Perceptual attributes for the comparison of head-related transfer functions

Laurent Simon, Nick Zacharov, Brian F.G. Katz

Perceptual attributes for the comparison of head-related transfer functions

Laurent S. R. Simon, Nick Zacharov, and Brian F. G. Katz

Citation: The Journal of the Acoustical Society of America **140**, 3623 (2016); doi: 10.1121/1.4966115
View online: https://doi.org/10.1121/1.4966115
View Table of Contents: https://asa.scitation.org/toc/jas/140/5
Published by the Acoustical Society of America

Perceptual attributes for the comparison of head-related transfer functions

Laurent S. R. Simon,[1] Nick Zacharov,[2] and Brian F. G. Katz[1,a)]
[1]*Audio Acoustics Group, LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France*
[2]*DELTA, SenseLab, Hørsholm, Denmark*

(Received 18 June 2016; revised 7 October 2016; accepted 11 October 2016; published online 11 November 2016)

The benefit of using individual head-related transfer functions (HRTFs) in binaural audio is well documented with regards to improving localization precision. However, with the increased use of binaural audio in more complex scene renderings, cognitive studies, and virtual and augmented reality simulations, the perceptual impact of HRTF selection may go beyond simple localization. In this study, the authors develop a list of attributes which qualify the perceived differences between HRTFs, providing a qualitative understanding of the perceptual variance of non-individual binaural renderings. The list of attributes was designed using a Consensus Vocabulary Protocol elicitation method. Participants followed an Individual Vocabulary Protocol elicitation procedure, describing the perceived differences between binaural stimuli based on binauralized extracts of multichannel productions. This was followed by an automated lexical reduction and a series of consensus group meetings during which participants agreed on a list of relevant attributes. Finally, the proposed list of attributes was then evaluated through a listening test, leading to eight valid perceptual attributes for describing the perceptual dimensions affected by HRTF set variations.
 [http://dx.doi.org/10.1121/1.4966115]

[JFL] Pages: 3623–3632

I. INTRODUCTION

Binaural hearing refers to the ability of humans to interpret sounds arriving at both ears into complex auditory scenes. This is made possible by the use of various cues, such as provided by the head-related transfer function (HRTF), sound reflections in the surrounding space, or dynamic modifications of these cues caused by head movements.

The HRTF is a set of filters describing the behavior of acoustic waves from a collection of points in space to both of an individual's ears. It can be used to simulate real-life hearing by filtering audio signals with an HRTF. Related to each individual's morphology (size of the head, shape of the pinna, etc.), HRTFs vary from one listener to another.

When listening to binaural audio that has not been synthesized or recorded with one's own HRTF (Seeber *et al.*, 2003; Wenzel *et al.*, 1993), i.e., non-individualized HRTF, front-back confusions, and distortions of spatial and timbral perception might occur, and sounds might not be externalized (Begault *et al.*, 2000). In order to reduce these effects, several approaches have been studied to select the "best" HRTF for a given participant.

Signal processing approaches compare signal-based features of different sets of HRTFs in order to select a number of HRTFs that have little redundancy one with each other (Bondu *et al.*, 2006). Such approaches aim to group similar HRTFs based on their signal characteristics. Other approaches for the reconstruction of individualized HRTFs have, for example, been proposed in Kistler and Wightman (1992) and Kirkeby *et al.* (2012). Such approaches may be used as an initial step for HRTF selection, as it reduces the number of HRTFs that should be used for subjective selection of HRTFs. However, the correlation between signal domain metrics and perceptual differences is still not well established.

Acoustical model approaches aim to model an individual's HRTF using morphological data or photographs (Guillon *et al.*, 2008; Iida *et al.*, 2014; Schönstein and Katz, 2010; Ziegelwanger *et al.*, 2015). However, the perceptual effect of a mismatch between a modeled HRTF and a measured one is still unknown.

In past studies, subjective tests have been used for HRTF selection through either localization (Begault *et al.*, 2000; Seeber *et al.*, 2003), preference (Katz and Parseihian, 2012), or externalization evaluation (Hur *et al.*, 2008) experiments. This assumes that if an individual can locate sounds where intended with a given HRTF set, then the HRTF is well matched to this person's own HRTF. Alternatively, preference ratings or rankings (Andreopoulou and Katz, 2016) are used when one wants to study quality of experience.

However, source position is only one perceptual attribute of sound, and preference is a general rating that gives little information about what a person perceives. Early applications of spatial and timbral attributes in the domain of binaural audio can be found in Huopaniemi *et al.* (1999) and Lorho *et al.* (2000). In order to evaluate in more detail the characteristics of audio recording and reproduction systems, other perceptual attributes have been listed in different contexts: room acoustics (Lokki *et al.*, 2012), sound reproduction over loudspeakers (Berg and Rumsey, 1999; Guastavino *et al.*, 2005; Zacharov and Koivuniemi, 2001), headphones

a)Electronic mail: brian.katz@limsi.fr

(Lorho, 2005a), or virtual acoustics (Lindau *et al.*, 2014). Despite these previous works, so far no attribute list has been defined specifically for binaural sound reproduction which involves a degree of complexity above and beyond these other methods that varies between individual listeners. The aim of this paper is therefore to propose a similar list designed for binaural listening, which can be used in future studies to investigate the subjective reasons that lead to why a person may prefer one HRTF set over another.

In order to find perceptual attributes that can be used to distinguish two different stimuli, multidimensional analysis techniques such as Multi-Dimensional Scaling (Colomes *et al.*, 2010; Martens and Zacharov, 2000) and Principal Component Analysis (Choisel and Wickelmaier, 2007) can be used. The main advantage of multidimensional analysis methods is the lack of bias, but the interpretation of the results is complex (Berg and Rumsey, 1999) and it is not possible with these methods to identify and describe the perceptual dimensions, which are therefore not re-usable for further experiments.

Alternatively, Descriptive Analysis aims to develop a list of perceptual attributes, definitions, and end points, that can be used for multiple experiments either by defining individual lists of attributes [Individual Vocabulary Protocol (IVP) (Delarue and Sieffermann, 2004; Kelly, 1955)], which requires to use the same individuals for all the experiments to the benefit of a lack of bias (Berg and Rumsey, 1999), or a single list of perceptual attributes that can be used by any number of trained individuals [Consensus Vocabulary Protocol (CVP) (Cairncross and Sjostrom, 2004; Stone *et al.*, 1974)]. This latter approach has been used in the current study, using a panel of sound engineer experts in spatial audio and binaural mixing.

