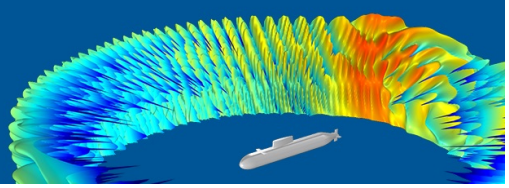
Impact of non-individualised head related transfer functions on speech-in-noise performances within a synthesised virtual environment

Maria Cuevas-Rodriguez, Daniel Gonzalez-Toledo, Arcadio Reyes-Lecuona, and Lorenzo Picinali

Citation: The Journal of the Acoustical Society of America **149**, 2573 (2021); doi: 10.1121/10.0004220
View online: https://doi.org/10.1121/10.0004220
View Table of Contents: https://asa.scitation.org/toc/jas/149/4
Published by the Acoustical Society of America



JASA ARTICLE

Impact of non-individualised head related transfer functions on speech-in-noise performances within a synthesised virtual environment

Maria Cuevas-Rodriguez,[1,a)] Daniel Gonzalez-Toledo,[1,b)] Arcadio Reyes-Lecuona,[1,c)] and Lorenzo Picinali[2,d)]
[1]*Departamento de Tecnología Electrónica, Universidad de Málaga, ETSI Telecomunicación, 29010 Málaga, Spain*
[2]*Dyson School of Design Engineering, Imperial College London, London SW7 2DB, United Kingdom*

ABSTRACT:
When performing binaural spatialisation, it is widely accepted that the choice of the head related transfer functions (HRTFs), and in particular the use of individually measured ones, can have an impact on localisation accuracy, externalization, and overall realism. Yet the impact of HRTF choices on speech-in-noise performances in cocktail party-like scenarios has not been investigated in depth. This paper introduces a study where 22 participants were presented with a frontal speech target and two lateral maskers, spatialised using a set of non-individual HRTFs. Speech reception threshold (SRT) was measured for each HRTF. Furthermore, using the SRT predicted by an existing speech perception model, the measured values were compensated in the attempt to remove overall HRTF-specific benefits. Results show significant overall differences among the SRTs measured using different HRTFs, consistently with the results predicted by the model. Individual differences between participants related to their SRT performances using different HRTFs could also be found, but their significance was reduced after the compensation. The implications of these findings are relevant to several research areas related to spatial hearing and speech perception, suggesting that when testing speech-in-noise performances within binaurally rendered virtual environments, the choice of the HRTF for each individual should be carefully considered. © *2021 Acoustical Society of America.*
https://doi.org/10.1121/10.0004220
(Received 28 October 2020; revised 18 March 2021; accepted 19 March 2021; published online 13 April 2021)
[Editor: James F. Lynch] Pages: 2573–2586

I. INTRODUCTION

An essential function of the human auditory system is the localisation of sound sources in different angular positions and at different distances around the listener. To do this, humans make use of direction-dependent interactions between the incident sound waves and the head and torso, which are at the basis of the so-called localisation cues. These have been widely studied in the past, from works carried out at the beginning of the past century (Rayleigh, 1907) to more recent research (Blauert, 1997). A sound source located on the right side of someone's head generates a sound wave that will reach the right ear (ipsilateral ear) before the left ear (contralateral ear), therefore causing a delay between the two [interaural time difference (ITD)]. Furthermore, the signal reaching the contralateral ear will be partly attenuated (typically in its high-frequency spectral components) by the head, generating a level difference between the two ears [interaural level difference (ILD)]. Other spectral cues are then used to determine whether the sound source is located above or below and in the front or at the back. Finally, distance perception relies on a mixture of these and other cues, including some generated by interactions with the surrounding environments, as the direct-to-reverberant energy ratio (Bronkhorst and Houtgast, 1999). Both interaural and spectral monoaural cues are described by the so-called head related transfer function (HRTF) (Møller *et al.*, 1995). Once a HRTF has been measured (or precisely estimated) for a given listener, immersive virtual reality (VR) audio systems can make use of it and process sounds so that, when presented over headphones, they are perceived as emanating from any position in the surrounding 3D space. This technique is referred to as binaural spatialization (Hammershøi and Møller, 2005). Because of individual characteristics such as diameter and shape of the head, size and shape of the pinna, etc., HRTFs vary from one individual to another.

When performing binaural spatialization, the HRTF should ideally be measured for each specific listener. However, the accurate measurement of a HRTF is complex and requires expensive equipment (Gardner and Martin, 1995). A common alternative is to use a HRTF measured from a dummy head mannequin or from another individual, since several HRTF datasets are currently freely available. Some examples are LISTEN (Warusfel, 2003), CIPIC (Algazi *et al.*, 2001), and ARI (Institute for Sound Research, 2013). The problem then becomes how to choose which, among the non-individual HRTFs, is the most appropriate

[a)]Electronic mail: mariacuevas@uma.es, ORCID: 0000-0002-4698-5170.
[b)]ORCID: 0000-0002-2698-2896.
[c)]ORCID: 0000-0002-3699-4065.
[d)]ORCID: 0000-0001-9297-2613.

for a specific listener. Several studies have been carried out in past years to investigate this issue, looking at a solution based either on physical measurement matching or on perceptual selection. The research belonging to the first category relies on comparing the anthropometric measurements of the pinnae corresponding to the individuals from which the HRTFs were measured with the same measurements taken from the target listener, aiming at finding a HRTF that minimises the differences (Geronazzo *et al.*, 2019, 2014; Iida *et al.*, 2014). The second category of work on non-individual HRTF selection is based instead on perceptual tests, relying on either objective (e.g., measuring localisation accuracy) (Härmä *et al.*, 2012) or subjective (e.g., using attributes such as preference, realism, and externalization) (Katz and Parseihian, 2012; Simon *et al.*, 2016) evaluations. These solutions are indeed rather promising but are still in need of further refinements and evaluations to improve effectiveness and repeatability (Andreopoulou and Katz, 2016). To summarise, while it has been shown that the use of non-individually measured HRTFs can have an impact on sound localisation and on the perceived quality of the virtual simulation, it is still unclear which is the best way to select the best matching HRTF for a specific listener.

HRTFs are also important for other functions besides sound localisation. Previous work demonstrates that some attentional processes use HRTF cues to support focusing auditory attention on a specific direction. Related to this is the cocktail party effect (Cherry, 1953). Although it was originally described by Cherry as the ability to "recognize what one person is saying when others are speaking at the same time," it has been extensively studied with multiple types of masking sounds (Bronkhorst, 2000; Culling *et al.*, 2004; Hawley *et al.*, 2004; Jones and Litovsky, 2011). Within this paper, we are referring to it as a phenomenon of selective attention that allows humans to focus on a single sound source when this is competing with other masking sources. This phenomenon is considerably enhanced when the target speech source is spatially separated from the masker noise sources. The advantage in terms of speech intelligibility gained from the spatial separation of masker and target is known as spatial release from masking (SRM) (Ching *et al.*, 2011). In that first work published mentioning the cocktail party problem (Cherry, 1953), it was reported that the effect is much more evident in binaural (i.e., involving both ears) than monoaural (i.e., involving only one ear) conditions. Bronkhorst and Plomp (1988) presented a seminal work about the contribution of interaural cues to SRM, defining and quantifying the benefit of spatial separation of maskers from the target to enhance speech intelligibility (Bronkhorst and Plomp, 1992; Hawley *et al.*, 1999; Koehnke and Besing, 1996).

At the beginning of the current century, Bronkhorst presented a review on the cocktail party problem (Bronkhorst, 2000), later revisited (Bronkhorst, 2015), and introduced models of binaural speech perception that allow estimation of the SRM for frontal targets and any configuration of noise maskers in the horizontal plane. This formula was further validated by (Culling *et al.*, 2004). Jones and Litovsky (2011) presented a new model based on Bronkhorst's work, which allows estimation of SRM with multiple speech and noise maskers. Hawley *et al.* (2004) showed how the contributions to SRM come from two independent components: (1) monaural advantages based on best-ear listening (Edmonds and Culling, 2006), which are related to the ILDs as the target-to-interferer ratio is better in one of the ears, and (2) binaural advantages, benefiting from ITDs. An extension of this work was carried out by Culling *et al.* (2004), who studied the individual role of ILDs and ITDs when measuring the speech reception threshold (SRT) for multiple interferers. A series of experiments were conducted to clarify the contribution of each individual interaural cue using manipulated HRTFs. Results revealed that, in the case of spatial separation between target speech and interferers, speech intelligibility improves when both ITD and ILD cues are present. Moreover, Culling *et al.* (2004) proposed a formula to estimate the binaural masking level difference (BMLD) based on the cross correlation between left and right head related impulse responses (HRIRs). This allowed the inclusion of the effects of room acoustics in the estimation of SRM. Based on that formula and other previous works, Lavandier and Culling (2010) developed a model to predict SRT with spatially separated interferers, also taking into account the room effects. This model processed the signals using two paths: the first one calculates the advantage caused by the binaural unmasking, predicting the BMLD using the equalization-cancellation theory and the formula from Culling *et al.* (2004). The second path predicts the effects of the better-ear listening, calculating the SNR as the target-to-interferer ratio at each frequency. Both paths integrate the signals across frequency using the speech intelligibility index (SII) weighting method (ANSI, 1997). This model was later revised by (Jelfs *et al.*, 2011) and tested with multiple, spatially distributed interferers, including anechoic and reverberant conditions. The revised model handles each interferer signal separately and operates directly on binaural impulse responses [HRIRs or binaural room impulse responses (BRIRs)]. In addition, it also introduces an improvement using gammatone filters, which are used in the second path of the model, where the effect of better-ear is estimated.

