Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

Free-Field Virtual Psychoacoustics and Hearing Impairment: Paper ICA2016-431

Speech perception by children in a real-time virtual acoustic environment with simulated hearing aids and room acoustics

Florian Pausch[a, b], Zhao Ellen Peng[a, b], Lukas Aspöck[a], Janina Fels[a, b]

[a]Institute of Technical Acoustics, RWTH Aachen University, Germany,
[b]Teaching and Research Area Medical Acoustics,
florian.pausch@akustik.rwth-aachen.de,
zhao.peng@akustik.rwth-aachen.de,
lukas.aspoeck@akustik.rwth-aachen.de,
janina.fels@akustik.rwth-aachen.de.

Abstract

Classrooms with demanding acoustic requirements for children fitted with hearing aids can be simulated effectively by real-time virtual acoustic environments. Physical accuracy is achieved using room impulse responses and a binaural reproduction system extended by research hearing aids. The generation of virtual sound sources is based on individualized head-related and hearing aid-related transfer functions. For the simulation of hearing aid algorithms, a software platform, which utilizes individual audiograms to generate fitting curves, processes the signals before being reproduced. In this study, a release from masking paradigm by Cameron and Dillon (2007) was adapted to assess speech intelligibility by children fitted with hearing aids in realistic reverberant environments. Speech reception thresholds are measured in the realistic acoustic scenes with room acoustics and compared to results from age-matched normal-hearing children.

Keywords: Speech reception threshold, release from masking, hearing loss, virtual acoustic environments, room acoustics

Speech perception by children in a real-time virtual acoustic environment with simulated hearing aids and room acoustics

Introduction

The number of children with hearing loss (HL) worldwide is estimated to be approximately 32 millions [1]. While hearing aids (HA) are used to apply a frequency-dependent amplification based on individual audiograms to partially restore normal hearing (NH), the clinical picture of HL is often accompanied by the diagnosis of spatial-processing disorder (SPD) [2]. This co-existing condition has been treated by Cameron and Dillon using an auditory training paradigm tested with NH children at the age between seven and eleven years old [3, 4]. The paradigm has been originally developed to train the usage of spatial cues in a "cocktail party" situation, where children diagnosed with SPD have to focus selectively on the target speaker while ignoring distractor speakers located at different spatial locations. The pre- and post-tests of their experiment showed a significant improvement of speech comprehension performance tested in an anechoic environment.

In the current project, the original paradigm is adopted in German and conducted while being immersed in a loudspeaker-based virtual acoustic environment (VAE) with simulated room acoustics. To grant access to children fitted with bilateral HA, the reproduction setup is extended by research hearing aids (RHA) which allows training under realistic reverberant conditions [5]. Speech training material is presented by the use of binaural technology and individualized head-related transfer function (HRTF) and hearing aid-related transfer function (HARTF) data sets.

This paper focuses on the pretest experiment where two groups of children, a group with HL and an age-matched NH control group, are asked to repeat sentences spoken by the target talker located at 0° azimuth in the horizontal plane while two distractor talkers are simultaneously telling irrelevant stories. Each child undergoes testing in a total of eight conditions, consisting of manipulations in spatial cues (target-distractor collocated vs. spatially separated at $\pm 90^\circ$), talker pitch cues (target-distractor sharing the same vs. different voice), and room acoustics ($0.4\,\mathrm{s}$ vs. $1.2\,\mathrm{s}$ reverberation time (RT)). The speech reception threshold (SRT) is measured adaptively at $50\,\%$ intelligibility under each condition. The spatial advantage, talker advantage, and total advantage is measured as release from masking (RM) and compared between reverberant conditions. Results from the pretest assessment prior to auditory training are presented and discussed.

Methodology

2.1 System description

As basis for a realistic listening scenario, a rectangular room with a volume of $270\,\mathrm{m}^3$ was simulated in Room Acoustics for Virtual Environments (RAVEN), a real-time framework for the auralization of interactive virtual environments [6]. The simulated room was equipped with furniture and different surface materials, such as carpet, plaster, and acoustic absorber, to match RT conditions in typical classrooms, as shown in Figure 1(a). Accordingly, the clarity values C_{50} are displayed in Figure 1(b).