The object of this study is to design a list of attributes that describe the perceptual dimensions associated with the choice of HRTF used in a binaural production. The current study was carried out in French, however, a translation of the attribute list is provided as it is expected to be equally applicable in other languages.

II. WHAT MAKES BINAURAL AUDIO QUALITY EVALUATION DIFFERENT

Binaural technology offers a solution for sound spatialization which is the closest to real-life listening. Binaural reproduction attempts to mimic all acoustic cues for the human localization of sounds, reproducing the corresponding acoustic pressure signal at the entrance of the two ear canals of the listener. These two signals should be a complete and sufficient representation of the sound scene, since they are the only information that the auditory system requires in order to identify the location of a sound source in three-dimensional (3D) space. Thus, binaural rendering of spatial information is fundamentally based on the production (either through recording or synthesis) of localization cues, namely, the interaural time difference (ITD), the interaural level difference (ILD), and spectral cues (Nicol, 2010). The combined effect of these different cues are represented by the HRTF which characterizes the spectro-temporal filtering of an incident acoustic wave due to the head, torso, and pinnae morphology of the listener.

The ILD and ITD as a function of source position are determined principally by the size and shape of the head, as well as the position of the ears on the head (Blauert, 1996). In a classic loudspeaker reproduction system using amplitude panning, the phantom source is generated by sending a coherent signal to two or more physical loudspeakers with varying ratios. These acoustic signals sum physically at the entrance to the auditory system, resulting in an ILD coherent with the phantom source position. However, the ITD cues are determined by the physical location of the loudspeakers, meaning that amplitude panning methods result in the creation of phantom sources having a phantom image with the ILD of this phantom source while the associated ITD is a superposition of the ITDs of the physical speakers as a function of the listener's position (inside or outside the "sweet-spot"). With regards to Spectral Cues, the same situation arises, where the Spectral Cues, created due to the directional filtering of the incident source waves by the listener's torso, head, and pinnae are those of the physical speakers, not the phantom source. These different combinations of conflicting auditory cues are the origin of many of the limitations of spatial audio rendering over loudspeakers (Pulkki and Karjalainen, 2001). However, in listening comparisons of a given reproduction system, for a given physical speaker configuration and listening position, one can expect that each listener will receive the same sound field, and therefore have the corresponding auditory cues for that system in combination with their individual morphology (head size, ear shape, etc.).

If we consider the case of binaural sound reproduction compared to loudspeaker reproduction methods we add a new layer of complexity. In binaural audio the ILD, ITD, and Spectral Cues are not determined by the actual listener's morphology but by the rendering system as the audio signal is presented directly at the ear canals, not to the listener as an acoustic wave field which then interacts with their morphology prior to entrance into the auditory system. As such, the binaural rendering system must impart all of the acoustic localization cues in the audio signal. Due to the necessity, and consequently the ability, to control all aspects of the perceived signal at the entrance of the ear canals, binaural reproduction is a presentation method well suited and often employed in psychoacoustics and sound perception studies (Blauert, 1996; Suzuki *et al.*, 2011; Xie, 2013).

When judging the quality of a binaural rendering using a non-individual HRTF set, the discrepancy between the listener's own HRTF and the HRTF employed for the binaural rendering will have a degradation impact. This degradation will vary between individuals, meaning that there is no means of establishing a global consensus perception for a given stimuli, contrary to previous works concerning loudspeaker spatial reproduction or audio-codec evaluations.

The goal of this study is therefore to establish a set of perceptual attributes which are directly connected to HRTF variations, based on a consensus approach, which is not limited by the individual nature of binaural audio perception. Attributes for which there is a global consensus in value

(e.g., stimuli *A* is louder than *B*) are not considered pertinent to the set of attributes relevant to HRTF effects.

III. CVP

According to previous works (ISO, 1994; Lawless and Heymann, 2010; Lawless and Civille, 2013; Lorho, 2010), the main steps of CVP are:

(1) Individual vocabulary elicitation for each of the participants of the study, performed by listening to audio stimuli. The aim is to build an initial list of attributes, the basis for the second step's group discussions.
(2) A series of group meetings to which some or all of the participants should take part. During these moderated group meetings, participants reduce the list of attributes by consensus under the guidance of a panel leader. At the end of the group discussions a number of *consensus* attributes will have been developed, with associated verbal definitions and scale labels. These attributes should enable the assessor to evaluate the key perceptual characteristics of the systems under test in an objective manner.
(3) Validation of the attribute list.

The final attributes should have several characteristics (Lawless and Heymann, 2010; Piggott, 1991). They should:

- be objective;
- have little overlap;
- allow discrimination between stimuli;
- be singular rather than a combination of several terms;
- not be a combination of sub-attributes;
- be precise, well defined, and unambiguous;
- generate consensus among participants;
- relate to reality;
- not use jargon;
- be specifiable by a reference.

IV. GENERAL EXPERIMENTAL CONDITIONS

A. Participants

A pool of 17 participants (13 male, 7 professional sound engineers, 9 students in sound engineering, and 1 researcher in binaural sound; aged 20–52) took part in the first step of this study, the individual vocabulary elicitation. Their experience with binaural audio varied from novice to experienced. From the original pool, 15 participants agreed to complete the second step of this study (6 sound engineers, 1 researcher, and 8 students) comprising group sessions that aimed to reduce the list of attributes. To conclude, due to availability, nine participants from the original pool (seven sound engineers, one student, and one researcher) agreed to complete the third step of this study, the validation of the list of attributes with one additional male participant, who had not been involved in the construction of the list of attributes.

