JANUARY 16 2024

Spatial release from masking in the median plane with non-native speakers using individual and mannequin head related transfer functions

Daniel González-Toledo ; María Cuevas-Rodríguez ; Thibault Vicente ; Lorenzo Picinali ; Luis Molina-Tanco ; Arcadio Reyes-Lecuona

J. Acoust. Soc. Am. 155, 284–293 (2024)
https://doi.org/10.1121/10.0024239

JASA ARTICLE

Spatial release from masking in the median plane with non-native speakers using individual and mannequin head related transfer functions

Daniel González-Toledo,[1] María Cuevas-Rodríguez,[1] Thibault Vicente,[2] Lorenzo Picinali,[2] Luis Molina-Tanco,[1] and Arcadio Reyes-Lecuona[1,a)]

[1]*Telecommunication Research Institute (TELMA), Universidad de Málaga, ETSI Telecomunicación, 29010 Málaga, Spain*

[2]*Audio Experience Design, Dyson School of Design Engineering, Imperial College London, London SW7 2DB, United Kingdom*

ABSTRACT:

Spatial release from masking (SRM) in speech-on-speech tasks has been widely studied in the horizontal plane, where interaural cues play a fundamental role. Several studies have also observed SRM for sources located in the median plane, where (monaural) spectral cues are more important. However, a relatively unexplored research question concerns the impact of head-related transfer function (HRTF) personalisation on SRM, for example, whether using individually-measured HRTFs results in better performance if compared with the use of mannequin HRTFs. This study compares SRM in the median plane in a speech-on-speech virtual task rendered using both individual and mannequin HRTFs. SRM is obtained using English sentences with non-native English speakers. Our participants show lower SRM performances compared to those found by others using native English participants. Furthermore, SRM is significantly larger when the source is spatialised using the individual HRTF, and this effect is more marked for those with lower English proficiency. Further analyses using a spectral distortion metric and the estimation of the better-ear effect, show that the observed SRM can only partially be explained by HRTF-specific factors and that the effect of the familiarity with individual spatial cues is likely to be the most significant element driving these results.

https://doi.org/10.1121/10.0024239

(Received 16 March 2023; revised 23 November 2023; accepted 12 December 2023; published online 16 January 2024)

[Editor: Pavel Zahorik] Pages: 284–293

I. INTRODUCTION

The ability to understand target speech in a multiple-talker scenario has attracted the interest of researchers over the last seven decades. In these conditions, often referred to as the "cocktail-party problem" (Cherry, 1953), a listener's auditory system uses complex mechanisms to improve speech intelligibility. These mechanisms have been widely studied considering many different characteristics of the acoustic signals, the conditions, and listeners.

According to Bronkhorst (2015), there are three main lines of research in the field. The first focuses on how simultaneous speech sounds interfere with each other at the peripheral level (see Bronkhorst, 2000, for a complete review). The second line focuses on *stream segregation*, the ability of the auditory system to segregate the target speech from other masking sounds, and to connect elements of the same sound stream across time into a group, which was defined by Bregman (1990) as sequential (*streaming*) and simultaneous (*segregation*) organisation. The third line addresses the process of *selective attention*, which can help in unmasking one of those streaming groups by focusing attention on the target speech, once it is segregated in a group.

At the peripheral level, the interfering sounds, or maskers, are considered to produce only *energetic masking*, caused by an overlap in the spectro-temporal domain between target and masker signals. However, an inability to segregate streams or focus attention on the target signal can produce an additional type of masking that is referred to as *informational masking*. The definition of informational masking is still elusive, but it can be considered as a failure in processing the target signal correctly even though it does not present spectro-temporal overlap with the masker signals (i.e., it cannot be explained by energetic masking). In that sense, energetic masking can be associated with masking occurring in the auditory periphery (low-level auditory processing), while informational masking can be referred to as any masking in the central auditory processing (high-level auditory processing) due to factors such as masker uncertainty or the resemblance between the target and masker (Culling and Stone, 2017; Durlach *et al.*, 2003; Kidd and Colburn, 2017) provide an extensive review of energetic and informational masking, including some suggestions of distinctions between the two.

When speech is spatially separated from the competing sources, a benefit in intelligibility can be observed with respect to a co-located setup. This is known as spatial release from masking (SRM, e.g., Plomb and Mimpen, 1981, or for a recent

a)Email: areyes@uma.es

https://doi.org/10.1121/10.0024239

review, Culling and Lavandier, 2021). When sources are separated in azimuth, differences in binaural cues contribute to SRM (Cuevas-Rodriguez *et al.*, 2021; Culling *et al.*, 2004; Hawley *et al.*, 2004; Jones and Litovsky, 2011; Lavandier and Best, 2020). In such situations, a difference in interaural level differences (ILD) between target and masker leads to a difference in signal-to-noise ratio (SNR) across ears, and therefore, to better ear listening (Brungart and Simpson, 2002; Shinn-Cunningham *et al.*, 2001; Usher and Martens, 2007; Westermann *et al.*, 2015). On the other hand, interaural time differences (ITDs) are also known to be useful for SRM, allowing the auditory system to partially cancel the masker according to the Equalisation-Cancellation theory (Durlach, 1963). This can be done through the application of delay and attenuation in one ear, aligning the masker signal with that received at the other ear, and allowing for subtraction between the two signals. The improvement due to this auditory mechanism is often referred to as binaural unmasking advantage.

Few studies have focused on the separation of sources at the same distance within sagittal planes. In these planes, binaural cues are the same along the surface, such as in the median sagittal plane. The only differences between source positions can be observed in spectral monaural cues. However, these still contribute to SRM (see Sec. II).

All the binaural and monaural spectral cues mentioned before change depending on the shapes of the human head, torso, and pinna. Such filters are called head related transfer functions (HRTFs). HRTFs are specific to each individual because they are based on morphological features. Individualising HRTFs (e.g., acoustically measuring them for each individual, Xie, 2013) when assessing speech intelligibility tasks leads to differences in SRM in the horizontal plane, as binaural cues vary from one HRTF to another (Ahrens *et al.*, 2021; Cuevas-Rodriguez *et al.*, 2021). Ahrens *et al.* (2021) applied the binaural speech intelligibility model from Jelfs *et al.* (2011) on different HRTFs to show that for a masker placed on the horizontal plane between 75° and 90°, SRM can vary 4 dB according to the HRTF, which suggests that energetic unmasking varies across HRTFs. This was further confirmed by Cuevas-Rodriguez *et al.* (2021) via a perceptual experiment using only non-individual HRTFs. However, this effect may no longer be significant when using speech maskers (i.e., with informational masking Drullman and Bronkhorst (2000); Zenke and Rosen (2022). Still, Zenke and Rosen (2022) involved only one SRM condition and Drullman and Bronkhorst (2000) did not provide a breakdown of the statistical analysis, thus some significant effects might have been omitted.

