tasks that can be rendered either in free-field using loudspeakers or through binaural virtual rendering with headphones, making use of Head-Related Transfer Functions (HRTFs) [13]. Individually-measured HRTFs are known to improve speech-in-noise (SIN) understanding in horizontal and median planes [5, 14], although benefits may diminish under informational masking conditions [4]. In the median plane, non-individual HRTFs struggle to replicate spectral monoaural cues [13], thus impacting SRM [5]. Improvements in SIN emerge with perceptual training for energetic masking, but the topic still suffers from the relatively unknown transferability of skills from binaural rendering to real life sound presentation, and poorly controlled studies [15, 16]. In order to create controllable and highly realistic environments, Virtual Reality (VR) has been used in the past to create audio-visual simulations and implement specific training and tasks routines to test and train auditory perception in individuals with or without hearing aids [17, 18]. The present study investigates SRM performance and listening effort (LE) in a speech-on-speech task (i.e. informational masking) within a VR environment. Participants experience free-field loudspeaker rendering (with loudspeakers in each source position, as well as with phantom sources) and binaural presentations (individual and non-individual HRTFs) in target-masker spatial configurations on horizontal and median planes while wearing a head mounted display (HMD).

1.2 Research questions and aims

Our goal is to address whether different types of sound renderings can change the behavioural and psychological response in an SRM task. Furthermore, we aim to explore LE as a valid measure for spatial hearing tasks.

To do so, we set 3 different research questions:

- RQ1: Do different HRTF (personalised vs non-personalised) induce different related effort?
- RQ2: Is the performance (i.e., accuracy) different between real and virtual rendering of sounds?
- RQ3: Is the cognitive effort different between real and virtual presentations?

2. METHODOLOGY

2.1 Technical set up

A custom computer was used to render the virtual environment (VE) in the Unity3D 6000.0.41f game engine [19]. The VE is displayed using a HTC VIVE FOCUS VISION headset. The VR device comes with a built-in eye tracker with a 120 Hz gaze data output frequency (binocular), 0.5° to 1.1° accuracy within 20° field of view, a 5-point calibration and automatic IPD adjustment. Sounds are rendered through the Max/MSP software [20]. They are played either by custom full-range loudspeakers (see [21] for more details) located at a distance of 1.5m, or through a pair of Sennheiser HD 650 headphones for the binaural stimuli. Binaural stimuli are synthesized using HRTFs from the SONICOM database [21] and spatialised using the VST release of the 3dti-Toolkit [22], also at a fixed distance of 1.5m.

2.2 Stimuli

The audio stimuli, sampled at 48kHz and 16bits, are presented at 65 dB A. We use the CRM corpus [23]. CRM is a corpus of recorded sentences designed for speech-on-speech intelligibility tests. It consists of sentences that follows the same structure: "Ready *call sign*, go to *colour number now*". There are eight call signs, four colours and eight numbers resulting in 256 standardized utterances recorded by four male and four female talkers. The sentences are presented simultaneously to the participants on each trial. Target sentences always used as a *call sign* the word *Baron* and a random combination of colour and number. Maskers always used another call sign, colour and number, different to the ones used for the target. Participants were asked to listen to the target and select, among all the options, which colour-number combination they thought they heard in the target sentence. The CRM corpus is particularly appropriate to highlight cognitive resource allocation, and estimate intelligibility performances in informational masking contexts [24]. Additionally, it is widely used for evaluating listening effort with spatialised talkers and maskers [25], or HRTFs evaluation in speech-related tasks [5].

2.3 Experiment design and task

The experiment includes two different stages, distinguished by the position of the masker sources either in the horizontal or on the median plane. Both follow a within-group design. Each participant undergoes 4 different conditions: loudspeaker, phantom source, individual HRTF and non individual HRTF (i. e., worst individual match following the methodology proposed in [26]) binaural renderings. The dependent variables include

both behavioural (i.e., response accuracy), physiological (i.e., pupillometry) and self-reported (i.e., questionnaires) measures. As shown in Figure 1, the interface does not include colors, in order to avoid influencing pupil dilation. Target stimuli are always played at 0° azimuth and 0° elevation. The two masker stimuli are either colocated with the target (always coming from 0° azimuth and 0° elevation), or at ±30° , ±60° and ±90° on azimuth. These locations were purposely chosen to provide steps from a condition without a SRM benefit (i.e., 0°), to a condition in which the benefit is supposed to be the largest (i.e., ±90°). Regarding elevation, masker stimuli are presented at 0°, ±20°, as well as in a front-back condition at elevations 60° and 120°.