B. Audio extracts

As discussed in Sec. I, the elicited attributes ideally need to be usable in a real context. Although noise bursts are a common type of stimuli for binaural evaluation, they may not reflect all perceptual attributes elicited by the comparison of sets of HRTFs on a musical recording. More ecological stimuli would represent examples of binaural audio content, highlighting the spatial distribution and sound scene complexity. For that reason, three audio extracts were considered, representing different audio content genres, each of them created by a different sound engineer:

(1) A 5-track surround sound mix, 5 s duration, radio documentary recording, featuring a male and a female voice in a kitchen, frying noise, and background kitchen noises.
(2) A 5-track, 8 s duration, multi-track electronic music recording.
(3) A 13-track, 15 s duration, multi-track radio fiction composition.

In each extract, individual sources were not all at the same sound level. Details of the audio context are provided in Table I.

As the purpose of the verbal elicitation procedures are to produce a set of generalizable attributes, the stimuli used for the generation of the set of attributes needs to be varied enough to cover a range of types of stimuli as wide as possible. The stimuli created for the current study covers radio fiction, music, speech, various reverberation times on the mono recordings used, virtual 5.0, and various 3D scenes. It

TABLE I. Descriptions of audio extract content and track positions assigned by the audio engineers for creating the binaural stimuli.

Channel/track	Azim (°)	Elev (°)
(#1) 5-ch surround documentary		
L—front left	30	0
C—center	0	0
R—front right	−30	0
Rs—rear right	−105	0
Ls—rear left	105	0
(#2) 5 track electronic music		
Bass synth	15	−30
Guira	90	30
Synth	0	60
Percussion	−135	0
Synth drum	−60	75
(#3) 13 track radio fiction		
Ambiance 1	30	0
Ambiance 2	−60	0
Thunderstorm 1	−120	30
Thunderstorm 2	0	90
Thunderstorm 3	90	45
Gunshot	−90	45
Rain and thunder	45	30
Water dripping and thunder	−45	30
River sound	180	60
Birds	0	60
Dog crying (close)	15	−30
Dog barking (distant)	135	−15
Grandfather clock	150	0

is therefore expected that the set of attributes will be generalizable to most types of binaural content.

C. Binauralization

The aim of the study is to identify a set of perceptual attributes that represent the qualitative differences associated to use of non-individual HRTFs, independent of the ITD. To that end, compared stimuli were binauralized with seven different HRTFs that were previously identified as a reduced optimized base in Katz and Parseihian (2012) and were part of the LISTEN database (Warusfel, 2003): $HRTF_{1008}$, $HRTF_{1013}$, $HRTF_{1022}$, $HRTF_{1031}$, $HRTF_{1032}$, $HRTF_{1048}$, and $HRTF_{1053}$.

In order to binauralize an audio sample with an HRTF, each sound signal produced for that audio sample was convolved with the impulse response corresponding to the HRTF for a given direction. The position of the binauralized sound sources was imposed by the sound engineers who mixed the audio extracts, and are given in Table I.

1. HRTF pre-processing

For each HRTF set, the average level across all directions was normalized between left and right filter sets. This was done to account for possible measurement gain errors between channels as have been observed in HRTF database analysis studies (Andreopoulou *et al.*, 2015).

In order to concentrate the study on spectral differences, ITD differences between the tested HRTFs were removed and a common ITD was assigned to each direction for all HRTF sets. The ITD of each HRTF was first estimated using the centroid IACC method (Katz and Parseihian, 2012). This ITD was then removed and replaced by the average ITD computed on the 51 HRTFs for each direction of the LISTEN database. This results in a reasonable estimate of the ITD for the various source positions leaving only spectral differences between tested HRTF sets.

2. Binauralized stimuli processing

Once an audio sample was binauralized with each of the 7 HRTFs by convolution, the 10% exceeded levels (L_{10}) of the different binauralized signals were normalized in order to ensure perceived level homogeneity. Level normalization was done in order to remove *loudness* as a potential attribute with regards to both HRTF and stimuli differences. *Loudness* would be an attribute of trivial importance, but it would be difficult for listeners to ignore. L_{10} was used as it is a robust estimate for time varying signals such as those used in the current study and metrics such as peak or root-mean-square values could vary slightly between HRTFs.

This produced three stimuli sets of seven binaural samples which were used for the individual and consensus vocabulary constructions. A subset of these samples was used for the final validation experiment of the attribute list.

Audio samples were played back over reference open circumaural headphones (Sennheiser HD600), with no specific headphone equalization employed. Playback level was calibrated and fixed for all participants at all stages of the study. The listening level for all HRTFs after normalization of the documentary mix was ≈58 dBA for the entire extract, the electronic music mix was ≈52 dBA for the entire extract, and the radio fiction level varied from 49 to 57 dBA due to the dynamics of the content during the extract.

3. Question of reference stimuli

Ideally in descriptive analysis methods it is beneficial to define references for each attribute elicited in order to communicate the concept of the attribute between assessors (and panels). Previous examples of attribute development and validation can be found in Pedersen and Zacharov (2015). Reference stimuli may be employed to illustrate the meaning and polarity of each attribute scale. However, this is not possible with binaural sound, as perception of binaural sounds varies from one individual to the other, being influenced by each individual's personal HRTF. As such, the question of a reference stimuli is a difficult one for binaural audio.

The use of free-field loudspeaker rendering as a reference poses a number of difficulties. First of all, it would require subjects to continually place and remove headphones when switching. One can also ask why should a loudspeaker rendering in a given *listening room* with a given *set of speakers* be employed as the "ideal reference." In the context of the validation listening test, the reference should ideally be hidden, and this is not possible with the loudspeaker reference. In addition to these concerns, given the variety of complex sound scenes that are used in the current study (≈20 different source positions), it would be difficult to construct such a loudspeaker installation.

This study is placed in the context of improving the selection of non-individual HRTFs when individual HRTF measurements are not available for a given listener. The proposed methodology of the study employs a set of HRTFs which have been shown to contain at least one very good HRTF match for a general subject. As the case of non-individual HRTF listening is much more likely than a subject having their individual HRTF, the current test methodology is a good approximation of the use case conditions of HRTF variability.

As the current study concerns the *identification* of attributes, not the *quantification* of one HRTF relative to another, the proposed methodology should be generalizable to other HRTF data sets. For example, we are not concerned with which HRTF provides the "correct" timbre, but only if "timbre/coloration" is a valid perceptual attribute for describing the differences between HRTFs. Absolute references were therefore not developed for this study.