Other approaches to predict speech intelligibility use room statistics instead of impulse responses, like the one proposed by Freyman and Zurek (2008), which considers the effect of room size, average absorption coefficient, and other parameters. Looking more at speech intelligibility modeling, van Wijngaarden and Drullman (2008) proposed a binaural version of the speech transmission index (STI), which is a well known and standardized method to objectively estimate speech intelligibility (IEC, 2003). A simplified binaural model based on interaural cross-correlograms was integrated to the standard monoaural STI method. A validation of the model further confirmed how interaural differences result in an overall improvement of speech-in-noise intelligibility when source and maskers are not co-located

https://doi.org/10.1121/10.0004220

and how this can be predicted with a certain accuracy using rather simple binaural models.

These previous works demonstrate that the cocktail party effect is enhanced with binaural listening and is related to both the spatial configuration of target and interferers as well as the acoustics of the room. Models have been proposed and validated to predict SRT in such situations, but these do not take into account how listeners can leverage the individual characteristics of their HRTF when trying to understand speech in cocktail party conditions. Considering the individual nature of HRTFs, we believe it is relevant to assess the fit of a non-individual HRTF to a specific subject by observing the performances in a VR-based cocktail party task.

A. Summary and aims of the present study

The main goal of this work is to study the impact on individual listeners of different non-individual HRTFs on speech intelligibility within a VR cocktail party context. If the human attentional system uses the listener's own experience with their individual HRTF to improve speech recognition, in an environment where target and maskers are located in different positions, we should be able to experimentally find an effect of the HRTF choice on speech intelligibility. Furthermore, if the HRTF is an idiosyncratic characteristic of each listener, we should also find that this effect is different for different subjects. Our hypotheses are as follows:

- H1: There is a significant effect of the HRTF choice on speech recognition within a virtual cocktail party context. That is to say that, for a given subject, different HRTFs provide different performances in terms of speech recognition of target words in diffuse masking conditions.
- H2: The effect of a given HRTF on speech recognition is different for different subjects; therefore, there are no individually measured HRTFs that are universally better or worse than others when evaluated on this specific task.

II. MATERIALS AND METHODS

A. Participants

Twenty-three participants were recruited among students and researchers of the School of Telecommunication Engineering in the University of Malaga. They received a USB stick as compensation for their participation. One of them withdrew from the experiment without completing all the sessions and was therefore discarded. Twenty-two participants (16 male and 6 female) were included, 17 of them with ages between 18 and 29 yrs and 5 of them with ages between 30 and 50 yrs. All of them self-reported normal hearing, and they were all native Spanish speakers. This number of participants was chosen based on a previous study where a similar HRTF set was analyzed (Katz and Parseihian, 2012). All procedures were reviewed and approved by the Ethical Committee for Research in Malaga (*Comité de Ética de la Investigación Provincial de Málaga*).

B. Stimuli

The target stimuli were a set of 221 two-syllable Spanish words spoken by a female voice. The words were extracted from a list used for logo-audiometry studies (de Cárdenas and Marrero Aguiar, 1994) and present small redundancy, phonetic and syllabic structure balance with Spanish language, similar difficulty, and similar familiarity. As maskers, filtered continuous noise with the same spectral density as the target word corpus was used. These maskers were included in the same database as the target words. Figure 1(a) shows the long-term average spectra of the target and maskers' signal, computed using the IoSR MATLAB toolbox (IoSR, 2017). The spectra were calculated using the average power spectral density (PSD) obtained from a series of overlapping fast Fourier transforms (FFTs) (Hann-windowed) of 4096 samples. The average PSD was then Gaussian-smoothed to 1/3-octave resolution.

The target sound consisted of one word virtually located in front of the listener (0° azimuth, 0° elevation), as it is shown in Fig. 1(b). Before each target word, a sentence, *"Por favor, escriba la palabra"* ("Please, type the word"), was always played back in the same virtual position as the target to help focus the attention on the target source direction. Two uncorrelated masker sources were used and virtually located at the right and left sides of the listener (±90° azimuth, 0° elevation); see Fig. 1. Due to its symmetrical nature, this is not the configuration that would have resulted in the highest SRM. The rationale behind this choice is that our requirements were to attempt to measure the effect of HRTF differences and minimise the better-ear effect [i.e., taking advantage of the ear with the better signal-to-noise ratio (SNR)]. Therefore, asymmetric configurations were discarded to minimise the effect of interaural cues and focus the study on the monaural spectral cues of the HRTF. A similar symmetric configuration was used by Culling and Mansell (2013). Due to the complexity of adding multiple spatial configurations to an already extensive test, after a series of initial pilot studies, we ultimately decided to use only the ±90° configuration. The power of maskers was fixed at 58 dB [sound pressure level (SPL)] in each ear before being filtered by the HRTF, which ensured a comfortable level. The power of targets was varied during the experiment.

C. Spatialized virtual sound and HRTF dataset

Sound source spatialization was purely anechoic. The binaural spatialization was performed using the 3DTI-Toolkit (Cuevas-Rodríguez *et al.*, 2019), a C++ open-source library for real-time binaural spatialization. Eight different HRTFs were used in the study. The first seven (named here as $HRTF_1$–$HRTF_7$) were taken from the LISTEN database (Warusfel, 2003). They were selected in a previous study by Katz and Parseihian (2012) to produce the best subjective spatialization for the most listeners, and they are identified in the database as IRC_1008, IRC_1013, IRC_1022, IRC_1031, IRC_1002, IRC_1048, and IRC_1053. Like Katz and

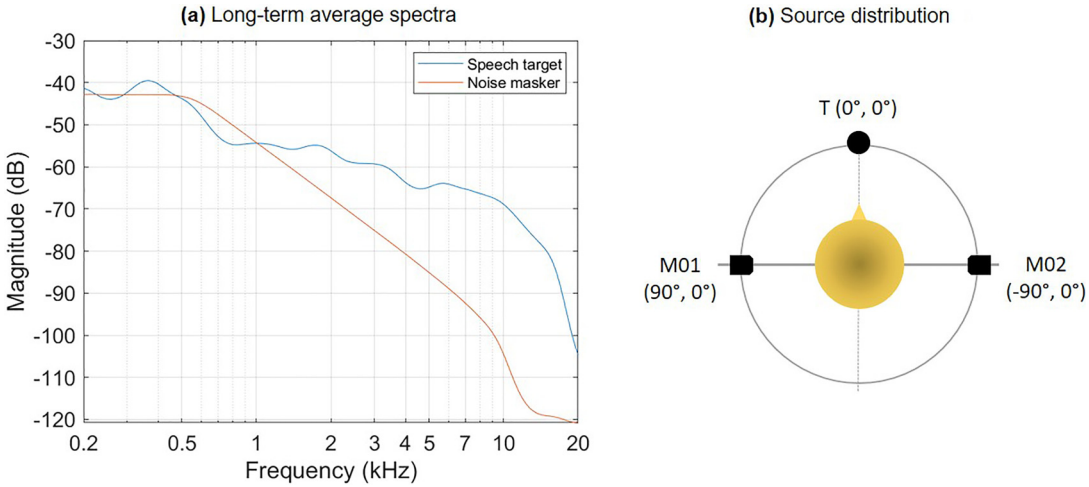


FIG. 1. (Color online) (a) Long-term average spectra of the target and maskers' signals, normalized at 1 kHz. (b) Target and masker configuration. The listener is in the middle. T is the target position, M01 is the masker on the left, and M02 is the masker on the right.

Parseihian (2012), we used the "raw" HRTF measurement data, not the diffuse field compensated. Due to an error during the preparation of the experiment, IRC_1002 was taken instead of IRC_1032. This corresponded to $HRTF_5$, and it was kept in the fifth position in all the reported results. Finally, the eighth HRTF was a synthetic spherical-head model used as an anchor condition (denoted as $HRTF_A$). The synthetic HRTF contained the two binaural cues, ITD and ILD. ILD was built as a simple one-pole one-zero model, based on the analytical model obtained by Lord Rayleigh (Rayleigh, 1907). ITD was modelled as a time delay function using Woodworths's formula (Woodworth *et al.*, 1954), with a head radius of 8.75 cm. The synthetic HRTF was normalized to have the same power as the power average of the LISTEN HRTF set in the front position (0° azimuth, 0° elevation).

A numerical analysis of the HRTFs used in the experiment has also been carried out. Figure 2 shows the magnitude of the HRTFs used in the experiment, for the target and masker positions. Looking at these positions, it is possible to see how each HRTF presents a noticeably different spectrum below 10 kHz, where both target and maskers have most of their signal energy [Fig. 1(a)]. This suggests that they could impact differently in terms of measured speech intelligibility using the chosen experimental configuration. This has been accounted for in the data analysis, as can be found in Sec. II F.

