Towards determining thresholds for room divergence: A pilot study on detection thresholds

Florian Klein*
Institute for Media Technology
TU Ilmenau
Ilmenau, Germany
florian.klein@tu-ilmenau.de

Sebastia V. Amengual Gari
Facebook Reality Labs Research
Facebook Inc.
Redmond, USA
samengual@fb.com

Johannes M. Arend*
Institute of Communications Engineering
TH Köln
Cologne, Germany
johannes.arend@th-koeln.de

Philip W. Robinson
Facebook Reality Labs Research
Facebook Inc.
Redmond, USA
philrob22@fb.com

***Abstract*—In binaural rendering, the room divergence effect refers to the decrease in perceived externalization due to a mismatch between the room acoustics of the virtual sounds and those of the listening space. However, it is currently unknown which specific acoustic differences cause this effect. In this work, we present a pilot study to determine detection thresholds between sound sources recorded under different acoustic conditions in a variable acoustics room. These results are intended to predict situations where divergence effects can be expected. The participants had to perform a triangle test where they could listen to three sound sources placed at different positions in the room. The test design was motivated by the fact that sound sources are not placed at the same position in real acoustic scenes. One sound source was recorded under different acoustic conditions than the other two, and the task for the participant was to detect the differing source. The test was conducted in the measured room using 3 DoF binaural reproduction and using a virtual reality (VR) headset to display a visual 360 capture of the room enabling the subjects to see the positions of the sources in the room. Detection rates are signal-dependent and increase with differences in reverberation time (RT). For the most critical signal in the test (castanets), an RT difference of 8% was detectable, while the difference was 15% across all conditions. Furthermore, we discuss the influence of sound source distance and absorption configuration (symmetric or asymmetric) on detection thresholds.**

***Index Terms*—Reveberation Time, Detection Threshold, Room Divergence, Binaural Synthesis**

I. Introduction

The goal of acoustic augmented reality is to add virtual sound sources into a real acoustic environment or to alter real sound sources, for example by attenuating them. For both scenarios knowledge about the real acoustics need to be incorporated into the sound rendering.

In comparison to anechoic conditions reflections from walls increase the apparent sound level, affect the apparent source width and apparent source position. They can also cause deviations in the perceived timbre and they increase the reverberance [1]. A mismatch between the acoustic room properties of the virtual object and the listening room [2], [3], can lead to in-head localization of audio images. Thus, to enable applications that require externalized sound images like audio presence or telepresence, it is important to minimize the acoustical divergence between the real room and the presented sound images. This leads to the question of how accurately room acoustic parameters need to be estimated in order to achieve a plausible acoustical illusion.

Reverberation time is believed to be an important parameter in this regard. In current standards [4], [5], a deviation of 5% of the reverberation time is considered the just noticeable difference (JND). However, as Blevins [6] points out, there are several studies which indicate a JND of 10% or even up to 25%. Seraphim [5], Blevins [6] and Frissen et al. [7] state that these differences are independent of sound stimuli.

When comparing rooms, reverberation time is obviously only one of many acoustical parameters. Their individual contribution in the process of recognizing a room or differentiating rooms is not well understood. Whether a room is perceived as plausible does not only depend on the signals which reach the ears at given time. Our perception is strongly influenced by the comparisons listeners can conduct. Both external and internal references can serve for comparison. Internal references are built upon previous experience and they give us a sense what kind of acoustics we can expect from the visual appearance of certain rooms. Changes of room acoustics which are not related to listener movement or the visual appearance of the room or sound sources might be unexpected and could be perceived as not plausible [8].

Especially in AR applications, listeners can compare to external references such as other sound sources or self-elicited sounds. Possibly, this could reveal differences between real and virtual sound sources. However, the detectability of differences is not necessarily a quality issue, because the listener would need to know which source is the real one and which is the virtual one. Depending on the availability of references

* Florian Klein and Johannes M. Arend performed the work while at Facebook Reality Labs Research.