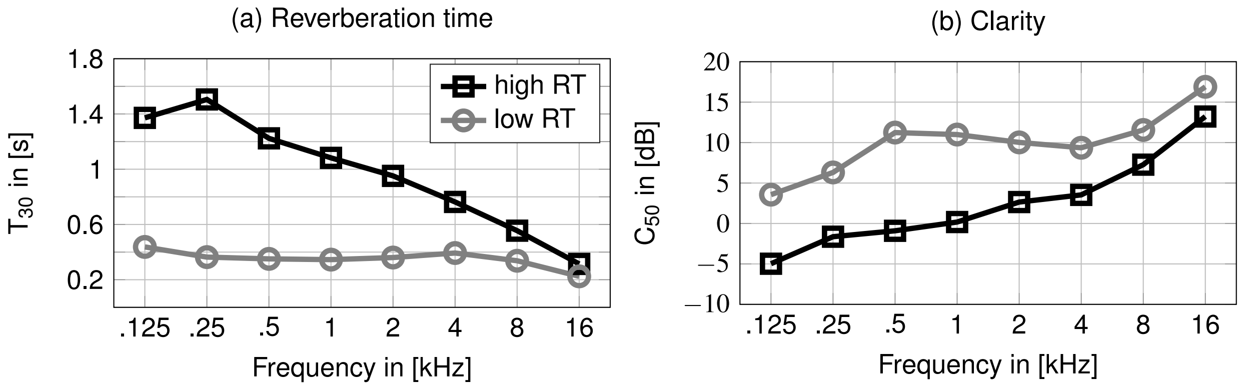


Figure 1: **(a) Reverberation time (RT) T_{30} and (b) clarity C_{50} of the simulated rectangular room. Simulated RT scenarios result in a mid-frequency RT of $0.4\,\mathrm{s}$ and $1.2\,\mathrm{s}$, calculated as arithmetic mean of 0.5 and $1\,\mathrm{kHz}$ octave bands. The mid-frequency clarity is $11\,\mathrm{dB}$ and $0\,\mathrm{dB}$, respectively.**

For the simulation of virtual sound sources, a set of generic HRTF [7] and HARTF with a spatial resolution of $1^\circ \times 1^\circ$ in azimuth and zenith angles, covering the whole sphere and measured at a radius of $1.86\,\mathrm{m}$, are used. For practical reasons (age of subjects, insurance issues), no individual data sets were measured but rather individualized ones are used. During the individualization routine, both interaural time differences (ITD) and spectral characteristics were modified in the HRTF data sets [8, 9], whereas only the ITD cues were modified in the HARTF, as these data do not contain pinna-related characteristics. A database based on the 15^{th}, 50^{th} and 85^{th} percentiles of typical children's head dimensions [10], i.e. head height, width, and depth [11], was generated. Subsequently, the best-fit data set was selected for each child depending on individual anthropometric measurements. For the simulation of room acoustics, these data sets are additionally utilized to generate binaural impulse responses (BRIR) in RAVEN. User movements are captured by an optical tracking system, running at a frame rate of $120\,\mathrm{Hz}$, which feeds back the child's current head orientation and position and accordingly updates the virtual scene in real-time.

The playback of the simulated binaural signals is realized over four loudspeakers with free-field equalization after filtering the input signals by a crosstalk cancellation (CTC) filter network [13], installed in an acoustically optimized hearing booth. For the group of children with HL, additional playback is realized over a pair of RHA with open fitting. The RHA input signals are processed on a Master Hearing Aid (MHA) platform [14] which utilizes individual audiograms and calculates frequency-dependent gains based on Cambridge formula [15]. Using this strategy, children with HL can be provided with uniform signals thus eliminating the potential bias from a variety of different signal processing strategies in their own commercial HA. To enable the use of residual hearing for this group of children, the loudspeakers are still in use to provide a more authentic listening situation. The fact that HA introduce an additional delay relative to the perceived external sound field was included in the system design by setting the relative delay between RHA and loudspeaker-based reproduction to $5\,\mathrm{ms}$ [16]. End-to-end latency of the combined binaural reproduction system was measured to be below $30\,\mathrm{ms}$, which enables a very reactive and natural listening situation.