Figure 3 shows the ITD and ILD values for the HRTFs used in the experiment, for both the target and masker position. ITDs were calculated using a modified threshold method similar to the one presented in Katz and Noisternig (2014), where a comparison of the left and right signals was carried out using a threshold detector to identify the first arrival time of the incident sound. A threshold of 5% of the maximum amplitude in each HRIR was chosen to detect the onset, visually checking that all HRIRs were aligned when the initial silence up to the threshold was removed. ILDs were calculated using the magnitude difference between left and right signals and then averaged by 30 uniformly spaced frequency bands between 1.5 and 20 kHz on an equivalent rectangular bandwidth (ERB) scale (Moore and Glasberg, 1983).

D. Apparatus

A software platform was developed specifically for these experiments, which included the 3DTI-Toolkit library, and it was used to automatically sequence the whole procedure of each session, without any intervention of the operator. The system allowed the participant to type each word using the keyboard, and it automatically recorded all the activities performed by the participants. A MOTU (Cambridge, MA) 896 mk3 audio interface was used to reproduce the sound, connected to the computer using an ASIO driver. Participants had to wear a pair of SONY (Tokyo, Japan) MDR-7506 headphones. Previous studies have shown that the transfer function between headphones and eardrums (HpTF) can play a role in terms of externalization and overall naturalness of the binaural rendering (Durlach *et al.*, 1992; Masiero and Fels, 2011). Nevertheless, strong evidence has not been found to support that HpTF can improve spatial hearing abilities, such as localisation accuracy (Engel *et al.*, 2019; Schonstein *et al.*, 2008). Furthermore, it has to be noted that HpTFs are not direction dependent and therefore do not vary depending on the position of the source and should not have an influence on HRTF-specific effects, which are the objects of this study. Finally, considering the fact that within this study we explicitly did not want to carry out any personalisation of the rendering and playback systems, and in line with other published research (e.g., Andreopoulou and Katz, 2016), no HpTF was measured and used in this study. To ensure consistency within each session and avoid potential spectral alterations due to repeated donning of the headphones, participants were instructed to wear the headphones at the beginning of each session and not to remove them until the

https://doi.org/10.1121/10.0004220

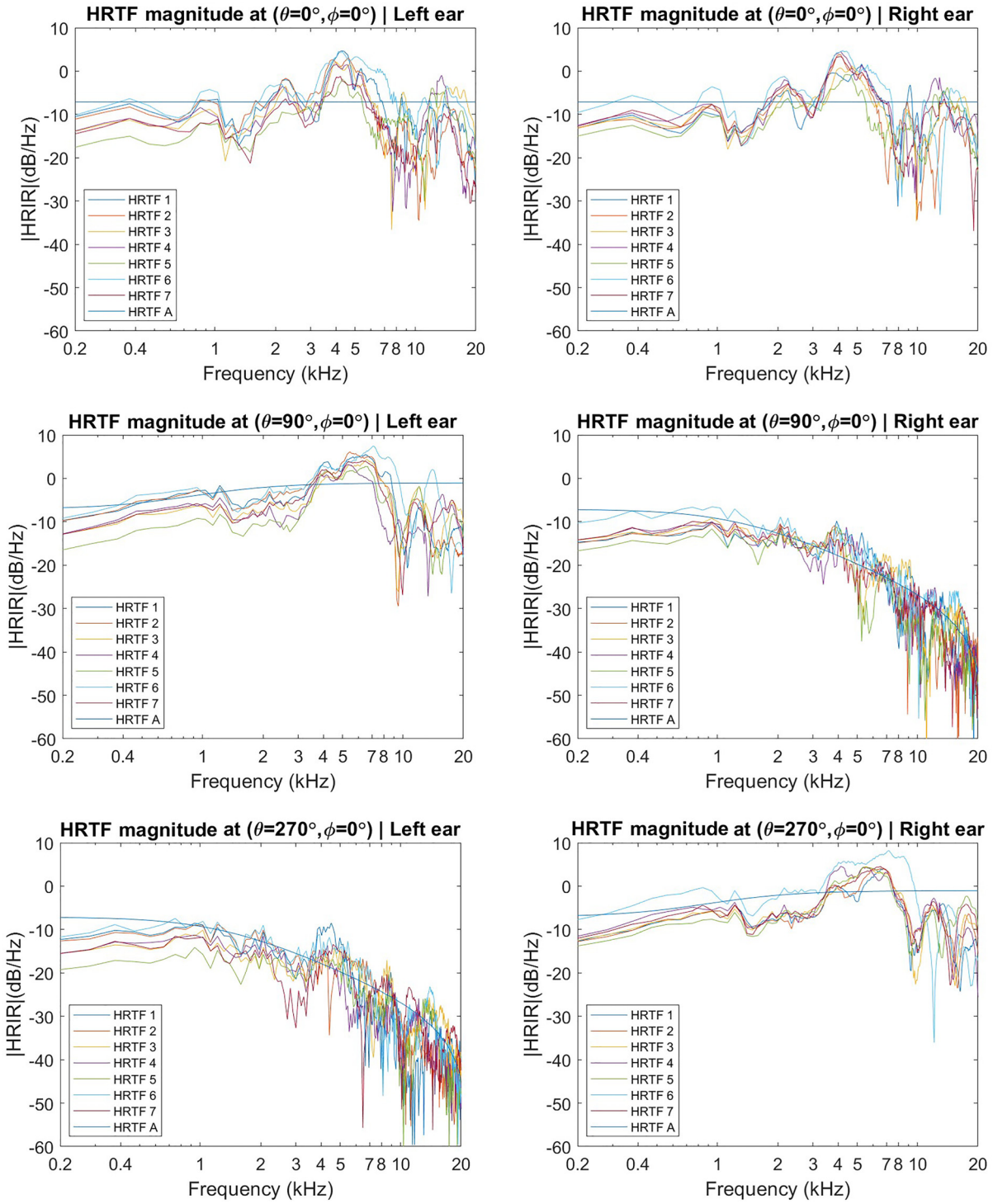


FIG. 2. (Color online) Power spectral density of the HRTFs used in the study, for the target position ($\theta = 0°, \phi = 0°$) and masker positions [($\theta = 90°, \phi = 0°$) and ($\theta = 270°, \phi = 0°$)] and left and right ear. θ is the azimuth and ϕ the elevation.

end. To our knowledge, they all complied with this requirement.

E. Procedure

During the experiment, participants were seated in a silent environment in front of a monitor, with a keyboard and a mouse. For each participant, and each session, a total of eight SRTs were measured, one for each HRTF. Each of these measurements was named a block, so that each session was composed of eight blocks, presented in a random order. Fifty percent SRT in noise, i.e., the SNR at which 50% of the speech material is repeated correctly, was measured using an adaptive up-down procedure, which required repeated presentations (trials) of different stimuli for each block (Levitt, 1971). During each trial, the participant

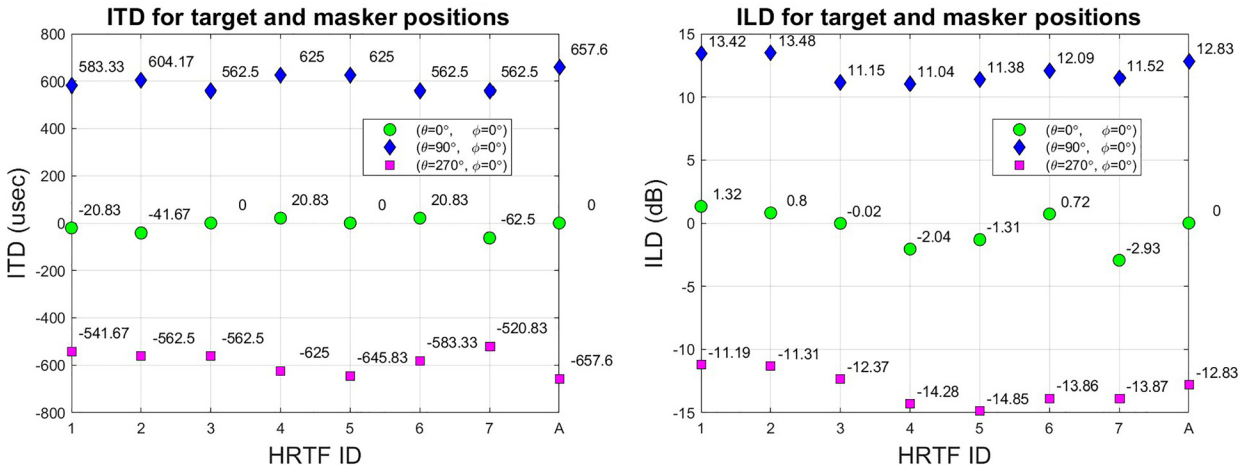


FIG. 3. (Color online) ITD (left) and ILD (right) of the HRTFs used in the study, for the target position ($\theta = 0°, \phi = 0°$) and masker positions [($\theta = 90°, \phi = 0°$) and ($\theta = 270°, \phi = 0°$)], where θ is the azimuth and ϕ the elevation.

listened to the target and typed the word using the computer keyboard. Participants were instructed to guess if they could not identify the word or leave the word-input space empty if they had no clue. No feedback about whether the typed word was correct or not was given.