(such as cues from other modalities), virtual sound sources might be easy to detect. When focusing on acoustic aspects, another fact comes into play: In real acoustic scenes, sound sources unlikely have the same audio content and they can not be placed a the exact same position in the room. Position dependent acoustics will therefore conceal potential differences in the rendering of sound sources.

The following study aims to investigate the perceptual tolerances with regard to reverberation differences. Discrimination thresholds within a room with variable acoustics were investigated. Different acoustic conditions are recorded and compared to determine which conditions can be discriminated from each other. A second study published in the proceedings of this conference focuses on the investigation of acceptable reverberation time mismatch, by comparing the externalization judgments of several versions of a re-synthesized room [9].

II. Experiment

A. Test System

To conduct this experiment a framework as shown in Figure 1 was used. The scenes for different room conditions and loudspeaker arrangements were measured acoustically using a purpose-built microphone array in order to create Binaural Room Impulse Respones (BRIRs) on the basis of the Spatial Decomposition Method [10]. KEMAR HRTFs (Head Related Transfer Functions) were used for BRIR synthesis. Visuals were created by using a 3D 360 camera at the listener position. These were presented along with the listening test interface using Unity and an Oculus Quest. By using OSC, head tracking data as well as all necessary audio controls were transmitted from Unity to the real-time BRIR convolution engine pyBinSim [11]. For the reproduction, Beyerdynamic DT990 Pro headphones were used with equalization filters based on KEMAR dummy head recordings.

This system allowed us to combine measured acoustics and visuals from real spaces and to create a head pose dynamic binaural synthesis. By embedding the listening interface within this framework, intuitive interaction and pointing method could be employed.

B. Test design and stimuli

Acoustic measurements with eight different panel arrangements in a variable acoustics room were conducted. Different panel arrangements resulted in reverberation times ranging from $0.39\,s$ to $0.62\,s$. RT30 was calculated in the frequency range from 200 Hz to 8 kHz and the values of all available sound source positions (see figure 2) were averaged. The measurements can be divided into uniform and non-uniform room conditions. For the uniform conditions, absorptive and reflective panels are distributed uniformly. These conditions range from all walls fully reflective to all walls fully absorptive. Figure 2 show the reverberation time for these conditions in gray. In between there are three intermediate conditions where 25%, 50% and 75% of the surface is absorptive. In the non-uniform conditions, whole walls were changed from absorptive to reflective (only right wall absorptive; right and back wall absorptive; right, back and left wall absorptive). These conditions are shown as colored lines. We can observe that the reverberation time is changing depending on the panel arrangements in the room. However, linearly increasing the absorptive surfaces is not linearly changing the reverberation time. Position depended differences are usually below 10% and thus in the magnitude of known JND values [12], [6], [4]. Because the basic geometry of the room was unchanged by the panel rearrangements, we assume that room modes and the pattern of the first reflection are almost constant along our test conditions. The difference in reverberation time is supposed to be the main difference between the conditions.