2.2 Experiment

In this study, the paradigm by Cameron and Dillon [3] to assess SPD was adopted to measure the speech perception of children when using speech material in German which is presented in reverberant conditions. In the original paradigm, a total of four conditions are tested, which consist of two spatial cues (target-distractor collocated vs. spatially separated at $\pm 90°$) $\times$ two talker pitch cues (target-distractor sharing the same vs. different voices) as illustrated in Figure 2. The original paradigm is additionally tested for each child in the $0.4\,\mathrm{s}$ vs. $1.2\,\mathrm{s}$ RT conditions to assess children's ability to utilize spatial and pitch cues in realistic room acoustic scenarios, which results in a total of eight test conditions. During the experiment, while being immersed in the VAE, the children had to orally repeat the sentence spoken by the target talker, while two distractors are simultaneously telling irrelevant stories.

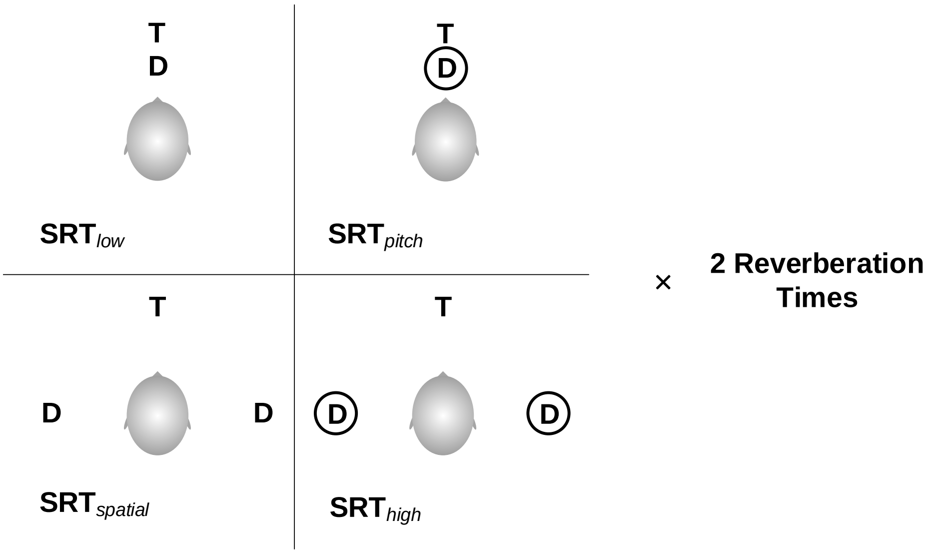


Figure 2: **Test conditions of the paradigm by Cameron and Dillon [3] extended by two reverberation time scenarios to assess spatial-processing disorder. The target talker (T) is always located at $0°$ azimuth in the horizontal plane. The distractors (D) are either collocated with the target talker (upper row) or spatially separated from the target talker (lower row) at $\pm 90°$ azimuth from the subject in the horizontal plane. The distractors either share the same voice as the target talker (left column) or different voices from two females other than the target talker (right column, cf. encircled distractors). In each condition, the speech reception threshold (SRT) is measured adaptively.**

2.3 Stimuli

A subset of the HSM sentences [17] containing only four- and five-word sentences was selected as the target stimuli and recorded with a native German-speaking female. Eight unfamiliar Grimm stories in German were recorded with the target talker, as well as two other females who are also native German-speakers. A summary of speaker characteristics is given in Table 1.

Table 1: **Speech characteristics of the talkers in the experiment.**

Speaker	Fundamental frequency [Hz]	Speech rate [syllables/s]
Target	213	3.4
Distractor 1	191	3.2
Distractor 2	198	3.3

All speech materials were recorded anechoically at a resolution of 16 bit and 44.1 kHz sampling rate with a Zoom H6 recording device and a Neumann TLM 170 condenser microphone. The speakers were instructed to speak in conversational style. All recordings were normalized per EBU-R128 standard [18], which provides normalization based on equal loudness instead of the conventional root-mean-square method of equal energy.