The effect of HRTF on SRM in the median plane is yet to be investigated. Specifically, it is still an open question whether personalising spatial cues provided by HRTF can improve SRM when sources are separated in the median plane. It can be speculated that HRTF individualisation should provide a better-perceived location of the sources resulting in a larger SRM, but this has to be verified experimentally.

II. PREVIOUS WORK ON SRM IN THE MEDIAN PLANE

Recently, Berwick and Lee (2020) have observed spatial unmasking in the median plane in energetic masking conditions. They placed a target speech coming from the front (0° elevation), and a speech-shaped noise at different elevations using loudspeakers. They found a significant effect of masker location with SRM up to 3.5 dB when the masker was at −30°, with respect to the co-located condition.

SRM in the median plane was already explored with non-individual HRTFs by Bolia *et al.* (1999), using the coordinate measure response (CRM) corpus (Bolia *et al.*, 2000). Later, McAnally *et al.* (2002) replicated this study using individual HRTFs with residual ITD explicitly removed in the impulse responses. They measured a SRM of about 1.3 dB when the target and masker were at different elevations in the median plane (two different same-sex speakers were used as stimuli), thus showing the relevance spectral cues for SRM in the median plane. Using a similar setup and the same CRM corpus, Best (2004) conducted a series of experiments including one target speech simulated in front of the listener, against two speech maskers placed together at different elevations. The importance of high frequency was studied by considering the CRM stimuli low-pass filtered at 8 kHz, and the original broadband corpus. They found a significant effect of separation in the median plane, as the percentage of correct words improved by 21% for the low-pass filter corpus, and by 28% for the broadband corpus on average across spatial and speaker conditions. These studies used either non-individual or individual HRTFs, but none of them considered HRTF as an independent variable to investigate its role in SRM.

Worley and Darwin (2002) investigated SRM in the median plane with two competing speech sources using loudspeakers in an anechoic chamber. They found that low-pass filtering of the stimuli degrades SRM, suggesting that there are relevant spatial cues for SRM in the median plane above 5 kHz. Such cues should therefore be sensitive to HRTF individualisation. To confirm that, they repeated the experiment using virtual sources rendered by convolving the stimuli with non-individualised HRTF. SRM was significantly lower in this setup, thus leading to a lower effect of low-pass filtering. These results suggest once again that HRTF plays an important role in SRM in the median plane.

In a later study, Martin *et al.* (2012) repeated the experiment reported by McAnally *et al.* (2002) to further investigate the effect of HRTF asymmetries on SRM in the median plane. In addition to the individual HRTF, other HRTFs with two left ears, two right ears, or an average of both ears were involved. The individual HRTF led to higher SRM compared to any of the manipulated HRTFs. They also took for each participant the best score between the two-left-ears and the two-right-ears HRTF conditions (better diotic condition in their paper). In this analysis, the average performance showed values similar to those of the full individual HRTF. This suggests the possibility of having a better ear that

provides more cues for SRM in the median plane. In other words, if the participant had the same score as the individual HRTF and the two-left-ears HRTF, then their left ear was the better ear. A personalised HRTF containing these individual asymmetries should therefore be beneficial for SRM in the median plane.

From these studies, it seems evident that, when individual HRTFs are deteriorated and SRM in the median plane is degraded. This suggests that HRTF individualisation might have an important role in speech-on-speech masking in the median sagittal plane, but to the best of our knowledge, the effect of HRTF individualisation on SRM has not been investigated yet in such conditions. The aim of the present study is to investigate SRM sensitivity to HRTF individualisation for sources located in the median plane. The experimental procedure is based on the one by Martin *et al.* (2012), but comparing individual vs non-individual HRTFs, thus investigating the importance of HRTF individualisation in SRM in the median plane. Our first hypothesis is that SRM in the median plane is improved when an individual HRTF is used to render the sources because the perceived locations of the sources should be more vivid.

Furthermore, differently from Martin *et al.* (2012), our study has been carried out with non-native English speakers, using an English corpus. Past studies have consistently indicated that speech perception performance is significantly impacted by various forms of masking, especially among non-native speakers. Lecumberri *et al.* (2010), in a comprehensive review of experimental studies on non-native speech perception in adverse conditions, highlighted that both energetic and informational masking exert a more pronounced influence on non-native listeners compared to their native counterparts. Moreover, Cooke *et al.* (2008) conducted an experiment using speech on speech masking involving native and non-native speakers. They found similar patterns across speaker combinations, both with native and non-native speakers having equal unmasking benefits from differences in $F0$, but with a lower percentage of correct recognition in the non-native ones. Based on these studies, our second hypothesis is that the overall performances in this speech-on-speech task of non-native English speakers are lower if compared with native ones. This work also complements previous work by the same team, which looked at SRM in the horizontal plane with non-individualised HRTFs (Cuevas-Rodriguez *et al.*, 2021), broadening our understanding of the impact of HRTFs in speech intelligibility amongst competing sources.

III. MATERIALS AND METHOD

A. Participants

Twenty-four participants were recruited among students and researchers of the School of Telecommunication Engineering at the University of Málaga. Three of them were coauthors of this report. All reported normal hearing and Spanish as their first language. One subject was discarded after reporting to have problems understanding the language during the experiment. During an initial training block without a masker, this subject presented seven wrong answers out of eighteen, with the other subjects presenting no wrong answers (18 participants), one wrong answer (four participants), or two wrong answers one participant). Thus, a total of 23 participants were included in the analysis, 18 of them with ages between 18 and 29 years and five of them with ages between 30 and 50 years. They received a USB stick as compensation for their participation. Participants reported a certified English level of: A1 (one participant), B1 (seven participants), B2 (nine participants), C1 (four participants), and C2 (two participants).

B. Apparatus

Head-related impulse responses (HRIRs) were measured using the sweep-sine technique at a sampling rate of 48 kHz (Farina, 2007) in an acoustically treated room. The recordings were made with a pair of Knowles FG-23329-P07 in-ear microphones (Knowles, Itasca, IL), placed at the entrance of the ear canals and embedded in foam earplugs. The interface to the computer was a MOTU (Cambridge, MA) 896 mk3. The participant's head was centred using a set of three coincident laser beams and asked to stay still during the measurement process. The HRIRs were measured at three different positions in the median plane (i.e., azimuth 0°), with 50°, 0°, and –50° elevations. The HRIRs of a KEMAR dummy head mannequin were also measured with the same setup and equipment as the individual ones but in a different acoustically treated room.

Given that the KEMAR and individual HRIRs were measured in different rooms, in order to minimise potential perceivable differences that would not have been related to HRTF individualisation per se, a variation of the frequency-dependent windowing (Karjalainen and Paatero, 2001) with eight half Hann windows, as proposed by Gutierrez-Parera (2020, Section 4.3.2), was applied to the HRIRs to remove any reflection.