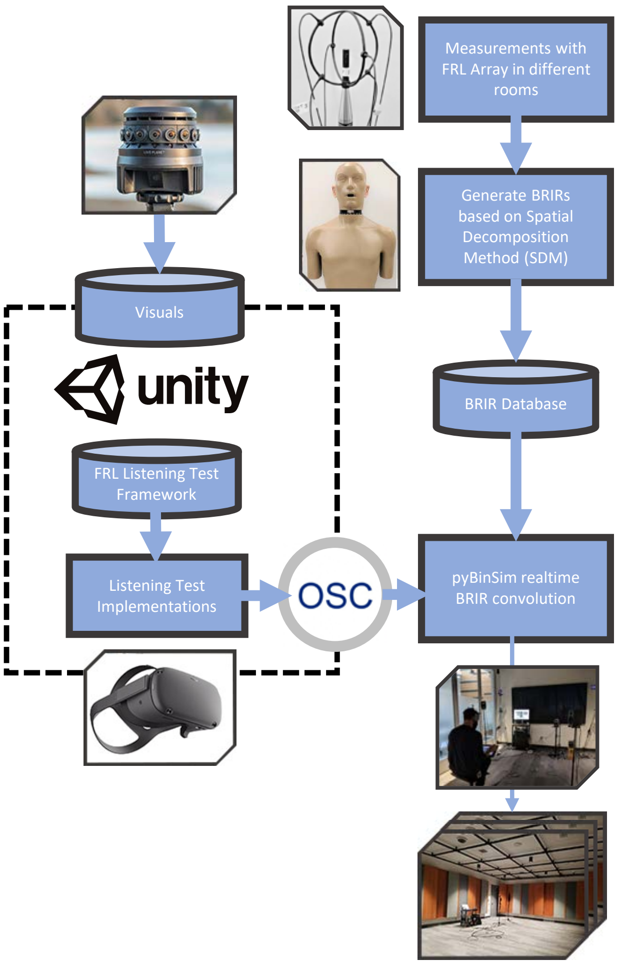
Measurements with FRL Array in different rooms
Generate BRIRs based on Spatial Decomposition Method (SDM)
BRIR Database
pyBinSim realtime BRIR convolution
Visuals
unity
FRL Listening Test Framework
Listening Test Implementations
OSC

Fig. 1. Creation of the audio-visual scenes for the listening test.

Loudspeakers were placed at 150 cm, 200 cm and 300 cm around the listener with a spacing of 30°. Corresponding visuals were captured to show the loudspeaker positions in accordance to the audio. To measure if the differences between the panel arrangements are perceptible a triangle test is performed. In a test trial the listener is confronted with three stimuli where one is different from the other two. For example: One stimuli is created based on a very reverberant

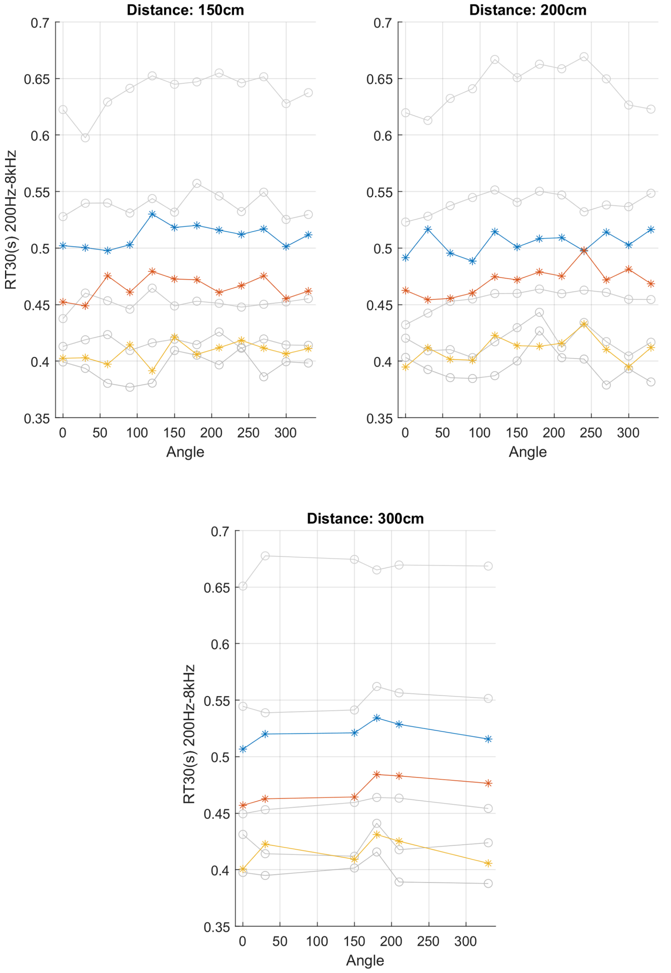

Fig. 2. Reverberation time for the different loudspeaker positions. Gray lines correspond to to 0%, 25%, 50%, 75% and 100% of uniformly absorptive surface. Colored lines correspond to one, two or three walls with absorptive material.

panel arrangement and the other two by a less reverberant arrangement.