In order to properly control and measure pupil dilation, between each trial a black cross appears in the center of the screen. The participant is asked to fix the cross. This remains visible for 10 seconds, after which the stimuli are played. Once over, the GUI is activated to allow participants for a response.

Figure 1. CRM Corpus inside the virtual Environment.

2.4 Procedure

The study is divided into sessions of approximately one and half hours in two separate days. At the beginning of the first session, a familarisation stage is proposed (i.e., procedural training). Participants are asked to take a break between each block; during the break they are given a self-report questionnaires and are asked to rest for 5 minutes. The eye tracking calibration is performed at the beginning and between each block. The order of presentation of the rendering conditions is randomized. Across both sessions, participants complete 12 blocks (4 conditions x 3 repetitions) each consisting in 4 trials per spatial location (i.e., azimuth 4x4, elevation 4x3) for a total of 12x16 trials in the horizontal plane, and 12x12 trials in the median plane.

3. EXPECTED RESULTS AND FUTURE WORK

Considering RQ1, we hypothesise differences in listening effort between non-individual HRTFs and all other conditions on the median plane [5] due to poor monaural cues replication. No significant differences are expected on the horizontal plane.

When it comes to RQ2 we expect similar outcomes for all conditions in the horizontal plane [18], mainly because of the fact that the task relies mainly on interaural cues, which are correctly reproduced by all rendering conditions. However, this is unlikely to be the case for the median plane, where we expect the reliance on monoaural cues to result in differences between free-field and binaural conditions, as well as between individual and non-individual HRTF conditions.

Finally, considering RQ3 we expect similar outcomes as to RQ2. In fact, we expect for binaural presentation to result in more effort than the loudspeaker conditions due to the amount of complex processes the signals go through before being rendered. Moreover, past research on the topic showed differences in performance and brain activation during a sound localisation task between binaural and free-field conditions, with the latter allowing for better precision [27].

Understanding SRM and LE differences in virtual and real-world settings can put a lens on the impact of understanding spatial hearing mechanisms not only at a cortical or sensory level of perception, but also looking at a higher, cognitive level. This may address potential insight for the implementation of a specific techniques over the others based on the purpose to achieve. Moreover, results can proof how LE can be address as a measure of HRTF fit, finally providing a good estimate for HRTFs evaluation and comparison. Results will be presented at the conference.

4. ACKNOWLEDGMENTS

This study was supported by the CherISH project, a European Doctorate Network funded by the European Union's Horizon Europe framework program under the Marie Sklodowska-Curie Grant Agreement No: 101120054.

5. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the acoustical society of America*, vol. 25, pp. 975–979, 1953.

[2] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. 791–802, 2014.

[3] R. L. Freyman, U. Balakrishnan, and K. S. Helfer, "Spatial release from informational masking in speech recognition," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2112–2122, 2001.

[4] K. Zenke and S. Rosen, "Spatial release of masking in children and adults in non-individualized virtual environments," *The Journal of the Acoustical Society of America*, vol. 152, no. 6, pp. 3384–3395, 2022.

[5] D. González-Toledo, M. Cuevas-Rodríguez, T. Vicente, L. Picinali, L. Molina-Tanco, and A. Reyes-Lecuona, "Spatial release from masking in the median plane with

non-native speakers using individual and mannequin head related transfer functions," *The Journal of the Acoustical Society of America*, vol. 155, no. 1, pp. 284–293, 2024.

[6] K. Ishikawa, S. Murgia, H. Li, E. Renkert, and P. Bottalico, "Cognitive load associated with speaking clearly in reverberant rooms," *Scientific Reports*, vol. 14, no. 1, p. 20069, 2024.

[7] C. Cherry and J. Bowles, "Contribution to a study of the "cocktail party problem"," *The Journal of the Acoustical Society of America*, vol. 32, no. 7, pp. 884–884, 1960.

