DOI: 10.61782/fa.2023.0540

forum acusticum 2023

TOWARDS THE USE OF SPATIAL RELEASE FROM MASKING AND SPATIAL STREAM SEGREGATION TESTS TO SELECT AN HRTF

Thibault Vicente[1] **Lorenzo Picinali**[1*]

[1] Imperial College London, Dyson School of Design Engineering, Imperial College Rd, South Kensington, SW7 2DB London, UK

ABSTRACT

In order to render 3D audio through headphones, sounds are spatialised using head-related transfer functions (HRTFs), which characterise how a body modifies an acoustic signal coming from a given spatial location to the ear. Each individual has a specific HRTF due to differences in terms of morphological features. Perceptual mismatches are observed when the sounds are spatialised using someone's else HRTF. Ideally, the audio should be customised to each individual using their own HRTF, but their acoustic measurement requires time and appropriate facilities. Hence, past research developed tests that allow to select an already-measured non-individual HRTF to perceptually improves the listener's immersion. These tests rely on subjective and/or objective evaluations. This contribution considers the possibility of selecting an HRTF using spatial release from masking (SRM) and spatial stream segregation (SSS) tests. An overview of the effect of HRTF choice on SRM is presented, specifically focussing on two experiments that were run to this purpose. Other test results assessing SSS are shown. The conclusion of each study supports the that the listeners are more sensitive to the familiarity to HRTF cues rather than the actual magnitude of such cues. The paper ends with some suggestions for future research to further verify these findings.

Keywords: *head-related transfer functions, spatial hearing, speech intelligibility, stream segregation.*

**Corresponding author*: *l.picinali@imperial.ac.uk.*

1. INTRODUCTION

Sounds that are travelling to the human ear canal have been absorbed, diffracted and reflected by several parts of the human body. These modifications can be captured through a head-related transfer function (HRTF), i.e., HRTFs characterise how a body modifies sound coming from one spatial location to the ear. This is useful when simulating realistic 3D audio scene through headphones. However, HRTFs relie on anthropomorphic features, which means that individuals have their own HRTFs. Using someone else's HRTF for the rendering can induce perceptual mismatch. Several techniques and approaches exist to personalise HRTFs when acoustic measurements are not a possibility (see [1] for extensive review). One of the most employed ones consists in selecting a non-individual HRTF from a set of already-measured ones [2]. Subjective tasks can consist in asking the participants to rate the quality/realism of the virtual audio scene. Objective tasks aim at either using performances in listening tests, such as localisation accuracy, to assess the quality of the HRTF for the listener (assuming the higher the accuracy, the better the HRTF); or finding an HRTF that matches the listeners' anthropomorphic features.

The current study looks at using other protocols involving spatial auditory mechanisms to select an HRTF. The first protocol measures spatial release from masking (SRM), which is the intelligibility benefit provided by moving the masking sound sources away from the target speech source location, and which has already been shown to be related with HRTF choices [3]. The second protocol assesses the ability of the listener to segregate interleaved sequences of sounds into distinct streams using spatial cues, so-called spatial stream segregation (SSS).

2. EXPERIMENT OVERVIEW

2.1 Speech intelligibility experiments

Two experiments were designed to study the effect of HRTF individualisation on SRM. They are inspired by experiments from literature. Both experiments involved a speech-on-speech paradigm using the coordinate response measures (CRM) corpus, in which sentences are always following the same structure with a call sign, colour and digit keywords. The target call sign was always "Baron". The target-to-masker ratio was always at 0 dB. The first experiment was inspired by [4] where one speech masker was simultaneously presented with the target. The target speaker was always different from the masking speaker, but both had the same sex. The target and masker could be simulated at 3 different locations: 0°, -50° and +50° of elevation (always at 0° of Azimuth), resulting in 9 target-masker spatial conditions. The second experiment (inspired by [5]) involved two speech maskers that were always simulated at the same location within a trial. The masking speakers were always different from the target speaker and could be from different sex. Three masker locations were also involved (0°, +45° and -45° of elevation) but the target was always simulated at 0° of elevation, resulting in 3 target-masker spatial conditions.Both studies from the literature that are replicated here involved only individual HRTFs and native English (N) speaker participants, while in the current experiment design we investigated the effect of HRTF by also considering a mannequin KEMAR HRTF and the effect of native language by recruiting also non-native English (NN) participants.

For the first experiment, 24 NN speakers were recruited at the University of Málaga (they were all native Spanish speaker), and 16 N speakers were recruited at Imperial College London. The 20 participants for the second experiment were recruited at Imperial College London, evenly split into N and NN speaking groups.