2.4 Procedure

During the experiment, the distractors are constantly speaking at 55 dB SPL (re 20 μPa). A one-up-one-down adaptive procedure is used to measure the SRT of 50% speech intelligibility by changing the target talker level, starting from 70 dB SPL. The step size is set at an initial 4 dB descending until the first reversal, at which point the signal-to-noise ratio (SNR) reduces below the child's SRT, and changed to steps of 2 dB thereafter. Each test block terminates once seven reversals are reached. The SRT is then calculated by averaging the SNR of the last four reversals. If the child is unable to accurately repeat at least half of the words in the initial sentence, the ascending step size is set at 2 dB. To ensure safe playback level, the test block terminates if the child fails to correctly identify at least half of the words in any of the first five sentences. Moreover, the target level never exceeds 80 dB SPL.

A nested Latin square was utilized to counterbalance the eight test conditions. The HSM sentences are randomly assigned to match each test condition, where the first 23 sentences in each test block are always 5-word sentences and 4-word sentences thereafter. The distractors' location (left vs. right) and story assignments are also counterbalanced. A 3-second leading of story playback is presented prior to the first target sentence in each test block. A leading beep (1 kHz sine) with a duration of 200 ms is provided 500 ms prior to each target sentence playback.

2.5 Subjects

All recruitment was performed after obtaining ethical approval by the Medical Ethics Committee at the RWTH Aachen University (EK 188/15). Two groups consisting of 12 children with HL, between seven and 13 years old, and 13 age-matched NH controls were recruited from the

David-Hirsch-Schule in Aachen through teacher-parent communications. Children with HL are all fitted with bilateral HA and have severe to profound HL in one or both ears. In the case of profoundly deaf children, who were unable to receive any stimulation from the MHA, they were asked to utilize their own HA. The HRTF data sets in the loudspeaker-based playback were replaced by HARTF data sets to account for the position offset from ear canal entrance to HA microphone position (approximately $0.025\,\mathrm{m}$) and the fundamental cue differences between these two data sets. An examination of the hearing threshold via pure tone audiometry for each child in the HL group was performed by an audiologist within three weeks of the experiment date. The measured audiograms were used to calculate the fitting curve in the MHA. Prior to experiment, all children were given a group introduction about the experimental set-up and procedure. Each child received a representation allowance for participating in the experiment.

Results

3.1 Speech reception threshold

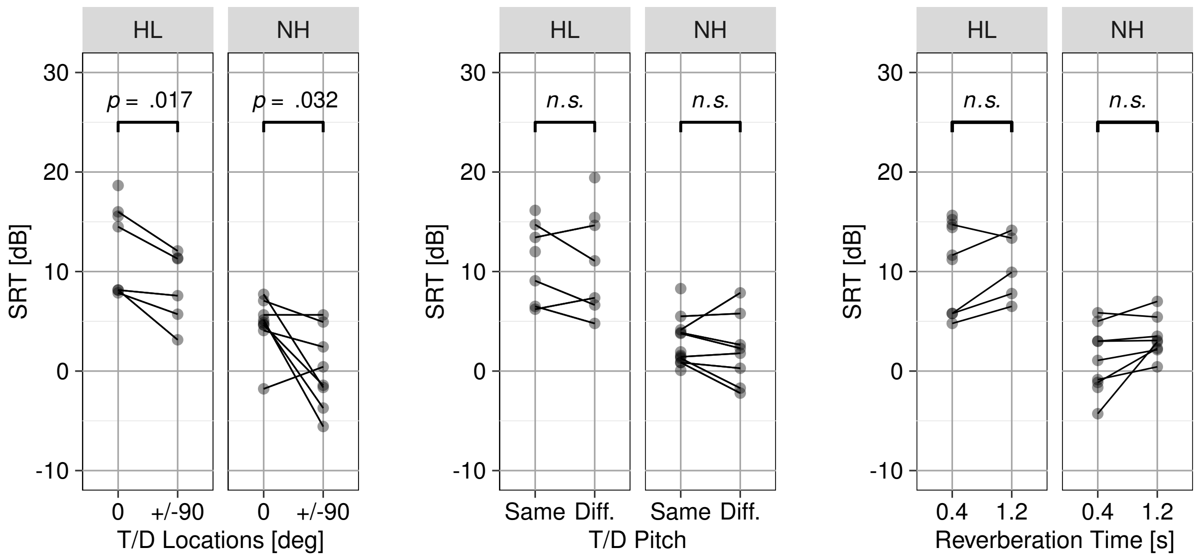
Dependent paired t-tests at a significance level of $\alpha = .05$ were conducted to compare the SRT for spatial cue, pitch cue, and RT conditions, as plotted in Figures 3(a)–(c). To prevent effects due to inflated degrees of freedom (pseudoreplication), SRT values are averaged per subject for different conditions. Associated subject responses in different conditions are linked via solid lines. As some subjects did not finish certain conditions, missing data points are present, which, however, are ignored during assessment. Effect sizes, calculated by $r = \sqrt{t^2/(t^2+df)}$, are given additionally (small effect, $r = .1$; medium effect, $r = .3$; large effect, $r = .5$).
There was a significant decrease in SRT for both groups when spatial cue was introduced, $t(4) = 3.91$, $p = .017$, $r = .89$ (HL), $t(7) = 2.68$, $p = .032$, $r = .71$ (NH). No significant difference in SRT was found for pitch cue, $t(4) = 1.11$, $p = .328$, $r = .48$ (HL), $t(7) = .81$, $p = .445$, $r = .63$ (NH), and for different RT scenarios, $t(4) = -2.01$, $p = .114$, $r = .71$ (HL), $t(7) = -2.18$, $p = .066$, $r = .64$ (NH).