Each stimuli is presented by a virtual loudspeaker. Figure 3 shows the arrangements of the virtual loudspeaker triplet. Corresponding visuals are provided through the head mounted display. Therefore, the listener could always see three loudspeakers at the same distance. The loudspeakers within a triple are not placed at the same position because real sound sources generally not overlap in position but rather co-exist at different positions. Different triplet positions are presented because we can not assume a fully diffuse sound field. Averaging over various positions helps to increase the generalizability of our results. The task of the participants is to select the stimulus which is different. It is important to note that the listeners were not instructed to listen to specific acoustic features. As audio signals, three different stimuli were used: Castanets, male speech and a speech mix from three speakers. Every time the listeners had to rate a triplet, each of the sound sources played the sound sample in a serial order. The mixed speaker signal was played back in such a way that each loudspeaker played back a different speaker. One presentation of a triplet took about $6\,s$ and each speaker played for $2\,s$. After two presentations of the stimuli for each speaker the playback stopped and the listener had to give a forced choice answer to the question “Which sound source is different?”. Overall, there were four triplet positions and three different audio samples. Table I summarizes all conditions in the test.

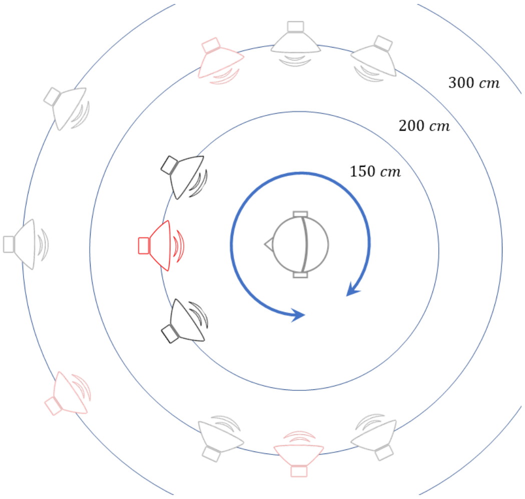

Fig. 3. All four triplet positions included in the test. Gray speakers are measured in the same acoustic room condition and the red speaker corresponds to a measurement from a different room condition. The listener is able to turn the head in order to face the current triplet.

TABLE I
OVERVIEW OVER ALL TEST CONDITIONS INCLUDED IN THE TRIPLET TEST. WITHIN A TRIPLET, AUDIO CONTENT AND DISTANCE WAS KEPT CONSTANT.

Room conditions	**Triplet positions**	**Audio samples**
Fully reflective	300 cm frontal	Castanets
25% uniformly absorptive	200 cm left	Female speech
50% uniformly abs.	200 cm right	Mixed speech
75% uniformly abs.	150 cm front	
Fully absorptive abs.		
Right wall absorptive		
Right and back wall abs.		
Right, back and left wall abs.		

To reduce the number of combinations for the test not all room conditions were compared to each other. Only the extreme conditions (fully reflective, fully absoptive) were compared to the other conditions. Overall, 168 ratings were given by each of the four expert listeners who took part in this preliminary study.

C. Hypotheses

- Non-uniform room conditions are easier to detect than uniform room conditions, because non-uniform room conditions can cause localization shifts or other additional cues.

- Detection rate will depend on distance with the highest detection rate at $300\,cm$, because a larger distance produces a smaller direct to reverberant ratio (DRR) and puts more attention to the reverberation tail.
- Detection rate will be the same for the speech sample and the castanets sample, but the speech mix sample will lead to higher thresholds, because the difference of the voices will conceal some of the room differences.