[8] J. F. Culling and M. Lavandier, "Binaural unmasking and spatial release from masking," *Binaural Hearing: With 93 Illustrations*, pp. 209–241, 2021.

[9] B. Dieudonné and T. Francart, "Redundant information is sometimes more beneficial than spatial information to understand speech in noise," *Ear and Hearing*, vol. 40, no. 3, pp. 545–554, 2019.

[10] J. Rennies, V. Best, E. Roverud, and G. Kidd Jr, "Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort," *Trends in Hearing*, vol. 23, p. 2331216519854597, 2019.

[11] M. K. Pichora-Fuller, S. E. Kramer, M. A. Eckert, B. Edwards, B. W. Hornsby, L. E. Humes, U. Lemke, T. Lunner, M. Matthen, C. L. Mackersie, *et al.*, "Hearing impairment and cognitive energy: The framework for understanding effortful listening (fuel)," *Ear and hearing*, vol. 37, pp. 5S–27S, 2016.

[12] R. Livingstone, C. Patro, and N. K. Srinivasan, "Spatial release from masking and pupillometry," *The Journal of the Acoustical Society of America*, vol. 155, no. 3_Supplement, pp. A308–A308, 2024.

[13] L. Picinali and B. F. G. Katz, *System-to-User and User-to-System Adaptations in Binaural Audio*, pp. 115–143. Cham: Springer International Publishing, 2023.

[14] A. Ahrens, M. Cuevas-Rodriguez, and W. O. Brimijoin, "Speech intelligibility with various head-related transfer functions: A computational modelling approach," *JASA Express Letters*, vol. 1, no. 3, 2021.

[15] J. P. Whitton, K. E. Hancock, J. M. Shannon, and D. B. Polley, "Audiomotor perceptual training enhances speech intelligibility in background noise," *Current Biology*, vol. 27, no. 21, pp. 3237–3247, 2017.

[16] H. Henshaw and M. A. Ferguson, "Efficacy of individual computer-based auditory training for people with hearing loss: a systematic review of the evidence," *PloS one*, vol. 8, no. 5, p. e62836, 2013.

[17] L. Picinali, G. Grimm, Y. Hioka, G. Kearney, D. J. C. Jin, and A. E. Design, "Vr/ar and hearing research: current examples and future challenges," 2023.

[18] M. Ramírez, A. Müller, J. M. Arend, H. Himmelein, T. Rader, and C. Pörschmann, "Speech-in-noise testing in virtual reality," *Frontiers in Virtual Reality*, vol. 5, p. 1470382, 2024.

[19] "Unity." Available from: https://unity.com/, 2025. cited 2025.

[20] "Cycling '74." Available from: https://cycling74.com/, 2025. cited 2025.

[21] I. Engel, R. Daugintis, T. Vicente, A. O. Hogg, J. Pauwels, A. J. Tournier, and L. Picinali, "The sonicom hrtf dataset," *journal of the audio engineering society*, vol. 71, no. 5, pp. 241–253, 2023.

[22] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3d tune-in toolkit: An open-source library for real-time binaural spatialisation," *PloS one*, vol. 14, no. 3, p. e0211899, 2019.

[23] R. S. Bolia, W. T. Nelson, M. A. Ericson, and B. D. Simpson, "A speech corpus for multitalker communications research," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 1065–1066, 2000.

[24] G. Kidd Jr, C. R. Mason, V. M. Richards, F. J. Gallun, and N. I. Durlach, "Informational masking," *Auditory perception of sound sources*, pp. 143–189, 2008.

[25] C. Lanzilotti, G. Andéol, C. Micheyl, and S. Scannella, "Cocktail party training induces increased speech intelligibility and decreased cortical activity in bilateral inferior frontal gyri. a functional near-infrared study," *Plos one*, vol. 17, no. 12, p. e0277801, 2022.

[26] R. Daugintis, R. Barumerli, L. Picinali, and M. Geronazzo, "Classifying non-individual head-related transfer functions with a computational auditory model: Calibration and metrics," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[27] N. Marggraf-Turley, M. Shiell, N. Pontoppidan, D. Cappotto, and L. Picinali, "Eeg-based decoding of sound location: Comparing free-field to headphone-based non-individual hrtfs," *arXiv preprint arXiv:2503.10783*, 2025.