3.2 Release from masking

As a measure of cue advantage, RM values are calculated by

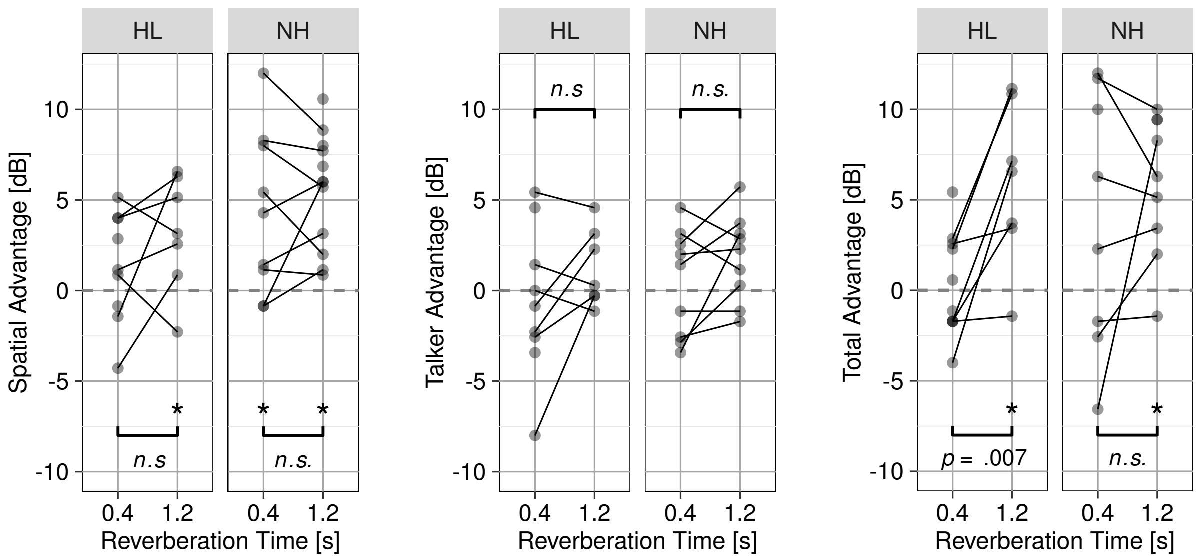
$$\mathrm{RM} = \mathrm{SRT}_{low} - \mathrm{SRT}_{cue} \qquad \text{[dB]}, \tag{1}$$

with indices as defined in Figure 2. The RM values for spatial cue (spatial advantage, spatial release from masking, SRM), pitch cue (talker advantage) and both spatial and pitch cue (total advantage) are assessed by dependent paired t-tests at $\alpha = .05$, checking for intra-group differences given two RT conditions (Δ_1) and for differences from zero mean (Δ_2), respectively, as plotted in Figures 4(a)–(c). Significant differences in Δ_1 are written in bold-face type, whereas significant differences in Δ_2 are indicated by asterisks. To facilitate reading, test statistics and effect sizes for all comparisons are summarized in Table 2.