V. CVP PROCEDURE AND RESULTS

A. IVP

As described in Sec. I, CVP methods involve several steps. The first step involves creating an individual vocabulary for each participant. The 17 participants were asked to freely qualify the differences they could perceive between pairs of the binauralized versions of the audio samples, using

TABLE II. List of attributes obtained after each group session (in French). Similar attributes are on the same rows. "*unnamed*" indicates a case where participants could not agree on the term for that category, despite having a definition and end points.

First group	Second group	Third group
Externalisation	Externalisation	Externalisation
Immersion	Immersion	Immersion
Crédibilité		Réalisme
Discrimination spatiale	Précision de la localisation	Précision de la localization
Ampleur	Relief latéral	
Élévation	Répartition verticale	Élévation
Position latérale	Répartition latérale	Position latérale
Position avant/arrière	Répartition avant arrière	Position avant / arrière
Stabilité		
Timbre	Coloration	Modifications spectrales
Crédibilité du timbre	Respect du timbre	
Percussif		
Phasing		
Relief	Relief avant arrière	Profondeur du champ sonore
Effet de salle	Sensation d'espace	Réverbération
Niveau sonore		Niveau sonore
Continuité spatiale		
Equilibre spatial	Profondeur latérale	
	Profondeur verticale	
	Relief vertical	
	Profondeur avant arrière	Profondeur
	un-named (similar to *crédibilité*)	
	un-named (incl. timbral attributes)	

an HTML 5/javascript interface based on BeaqleJS (Kraft and Zolzer, 2014), adapted to the needs of the current study.

For each audio sample, all possible pairs of HRTFs were compared, which led to a total of 63 pairs of stimuli for which differences were freely described using a text box in the web interface. Participants took approximately 1 h for the elicitation step.

The concatenation of all participants' responses resulted in 8984 words. This word collection was then analyzed with the semantic analysis software, Tropes.[1]

B. Semi-automatic semantic reduction of the list of attributes

The semantic analysis software uses pre-defined sets of rules to analyze the semantic content of a given text file. To do that the software groups words into grammatical and semantic categories. A set of rules, termed a *scenario*, is a combination of a list of words and semantic categories. Depending on the context, a word may have several very different meanings. For example, in English, "clear" might refer to a meteorologic attribute, a differentiation attribute, a visual attribute, a juridic term, etc. In the current study, most terms used by participants were audio attributes. However, they were not recognized as such by the default scenario. "Clair" and "clarté," the French for "clear" and "clarity," were grouped, though the first one is generally considered as a timbre attribute and the second one as a room acoustics or intelligibility attribute. The default *scenario* therefore had to be extensively modified to properly classify acoustic terms.

This analysis reduction produced a list of 162 attributes. The list included some attributes that could easily be manually grouped or dismissed. An explanation of this process allowed the beginning of the group sessions to function as a tutorial for participants on how to correctly perform their task.

C. Consensus reduction of the list of attributes

The third step of the attribute development process consisted in three group meetings, during which a subset of the participants would discuss the list of elicited attributes in order to form a concise and yet complete list of attributes that can be used to describe the perceived differences between the seven binaural versions of each extract described in Sec. IV, ensuring that they possessed the functional characteristics listed in Sec. I.

The first author of this article had the role of panel leader/moderator (ISO, 2006) for the meetings, ensuring every participant was allowed equal time to speak, and reminding participants of the desired characteristics of the attributes which were presented to the participants at the outset of each session. All meetings of the French native panel were run by the native French panel leader. Participants were asked to remove any attribute that did not comply and to group remaining attributes that were synonyms. At the end of each group meeting, a list of perceptual attributes, as well as definitions and endpoint of these attributes, was obtained.

During the meetings, participants were allowed to listen to the stimuli used in the individual elicitation step. Each group session lasted approximately 4 h.

Participant members of the first and second group sessions did not overlap. They were grouped so that each panel

was as varied as possible in terms of previous experience with binaural audio. The third group meeting session comprised participants from both panels in order to produce a consensus list of attributes. The third group meeting started from the list of attributes and definitions obtained at the end of the previous two separate group meetings. The lists of French attributes obtained at the end of each group session are shown in Table II.

During the third session, for each attribute, participants were also asked to find pairs of stimuli for which they could perceive a large difference according to that attribute. This could be used, if not to produce references for each attribute, for the training of these participants.

D. Validation of the list of attributes

1. Experimental procedure

A total of 12 attributes were produced at the end of the final CVP session. The hypothesis on which the list was based is that these attributes can explain the primary differences listeners could meaningfully perceive as a function of the HRTF set employed for binauralization. As such, the attributes should describe the perceptual dimensions of binaural rendering linked to HRTF variations.

The fourth step of the verbal elicitation study consists in validating this hypothesis through a listening test. During this listening test, participants were required to evaluate 7 binaural renderings according to preference followed by an evaluation according to the 12 attributes obtained from the third and final group in Sec. V C.

Ten participants took part in this test, 9 of whom took part in the previous steps of the study. Audio extracts #2 and #3, shortened to 8 s for temporal homogeneity, from Sec. IV C where used. Audio extract #1 was removed to maintain a reasonable experimental duration. Processing of these recordings was as described in Sec. IV. Playback level and calibration were identical to Sec. V A.

Participants used an interface comprising several evaluation *pages*. On each page, participants were presented a single attribute according to which they had to rate the stimuli, the definition, and endpoints to the scale of that attribute, and seven continuous scale sliders to rate the seven binauralized versions of the audio extract. An additional slider was present to inquire on the "ideal" value, but these results are not discussed in the current study. Participants could replay each stimuli as desired and switch between the seven binaural stimuli without interrupting playback, although they were advised to give their responses as spontaneously as possible.

Each evaluation page for a given attribute and stimuli was repeated 3 times, presented in random order. For each audio extract attribute evaluation repetition, the "preference" rating was carried out prior to the attributed ratings.

All attributes for a given audio extract were presented before progressing to the next audio extract. Attributes were presented in random order. The order in which the seven binaural stimuli (i.e., the HRTFs) were presented was randomized between each page.