A specific application was developed for managing the whole experimental process; it provided the user interface, performed binaural rendering (using the 3DTI Toolkit, see Cuevas-Rodríguez *et al.*, 2019), a C++ open-source library for real-time binaural spatialisation), allowed users to respond to each trial and saved the responses to file. A standard computer without specialised digital signal processing hardware was used to run the application. Participants interacted with the application using a mouse. The procedure was automatically sequenced for the whole experiment, without any intervention of the operator. All the activities performed by the participants were logged. An RME Babyface Pro USB audio interface (RME, Ft. Lauderdale, FL) was used to play back the stimuli, which were presented binaurally via a pair of Sennheiser HD600 headphones (Sennheiser, Wedemark, Germany). The headphone transfer function (HpTF) was measured for each individual. Once the microphones where placed at the entrance of the ear canals, participants were asked to put the headphones on.

https://doi.org/10.1121/10.0024239

Then, the impulse response on each side was measured using the same protocol as for the HRIR. This procedure was repeated five times asking the participants to take off and put headphones back on their ears each time. The spectra were then averaged and the minimum phase individual HpTF was obtained. An equalization filter was calculated using the technique presented by Engel *et al.* (2022) in order to remove significant peaks and notches from the frequency response.

C. Stimuli

The CRM corpus (Bolia *et al.*, 2000) was used for the experiment. It consists of sentences following the same structure: "Ready *call sign*, go to *colour number* now." There are eight call signs, four colours, and eight numbers resulting in 256 sentences when all combinations are considered. The corpus is recorded with four male speakers and four female speakers. As in Martin *et al.* (2012), the original CRM recordings were used, in which sentences are band-pass filtered from 200 Hz to 18 kHz (the publicly available version being low-pass filtered at 8 kHz) and adjusted to have the same root mean square amplitude. Both target and masker sentences were taken from the CRM corpus, ensuring that they were always uttered by different speakers of the same sex. Sentences were presented at a sound level of approximately 60 dBA and the target-to-masker ratio was 0 dB. Sound source spatialisation was purely anechoic.

D. Experimental conditions

The spatial sound was simulated using either the individual HRTF or the KEMAR HRTF. Maskers and targets were simulated at three different positions (expressed in polar coordinates, azimuth and elevation): (0°, 50°), (0°, 0°), (0°, –50°). This made a total of 18 different conditions: 2 (HRTFs) × 3 (target positions) × 3 (masker positions).

E. Procedure

Participants came to the lab on two different days. Their HRTFs were measured on the first day, and they performed the speech-on-speech masking test on the second.

During the experiment, participants were seated in the same room where the individual HRIRs were measured, in front of a monitor, with a mouse, and wearing headphones. The two spatialised sentences (one for the target and another one for the masker) were presented simultaneously to the participants on each trial. Both sentences, uttered by a different person but of the same sex, could come from the same or different directions. The sex of the speakers was randomly selected at the beginning of each trial. From among the speakers of that sex, two were randomly selected, one as a target and one as a masker. In the following trial, the process was repeated without any influence from the previous one. Target sentences always used the word *Baron* as a call sign and a random combination of colour and number. Maskers always used another call sign, colour, and number than 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 interface used to select the target combination is shown in Fig. 1. The target position was displayed on the left side of the screen before the trial started, and then, participants triggered the trial by pressing an OK button on the interface, and the sentences were played back. Participants were instructed to focus in advance their attention on the target position and ignore the masker. It was strongly emphasised that by focusing on the target direction participants could take advantage of selective attention based on spatial cues. Once the participant responded, their answer was recorded and a new trial started.

Trials were grouped into blocks, and blocks into sessions. In each block, the target is always in the same

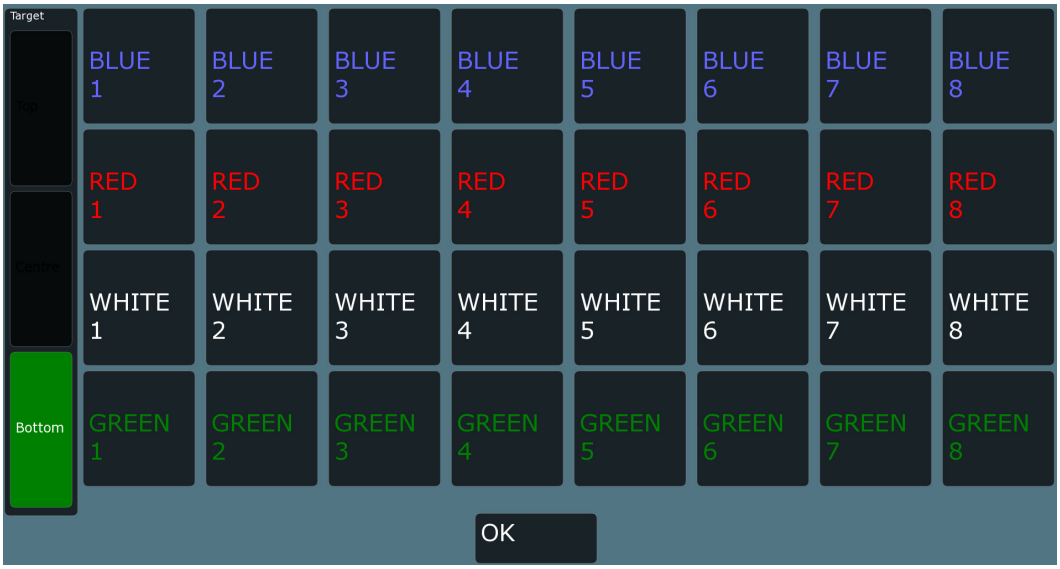

FIG. 1. (Color online) Interface to select the colour-number combination of the target sentence. Note the left part of the screen, where information about the position of the target is clearly displayed to the participant.

position. In this way, each block presents six different conditions (2 HRTFs × 3 masker positions). These conditions are repeated three times within each block, making a total of 18 trials per block, which were presented in a random order using Latin squares. The blocks were grouped by three (so that the three target positions were simulated in each session), forming a session of 54 trials (three blocks of 18 trials each) randomised using Latin squares as well. There were a total of six sessions plus a training session at the beginning. Participants were asked to leave the room and rest after each session since we needed them to be able to maintain their ability to focus their attention throughout all sessions.

The training session was the first; it consisted of three blocks, but differed from the others in that feedback was provided to the participants. After each participant's answer, the button corresponding to the correct answer was highlighted and the correct answer was displayed in text, together with the answer provided by the user. In addition, in the first block of this training session (i.e., the first eighteen trials) only targets were played. The aim was to give the participants the possibility to familiarise themselves with the voices and sentences.