2.2 Stream segregation experiment

A third experiment has been designed based on the experiment of [6], which used a rhythmic-masking release task to assess SSS thresholds. This task involved two streams that were designed to play two rhythmic patterns. When the streams were co-located, both patters sounded as part of a single stream. Then the higher the thresholds the lower the segregation ability. An adaptive procedure was used to measure the minimum angle between the streams (by fixing one stream and moving the other stream) required to correctly identify the two patterns 50% of the time. Regarding the HRTF conditions, the individual HRTF of the participant was used as well as a mannequin HRTF (KEMAR) and the HRTF of another individual (non-individual HRTF), always the same for each participant. The interaural time differences (ITD) of the non-individual and KEMAR HRTFs were modified to match the ones of the individual's ITD. Three spatial conditions were considered: the streams were segregated either along the horizontal plane (fixed stream at -10° or -80°) or along the median plane (fixed stream at -30° elevation). Participants repeated twice each condition they took part in.

Participants could take part to 1, 2 or 3 spatial conditions because the motivation of this experiment was about assessing the effect of HRTF on SSS rather than assessing SSS at different spatial locations. Ten participants were recruited for each condition involving horizontal plane separation, while 15 were recruited for the median plane condition. Participants who took part in at least 2 conditions were 11, 8 of which performed the 3 conditions.

2.3 Numerical metrics analysis

In order to investigate whether the perceptual data were relying on differences in spectral cues between HRTFs, two objectives metrics were computed for a correlation analysis. The spectral distortion factor (SD) was used to assess the deviation between two spectra. The computation consisted in integrating the HRTF spectra in a logarithmic scale (equivalent rectangular bandwidth scale) and then assessing the root-mean-square difference between both spectra to obtain a broadband value. To provide a binaural interpretation of this monaural metric, the maximum across ears was considered. Only spectra from the same HRTF were compared in the computation, thus assessing the amount of spectral cues provided by this HRTF. In addition, the better-ear SNR was also computed. The HRTF spectra used for the computation were the same as for the SD computation, then they were subtracted to each other, and within each frequency band the maximum value between both ears was considered. Finally, the broadband value was obtained by averaging the frequency-dependent better-ear SNR. These values were then used to perform a correlation analysis, where the differences in perceptual data between HRTF conditions were correlated to the differences in objective metrics. The frequency range of the computation was from 0.2 to 20 kHz except for the better-ear SNR in the SRM experiments, which was restricted

to 0.1 to 10 kHz because this is the frequency range relevant for speech intelligibility. For the SSS experiment, the better-ear SNR was computed considering the fixed stream as target. Given the spatial separation between the streams was adapted within spatial conditions, the computation of the metrics consisted in averaging the metric values obtained at each possible stream separation.

3. RESULTS

In order to analyse the results, linear mixed-effects models were designed to describe the data and find the significant factors and interactions. For the first experiment assessing the effect of HRTF on SRM, the fixed effects were target elevation, masker elevation, HRTF individualisation and native language. For the second experiment on SRM, maskers elevation, HRTF individualisation and native language were set as fixed effects. Regarding the experiment measuring SSS, HRTF conditions and spatial separation were set as fixed effects. The factor *listener* was set as random intercept for all the experiments.

The results of two experiments measuring SRM are shown in Fig. 1, as a function of the target-masker absolute elevation difference and HRTF conditions. SRM scores (in percentage of correct answers) were computed by subtracting the co-located condition scores from the separated condition scores. The higher the SRM scores the better the performances, therefore the more someone can achieve intelligibility improvement by spatial separation. The absolute difference of 45 corresponds to the experiment involving two maskers and the other (50 and 100) correspond to the experiment with only one masker. The blue and orange symbols refer to the data collected with N and NN speakers, respectively. To distinguish relative separation between target and masker, upward triangles are used for the conditions involving a masker above the target (always on the left-hand side of the abscissa) and downward triangles for the opposite set up. Note that the highest SRM scores were measured in the second experiment because there were two maskers instead of one.

Regarding the ANOVA outcomes, in the experiment involving 1 masker the four main factors (masker locations, target location, native language and HRTF individualisation) were significant, but no interaction was reported significant. Interestingly, SRM was higher when the sources were spatialised individual HRTFs rather than the KEMAR HRTF. For the experiment involving two maskers, the HRTF and masker location factors were reported as significant but not the native language factor. As for the other experiment, SRM scores were higher when the sources were spatialised with individual HRTFs. The factor native language was significant in experiment one only, which might be explained by either the number of maskers or by the proficiency in English language of NN speakers in experiment, who were all living in the United-Kingdom.

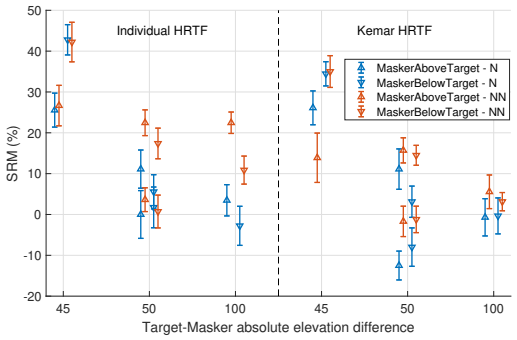

Figure 1. SRM scores average across participants. Error bars indicate standard errors.