Participants were told that the target was played in front of them and that the maskers were located on the left and on the right of their heads. No information about distance was given. They were instructed to face straight ahead and focus their attention on the target, trying to ignore the surrounding noise. The structure of each trial is presented in Fig. 4. Participants listened to the prompt, virtually located in the same position as the target, and, after a short silence (randomly selected with a uniform distribution between 500 and 700 ms), the maskers started. A few hundred milliseconds later (also randomly selected with a uniform distribution between 200 and 800 ms), the target word started. The maskers stopped 600 ms after the target finished.

For the first trial, the target was played at an initial SNR of −3 dB (SNR was defined as the ratio between target and one masker). If the word was correctly identified, the level of the next target decreased 2 dB; if not, it increased 2 dB. After a pilot study focused on finalising the testing procedure, to avoid interpreting common spelling mistakes as errors and robustly discriminate between intelligible and non-intelligible words, it was decided to consider a word as correct when it matched the target or when there was a spelling error of one single letter. The level of the masker remained unvaried. Target and maskers were randomly selected in each trial, ensuring that both maskers were uncorrelated. This procedure was repeated until four up-down reversals (changes between increase and decrease in SNR) occurred. The four SNRs obtained in this way were averaged to yield the SRT value for the block. The number of trials within a block depended on the participant's performance.

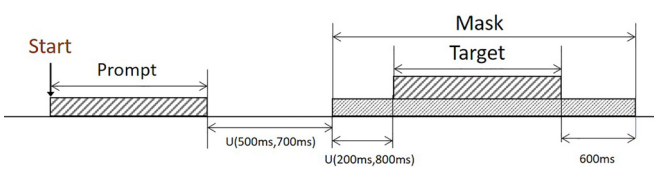


FIG. 4. (Color online) Structure for one trial. Each trial started with a prompt, followed by the two maskers and, after a short delay, the actual target word.

This procedure was carried out for each participant in each session. Taking into account that each participant can be considered as an independent experiment, a certain number of sessions must be performed by each participant. As mentioned above, a pilot experiment with the same experimental design was carried out. From that pilot experiment, we obtained the variances to be used by the GPower statistical analysis tool (Mayr *et al.*, 2007) to estimate the sample size for a test power of 95% with $p < 0.01$. This yielded to 20 repetitions (sessions) per participant.

Participants were received the first day, and they were informed about the purpose of the experiment, all of them gave their written consent. As the experiment consisted of 20 sessions, participants could choose the days they came to the laboratory. They were allowed to carry out a maximum of three sessions per day, keeping a break between sessions of at least 10 min. Among the 440 conducted sessions (20 sessions for each of the 22 participants), the average duration was 10 min, 58 s, with a standard deviation of 2 min, 38 s. The shortest session lasted 7 min, 23 s and the longest 28 min, 58 s. Participants were informed that they could stop and rest between blocks.

F. Data collection and analysis

A total of 3520 SRTs were measured at the end of the experiment (one SRT per HRTF for each session and for each participant, 8 × 20 × 22). The unprocessed data are referred to here as raw SRT.

It is important to take into account that there might be some characteristics of the HRTFs, such as differences in the power ratio between sides and front, that could make

https://doi.org/10.1121/10.0004220

TABLE I. Factors used to compensate the raw SRT data, calculated using LCJ model (Jelfs *et al.*, 2011).

HRTF ID	$HRTF_1$	$HRTF_2$	$HRTF_3$	$HRTF_4$	$HRTF_5$	$HRTF_6$	$HRTF_7$	$HRTF_A$
SRM_factor (dB)	−0.44	−0.58	−0.98	−0.78	−1.73	−0.89	−0.79	−1.71

some HRTFs worse or better for the overall sample of participants, regardless of individual differences. For instance, in the case of a HRTF with an increased attenuation within the speech spectral bands for sources at azimuth = 0°, elevation = 0°, the target would be attenuated more than when using other HRTFs, yielding a higher SRT. Considering the second hypothesis of the current study (H2) and the aim to identify subject-specific differences between performances using different HRTFs, it is important to quantify the potential HRTF-specific advantages affecting all participants in the same way. To do this, we used the model developed by Lavandier and Culling (2010) and later revised by Jelfs *et al.* (2011), included in the MATLAB Auditory Modeling Toolbox (Auditory Modeling Toolbox, 2011; Sondergaard and Majdak, 2013), from here on referred to as the "LCJ model." This model predicts the total benefit of the SRM for a given HRTF (or BRIR). As described in Sec. I, this SRM is calculated as the sum of the SNR, which is the component of SRM due to the better-ear listening, and the BMLD, which is the component of SRM due to binaural unmasking. Using the HRIRs correspondent to the target and masker positions as input for the model, we obtained the total benefit of the SRM in dB for each HRTF. We then used these to compensate the raw SRT values, to remove the overall HRTF-specific benefit, which is to be considered as common for every subject. We defined our measured SRT as the SNR between one target and one masker. Therefore, we reduced the SRM value estimated with the LCJ model by 3 dB, as our experimental condition consisted of one target and two uncorrelated maskers. Then the compensated SRT was calculated as $CompensatedSRT = RawSRT + SRM\_factor$. SRM_factors are shown in Table I. In Sec. III, the analysis is performed separately for the raw and the compensated SRTs.

III. RESULTS

A. Overall analysis

To initially explore the data regardless of the individual subject differences, an overall analysis was carried out pooling all participants together. First, the collected data were averaged for each participant (using the SRT from the 20 sessions), obtaining one mean SRT per HRTF per participant. The distribution of these data is shown in the box plot in the left part of Fig. 5 for both raw and compensated SRT. In addition, the mean SRT was calculated across all participants, and it is shown in the right graph of Fig. 5, together with the 95% confidence interval (CI) for each HRTF used in the experiment.

While looking at the raw data, relevant differences can be identified between the various HRTFs, in particular $HRTF_5$ and $HRTF_A$. It is evident that, once the compensation is applied, these differences become less marked. This is corroborated by a one-way analysis of variance (ANOVA). Raw SRTs showed a significant effect of HRTF on SRT when the $HRTF_A$ was included [$F(7, 168) = 16.7861, p < 0.001$] and also when it was removed from the dataset [$F(6, 147) = 6.3972, p < 0.001$]. Compensated SRTs showed a significant effect when the $HRTF_A$ was included [$F(7, 168) = 4.1892, p < 0.001$] but not when it was removed from the dataset [$F(6, 147) = 0.76083, p = 0.602$]. *Post hoc* pairwise comparison using Bonferroni correction indicates that, using the raw SRT, only $HRTF_5$ and $HRTF_A$ are significantly

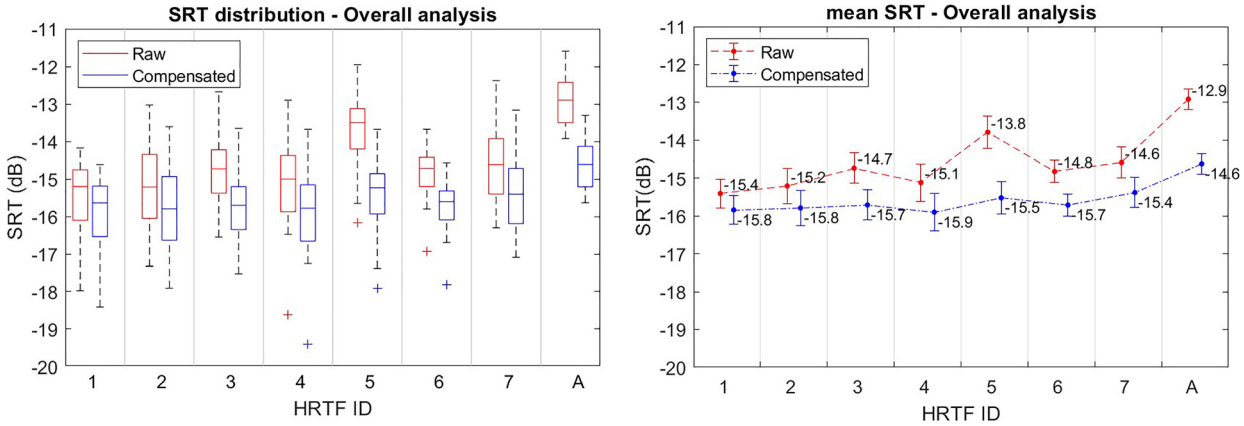


FIG. 5. (Color online) The left graph shows the distribution of the SRT means for each participant across sessions. On each box, the central horizontal mark indicates the median, and the bottom and top edges indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, which are plotted individually (+). The right graph shows the mean SRT across sessions and participants and the 95% CIs. The vertical axis indicates the SRT value in dB, and the horizontal axis indicates the HRTF condition ID. Both show the results of the overall analysis of the raw and compensated SRT data.

different from each of the rest of the other conditions, but not from each other ($p = 0.075$). Analysing the compensated SRT, only $HRTF_A$ results as significantly different from the rest of the HRTFs, except $HRTF_5$ ($p = 0.057$).