In total, each participant completed (2 extracts × (preference + 12 attributes) × 3 repetitions) 78 rating pages containing the 7 HRTFs. The test was carried out in two sessions of approximately 1 h each, in order to limit fatigue. Participants were free to give either written or verbal feedback at the end of the session.

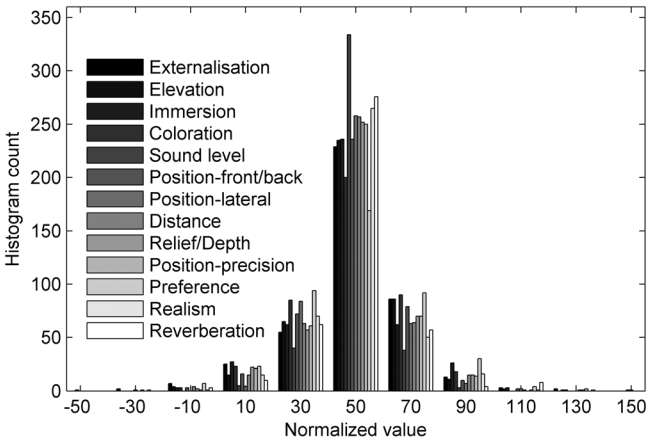

FIG. 1. Histogram of normalized response data for each attribute.

2. Analysis of data distribution and consistency

Data were normalized in order to reduce the variances in the use of the scale between participants and as a function of order across the test. Normalization consisted in setting the mean response value on each test page to the center of the scale; participants were free to use the scale as a relative scale, in which case from one repetition to the other, they may not have used the same portion of the scale. Figure 1 shows the distribution of the normalized subjective ratings for each attribute across all subjects.

A preliminary examination of normalized participant responses was carried out to verify that responses were both consistent over repetitions and that any observed variances were of a smaller magnitude than the differences between HRTFs for the different attributes. Participant consistency was quantified by taking the mean of the standard deviation (stdev) of the normalized ratings across the three repetitions for each test page (for each Extract across the seven HRTFs), providing one value for each attribute/participant.

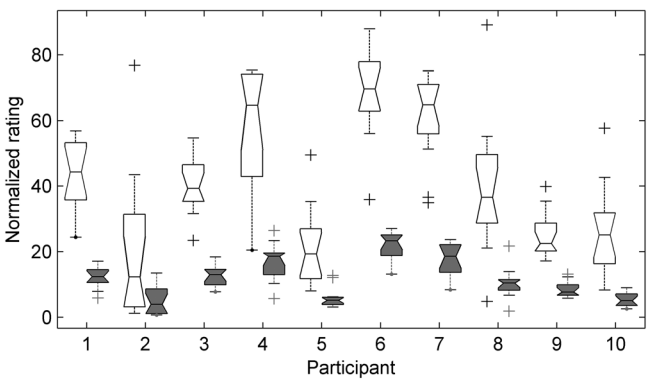

FIG. 2. Boxplot distribution comparison of *scale range* across the seven HRTFs {*mean*[*max*(HRTF rating)–*min*(*HRTF rating*)] across Extract and Repetition, per attribute; white boxes} versus *consistency* of responses across the three repetitions [mean(stdev) across Extract and HRTF, per attribute; gray boxes] for each participant.

The range of perceived differences was quantified by taking the mean difference between the *maximum* and *minimum* normalized rating of the seven HRTFs for each test page (for each Extract across the three repetitions), providing again one value for each attribute/participant. Figure 2 presents the distribution of these results, with the mean consistency metric across participants being a stdev of 11.8 ± 6.6, while the mean range size across participants was 40.6 ± 22.0. It can be clearly seen that the variance between repetitions is small or at worst comparable to the range of ratings used in judging the different HRTFs for the tested attributes. In addition, as seen in Fig. 1, the full range of rating values is even greater than this mean range scale metric. As such, we can have confidence that while the perceived differences may have been small, the range of perceptual responses was larger than repetition variances for the different participants. It can also be noted that the scale range of responses varies between participants, and that some participants may be more discerning with regards to the perceived differences between HRTFs.

3. ANOVA analysis of the results

Results of the experiment were analyzed in R, using FactoMineR and SensoMineR (Le and Worch, 2014), and in PanelCheck.[2] The data normalization employed should not have any influence on analysis of variance (ANOVA) results regarding significance of the HRTF independent variable nor on any interaction between independent variables that include the HRTF. However, such normalization does cause any other variable or interaction between variables to be non-significant.

An initial ANOVA was performed using both audio extracts. It was found that the audio extract had a significant effect on 7 of the 12 attributes as well as *Preference*. Due to these initial results and to content and positional variations, it was decided to analyze the two audio extracts separately. A repeated-measures (RM) interaction ANOVA was carried out in order to evaluate the validity of each attribute for rating perceptual differences due to HRTF selection. A single factor RM-ANOVA analyzed the influence of **HRTF** while a 2-way interaction RM-ANOVA analyzed the influence of the interaction between **HRTF** and **Participant** for each attribute.

[HRTF]: A significant effect of HRTF across Participants implies general agreement in attribute ratings. Lack of differences across participants implies a consensus judgment, indicating a non-individual factor. For example, if one HRTF was louder than all the others, this would be judged in a global sense, and not an individual sense as a valid HRTF attribute affecting individual spatial perception, and thus potentially a *poor* attribute.

[HRTF × Part.] A significant effect of HRTF and Participant interaction indicates that for that attribute the HRTFs were perceived differently, and individually different, by Participants. This is the indication of a *good* attribute.

Table III show the *p*-values of these ANOVAs. Additional ANOVA analysis conducted on the single variables Participant and Repetition or the interaction Participant × Repetition would not be meaningful given the applied data normalization.

These results show several points:

- For both stimuli, both the HRTF independent variable and the HRTF × Part. interaction had a significant effect on *Coloration*, *Externalization*, *Position-front/back*, and *Sound level*. This means that there was consensus on the judgment of these attributes, both across audio extracts and participants. However, no attribute showed a significant effect for the HRTF independent variable and not the HRTF × Part. interaction.
- The HRTF × Part. interaction had a significant effect on *Realism* for both audio extracts.
- The HRTF × Part. interaction had a significant effect on *Elevation*, *Immersion*, *Position-lateral*, and *Relief* for one of the audio extracts.
- For *Position-precision*, *Reverberation*, and *Distance*, neither the HRTF independent variable nor the HRTF × Part. interaction had a significant effect.
- There was a consensus on *Preference* for the audio extract #3 while this was not the case for audio extract #2.