F. Objective metrics

Two objective metrics were computed in order to assess whether the difference in perceptual data between HRTF conditions can be explained by differences in HRTF-specific numerical attributes. In other words, the metrics are here to control if the variation in the perceptual data is due to a spatial cue or to an improvement of the perceived spatial location of the sources. The two metrics are the spectral distortion metric (SD, Andreopoulou and Katz, 2022) and SRM predicted by the Jelfs model (Jelfs SRM, Jelfs *et al.*, 2011).

The computation of SD consists of smoothing the HRTF spectra at both ears into bands using equivalent rectangular bandwidth to approximate human perception. The metric is defined as the root mean square across the auditory frequency range of the difference between two HRTF spectra, and it quantifies how much one HRTF spectrum deviates from the other. Here, the SD is computed at each ear for each target and masker HRTF spectra involved in the experiment. To provide a binaural version of this monaural metric, the maximum (SD_{Max}) and the average (SD_{Mean}) across ears are considered.

The Jelfs model takes as input a target HRIR and a masker HRIR, which are then passed through a gammatone filterbank. A first component is predicting binaural unmasking advantage by using the formula proposed by Culling *et al.* (2004, 2005). The second model component is predicting better-ear listening by taking the higher SNR between the left- and right-ear SNRs. The frequency-specific values from the two model components are summed, and then integrated across frequency bands using a Speech Intelligibility Index-weighting, which provides more weight to the frequency bands relevant for speech intelligibility, i.e., between 100 Hz and 10 kHz (ANSI, 1997). The model output can be compared to SRM.

The Jelfs model deviates from the SD computation in major aspects, which makes both metrics worth considering. First, the Jelfs model computation determines the better ear (the ear having the higher SNR) per frequency band, while the SD computation determines the ear providing the larger broadband spectral difference (SD_{Max}), or just averages the broadband spectral differences across ears (SD_{Mean}). Thus, the influence of the broadband asymmetry versus by-band asymmetry on the perceptual data will be explored. This has the potential to complete the finding of Martin *et al.* (2012) (see their Sec. I) showing that participants could reach the scores obtained with the non-processed individual HRTF using either two-left-ears or two-right-ears HRTF. Second, due to the use of a root mean square, the SD computation leads to the same value when target and masker locations are swapped, which is not the case for the Jelfs model. In addition, Brungart (2001) showed that for two competitive talkers the absolute intensity difference matters more than the relative SNR; the consideration of the SD metric will further investigate this aspect. Third, the Jelfs model also takes into account the effect of ITDs, which might be relevant, but rather limited in this specific case. Finally, SD has been used in different studies to assess the differences between HRTFs (e.g., Andreopoulou and Katz, 2022; Jo *et al.*, 2009; Takane, 2016). Then, it is of interest to investigate the relevance of this metric for speech intelligibility.

Both metric values are used to assess whether the variance in SRM between HRTF conditions can be explained by the variance in HRTF-specific numerical attributes. To do so, the differences in SRM between individual and KEMAR HRTF are compared to the changes in SD and Jelfs SRM.

Given the fact that the SD is a symmetric function, SRM measured for one particular target-masker setup is averaged with SRM measured swapping target and masker locations. In other words, the SD relies on the location of the sources regardless the specific locations of the target and masker. For instance, the source locations are the same when the masker is placed at 0° and the target at 50° or when the masker is placed at 50° and the target at 0°. However, SRM relies on the actual locations of the target and masker, i.e., it is very unlikely to obtain the same SRM when you swap target and masker positions. Then, for the correlation analysis between SRM and SD, SRM measured in a specific target-masker setup (e.g., target at X°, masker at Y°) is averaged with SRM measured in the mirror masker-target setup (target at Y°, masker at X°). By doing so, this average SRM is dependent on the pair of sources location regardless of the specific positions of the target and masker, such as the SD.

Thus, 5 correlation coefficients are computed per method to obtain SD (i.e., SD_{Max} or SD_{Mean}): one for each of the 3 source elevation sets (0°/50°; 0°/ –50°; 50°/ –50°), one concatenating the values across source elevation sets and a last one averaging across those values. Regarding the correlations with Jelfs SRM, 8 coefficients are computed: one for each of the 6 target-masker elevation sets (0°/50°; 50°/0; 0°/ –50°; –50°/0°; 50°/ –50°; –50°/50°), one concatenating the

https://doi.org/10.1121/10.0024239

values across target-masker sets, and a last one averaging across those values.

IV. RESULTS

A. Perceptual data

The percentages of correct answers and SRM averaged across participants are shown in Tables I and II from the Appendix, respectively. We pooled the results of all trials for each participant and condition, excluding the training sessions. Thus, the average of each condition for each participant is calculated over 18 trials. SRM scores are computed as the difference in terms of % of correct answers between the separated conditions and the co-located conditions for a given target elevation. Figure 2 displays SRM as blue symbols as a function of target elevation, target-masker separation and HRTF conditions. The two HRTF conditions are separated into two panels, left panel for individual HRTF condition, and the right panel for the KEMAR HRTF condition. The relative target-masker separation conditions change along the horizontal axis, and the different target elevation conditions are reported using different symbols.

SRM is generally positive, meaning that participants benefited from the spatial separation between the target and masker locations to improve intelligibility. SRM seems to be lower when the stimuli were spatialised with the KEMAR HRTF if compared with the individual HRTF. The highest scores were obtained for –50° Target Elevation spatialised using individual HRTF.

Two linear mixed-effect models were designed to analyse the datasets, one for the percentage of correct answers and another for the SRM. The fixed effects were the three factors involved in the experiment, namely, target elevation, masker elevation, and HRTF individualisation. The factor listener was set as a random intercept. The equation of the model can be found in Eqs. (1) and (2). The models were designed using the *lmer* function from the *car* R package, and the significance of the factors in the different models was assessed using the anaylsis of variance (*ANOVA*) function from the same package,

$$\begin{aligned} Answers \sim\ & HRTF\_individualisation \\ & * Target\_Elevation * Masker\_Elevation \\ & + (1|Listener), \end{aligned} \tag{1}$$

$$\begin{aligned} SRM \sim\ & HRTF\_individualisation * Target\_Elevation \\ & * Masker\_Elevation + (1|Listener). \end{aligned} \tag{2}$$

The main significant factors reported by the model on the percentage of correct answers were target elevation [$F(2, 374) = 25.49$, $p < 0.0001$], masker elevation [$F(2, 374) = 23.07$, $p < 0.0001$], and HRTF individualisation [$F(1, 374) = 17.97$, $p < 0.0001$]. The interactions between target elevation and masker elevation [$F(4, 374) = 14.98$, $p < 0.0001$] or HRTF individualisation [$F(2, 374) = 4.28$, $p = 0.01$] were also found to be statistically significant. The first interaction is a proof of SRM, as already observed in Fig. 2. Interestingly, the interaction between the three main factors was also significant [$F(4, 374) = 3.17$, $p = 0.01$], which means that there was most likely an effect of HRTF individualisation on SRM.