It is known that repeated exposure to a given task can result in a certain proportion of improvement due to procedural and perceptual training (Ortiz and Wright, 2009). This has been observed also in auditory tasks (Musiek *et al.*, 2014) and specifically in speech-related ones (Fu and Galvin, 2003). Considering the extensive duration of this study and the fact that each participant went through 20 separate test sessions lasting an average of 11 min per session, an analysis of the participant's overall SRT improvements across the sessions was carried out. A linear regression model was calculated to predict the SRT as a function of the session number (n), obtaining $SRT = -0.0844n - 13.686$. While a slight but significant SRT improvement can be noted across sessions $[F(1, 3518) = 103.813, p < 0.001]$, the effect of learning is very small and accounts for only 2.87% of the SRT variations.

B. Individual analysis

As this study is dealing with individual characteristics of the listeners, an individual analysis of both raw and compensated SRT data was carried out for each participant, considering each of them as an independent experiment.

To assess the effect on SRT of the different HRTF conditions, a one-way ANOVA was performed for each participant. Results are shown in Table II. For raw SRT, when all of the eight conditions ($HRTF_{1-7} + HRTF_A$) are included in the analysis, 18 of 22 participants show a significant effect of the HRTF. If the $HRTF_A$ is excluded, this number decreases to 9 of 22. For the compensated SRT, analysis including all HRTF conditions shows five participants with significant differences. Removing the $HRTF_A$, only one participant shows significant differences between HRTFs.

SRT values for the worst and best measured HRTF obtained for each participant are shown in Table III, together with the difference in dB between the SRT of the best HRTF and the SRT of the worst.

Post hoc pairwise comparisons for the different HRTF conditions were carried out using Fisher's least significant difference (LSD) test. Figure 6 shows the number of participants with significant differences in each pairwise comparison ($p < 0.05$) for both raw and compensated SRT data. Tables indicate the number of participants, with significant differences between the HRTF condition indicated in the header and one in the very left column. Graphs indicate the number of participants with significant differences between the HRTF condition indicated on the horizontal axis and the one corresponding with the color in the legend. When using the LSD test, no mathematical correction is made for multiple comparisons, as is recommended by some authors

TABLE II. One-way ANOVA outputs when looking at the differences between HRTFs for each participant. Results are presented including all the HRTF conditions ($HRTF_{1-7} + HRTF_A$) and only measured HRTFs (i.e., excluding $HRTF_A$) and separate for raw and compensated SRTs. The first column shows the ID of the participants. Asterisks indicate significant differences.

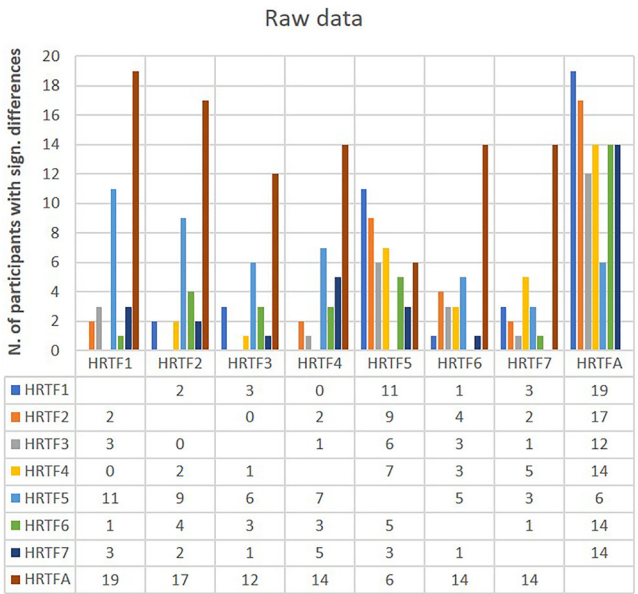
	Raw				Compensated			
	$HRTF_{1-7} + HRTF_A$		$HRTF_{1-7}$		$HRTF_{1-7} + HRTF_A$		$HRTF_{1-7}$	
ID	F(7,152)	p-value	F(6,133)	p-value	F(7,152)	p-value	F(6,133)	p-value
#1	3.62	0.001**	2.89	0.011*	1.78	0.095	1.68	0.130
#2	2.46	0.021*	0.62	0.711	1.16	0.331	0.39	0.882
#3	4.41	<0.001***	2.62	0.020*	1.84	0.084	1.06	0.389
#4	1.31	0.248	0.70	0.654	0.83	0.565	0.75	0.610
#5	2.44	0.021*	2.26	0.042*	0.97	0.453	1.06	0.387
#6	3.49	0.002**	2.72	0.016*	1.60	0.138	1.54	0.169
#7	3.20	0.003**	1.99	0.071	1.36	0.224	0.89	0.505
#8	3.53	0.002**	2.05	0.064	1.70	0.113	1.18	0.321
#9	3.57	0.001**	3.07	0.008**	2.10	0.047*	2.16	0.051
#10	1.73	0.107	0.67	0.677	1.13	0.345	0.84	0.540
#11	2.43	0.022*	1.53	0.172	1.42	0.200	1.25	0.285
#12	2.51	0.018*	1.76	0.113	1.51	0.169	1.42	0.212
#13	0.80	0.591	0.71	0.641	0.17	0.991	0.16	0.986
#15	5.85	<0.001***	2.53	0.024*	3.27	0.003**	1.51	0.181
#16	2.61	0.014*	0.92	0.486	1.46	0.186	0.80	0.573
#17	1.57	0.147	1.56	0.163	0.53	0.809	0.59	0.741
#18	3.65	0.001**	3.56	0.003**	2.23	0.034*	2.55	0.023*
#19	2.80	0.009**	1.00	0.429	1.96	0.064	1.23	0.295
#20	3.57	0.001**	1.79	0.106	2.47	0.020*	1.77	0.109
#21	2.67	0.012*	1.50	0.182	1.05	0.399	0.72	0.632
#22	4.39	<0.001***	2.65	0.019*	1.91	0.071	1.16	0.330
#23	5.92	<0.001***	2.48	0.026*	3.54	0.001**	1.66	0.135

https://doi.org/10.1121/10.0004220

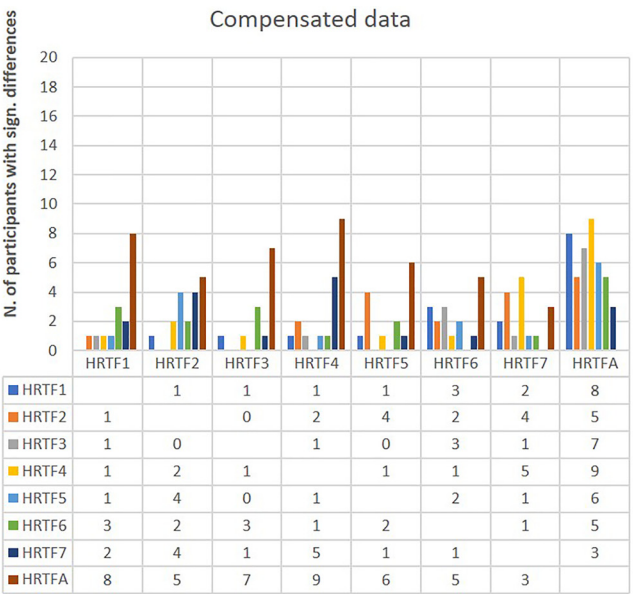
TABLE III. SRT values (dB) and IDs for the best and worst measured HRTFs (i.e., excluding $HRTF_A$), as well as the difference between them. Values are displayed for both raw and compensated SRT data. The first column shows the ID of the participants.

	Best HRTF				Worst HRTF			
	Raw		Compensated		Raw		Compensated	
ID	SRT (dB)	HRTF ID	SRT (dB)	HRTF ID	SRT (dB)	HRTF ID	SRT (dB)	HRTF ID
#1	−16.27	1	−13.71	1	−13.22	5	−11.59	7
#2	−15.40	7	−13.19	7	−14.20	5	−12.08	2
#3	−16.38	2	−13.96	2	−13.53	5	−12.26	5
#4	−16.05	2	−13.63	2	−14.20	6	−12.09	6
#5	−15.55	6	−13.44	6	−13.13	5	−11.78	3
#6	−15.70	7	−13.49	7	−12.95	5	−11.48	4
#7	−16.20	1	−13.80	4	−13.40	5	−12.13	5
#8	−15.40	6	−13.29	6	−12.80	5	−11.48	3
#9	−15.13	1	−12.56	1	−12.38	7	−10.17	7
#10	−16.30	7	−14.38	5	−14.75	4	−12.53	4
#11	−16.55	3	−14.53	3	−14.75	7	−12.54	7
#12	−15.88	4	−13.65	4	−13.75	7	−11.54	7
#13	−14.40	1	−11.88	4	−12.95	5	−11.27	7
#15	−16.35	2	−13.93	2	−13.47	5	−11.49	7
#16	−15.80	6	−13.69	6	−14.03	5	−11.89	1
#17	−15.78	1	−13.21	1	−13.38	5	−11.77	7
#18	−15.20	6	−13.09	6	−11.95	5	−10.61	2
#19	−15.88	1	−13.61	5	−13.95	2	−11.53	2
#20	−16.48	4	−14.25	4	−14.03	5	−11.81	2
#21	−15.25	3	−13.23	3	−13.25	5	−11.83	4
#22	−16.82	2	−14.41	2	−13.78	5	−12.51	5
#23	−18.63	4	−16.40	4	−16.18	5	−14.07	7