The thresholds measured when assessing the effect of HRTF on SSS are shown in Fig. 2. Each color represents a tested spatial condition and abscissa shows the HRTF conditions. The lower the thresholds, the higher the performances, i.e., participants can correctly guess the pattern with a smaller separation angle. The effect of HRTFs seems to be more pronounced for the median plane condition rather than in the two horizontal plane ones. This was confirmed by the ANOVA, the HRTF and spatial separation factors were significant as well as the interaction between both factors. A pairwise comparison with a Bonferroni correction reveals significant differences in HRTF conditions only for the separation along the median plane. Individual HRTF threshold was 6.1° higher (i.e., worse) than KEMAR HRTF threshold; the latter being 9.9° higher than non-individual HRTF. Moreover, for the Individual and KEMAR HRTFs the three spatial conditions were significantly different while only the difference between the Horizontal/-10° and Median conditions was found significant for the non-individual HRTF.

No correlation was significant in any of the SRM experiments. For the experiment measuring SSS, the horizontal plane conditions were discarded from the correlation analysis because no significant differences were found between HRTFs. The correlations considering individual and non individual HRTFs were significant ($r = -0.63$, $p = 0.01$ for SD; $r = 0.61$, $p = 0.02$ for better-ear

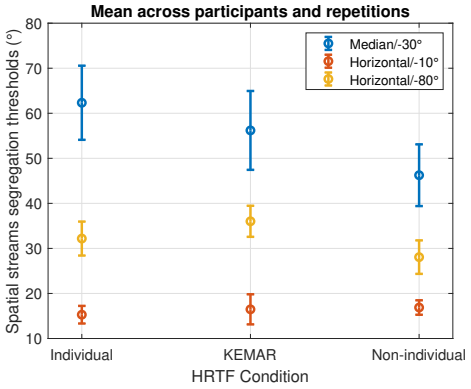

Figure 2. Segregation thresholds as function of the spatial separation and HRTF conditions. Error bars indicate standard errors.

SNR) as well as the correlation comparing Individual and KEMAR HRTFs and considering SD as objective metric ($r = -0.55, p = 0.03$).

4. DISCUSSION AND CONCLUSION

The effect of HRTFs seems to have different (even opposite) behaviours according to the auditory mechanism being assessed, especially in the median plane. In the SRM experiments, performances were better with individual HRTFs, while for the SSS experiment the non-individual and KEMAR HRTFs led to better performances due to their larger spectral cues. The subjective/numerical correlation analysis confirms these opposite behaviours. On one side, the absence of correlation for the SRM experiments suggest that participants relied on something else rather than the magnitude of the spectral cues. On the other side, the participants involved in the SSS experiment performed better with the non-individual and KEMAR HRTFs because these provided more spectral cues, as shown by the significant correlations. It is quite unclear why this opposite behaviours are observed between SRM and SSS, meaning that further investigations are required. Note that this opposite behaviour can be due to differences in the test designs (e.g., ITD individualisation, speech vs white noise bursts, etc).

To conclude, the SRM protocol seems to be more sensitive to HRTF choice rather than the SSS protocol, especially having a prior knowledge of the HRTF spectral cues was rather relevant for SRM but not for SSS, and this was particularly evident in the median plane. However, the SSS protocol shows some threshold differences between non-individual HRTF and KEMAR HRTF that cannot be explained by spectral cues, which means that the protocol might still be relevant for selecting a best/worst fitting HRTF from a database. Further investigation are required to strengthen these conclusions; specifically, it is yet unclear whether better SRM and SSS performances in the experiments are somehow related with a better realism of the rendered audio scene.

5. ACKNOWLEDGMENTS

The authors are grateful for the contribution of Martha Shiell (Eriksholm Research Centre) and of the DIANA Team at the University of Malaga. This study has been supported by SONICOM, a project funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017743.

6. REFERENCES

[1] L. Picinali and B. F. Katz, "System-to-user and user-to-system adaptations in binaural audio," in *Sonic Interactions in Virtual Environments*, pp. 115–143, Springer International Publishing Cham, 2022.

[2] C. Kim, V. Lim, and L. Picinali, "Investigation into consistency of subjective and objective perceptual selection of non-individual head-related transfer functions," *JAES*, vol. 68, no. 11, pp. 819–831, 2020.

[3] M. Cuevas-Rodriguez, D. Gonzalez-Toledo, A. Reyes-Lecuona, and L. Picinali, "Impact of non-individualised head related transfer functions on speech-in-noise performances within a synthesised virtual environment," *JASA*, vol. 149, no. 4, pp. 2573–2586, 2021.

[4] R. L. Martin, K. I. Mcanally, R. S. Bolia, G. Eberle, and D. S. Brungart, "Spatial release from speech-on-speech masking in the median sagittal plane," *JASA*, vol. 131, pp. 378–385, 2012.

[5] V. A. Best, *Spatial hearing with simultaneous sound sources a psychophysical investigation*. University of Sydney, 2004.

[6] M. Shiell and E. Formisano, "Acuity of spatial stream segregation along the horizontal azimuth with non-individualized head-related transfer functions," in *Proc. of the 23rd ICA*, (Aachen, Germany), pp. 6644–6649, 2019.