The second model (SRM), confirmed the results of the first model, i.e., target elevation [$F(2, 248) = 11.00$,

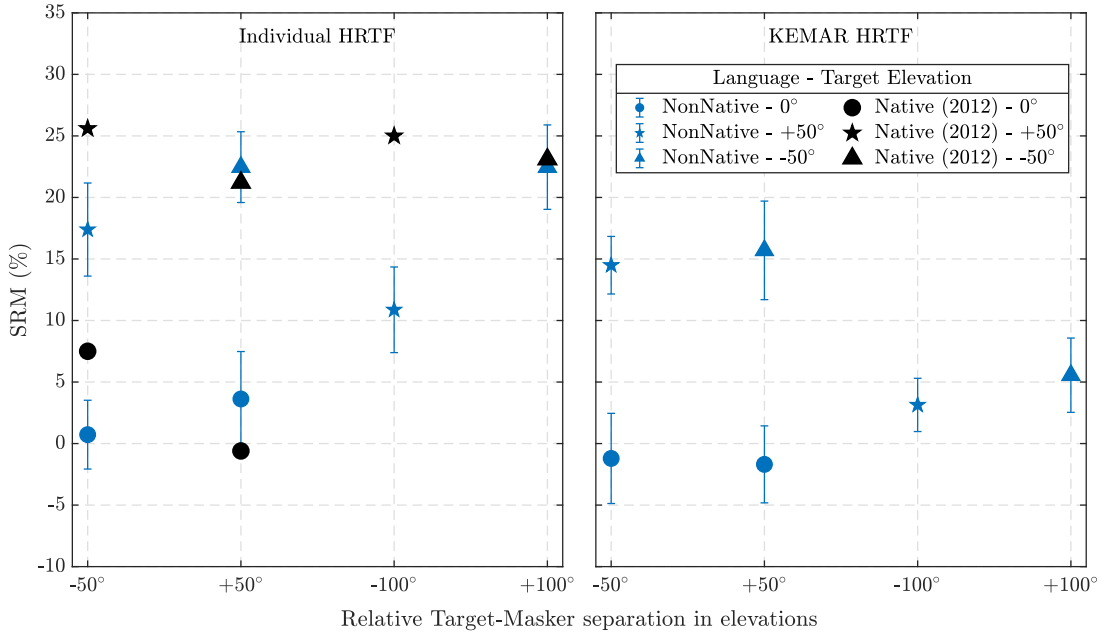

FIG. 2. (Color online) SRM in terms of % of correct answers between the separated and the co-located conditions for a given target elevation, as a function of the relative Target–Masker separation in elevation. Each symbol is attributed to a target elevation. The left and right panels show the data obtained with Individual and KEMAR HRTF, respectively. For comparison purposes, the data of Martin *et al.* (2012, black symbols) are plotted next to the current data (blue error bars showing standard errors). The sample size of the Martin *et al.* study was eight participants. Error bars on SRM cannot be reported for the Martin *et al.* data because they were not available in their publication.

$p < 0.0001$], masker elevation [$F(2, 248) = 5.00$, $p < 0.01$], and HRTF individualisation [$F(2, 248) = 13.75$, $p = 0.0003$] were significant. *Post hoc* pairwise comparisons with Bonferroni adjustment (designed for multiple comparisons) revealed that SRM was 6.9% larger [$t(248) = 3.71$, $p = 0.0003$] when the stimuli were spatialised with individual HRTFs. Regarding target elevation, SRM was lower when the target was spatialised at 0°, by 12.1% compared to –50° [$t(248) = 4.603$, $p < 0.0001$] and by 8.1% compared to +50° [$t(248) = -3.08$, $p = 0.007$]. Only one comparison found significant differences in SRM for masker elevation: SRM was lower (by 8.1%) when the masker was at –50° compared to 0° [$t(248) = -3.05$, $p = 0.008$].

The number of correctly identified digits and colours was also analysed independently. In general, the same trends were observed but the absolute percentages were higher compared to the ones presented previously. Regarding target digits, SRM was approximately 4% for the KEMAR HRTF and 11% for the individual HRTF. The analysis on target colours shows lower SRM of around 4% for KEMAR HRTF and 8% for the individual HRTFs. In order to analyse whether the participants were confounding the target and masking speakers, a further analysis was made. This revealed that participants answered a number or a colour that had not been told by the speakers only in 2.5% of the trials. In other words, if the participants did not get the answer right it was very likely that they picked up at least one keyword from the masking speaker.

To further investigate the influence of the different levels of English proficiency, we considered three groups of participants: basic (A1, A2, and B1; eight participants), intermediate (B2; nine participants), and advanced (C1 and C2; six participants). Then the linear mixed-effect model for SRM was modified to include English proficiency as an extra factor. No significant effect of the latter alone was found but it revealed a significant interaction between HRTF individualisation and English proficiency [$F(2, 236) = 4.153, p = 0.017$]. Pairwise comparisons between mannequin HRTF and individual HRTF for the three groups showed significant differences for basic and intermediate levels [$t(236) = 3.1, p = 0.02$; $t(236) = 3.74, p = 0.0021$], but not for the advanced level. This could suggest that for a high level of language proficiency, HRTF individualisation may have a lesser effect. However, the number of participants per group is too low, and further research is needed in order to explore this hypothesis. A three-factors interaction was reported as significant between English proficiency, target location, and HRTF individualisation. However, we are not able to provide a relevant explanation of that interaction.

B. Spectral distortion and predicted SRM

Figure 3 shows the relationship between the difference in SRM between HRTF conditions and the difference in the two HRTF-based metrics, therefore SD and Jelfs SRM, each dot representing one participant per metric computation method. The values shown here are the ones obtained by averaging the results across spatial conditions (see Sec. III F). Quadrants 1 and 3 are coloured in green to show that the perceptual outcomes seem to be somehow generally proportional with the HRTF-based metrics. In other words, they had more Jelfs SRM or SD with their own HRTF, and they had better measured SRM (Quadrant 1); inversely, if they had lower Jelfs SRM or SD with their own HRTF, they had lower measured SRM (Quadrant 3). In Quadrant 2 (blue Quadrant), participants had smaller spatial cues with their own HRTF if compared with the KEMAR, but still showed higher SRM. In this case, they could not use the cues of the KEMAR as successfully as their own, even though KEMAR's cues were larger. In Quadrant 4 (red Quadrant) are the participants for which lower SRM was observed with their own HRTF even though these had bigger cues.