(Rothman, 1990; Saville, 2015). We therefore have to consider the probability of false positives with our chosen significance threshold ($\alpha = 0.05$). In this case, we have 28 comparisons per participant, and with 22 participants, we have 616 comparisons in total. With such a number of comparisons, we can expect that an average of 5% of them are false positives, which means approximately one false positive per paired comparison. It can be noted that, for raw SRT analysis, the $HRTF_A$ and $HRTF_5$ show a larger number of participants with significant differences. However, for



	HRTF1	HRTF2	HRTF3	HRTF4	HRTF5	HRTF6	HRTF7	HRTFA
HRTF1		2	3	0	11	1	3	19
HRTF2	2		0	2	9	4	2	17
HRTF3	3	0		1	6	3	1	12
HRTF4	0	2	1		7	3	5	14
HRTF5	11	9	6	7		5	3	6
HRTF6	1	4	3	3	5		1	14
HRTF7	3	2	1	5	3	1		14
HRTFA	19	17	12	14	6	14	14	



	HRTF1	HRTF2	HRTF3	HRTF4	HRTF5	HRTF6	HRTF7	HRTFA
HRTF1		1	1	1	1	3	2	8
HRTF2	1		0	2	4	2	4	5
HRTF3	1	0		1	0	3	1	7
HRTF4	1	2	1		1	1	5	9
HRTF5	1	4	0	1		2	1	6
HRTF6	3	2	3	1	2		1	5
HRTF7	2	4	1	5	1	1		3
HRTFA	8	5	7	9	6	5	3	

FIG. 6. (Color online) *Post hoc* pairwise comparisons using LSD. The vertical axis indicates the number of participants with significant differences in the pairwise comparison between the HRTF IDs indicated by the color and the horizontal axis. In addition, this information is also shown in the tables below the graphs.

compensated SRT analysis, the $HRTF_A$ is the only one showing a large number of participants with significant differences when comparing with other conditions. Nevertheless, also when considering only the measured HRTFs (i.e., $HRTF_{1-7}$), the number of pairs with significant differences (37) is clearly higher than the expected number of false positives by chance (23.1), showing a modest sign that individuals perform differently with different measured HRTFs.

The distribution of the SRT measurements for each HRTF is presented in Fig. 7 for four selected participants: two showing overall significant differences in the ANOVA analysis (#15 and #23) and two with no significant differences (#4 and #17) for both raw and compensated SRTs.

IV. DISCUSSION

Previous studies have investigated the binaural loudness phenomenon and the fact that sources located in lateral positions are perceived as louder than sources located in frontal positions due to the specific shape of the human head (Lokki and Pätynen, 2011). The benefit caused by, among other things, the power ratio between sides and front for each HRTF can be considered as the same for every participant. Considering the aims and hypotheses of the present study (H1 and H2), an argument could be made for quantifying this HRTF-related benefit and using it for compensating the SRT results. This would result in minimising the HRTF-specific differences, focusing the analysis on the monoaural spectral nature of the HRTFs, on the relationship between each HRTF and each subject, and, possibly, also on the impact of cognitive processes when completing SRT tasks using different HRTFs (this is, however, beyond the scope of the current study).

The Lavandier–Culling–Jelfs model was used to estimate the HRTF-specific benefit, and adjustments have been carried out on the SRTs for every HRTF. Larger compensations were needed for $HRTF_5$ and $HRTF_A$ (see Table I). Looking specifically at this result, it is evident that the model predicted very well the observed data, with a correlation coefficient of 0.9547 ($p = 0.0008$) when comparing across measured HRTF conditions (i.e., $HRTF_{1-7}$). These correlation coefficients are in line with the ones obtained by Jelfs *et al.* (2011), where they compare their predictions with results from previous studies. The current study can be considered as a further validation of the LCJ model, extending its use (and validity) to the comparison of SRT outcomes between different HRTFs (while previous comparisons focused mainly on different acoustic environments and source/receiver configurations). The analysis of the

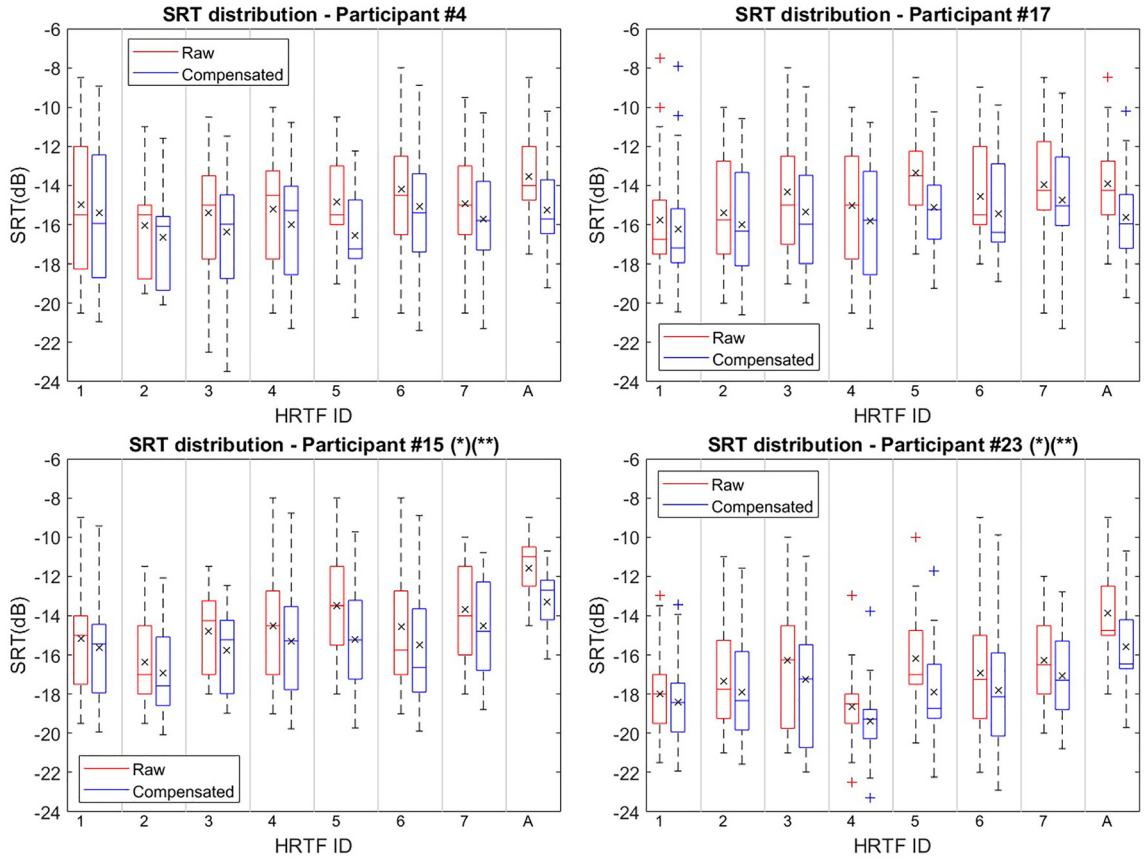


FIG. 7. (Color online) SRT distribution box plots for four specific participants. In each box, the central horizontal mark indicates the median, black crosses the mean, and the bottom and top edges the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, which are plotted individually (+). Asterisks in the title of each graph indicate participants with significant differences in the ANOVA for raw (*) and compensated (**) SRT data.

https://doi.org/10.1121/10.0004220

results and the discussion that follows are carried out for both raw and compensated SRT data.

The initial analysis focused on the whole dataset, pooling the individuals' data together and looking at the overall differences between HRTFs. The overall measured SRT values are in line with findings from previous studies carried out in similar conditions (e.g., Bronkhorst and Plomp, 1988). Looking at the raw data, $HRTF_5$ and $HRTF_A$ are the only two conditions showing significantly worse (i.e., higher dB value) SRT values if compared with the other HRTFs. Regarding $HRTF_A$, a plausible explanation can be made considering that it does not include any monoaural spectral cue, as the model does not consider the pinnae or other relevant anthropometric features beyond an approximated spherical head. The same explanation is clearly not valid for $HRTF_5$; although $HRTF_{1-7}$ all come from the same database, $HRTF_5$ is the only one that was not included in the subset obtained by Katz and Parseihian (2012), which was used to select our original sample. As explained in Sec. II, this group of HRTFs represents an optimization of the LISTEN database, selected to produce the best subjective spatialization for the most listeners. The fact that the $HRTF_5$ was selected by mistake and it is not included in the subset can serve as a potential explanation of its lower SRT performances. The outcome of this study can therefore be considered as a further validation of the method used by (Katz and Parseihian, 2012). However, as it is shown in Fig. 2, the spectrum of the $HRTF_5$ does not present any relevant difference if compared with the rest of the HRTFs, which means that it is appropriate for it to be used in the present study together with the other HRTFs.