III. Pre-test results

All of the result figures follow the same principles: Each figure shows the percentage of correct answers for the given room condition when compared to the reference in the triplet test. The reference is either the fully reflective condition (top plot) or the fully absorptive condition (bottom plot). The conditions are ordered according to their reverberation time. The chance level to select the correct source in one trial was 33%. The dotted line in each plot is calculated on the basis of a 5% significance level. In other words, results above this line have a guessing level below 5%.

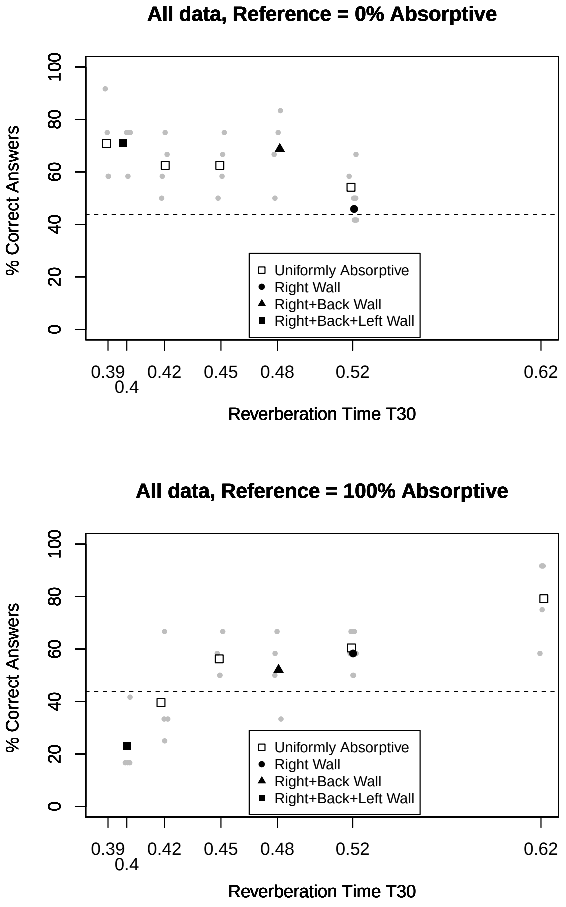

Fig. 4. Number of correct detections in percent. Results for all audio samples combined. Individual results are presented by small gray dots.

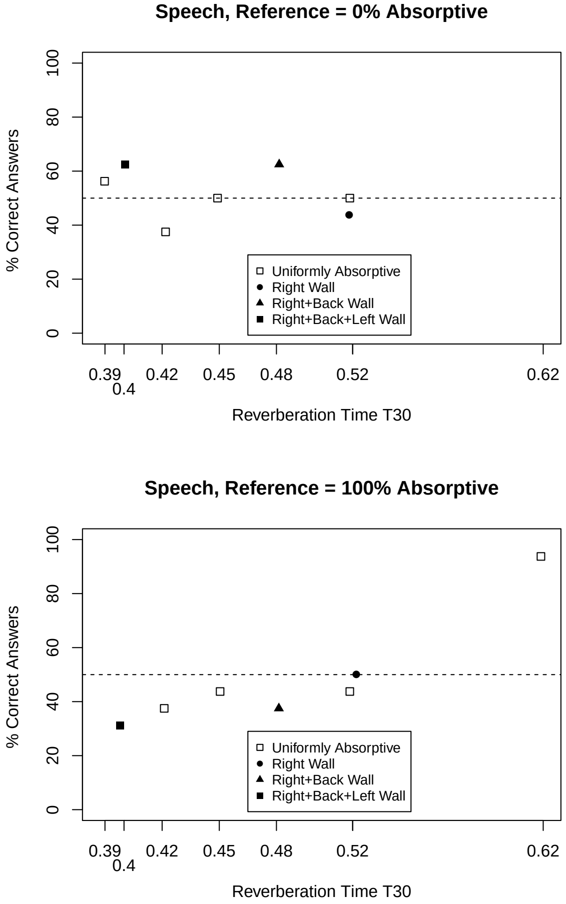

Fig. 5. Number of correct detections in percent. Results for the speech signal.