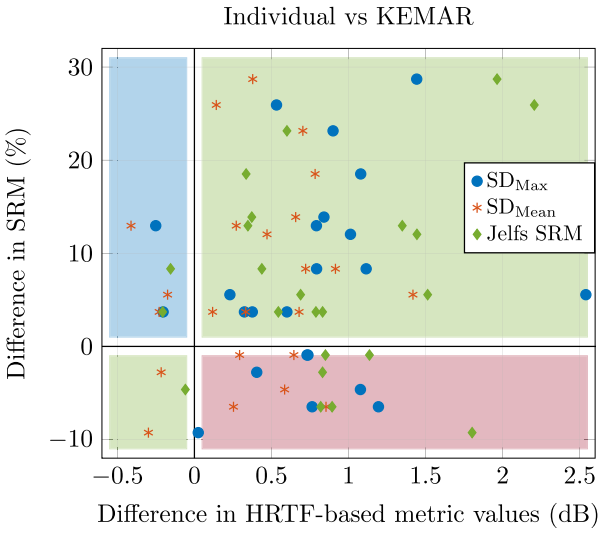

FIG. 3. (Color online) Participants' performance differences in terms of SRM between individual and KEMAR HRTFs, plotted in relation to differences in SD and Jelfs SRM (HRTF-based metrics). Each dot represents one participant per metric computation method.

Generally, the participants had larger cues with their own HRTFs. This can be explained by the fact that the participants' HRTFs are asymmetric, while the KEMAR HRTF is rather symmetric (regardless of the variations in the microphone position). Furthermore, the morphology of real heads is more complex, both in terms of geometry and acoustical properties, than the one of a KEMAR dummy head. Hence, it is likely that at least one ear shape provides more spectral cues than the KEMAR ear shape (i.e., bigger SD). This is confirmed by the method to compute the SD. When the average across ears is used (orange stars), five participants have a lower SD compared to the KEMAR, however, when considering the maximum across ears (blue dots) only two participants have a lower SD with their own HRTFs. Moreover, Jelfs SRM also emphasizes this asymmetry: three participants had lower SRM with their own HRTFs (green diamonds). Regarding the correlation analysis results, none of the correlation coefficients were found to be significant.

V. DISCUSSION

This study was designed to extend the results of Martin *et al.* (2012), where SRM was measured in the median plane

using different versions of individual HRTFs, by investigating the effect of non-individual versus individual HRTFs. Both studies simulated the target and masker at the same elevation angles. Using individual HRTFs, they got a percentage of correct answers of 62.3% on average across co-located conditions and 79.3% on average separated conditions. In our case, 50.7% and 63.4% of correct answers were measured, respectively. Thereby, overall SRM is comparable across studies (15% versus 12.7%) even though the absolute percentages of correct answers obtained in the current study are lower, which might be explained by the difference in native language.

As mentioned in Sec. III A, all participants were Spanish native speakers, while the CRM corpus is in English. It is well known that speech perception performance is more affected by different types of masking when participants are non-native speakers. Lecumberri *et al.* (2010) did a complete review of experimental studies about non-native speech perception in adverse conditions and identified that both energetic and informational masking have a greater impact on non-native listeners than on native ones. Moreover, Cooke *et al.* (2008) conducted an experiment measuring speech-on-speech intelligibility using the GRID corpus (Cooke *et al.*, 2006), which is in English as well, and similar to the CRM corpus (different speakers, same structure across sentences and two keywords per sentence, being a number and a letter), involving both native English and Spanish speakers in several tasks with different combinations of target and maskers. They found similar patterns across speaker combinations in both native and non-native speakers, but with a lower percentage of correct recognition in the latter group (around 10%–20% difference). This is in line with our results, and it provides an explanation of the overall lower values found in the current study as opposed to Martin *et al.* (2012).

However, Cooke *et al.* (2008) found that both native and non-native speakers drew equal unmasking benefits from differences in $F0$ between target and masker. In line with that, Lee *et al.* (2010) showed that stimuli blocked by a speaker yielded higher accuracy and shorter reaction time compared to mixing them within blocks, but that effect was similar for native and non-native listeners. This suggests that processes which make use of these low-level cues are independent of the language. In a similar way, and more specifically related to our study, spatial separation of target and masker in the horizontal plane seems to provide equivalent benefits in releasing from masking for native and non-native listeners (Ezzatian *et al.*, 2010). This would support the fact there was no significant effect of English proficiency alone on SRM. However, the interaction between English proficiency and HRTF individualisation was significant suggesting that the perceived spatial location mattered more for non-native speaker with low proficiency. This must be further tested to draw a conclusion.

Our outcomes, therefore, reinforce those in the literature, supporting the need for taking into account a systematic offset in performance in favour of native speakers, but we consider that the validity of our results is not compromised because of using a closed-set sentence corpus for speech-on-speech experiment in a language different to the participants' mother tongue. Nevertheless, further studies comparing both native and non-native listeners should be done to confirm this.

On the other hand, the benefit of spatial separation is observed not only when the sources are physically separated, but also when they are *perceived* as spatially separated. The auditory system tends to integrate a signal and its reflection into a single auditory object at a location near the first-arriving signal. This is known as the precedence effect (see Zurek, 1987) and it can be used to create an illusion of a spatial separation between target and masker sources while keeping the spatial cues unchanged. Freyman *et al.* (1999) investigated the effect of perceived spatial separation on target unmasking using loudspeakers and the precedence effect. The results showed that the perceived spatial separation provided an advantage for both informational and energetic maskers, but more so for the first (8 dB versus 1 dB at 60% of intelligibility, respectively). This indicates that the representation of speech masker location in the auditory system is essential for SRM.

In this sense, the effect of HRTF individualisation was investigated by considering a non-individual HRTF measured from a KEMAR dummy head and torso in addition to the individual HRTFs. The results suggest that there might be an effect of being familiar with someone's own spatial cues. In other words, the auditory system seems to be able to take better advantage of the spectral cues present in the individual HRTF for unmasking target speech among concurrent speech when they are separated in the median plane. This is likely due to a better perceived spatial separation of the sources, which is supported by the correlation analysis that reported no significant interaction between the perceptual data and the numerical metrics. This suggests that neither the Jelfs SRM model nor SD are sufficient to explain the changes in the SRM data, meaning that the SRM improvement when HRTFs are individualised is not necessarily due to the increase in the spectral cues. Hence, the current finding seems to support and extend the work of Freyman *et al.* (1999) and refutes the ones of Drullman and Bronkhorst (2000) and Zenke and Rosen (2022), whose interest was focused on the horizontal plane. Further research is needed to confirm this speculation; for example, in order to reduce the test time, the amount of energetic masking was evaluated by means of objective metrics, but ideally, the involvement of a speech-shaped noise or a non-intelligible noise-vocoded speech (to preserve some speech-like modulations) would provide a better assessment.