The analysis of the compensated data gives us some additional cues for understanding this result, as both $HRTF_5$ and $HRTF_A$ resulted in a significantly higher compensation factor (Table I), underlying how those two HRTFs are "universally" worse in terms of SRT if compared with the others. This partly disproves our second hypothesis (H2), therefore, that there are no individually measured HRTFs that are universally better or worse than others on this specific task. But after the compensation has been taken into account, no significant differences could be found anymore between the measured HRTFs, while a significant difference still appears between all the measured HRTFs and $HRTF_A$, except for $HRTF_5$. This seems to underline that, once the SRT values have been compensated in terms of frontal/sides power ratio (and other differences due mainly to interaural cues), no measured HRTF results in being better or worse than the others. On the other side, the synthesised HRTF, which does not include any monoaural spectral cue, results in significantly worse SRT performances, possibly underlying the importance of the direction dependent filtering caused by the pinnae and other relevant elements not included in the spherical head model (e.g., torso, shoulders, etc.).

Considering the modest learning effect that was measured across the various sections, accounting for 2.87% of the SRT variation, this can be due to the fact that each word appears several times across the whole experiment for each subject, making the recognition task potentially easier as the participant progresses through sessions. It is important to consider that feedback is an essential mechanism for both procedural and perceptual learning (Ortiz and Wright, 2009), and during the tests, participants were not given any feedback regarding whether they correctly identified the various words. It is therefore not surprising that only a small effect of learning was found in the SRT data, which could be attributed to improvements in the participants' understanding of the task, their ability to focus attention, and/or, as mentioned before, familiarization with the target words. We can therefore consider that, even though it is significant, the effect of learning can be disregarded for the purpose of this analysis.

SRT data have then been analyzed separately for each individual. Looking at the raw data, for 82% of the participants (18 of 22 participants) a significant effect of the HRTF was found on the SRT score (see the left part of Table II, column $HRTF_{1-7} + HRTF_A$). The choice of HRTF seems therefore to have a significant impact on the SRT scores for the large majority of the tested participants, confirming the first of our initial hypotheses (H1). We have already discussed above the nature of $HRTF_A$ and the fact that its overall worse performances in terms of SRT seem to indicate that monoaural spectral cues have a significant impact on SRT in cocktail party conditions. This improvement is smaller than the one generated by interaural cues (Culling *et al.*, 2004) but nevertheless significant for several of the tested subjects. A reduction of the number of participants showing a significant effect of HRTF is therefore expected when removing $HRTF_A$ from the comparison (from 18 to 9 of 22), but it is important to notice how a significant effect can still be found for 41% of the participants (see column $HRTF_{1-7}$). Albeit less strongly, this result is still supporting H1, therefore, that for a given subject, different HRTFs provide different performances in terms of speech recognition. Looking at the raw data displayed in Table III, the differences between SRTs using the "best" and "worst" measured HRTFs (i.e., excluding $HRTF_A$) for the various participants are between 3.25 and 1.2 dB. These are comparable with the ranges found in previous studies when looking at BMLD and at the impact of interaural differences on SRM (e.g., Culling *et al.*, 2004). Looking at the comparison between the ITD and ILD estimates for the different HRTFs (shown in Fig. 3), no major differences could be found. In terms of ITD, the maximum observed difference is 136.7 μs between $HRTF_7$ and $HRTF_A$. Looking at the ILD, the larger difference is of 4.24 dB between $HRTF_1$ and $HRTF_7$. In addition, no correlation can be seen with the results obtained in the perceptual study (i.e., the SRT values). Also, looking at the spectral representation of each HRTF (see Fig. 2), while overall magnitude difference can be found for frequencies below 10 kHz, these are generally similar for frontal and lateral positions and therefore should not result in altered SNR beyond the improvement already accounted for by the proposed compensation (see Sec. II F). These observations

reinforce the concept that the influence of the HRTF on the SRT, within the experimental conditions of this study, is due to the monaural spectral cues of each HRTFs and the potential match with those of each specific individual.

The analysis of the compensated SRT data gives rather different results (see the right part of Table II). The number of participants showing a significant effect of the HRTF decreases to five, and to only one when excluding $HRTF_A$. Interestingly, when looking at data in Table III, it can be observed that while the best HRTFs remain more or less the same between raw and compensated data (this is the case for 82% of the participants), the worst HRTFs change significantly, with $HRTF_5$ being the worst for 16 participants for the raw data and for only 6 participants for the compensated data. A similar situation is found when looking at the *post hoc* analyses in Fig. 6. It is evident that no single HRTF achieved the best SRT values and that the differences between HRTF pairs happen across the whole corpus, with higher occurrences including $HRTF_A$ and $HRTF_5$. Even in this case, the number of significantly different HRTF pairs is reduced when looking at the compensated data if compared with the raw data, albeit remaining above the expected number of false positives by chance. The fact that the $HRTF_5$ performs as the worst for the 77% of the participants, when looking at the raw data, can be considered again as a potential validation of the method and set proposed by Katz and Parseihian (2012), since it is the only one not belonging to the subset and the one that performed significantly worse if compared with the others.

It is evident that the compensation of the SRT values caused a significant change in the results and consequently their interpretation. When looking at H1, while the raw data support the fact that there is a significant effect of the HRTF choice for a large number of individuals, this cannot be so clearly evinced when looking at the compensated data. It is, however, true that also in the latter case, a certain number of significant pairwise differences (higher than the estimate of false positives by chance) can still be found (see Fig. 6). A symmetrically different situation is found when looking at H2; in this case, the compensated data offer better results, if compared with the raw data, in supporting the hypothesis that there are no individually measured HRTFs that are universally better or worse than others. When making these considerations, it is, however, important to take into account the nature of the compensation, which aimed at balancing those differences that could make some HRTFs worse or better for the overall sample of participants, regardless of individual differences. There are clearly some HRTFs that are generally better or worse than others when looking at performances in virtual speech-in-noise recognition. At the same time, there are individual features of HRTFs that allow certain subjects to perform better with them and others to perform worse.

One further element outlined by the analyses of our experimental data is the variance of the measurement across participants, HRTFs, and test sessions. Using the same HRTF, the standard deviation of the SRT values for individuals across the different sessions spans between 1.5 and 3.5 dB, with a mean value of 2.8 dB. Several studies have been carried out in the past by other researchers using the same two-syllable Spanish word dataset (de Cárdenas and Marrero Aguiar, 1994), but none of these looked specifically at test-retest reliability. Other studies using different speech material have looked at this and have reported standard deviation values between 1.5/1.8 dB (Hagerman and Kinnefors, 1995) and 2.2 dB (Saleh, 2013), slightly lower than the values found in the current study.

The overall variability of the SRTs is evident also when looking at individual cases, as reported in the box plots in Fig. 7. It is evident that for specific participants (e.g., #15 and #23), one HRTF is achieving overall better SRT values, and similarly another HRTF is achieving worse ones. For these HRTFs (and these participants), the variance of the values across sessions is markedly smaller if compared with the other HRTFs and with the other participants. This can be noted by looking at the box plots for participants #4 and #17, which show a higher overall variance and no best or worst HRTFs.

Despite the fact that subjects were not asked to rate different HRTFs based on subjective perceptual attributes, it was nevertheless of interest to attempt a comparison between the outcomes of the current study and those looking at qualitative HRTF ratings. It is in fact true that the SRT results seem to be in line with the ones from previous research looking at the repeatability of HRTF ratings (Andreopoulou and Katz, 2016), where only a certain number of participants (categorised as expert assessors) were able to rate a certain number of HRTFs repeatably across different sessions. An attempt was made to assess the ranking of the HRTF collection for each subject using other perceptual metrics, with the aim to compare these results with the SRT measurements. The trajectory-based HRTF selection procedure developed by Katz and Parseihian (2012) was used. This test aims at subjectively ranking a set of HRTFs in terms of how well the rendering generated using each of them corresponds to a described trajectory of the sound source around the listener's head. Eighteen individuals who took part in the SRT experiment also performed the other test. Only for one subject was the best HRTF selected with the trajectory-based method the same as the one yielding the best SRT. For the same subject, also the worst performing HRTFs were a match across the two experiments. No other match was found for any of the remaining subjects. When trying to explain the reasons behind this result, it is important to consider a further study from Andreopoulou and Katz (2016), which outlined how the trajectory-based test is significantly less reliable and repeatable when performed by non-expert assessors if compared with expert ones. All the subjects who took part in this experiment would have been classified in the former category, as they did not have any previous experience with binaural audio content and experiments. Furthermore, it needs to be noted that several participants reported, after the trajectory-based test, selecting the best and/or worst HRTF

just because they were forced to do so by the test procedure, while they could not actually hear any significant difference in the quality of spatialisation between the various HRTFs in the set. The attempt to match the SRT measurements with a HRTF ranking made using other perceptual metrics was therefore unsuccessful.

In summary, we have demonstrated that, within the tested conditions and looking at both raw and compensated SRT data, there can be a significant effect of HRTF choice on speech recognition, and this effect can be different for different subjects. The implications of these findings could be relevant to several research areas. For example, when modeling binaural speech-in-noise perception, monoaural cues should be accounted for as well as binaural ones; when assessing speech-in-noise performances within binaurally rendered virtual cocktail party scenarios, the choice of the HRTF should be carefully considered. Furthermore, this research opens new questions that require further investigations, such as

- How do HRTF-specific SRT performances compare with qualitative and quantitative HRTF selection for an expert assessor?
- Could the SRT differences between HRTFs for a given subject be even larger when the maskers are positioned on the medial plane, therefore when source separation can be performed only relying on monoaural cues?
- Would these results be different if the masker signals were more similar to speech, i.e., presenting more complex and variable spectral envelopes?
- Could there be an effect of accommodation/adaptation to a specific HRTF, which could result in improvements in SRT as well as sound source localisation and overall quality of the spatialization when using the trained HRTF?