In general we can see that the detection rate increases with T30 difference (current condition compared to the reference). Figure 4 shows the results for all available data points. Individual results of the four participants are presented by small gray dots. In the top plot, conditions with a T30 lower than $0.52\,s$ (25% uniformly absorptive and right wall absorptive) are above the detection threshold. In the bottom plot the 50% uniformly absorptive condition with a T30 of $0.45\,s$ is the first above the threshold. These conditions relate to a relative T30 difference of -15% (top) and +15% (bottom). These values are in the order of magnitude of knowm RT JNDs [12], [6]. Even for the most different conditions, the highest detections rates are around 80%. The reason for this could be the concealment of the room differences by the positional differences and different sound spectra. This hypothesis got backed up by the verbal feedback of one participant, who stated, that in some cases all three sound sources in one trial sounded differently. For data in figure 4 it does not seem to matter which conditions served as the reference. Also, we cannot observe any obvious difference between the uniform and non-uniform conditions.

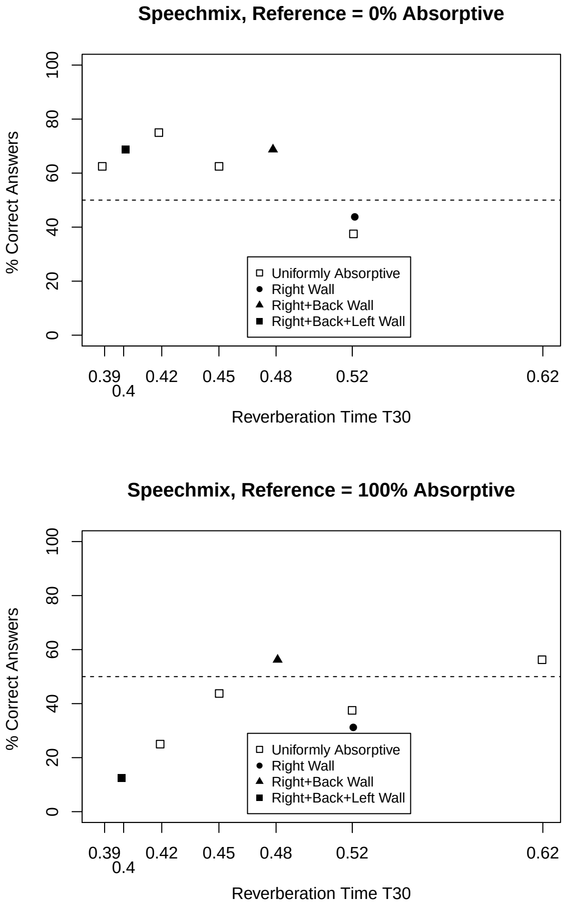

Fig. 6. Number of correct detections in percent. Results for the mixed speech signal.