As discussed above, attributes for which neither the HRTF variable nor the HRTF × Part. interaction had a

TABLE III. *p*-values of the ANOVAs conducted on the results obtained for audio extracts #3, 13-channel fiction, and #2, 5-channel electronic music recording. $\epsilon = 0.001$. Instances of $p < 0.05$ are **indicated**. Attributes considered as valid regarding HRTF variations are indicated (✓).

Attribute		Audio extract #3		Audio extract #2		Valid
(English translation)	(Original *French*)	HRTF	HRTF × Part.	HRTF	HRTF × Part.	
Coloration	*Modifications spectrales*	$<\epsilon$	$<\epsilon$	$<\epsilon$	$<\epsilon$	✓
Distance	*Profondeur*	0.094	0.051	0.185	0.122	
Elevation	*Elévation*	0.205	0.853	$<\epsilon$	**0.022**	✓
Externalization	*Externalisation*	**0.016**	**0.048**	$<\epsilon$	**0.016**	✓
Immersion	*Immersion*	0.302	0.581	0.216	$<\epsilon$	✓
Position-front/back	*Position avant arrière*	$<\epsilon$	$<\epsilon$	$<\epsilon$	**0.005**	✓
Position-lateral	*Position latérale*	0.798	0.269	$<\epsilon$	$<\epsilon$	✓
Position-precision	*Precision*	0.385	0.351	0.922	0.625	
Realism	*Réalisme*	0.934	**0.021**	**0.006**	**0.006**	✓
Relief/Depth	*Profondeur du champs sonore*	0.924	0.941	0.404	**0.003**	✓
Reverberation	*Réverbération*	0.449	0.353	0.362	0.793	
Sound level	*Niveau sonore*	$<\epsilon$	**0.001**	**0.007**	**0.005**	
Preference	*Préférence*	**0.001**	**0.006**	0.098	$<\epsilon$	

TABLE IV. List of the validated attributes, definitions, and endpoints (English translation).

Attribute	End points	Definition
Coloration	More high frequency content More low frequency content	Feeling of a sound richer in high/medium/low frequencies
Elevation	More toward the top More toward the bottom	*Self-explanatory*
Externalization	Inside the head Outside the head	Perception of sounds located outside the head
Immersion	Immersive Non-immersive	Feeling of being located in the middle of the audio scene
Position-front/back	Front Back	*Self-explanatory*
Position-lateral	More toward the left More toward the right	*Self-explanatory*
Realism	Realistic Non-realistic	Sounds seem to come from real sources located around you
Relief	Compact Spread out	Distance between the closest sound objects and the farthest

significant influence on the ratings for both stimuli are poor attributes to describe perceived differences between the HRTFs. *Distance*, *Position-precision*, and *Reverberation* were therefore dismissed.

The presence of the *Sound level* attribute underlines the difficulty in level equalizing complex binaural scenes. It was therefore decided to dismiss the *Sound level* attribute as well. Attributes which respond to the criteria of a valid attribute for describing the qualitative perceptual differences due to binauralization using different HRTF datasets, indicated in Table III, are the following: *Coloration*, *Elevation*, *Externalization*, *Immersion*, *Position-front/back*, *Position-lateral*, *Realism*, and *Relief/Depth*. The English translation of the validated attributes, their definitions, and end points are listed in Table IV.

VI. DISCUSSION

One of the difficulties of the initial group meetings was to force participants to list only the attributes that they perceived in the supplied examples. Being trained sound engineers, some of them had pre-conceived ideas about the attributes they should focus on when comparing audio recordings or previous experience with binaural renderings. Some of these attributes were dismissed early on in the group meetings. Others were kept at the end of the final group meeting. For example, *reverberation* was maintained despite the absence of any reverberation processing in the stimuli and the difficulty participants had in finding an example of a large difference of *reverberation*. Other examples of pre-conceived attributes that were abandoned because participants could eventually not find examples of such attributes when comparing HRTFs were *phasing* and *fatigue*. In general, participants commented that differences between HRTFs were quite subtle at times, likely due to the fact that they were taken from the same HRTF database, thereby exhibiting less variations than inter-database variations due to acquisition protocol (Andreopoulou *et al.*, 2015).

Most participants during the validation of the results mentioned that the *Position-lateral* attribute was difficult to rate, as was *Sound level*. Some participants reported that instead of evaluating *Position-lateral* for the whole auditory scene, they focused on single sources to accomplish the comparison. This means that *Position-lateral* may not be suitable for the evaluation of complex scenes, but may only be meaningful for single sources, which was not the purpose of this study. Alternatively, *Position-lateral* could imply more global concepts such as extent or width. It should be repeated however that a uniform ITD was imposed across all HRTFs.

Comparing the validated attributes with those obtained in previous studies concerning headphone spatial audio enhancement and virtual acoustics environments shows that for HRTF comparisons, timbral attributes are less prominent [only one timbral attribute, *coloration*, was validated, in opposition to seven such attributes in Lorho (2005b) and eight in SAQI (Lindau *et al.*, 2014)]. The remaining attributes are similar to some of the spatial and general attributes of SAQI: *immersion* is similar to SAQI's *presence*, with the other attributes found under the same name in both studies.

However, a large number of attributes found in SAQI were not validated in the current study. This is the case for the room attributes, temporal attributes, dynamics, and artifacts. This is to be expected as the current study focused solely on HRTF comparisons and no other processing or degradation effects. None of the attributes in these categories were found useable by the participants for the required task. *Reverberation* was initially elicited, but has been considered more as pre-conceived expectations than as an actual perceived difference between the presented HRTFs. *Distance* and *Position-precision* were also dismissed, showing again a difference between the participant's expectations and their perception using a well-controlled stimuli set.

It should be noted that while a varied set of stimuli were used in the study, they did not include any dynamic binaural rendering of moving sources. In addition, the binaural rendering were all rendered using full-phase HRTF convolution. As such, these attributes should be valid for other types of

binaural comparisons, but additional attributes might be necessary to describe potential artifacts that may arise from HRTF processing, like minimum-phase or IIR modeling, or from the use of moving sources.