(a) **SRT Spatial Cue** (b) **SRT Pitch Cue** (c) **SRT Reverb Cue**

Figure 3: **Speech reception thresholds (SRT) at 50 % intelligibility under various conditions.**



(a) **Spatial Advantage** (b) **Talker Advantage** (c) **Total Advantage**

Figure 4: **Release from masking for different cues under different reverberation time conditions. Asterisks denote significant differences from zero mean (dashed lines) at a significance level of α = .05.**

For Δ_1, a significant increase in total advantage is present for HL group. No significant change in spatial or talker advantage was found for both groups, and no significant change in total advantage for NH group was observed.
Comparing the mean with zero mean (Δ_2) resulted in a significant spatial advantage in high RT for HL group, and a significant spatial advantage in both RT scenarios for NH group. A total advantage was present for both groups in high RT. Neither spatial advantage in low RT for HL group, nor talker advantage for both groups in both RT scenarios was observed. Additionally, no significant effect was found in total advantage for both groups in low RT.

Table 2: **Summary of test statistics related to relase from masking (RM).**

		Δ_1				Δ_2 (low RT)				Δ_2 (high RT)			
	Advantage	*t*	*df*	*p*	*r*	*t*	*df*	*p*	*r*	*t*	*df*	*p*	*r*
HL	Spatial	-1.26	6	.255	.46	1.24	8	.25	.40	**2.65**	**6**	**.038**	**.73**
	Talker	-1.7	6	.141	.57	$-.46$	8	.657	.16	1.53	6	.178	.53
	Total	**-3.96**	**6**	**.007**	**.85**	.38	9	.715	.12	**3.52**	**6**	**.013**	**.82**
NH	Spatial	$-.44$	11	.665	.57	**6.14**	**11**	**$<.001$**	**.88**	**2.65**	**6**	**.038**	**.73**
	Talker	-1.55	8	.159	.48	.42	8	.686	.48	2.27	8	.053	.63
	Total	$-.19$	7	.858	.07	2.02	8	.078	.58	**4.47**	**8**	**.002**	**.84**

Values in bold-face type denote significant differences at $\alpha = .05$.

3.3 Discussion

The performance of the two tested groups of children is summarized in Table 3. In general, mean RM values in HL group tend to be lower than those in NH group. Results suggest that the tested children were able to benefit from spatial advantage in low RT (NH) and high RT (both groups), but not from talker advantage in both RT scenarios (both groups). The latter result may be explained by the fact that the fundamental frequencies of the distractors insufficiently differ from the target talker's pitch, cf. Table 1. For total advantage, both groups benefited under high RT and, for the HL group, the increase of total advantage was found to be significant in high RT compared to low RT.

Table 3: **Summary of cue benefits (RM) for the two tested groups, given as geometric mean and standard deviation (SD).**

Advantage	**HL**		**NH**	
	Mean [dB] (low / high RT)	**SD [dB] (low / high RT)**	**Mean [dB] (low / high RT)**	**SD [dB] (low / high RT)**
Spatial	1.3 / 3.2*	3.1 / 3.2	4.7* / 5.6*	7.2 / 3.1
Talker	-0.6 / 1.2	3.0 / 2.4	0.4 / 1.8	3.0 / 2.4
Total	**0.3 / 5.9***	2.9 / 4.4	5.5 / 5.8*	8.1 / 3.9

Values in bold-face type and asterisks denote significant difference after intra-group comparison for different RT scenarios and from zero mean, respectively, at $\alpha = .05$.

In general, large individual differences in both groups were observed in responding to the negative effect of increased RT. Further experiments with a higher number of different RT scenarios will be necessary to investigate the development of RM over RT.