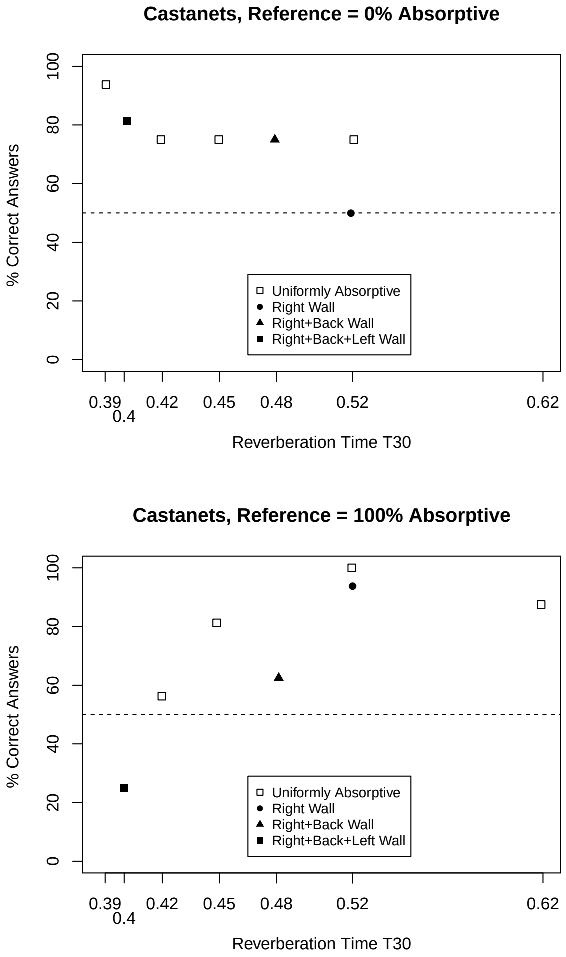

Fig. 7. Number of correct detections in percent. Results for the castanets signal.

For further analysis, the data were split according to the test signal. Figures 5, 6, 7 show the results for the signals speech, speechmix and castanets respectively. Individual results are omitted for these plots, because only four ratings would be available per participant for each individual data point. Overall, detection rates for castanets are higher than for the speech signals. For certain conditions with the castanets signal, detection rates are close to 100%. For these conditions we can conclude, that the difference in reverberation time was more obvious than the position depended acoustics. For the speech signals we observe most detection rates closer to the guessing rate. There are several possible reasons for this.

Compared to the castanets signal, the speech signals are less repetitive which makes it more difficult to perceive the differences between the conditions. Figure 8 shows the frequency spectra of the audio samples. The speech signals show more energy towards the lower frequencies than the castanets signal which could put greater emphasis on the position depended acoustic differences in the case of the speech signals. Again, this would lead to a masking of condition differences. The speechmix signal contained different voices for each sound source. This was expected to lower the detection rate over the normal speech signal further. But looking at the results in figures 5, 6 and table II, we only see small differences. Table II shows the percentage of the just noticeable reverberation time differences for each plot. For the speechmix signal we can read the same threshold from both plots, but for the speech signal, the values differ greatly. Therefore we cannot state a threshold difference between the signals, which is unexpected giving the characteristics of the audio samples. For the castanets signal we observed a very low threshold of 8%. Compared to the plot showing all audio samples combined, results seem to differ depending on the reference condition. In the study of Blevins [6] a similar effect was measured.

In the next step we investigated, wether the triplet distance has any influence on the detection threshold. Because the sound sources of each triplet have always the same angle spacing, sound sources are further apart at greater triplet distance. Position dependent acoustic features should therefore become greater, too. Another reason for this hypothesis was that the DRR reduces with distance. This could lead to a greater