This idea is also in line with Oh *et al.* (2022), where correlations were found between SRM and localisation acuity in the horizontal plane, although not statistically significant ($p > 0.07$), and also with Srinivasan *et al.* (2022), where it was found that localisation acuity for 1/3-octave-wide Gaussian noise centred at 500 Hz was a significant predictor of SRM. In order to get a deeper insight into this link between the use of spatial cues to release from masking or

for localisation tasks, further studies are needed and new potential areas of future research can be identified. For example, it is known that sound localisation performances using non-individual HRTFs can improve with perceptual training (Picinali and Katz, 2023; Steadman *et al.*, 2019). Would such acquired perceptual training **transfer** to a speech-on-speech SRM task, therefore resulting in a significant improvement of performances with the trained non-individual HRTF? Would these performances be comparable with the ones achieved using individual HRTFs?

Finally, it is worth noting that this study has used a non-individual HRTF an HRTF measured with a KEMAR mannequin, and not an HRTF measured from another subject. Considering this, a question arises as to whether the SRM differences found between individual and mannequin HRTFs may be due to the similarities or differences between the two, or again to a KEMAR-specific issue (e.g., symmetry of the head). To resolve this question, more research is needed to adequately control for the variability of non-individual HRTFs, for example using objective metrics for selecting an "equally-distant" non-individual HRTF for each participant (see also Daugintis *et al.*, 2023).

ACKNOWLEDGMENTS

We want to thank Dr. Virgina Ann Best for providing a copy of the CRM corpus, which has been used in this experiment. This study has been supported by SONICOM (www.sonicom.eu), a project funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017743, and the Spanish National Project SAVLab, under Grant No. PID2019-107854GB-I00.

APPENDIX: DATA TABLES

This appendix contains two tables showing the experimental data for the sake of reproducibility. Table I presents the percentages of correct answers averaged across participants per target-masker spatial condition and HRTF condition. Numbers in parentheses are the standard errors of the means. Table II is similar to Table I but for SRM, computed by subtracting the co-located score from the separated scores for a given target position.

TABLE I. Mean percentages of correct answers for all target and masker positions for individual and KEMAR HRTF. Values in parentheses are standard errors of the means.

Individual HRTF (% correct)			
	Masker elevation		
Target elevation	−50°	0°	50°
−50°	51.2 (3.3)	73.7 (3.5)	73.7 (3.3)
0°	53.9 (4.7)	53.1 (3.9)	56.8 (3.7)
50°	58.2 (4.4)	64.7 (3.5)	47.3 (3.9)
KEMAR HRTF (% Correct)			
	Masker elevation		
Target elevation	−50°	0°	50°
−50°	50.5 (3.7)	66.2 (3.7)	56 (4.1)
0°	48.6 (3.3)	49.8 (4.0)	48.1 (3.8)
50°	53.6 (3.5)	65 (3.8)	50.5 (3.3)

TABLE II. Same as Table I but for SRM, computed as the difference in mean percentages of correct answers between separated conditions and co-located conditions. Values in parentheses are standard errors of the means.

Individual HRTF (SRM % difference)			
	Masker elevation		
Target elevation	−50°	0°	50°
−50°	N/A	22.5 (3.2)	22.5 (2.6)
0°	0.7 (4.0)	N/A	3.6 (2.9)
50°	10.9 (3.4)	17.4 (3.8)	N/A
KEMAR HRTF (SRM % Difference)			
	Masker elevation		
Target elevation	−50°	0°	50°
−50°	N/A	15.7 (3.1)	5.6 (4.1)
0°	−1.2 (3.2)	N/A	−1.7 (3.7)
50°	3.1 (2.2)	14.5 (2.4)	N/A

Ahrens, A., Cuevas-Rodriguez, M., and Brimijoin, W. O. (**2021**). "Speech intelligibility with various head-related transfer functions: A computational modelling approach," JASA Express Lett. **1**(3), 034401.
Andreopoulou, A., and Katz, B. F. (**2022**). "Perceptual impact on localization quality evaluations of common pre-processing for non-individual head-related transfer functions," J. Audio Eng. Soc. **70**(5), 340–354.
ANSI (**1997**). ANSI/ASA S3.5-1997 (R2017)—*Methods for Calculation of the Speech Intelligibility Index* (ANSI, New York).
Berwick, N., and Lee, H. (**2020**). "Spatial unmasking effect on speech reception threshold in the median plane," Appl. Sci. (Switzerland) **10**(15), 5257.
Best, V. (**2004**). "Spatial hearing with simultaneous sound sources: A psychophysical investigation," Ph.D. thesis, The University of Sydney, Sydney, Australia.
Bolia, R. S., Ericson, M. A., Nelson, W. T., McKinley, R. L., and Simpson, B. D. (**1999**). "A cocktail party effect in the median plane?," J. Acoust. Soc. Am. **105**(2), 1390–1391.
Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**(2), 1065–1066.
Bregman, A. S. (**1990**). *Auditory Scene Analysis* (The MIT Press, Cambridge, MA).
Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust. united Ac. **86**(October), 117–128.
Bronkhorst, A. W. (**2015**). "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," Atten. Percept. Psychophys. **77**, 1465–1487.
Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.
Brungart, D. S., and Simpson, B. D. (**2002**). "The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal," J. Acoust. Soc. Am. **112**(2), 664–676.
Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**(5), 975–979.
Cooke, M., Barker, J., Cunningham, S., and Shao, X. (**2006**). "An audio-visual corpus for speech perception and automatic speech recognition," J. Acoust. Soc. Am. **120**(5), 2421–2424.
Cooke, M., Lecumberri, M. L. G., and Barker, J. (**2008**). "The foreign language cocktail party problem: Energetic and informational masking