It is also understood that the different validated attributes resulting from this study could be refined, or decomposed in more detail (e.g., low-freq-coloration, mid-freq-coloration, high-freq-coloration) if one was to try to quantify this attribute. The result of this study simply validates such variations (e.g., timbral) as perceptually significant in HRTF comparisons. Dissection and exploration of the nature of these attributes (perceptually or acoustically) should be the goal of future studies, which can be founded on the results of this study as providing the validated attribute list.

VII. CONCLUSION

Binaural audio is a technology on the rise, brought about by the prevalence of portable music and video players, smart phones, and the emergence of low cost virtual reality through these devices as well as gaming consoles. The quality of binaural audio has also improved over recent decades, thanks to improved signal processing power and real-time rendering techniques. These advances have allowed binaural audio, or virtual auditory simulations, to be used for complex studies such as spatial cognition (Afonso *et al.*, 2010), comprehension of virtual architectures (Picinali *et al.*, 2014), plasticity of the auditory system (Parseihian and Katz, 2012), or peripersonal space object localization (Parseihian *et al.*, 2014). Fundamental studies in spatial audition through the use of binaural simulations also helps improve our understanding of the mechanisms of the auditory system and the creation of reliable models (Baumgartner *et al.*, 2014).

For all such studies, it is necessary to select a suitable HRTF for the listener, whether it be an individually measured HRTF, a personalized HRTF adapted from an existing database, or an identified HRTF from a database. While previous studies have clearly shown the impact of HRTF choice on the localization precision of binaural audio, there has been little investigation concerning additional perceptual impacts. In many situations, precise localization position is not the predominant factor in the design or evaluation of a spatial sound scene. In such cases, the quality of the binaural audio rendering can be affected by a variety of other attributes. Unlike the assessment of loudspeaker rendering, the perceptual attributes affected by binaural rendering parameters, specifically the choice of HRTF, are heretofore unknown.

This study derived the perceptual attributes elicited by the comparison of HRTF sets, using complex scenes created by professional sound engineers. The list of attributes identified as describing the qualitative differences between HRTFs for binaural rendering are *Coloration*, *Elevation*, *Externalization*, *Immersion*, *Position-front/back*, *Position-lateral*, *Realism*, and *Relief/Depth*. These attributes go beyond the simple issue of localization, commonly used to evaluate HRTFs, but which does not cover the variety of perceptual aspects affected by non-individual HRTFs.

This list is aimed at audio experts and sound engineers, with precise definition usage, and may not be viable for naive participant evaluations. The validated attributes, and the protocol used for their validation, can be used for the evaluation of binaural stimuli. Carrying out a comparison of 7 HRTF sets, 1 audio extract, and 3 repetitions of each condition, following the same protocol as the one used for the validation of this experiment, took approximately 40 min, making the evaluation of HRTF processing or HRTF individualization methods relatively quick.

ACKNOWLEDGMENTS

This work was funded in part by the French project BiLi ("Binaural Listening," http://www.bili-project.org, FUI-AAP14) and The Danish Council for Technology and Innovation, under The Ministry of Science, Technology and Innovation.

[1]http://www.tropes.fr/.
[2]http://www.panelcheck.com/.

Afonso, A., Blum, A., Katz, B., Tarroux, P., Borst, G., and Denis, M. (**2010**). "Structural properties of spatial representations in blind people: Scanning images constructed from haptic exploration or from locomotion in a 3-D audio virtual environment," Memory Cognit. **38**, 591–604.

Andreopoulou, A., Begault, D. R., and Katz, B. F. (**2015**). "Inter-laboratory round robin HRTF measurement comparison," IEEE J. Sel. Topics Signal Process. **9**, 895–906.

Andreopoulou, A., and Katz, B. F. G. (**2016**). "Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees," J. Multimodal User Interfaces **10**(3), 259–271.

Baumgartner, R., Majdak, P., and Laback, B. (**2014**). "Modeling sound-source localization in sagittal planes for human listeners," J. Acoust. Soc. Am. **136**, 791–802.

Begault, D. R., Lee, A. S., Wenzel, E. M., and Anderson, M. R. (**2000**). "Direct comparison of the impact of head tracking, reverberation, and individualized Head-Related Transfer Functions on the spatial perception of a virtual speech source," in *Audio Engineering Society Convention*, San Jose, CA, paper number 5134.

Berg, J., and Rumsey, F. (**1999**). "Identification of perceived spatial attributes of recordings by Repertory Grid Technique and other methods," in *Audio Engineering Society Convention*, Münich, Germany, paper number 4924.

Blauert, J. (**1996**). *Spatial Hearing* (MIT Press, Cambridge, 1996), 508 pp.

Bondu, A., Busson, S., Lemaire, V., and Nicol, R. (**2006**). "Looking for a relevant similarity criterion for HRTF clustering: A comparative study," in *Audio Engineering Society Convention*, Paris, France, paper number 6629.

Cairncross, S., and Sjostrom, L. (**2004**). "Flavor profiles: A new approach to flavor problems," in *Descriptive Sensory Analysis in Practice*, edited by M. C. Gacula, Jr. (Food & Nutrition Press, Inc., Trumbull, CT), pp. 15–22.

Choisel, S., and Wickelmaier, F. (**2007**). "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," J. Acoust. Soc. Am **121**, 388–400.

Colomes, C., Le Bagousse, S., and Paquier, M. (**2010**). "Families of sound attributes for assessment of spatial audio," in *Audio Engineering Society Convention*, San Francisco, CA, paper number 8306.

Delarue, J., and Sieffermann, J.-M. (**2004**). "Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products," Food Qual. Preference **15**, 383–392.

Guastavino, C., Katz, B. F., Polack, J.-D., Levitin, D. J., and Dubois, D. (**2005**). "Ecological validity of soundscape reproduction," Acta Acust. Acust. **91**, 333–341.

Guillon, P., Guignard, T., and Nicol, R. (**2008**). "Head-Related Transfer Function customization by frequency scaling and rotation shift based on a new morphological matching method," in *Audio Engineering Society Convention*, San Francisco, CA, paper number 7550.