Conclusions

A release from masking paradigm by Cameron and Dillon [3] has been adapted to assess speech perception performance by children fitted with bilateral hearing aids and age-matched normal-hearing controls in realistic room acoustic conditions simulated in a real-time virtual acoustic environment. Individual speech reception thresholds and according release from masking values have been measured in eight conditions in low and high reverberation. Measured data showed a significant decrease in speech reception thresholds in both groups as a result of the spatial separation between target and distractors. Release from masking was observed for the normal-hearing controls if target talker and distractors were spatially separated in both reverberation time scenarios, whereas the group of children with hearing loss was able to benefit from spatial advantage in high reverberation only. No talker advantage was found for both groups in both reverberation time scenarios. When combining spatial and talker advantage, a release from masking was present in both groups in high reverberation only. The increase in total advantage was significantly different for the group of children with hearing loss comparing the two reverberation time scenarios. The individual children's release from masking values will be used as baseline measures for comparisons after they complete the auditory training in three-month's time.

Acknowledgements

This work received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. ITN FP7-607139: Improving Children's Auditory Rehabilitation (iCARE). The authors would also like to acknowledge Ramona Bomhardt for providing support in the individualization procedure, as well as student research assistants Arndt Brandl, Karin Loh and Alokeparna Ray for assisting paradigm design and data collection.

References

[1] World Health Organisation 2012. URL: http://www.who.int/pbd/deafness/estimates, accessed on 2016-05-21.

[2] Glyde, H.; et al. The effects of hearing impairment and aging on spatial processing. Ear Hear, Vol 34 (1), 2012, pp. 15–28.

[3] Cameron S.; Dillon H. Development of the Listening in Spatialized Noise-Sentences Test (LISN-S). Ear and hearing Vol 28 (2), 2007, pp. 196–211.

[4] Cameron S.; Glyde H.; Dillon H. Eficacy of the LiSN & Learn auditory training software: randomized blinded controlled study. Audiology research Vol 2 (1), 2012, p. e15.

[5] Aspöck L.; Pausch F.; Vorländer M.; Fels J. Dynamic real-time auralization for experiments on the perception of hearing impaired subjects, DAGA 2015 Conference on Acoustics, Nuremberg, Germany, March 16-19, 2015.

[6] Pelzer S.; et. al. Interactive Real-Time Simulation and Auralization for Modifiable Rooms, Building Acoustics, Vol 21 (1), 2014, pp. 65–74.

[7] Schmitz A. Ein neues digitales Kunstkopfmeßsystem, Acta Acustica united with Acustica, Vol 81 (4), 1995, pp. 416–420.

[8] Middlebrooks J. C. Individual differences in external-ear transfer functions reduced by scaling in frequency. The Journal of the Acoustical Society of America, Vol 106 (3) Pt 1, 1999, pp. 1480–1492.

[9] Bomhardt R. Analytical Interaural Time Difference Model for the Individualization of Arbitrary Head-Related Impulse Responses. 137th Audio Engineering Society Convention, Los Angeles, CA, October 9-12, 2015, pp. 1–7 2015.

[10] Fels J. From children to adults: How binaural cues and ear canal impedances grow. Logos Verlag, Berlin (Germany), ISBN: 978-3-8325-1855-4, 2008.

[11] Algazi V.; et. al. The CIPIC HRTF database, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, NY, October 21–24, 2001, pp. 99–102.

[12] Lentz T.; et. al. Virtual reality system with integrated sound field simulation and reproduction, Eurasip Journal on Advances in Signal Processing, Vol 2007, pp. 1–19.

[13] Masiero B.; Vorländer M. A Framework for the Calculation of Dynamic Crosstalk Cancellation Filters, IEEE ACM transactions on audio, speech, and language processing, Vol 22 (9), 2014, pp. 1345–1354.

[14] Grimm, G.; et. al. The master hearing aid-a PC-based platform for algorithm development and evaluation, Acta Acustica united with Acustica, Vol 92 (4), 2006, pp. 618–628.

[15] Moore B. C.; Glasberg B. R. Use of a loudness model for hearing-aid fitting. I. Linear hearing aids, British journal of audiology, Vol 32 (5), 1998, pp. 317–335.

[16] Stone M.; et. al. Tolerable hearing aid delays. V. Estimation of limits for open canal fittings, Ear and hearing, Vol 29 (4), 2008, pp. 601–17.

[17] Hochmair-Desoyer, I.; et. al. The HSM sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users, The American Journal of Otology, Vol 18 (6 Suppl), 1997, p. 83.

[18] EBU-R128 EBU Recommendation: Loudness normalisation and permitted maximum level of audio signals, Geneva, 2014.