effects in non-native speech perception," J. Acoust. Soc. Am. **123**, 414–427.
Cuevas-Rodriguez, M., Gonzalez-Toledo, D., Reyes-Lecuona, A., and Picinali, L. (**2021**). "Impact of non-individualised head-related transfer functions on speech-in-noise performances within a synthesised virtual environment," J. Acoust. Soc. Am. **149**, 2573–2586.
Cuevas-Rodríguez, M., Picinali, L., González-Toledo, D., Garre, C., de la Rubia-Cuestas, E., Molina-Tanco, L., and Reyes-Lecuona, A. (**2019**). "3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation," PLoS One **14**(3), e0211899.
Culling, J., Hawley, M., and Litovsky, R. (**2004**). "The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources," J. Acoust. Soc. Am. **116**(2), 1057–1065.
Culling, J., Hawley, M., and Litovsky, R. (**2005**). "Erratum: The role head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. **116**, 1057 (2004)]," J. Acoust. Soc. Am. **118**(1), 552.
Culling, J. F., and Lavandier, M. (**2021**). "Binaural unmasking and spatial release from masking," in *Binaural Hearing*, edited by R. Y. Litovsky, M. J. Goupell, R. R. Fay, and A. N. Popper (Springer International Publishing, Cham), pp. 209–241.
Culling, J. F., and Stone, M. A. (**2017**). "Energetic masking and masking release," in *The Auditory System at the Cocktail Party*, edited by J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay (Springer International Publishing, New York), pp. 41–73.
Daugintis, R., Barumerli, R., Picinali, L., and Geronazzo, M. (**2023**). "Classifying non-individual head-related transfer functions with a computational auditory model: Calibration and metrics," in *Proceedings of the ICASSP 2023*, June 4–10, Rhodes Island, Greece.
Drullman, R., and Bronkhorst, A. W. (**2000**). "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," J. Acoust. Soc. Am. **107**(4), 2224–2235.
Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking—Level differences," J. Acoust. Soc. Am. **35**(8), 1206–1218.
Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd, G. (**2003**). "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**(1), 368–379.
Engel, I., Alon, D. L., Scheumann, K., Crukley, J., and Mehra, R. (**2022**). "On the differences in preferred headphone response for spatial and stereo content," J. Audio Eng. Soc. **70**(4), 271–283.
Ezzatian, P., Avivi, M., and Schneider, B. A. (**2010**). "Do nonnative listeners benefit as much as native listeners from spatial cues that release speech from masking?," Speech Commun. **52**(11), 919–929.
Farina, A. (**2007**). "Advancements in impulse response measurements by sine sweeps," in *Proceedings of the 122nd AES Convention*, May 5–8, Vienna, Austria, pp. 1–21.
Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.
Gutierrez-Parera, P. (**2020**). "Optimization and improvements in spatial sound reproduction systems through perceptual considerations," Ph.D. thesis, Universidad Politécnica de Valencia, Valencia, Spain.
Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**(2), 833–843.
Jelfs, S., Culling, J. F., and Lavandier, M. (**2011**). "Revision and validation of a binaural model for speech intelligibility in noise," Hear. Res. **275**(1–2), 96–104.
Jo, H., Park, Y., and Park, Y.-S. (**2009**). "Analysis of individual differences in head-related transfer functions by spectral distortion," in *Proceedings of the 2009 ICCAS–SICE*, August 18–21, Fukuoka, Japan, pp. 1769–1772.
Jones, G. L., and Litovsky, R. Y. (**2011**). "A cocktail party model of spatial release from masking by both noise and speech interferers," J. Acoust. Soc. Am. **130**(3), 1463–1474.
Karjalainen, M., and Paatero, T. (**2001**). "Frequency-dependent signal windowing," in *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, October 21–24, New Paltz, New York, pp. 35–38.
Kidd, G., and Colburn, H. S. (**2017**). "Informational masking in speech recognition," in *The Auditory System at the Cocktail Party*, edited by J. C. Middelbrooks, J. Z. Simon, A. N. Popper, and R. R. Fay (Springer International Publishing, Cham), pp. 75–109.
Lavandier, M., and Best, V. (**2020**). "Modeling binaural speech understanding in complex situations," in *The Technology of Binaural Understanding*, edited by J. Blauert and J. Braasch (Springer International Publishing, Cham), pp. 547–578.
Lecumberri, M. L. G., Cooke, M., and Cutler, A. (**2010**). "Non-native speech perception in adverse conditions: A review," Speech Commun. **52**(11), 864–886.
Lee, C.-Y., Tao, L., and Bond, Z. S. (**2010**). "Identification of multi-speaker Mandarin tones in noise by native and non-native listeners," Speech Commun. **52**(11), 900–910.
Martin, R. L., McAnally, K. I., Bolia, R. S., Eberle, G., and Brungart, D. S. (**2012**). "Spatial release from speech-on-speech masking in the median sagittal plane," J. Acoust. Soc. Am. **131**(1), 378–385.
McAnally, K. I., Bolia, R. S., Martin, R. L., Eberle, G., and Brungart, D. S. (**2002**). "Segregation of multiple talkers in the vertical plane: Implications for the design of a multiple talker display," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, September 1, Los Angeles, CA, pp. 588–591.
Oh, Y., Hartling, C. L., Srinivasan, N. K., Diedesch, A. C., Gallun, F. J., and Reiss, L. A. J. (**2022**). "Factors underlying masking release by voice-gender differences and spatial separation cues in multi-talker listening environments in listeners with and without hearing loss," Front. Neurosci. **16**, 1059639.
Picinali, L., and Katz, B. F. G. (**2023**). "System-to-user and user-to-system adaptations in binaural audio," in *Sonic Interactions in Virtual Environments*, edited by M. Geronazzo and S. Serafin (Springer International Publishing, Cham), pp. 115–143.
Plomb, R., and Mimpen, A. M. (**1981**). "Effect of the orientation of the speaker's head and azimuth of a noise source on the speech reception threshold for sentences," Acta Acust. united Ac. **48**(5), 325–328.
Shinn-Cunningham, B. G., Schickler, J., Kopčo, N., and Litovsky, R. (**2001**). "Spatial unmasking of nearby speech sources in a simulated anechoic environment," J. Acoust. Soc. Am. **110**(2), 1118–1129.
Srinivasan, N. K., Staudenmeier, A., and Clark, K. (**2022**). "Effect of gap detection threshold and localisation acuity on spatial release from masking in older adults," Int. J. Audiol. **61**(11), 932–939.
Steadman, M. A., Kim, C., Lestang, J. H., Goodman, D. F., and Picinali, L. (**2019**). "Short-term effects of sound localization training in virtual reality," Sci. Rep. **9**(1), 18284.
Takane, S. (**2016**). "Effect of domain selection for compact representation of spatial variation of head-related transfer function in all directions based on spatial principal components analysis," Appl. Acoust. **101**, 64–77.
Usher, J., and Martens, W. L. (**2007**). "Perceived naturalness of speech sounds presented using personalized versus non-personalized HRTFs," in *Proceedings of the 13th International Conference on Auditory Display*, June 26–29, Montreal, Canada, pp. 10–16.
Westermann, A., Buchholz, J. M., Org, J., and Buchholz, M. (**2015**). "The effect of spatial separation in distance on the intelligibility of speech in rooms," J. Acoust. Soc. Am. **137**(2), 757–767.
Worley, J. W., and Darwin, C. J. (**2002**). "Auditory attention based on differences in median vertical plane position," in *Proceedings of the 2002 International Conference on Auditory Display*, July 2–5, Kyoto, Japan, pp. 1–5.
Xie, B. (**2013**). *Head-Related Transfer Function and Virtual Auditory Display*, 2nd ed. (InTech, London).
Zenke, K., and Rosen, S. (**2022**). "Spatial release of masking in children and adults in non-individualized virtual environments," J. Acoust. Soc. Am. **152**, 3384–3395.
Zurek, P. M. (**1987**). "The precedence effect," in *Directional Hearing*, edited by W. A. Yost and G. Gourevitch (Springer, New York), pp. 85–105.