Huopaniemi, J., Zacharov, N., and Karjalainen, M. (**1999**). "Objective and subjective evaluation of head-related transfer function filter design," J. Audio Eng. Soc. **47**, 218–239.
Hur, Y., Lee, S.-P., Park, Y.-C., and Youn, D.-H. (**2008**). "Efficient individualization of HRTF using critical-band based spectral cues control," in *Audio Engineering Society Convention*, Amsterdam, paper number 7447.
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**, 317–333.
ISO (**1994**). ISO 11035.1994, "Sensory analysis—Identification and selection of descriptors for establishing a sensory profile by a multidimensional approach" (International Organization for Standardization, Geneva, Switzerland).
ISO (**2006**). ISO 13300-2.2006, "Sensory analysis—General guidance for the staff of a sensory evaluation laboratory—Part 2: Recruitment and training of panel leaders" (International Organization for Standardization, Geneva, Switzerland).
Katz, B. F. G., and Parseihian, G. (**2012**). "Perceptually based head-related transfer function database optimization," J. Acoust. Soc. Am **131**, EL99–EL105.
Kelly, G. A. (**1955**). *The Psychology of Personal Constructs* (Norton, New York), 208 pp.
Kirkeby, O., Lorho, G., Virolainen, J. K., Shenoy, R., and Patwardhan, P. P. (**2012**). "Multi-way analysis for audio processing," U.S. patent 20120207310 A1.
Kistler, D. J., and Wightman, F. L. (**1992**). "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," J. Acoust. Soc. Am **91**, 1637–1647.
Kraft, S., and Zolzer, U. (**2014**). "BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference* (Karlsruhe).
Lawless, H. T., and Heymann, H. (**2010**). *Sensory Evaluation of Food*, Food Science Text Series (Springer, New York), 471 pp.
Lawless, L. J. R., and Civille, G. V. (**2013**). "Developing lexicons: A review," J. Sensory Studies **28**, 270–281.
Le, S., and Worch, T. (**2014**). *Analyzing Sensory Data with R* (Chapman and Hall/CRC, Boca Raton, FL), 374 pp.
Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S. (**2014**). "A spatial audio quality inventory (SAQI)," Acta Acust. Acust. **100**, 984–994.
Lokki, T., Patynen, J., Kuusinen, A., and Tervo, S. (**2012**). "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," J. Acoust. Soc. Am **132**, 3148–3161.
Lorho, G. (**2005a**). "Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating," in *Audio Engineering Society Convention*, Barcelona, Spain, paper number 6514.
Lorho, G. (**2005b**). "Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction," in *Audio Engineering Society Convention*, New York, paper number 6629.
Lorho, G. (**2010**). "Perceived quality evaluation: An application to sound reproduction over headphones," Ph.D. thesis, Aalto University, Finland.
Lorho, G., Huopaniemi, J., Zacharov, N., and Isherwood, D. (**2000**). "Efficient HRTF synthesis using an interaural transfer function model," in *European Signal Processing Conference, EUSIPCO*, pp. 1–4.
Martens, W. L., and Zacharov, N. (**2000**). "Multidimensional perceptual unfolding of spatially processed speech I: Deriving stimulus space using INDSCAL," in *Audio Engineering Society Convention*.
Nicol, R. (**2010**). *Binaural Technology*, AES Monograph (Audio Engineering Society, New York), 77 pp.
Parseihian, G., Jouffrais, C., and Katz, B. F. (**2014**). "Reaching nearby sources: Comparison between real and virtual sound and visual targets," Front. Neurosci. **8**(269), 1–13.
Parseihian, G., and Katz, B. F. G. (**2012**). "Rapid head-related transfer function adaptation using a virtual auditory environment," J. Acoust. Soc. Am **131**, 2948–2957.
Pedersen, T. H., and Zacharov, N. (**2015**). "The development of a sound wheel for reproduced sound," in *Audio Engineering Society Convention*.
Picinali, L., Afonso, A., Denis, M., and Katz, B. F. (**2014**). "Exploration of architectural spaces by the blind using virtual auditory reality for the construction of spatial knowledge," Int. J. Human-Comput. Stud. **72**, 393–407.
Piggott, J. R. (**1991**). "Selection of terms for descriptive analysis," in *Sensory Science Theory and Application in Food* (Dekker, New York), pp. 339–351.
Pulkki, V., and Karjalainen, M. (**2001**). "Localization of amplitude-panned virtual sources I: Stereophonic panning," J. Audio Eng. Soc. **49**, 739–752.
Schönstein, D., and Katz, B. (**2010**). "HRTF selection for binaural synthesis from a database using morphological parameters," in *International Congress on Acoustics* (Sydney, Australia), pp. 1–6.
Seeber, B. U., Fastl, H., Akustik, A. T., and München, T. (**2003**). "Subjective selection of nonindividual head-related transfer functions," in *International Conference on Auditory Display* (Boston, MA), pp. 259–262.
Stone, H., Sidel, J., Oliver, S., Woolsey, A., and Singleton, R. C. (**1974**). "Sensory evaluation by quantitative descriptive analysis," Food Technology **28**, 22–43.
Suzuki, Y., Brungart, D., Kato, H., Iida, K., Cabrera, D., and Iwaya, Y., eds. (**2011**). *Principles and Applications of Spatial Hearing* (World Scientific Publishing Company, Hackensack, NJ), 520 pp.
Warusfel, O. (**2003**). "Listen HRTF database," http://recherche.ircam.fr/equipes/salles/listen/ (Last viewed October 24, 2016).
Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (**1993**). "Localization using nonindividualized head-related transfer functions," J. Acoust. Soc. Am **94**, 111–123.
Xie, B. (**2013**). *Head-Related Transfer Functions and Virtual Auditory Display*, 2nd ed. (J. Ross Publishing, Plantation, FL), 504 pp.
Zacharov, N., and Koivuniemi, K. (**2001**). "Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training," in *Audio Engineering Society Convention* (New York), paper number 5424.
Ziegelwanger, H., Majdak, P., and Kreuzer, W. (**2015**). "Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization," J. Acoust. Soc. Am **138**, 208–222.