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 644051 and Spanish National Project Grant No. PID2019-107854GB-I00. The authors would like to thank Dr. Brian F. G. Katz and his team for granting permission to use their trajectory-based HRTF selection tool.

Algazi, V., Duda, R., Thompson, D., and Avendano, C. (**2001**). "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, Catalog No. 01TH8575, October 24, New Platz, NY, pp. 99–102.

Andreopoulou, A., and Katz, B. F. G. (**2016**). "Investigation on subjective HRTF rating repeatability," in *Proceedings of Audio Engineering Society Convention 140*, June 4–7, Paris, Paper 9597.

ANSI (**1997**). ANSI/ASA S3.5-1997 (R2017), *Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).

ARI (**2013**). Acoustics Research Institute of the Austrian Academy of Sciences, HRTF-Database, https://www.oeaw.ac.at/en/isf/das-institut/software/hrtf-database (Last viewed April 2, 2021).

Auditory Modeling Toolbox (**2011**). "JELFS2011—Predicted binaural advantage for speech in reverberant conditions," http://amtoolbox.sourceforge.net/amt-0.9.9/doc/models/jelfs2011.php (Last viewed April 2, 2021).

Blauert, J. (**1997**). *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA), p. 494.

Bronkhorst, A. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust. United Acust. **86**, 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.

Bronkhorst, A. W., and Houtgast, T. (**1999**). "Auditory distance perception in rooms," Nature **397**(6719), 517–520.

Bronkhorst, A. W., and Plomp, R. (**1988**). "The effect of head-induced interaural time and level differences on speech intelligibility in noise," J. Acoust. Soc. Am. **83**, 1508.

Bronkhorst, A. W., and Plomp, R. (**1992**). "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," J. Acoust. Soc. Am. **92**(6), 3132–3139.

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.

Ching, T. Y., Van Wanrooy, E., Dillon, H., and Carter, L. (**2011**). "Spatial release from masking in normal-hearing children and children who use hearing aids," J. Acoust. Soc. Am. **129**(1), 368–375.

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. F., Hawley, M. L., and Litovsky, R. Y. (**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. F., and Mansell, E. R. (**2013**). "Speech intelligibility among modulated and spatially distributed noise sources," J. Acoust. Soc. Am. **133**(4), 2254–2261.

de Cárdenas, M. R., and Marrero Aguiar, V. (**1994**). *Cuaderno de Logoaudiometría. Guía de Referencia Rápida* (Universidad Nacional de Educación a Distancia, Madrid).

Durlach, N. I., Rigopulos, A., Pang, X., Woods, W., Kulkarni, A., Colburn, H., and Wenzel, E. (**1992**). "On the externalization of auditory images," Presence (Camb.) **1**(2), 251–257.

Edmonds, B. A., and Culling, J. F. (**2006**). "The spatial unmasking of speech: Evidence for better-ear listening," J. Acoust. Soc. Am. **120**(3), 1539–1545.

Engel, I., Alon, D. L., Robinson, P. W., and Mehra, R. (**2019**). "The effect of generic headphone compensation on binaural renderings," in *Proceedings of the Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, March 27–29, York, UK.

Freyman, R. L., and Zurek, P. M. (**2008**). "Effects of reverberation on spatial release from masking," J. Acoust. Soc. Am. **123**(5), 2977–2977.

Fu, Q.-J., and Galvin, J. J. III (**2003**). "The effects of short-term training for spectrally mismatched noise-band speech," J. Acoust. Soc. Am. **113**(2), 1065–1072.

Gardner, W. G., and Martin, K. D. (**1995**). "HRTF measurements of a KEMAR," J. Acoust. Soc. Am. **97**(6), 3907–3908.

Geronazzo, M., Peruch, E., Prandoni, F., and Avanzini, F. (**2019**). "Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization," J. Audio Eng. Soc. **67**(6), 414–428.

Geronazzo, M., Spagnol, S., Bedin, A., and Avanzini, F. (**2014**). "Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 4–9, Florence, Italy, pp. 4463–4467.

Hagerman, B., and Kinnefors, C. (**1995**). "Efficient adaptive methods for measuring speech reception threshold in quiet and in noise," Scand. Audiol. **24**(1), 71–77.

Hammershøi, D., and Møller, H. (**2005**). "Binaural technique-basic methods for recording, synthesis, and reproduction," in *Communication Acoustics* (Springer, New York), pp. 223–254.

Härmä, A., van Dinther, R., Svedström, T., Park, M., and Koppens, J. (**2012**). "Personalization of headphone spatialization based on the relative localization error in an auditory gaming interface," in *Proceedings of*

Audio Engineering Society Convention 132, April 26–29, Budapest, Hungary.
Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (**1999**). "Speech intelligibility and localization in a multi-source environment," J. Acoust. Soc. Am. **105**(6), 3436–3448.
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.
IEC (**2003**). "Sound system equipment—Part 16: Objective rating of speech intelligibility by speech transmission index," Technical Report, https://webstore.ansi.org/standards/iec/iec6026816eden2003 (Last viewed April 2, 2021).
Iida, K., Ishii, Y., and Nishioka, S. (**2014**). "Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae," J. Acoust. Soc. Am. **136**(1), 317–333.
IoSR (**2017**). MatlabToolbox/+iosr at master · IoSR-Surrey/MatlabToolbox. https://github.com/IoSR-Surrey/MatlabToolbox (Last viewed April 2, 2021).
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), 96–104.
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.
Katz, B. F. G., and Noisternig, M. (**2014**). "A comparative study of interaural time delay estimation methods," J. Acoust. Soc. Am. **135**(6), 3530–3540.
Katz, B. F. G., and Parseihian, G. (**2012**). "Perceptually based head-related transfer function database optimization," J. Acoust. Soc. Am. **131**(2), EL99–EL105.
Koehnke, J., and Besing, J. M. (**1996**). "A procedure for testing speech intelligibility in a virtual listening environment," Ear Hear. **17**(3), 211–217.
Lavandier, M., and Culling, J. F. (**2010**). "Prediction of binaural speech intelligibility against noise in rooms," J. Acoust. Soc. Am. **127**(1), 387–399.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Lokki, T., and Pätynen, J. (**2011**). "Lateral reflections are favorable in concert halls due to binaural loudness," J. Acoust. Soc. Am. **130**(5), EL345–EL351.
Masiero, B., and Fels, J. (**2011**). "Perceptually robust headphone equalization for binaural reproduction," in *Proceedings of Audio Engineering Society Convention 130*, May 13–16, London.
Mayr, S., Buchner, A., Erdfelder, E., and Faul, F. (**2007**). "A short tutorial of GPower," Tutor. Quant. Methods Psychol. **3**(2), 51–59.
Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (**1995**). "Head-related transfer functions of human subjects," J. Audio Eng. Soc. **43**(5), 300–321.
Moore, B., and Glasberg, B. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**(3), 750–753.
Musiek, F. E., Chermak, G. D., and Weihing, J. (**2014**). *Auditory Training* (Plural Publishing Inc., San Diego, CA).
Ortiz, J. A., and Wright, B. A. (**2009**). "Contributions of procedure and stimulus learning to early, rapid perceptual improvements," J. Exp. Psychol. **35**(1), 188–194.
Rayleigh, L. (**1907**). "XII. on our perception of sound direction," Lond. Edinb. Dubl. Philos. Mag. J. Sci. **13**(74), 214–232.
Rothman, K. J. (**1990**). "No adjustments are needed for multiple comparisons," Epidemiology **1**(1), 43–46.
Saleh, S. M. I. (**2013**). "The efficacy of fitting cochlear implants based on pitch perception," Ph.D. thesis, University College London.
Saville, D. J. (**2015**). "Multiple comparison procedures—cutting the Gordian knot," Agron. J. **107**(2), 730–735.
Schonstein, D., Ferré, L., and Katz, B. F. (**2008**). "Comparison of headphones and equalization for virtual auditory source localization," J. Acoust. Soc. Am. **123**(5), 3724.
Simon, L. S., Zacharov, N., and Katz, B. F. (**2016**). "Perceptual attributes for the comparison of head-related transfer functions," J. Acoust. Soc. Am. **140**(5), 3623–3632.
Sondergaard, P. L., and Majdak, P. (**2013**). "The auditory modeling toolbox," in *The Technology of Binaural Listening* (Springer, Berlin), pp. 33–56.
van Wijngaarden, S. J., and Drullman, R. (**2008**). "Binaural intelligibility prediction based on the speech transmission index," J. Acoust. Soc. Am. **123**(6), 4514–4523.
Warusfel, O. (**2003**). "LISTEN HRTF DATABASE," http://recherche.ircam.fr/equipes/salles/listen/ (Last viewed April 2, 2021).
Woodworth, R. S., Barber, B., and Schlosberg, H. (**1954**). *Experimental Psychology* (Oxford, London).