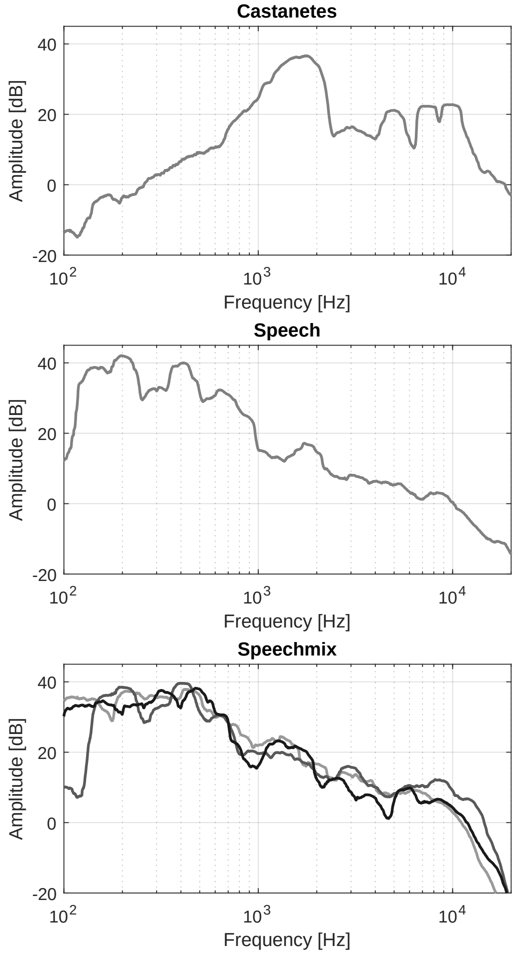

Fig. 8. Magnitude spectra of the audio samples. The plot for speechmix shows the spectra for each of the three speakers separately.

TABLE II
PERCENTAGE OF REVERBERATION TIME DIFFERENCE, WHICH WAS DETECTABLE. RESULTS ARE DIVIDED ACCORDING TO THE REFERENCE CONDITION IN THE TRIPLET TEST.

Reference	**All data**	**Castanets**	**Speech**	**Speechmix**
0% absorptive	-15%	-15%	-15%	-23%
100% absorptive	15%	8%	33%	23%

emphasis on the reverberation during the comparison. Table III briefly summarizes the found detection thresholds. Except for one combination ($150\,cm$ and reference: 0% absorptive), all detection thresholds are the same. Apart from this slight trend we found no evidence which supports this hypothesis. Detection thresholds seem not to be affected by triplet distance for our room conditions.

IV. DISCUSSION

The aim of the study was to investigate a method to measure perceptual tolerances with regard to reverberation differences. The test procedure was designed to consider real-life communication scenarios. We assumed that detectable reverberation differences could be considerably higher in such scenarios. This would give is insight, if such scenarios are less prone to the appearance of the room divergence effect.

TABLE III
PERCENTAGE OF REVERBERATION TIME DIFFERENCE, WHICH WAS DETECTABLE. RESULTS ARE DIVIDED ACCORDING TO THE REFERENCE CONDITION IN THE TRIPLET TEST AND THE DISTANCE OF THE TRIPLETS

Reference	**150 cm**	**200 cm**	**300 cm**
0% absorptive	-23%	-15%	-15%
100% absorptive	15%	15%	15%

Generally, our test approach is able to deliver reasonable results, since the just noticeable differences we observed were well within the range of JNDs reported by several studies. However, the preliminary results are quite surprising, because most of our initial hypotheses could not be confirmed. Non-uniform room conditions did not lead to lower detection thresholds. This could be because the expected effects (localization shifts, coloration) were simply to small in comparison to the change of reverberation. Since we used directional speakers pointed towards the listener, reflections from the walls were likely too low in level in comparison to the direct sound in order to influence the localization. Also, no evidence for our second hypothesis was found: Detection thresholds seem not to increase with distance of the triplet. An explanation is hard to find, but maybe the effect could not be uncovered, because the reverberation difference between the room conditions were too big. The third hypothesis turned out false, too. Thresholds for the castanets signals were lower than for the two speech signals and from our preliminary data we could not observe a difference between the speech signals. At first this was unexpected because the reverberation time JND is believed to be signal independent. Also, the speechmix signal contained different voices for each sound source, and thus we expected an impairment of the detection thresholds. However, other factors could have been more important: The frequency spectra of the speech signals have more energy in the lower frequencies ($< 1\,kHz$) than the castanets signal. This could emphasize position dependent acoustic differences (e.g. related to room modes) and thus become a more distinct feature than the reverberation difference. Another, more simple explanation could be that the differences were just easier to detect in case of the castanets signals, because the signal was more transient and repetitive than the speech signals.

Due to the limited number of participants in this preliminary study, our results should be taken with caution. Especially the analysis related to the audio samples and triplet distance are prone to error, since an already small data set was split further. However, the chosen test approach seems to be suited to measure just noticeable reverberation time differences. More room conditions with a finer gradation of the reveberation time, along with more participants could deliver more robust results. Of course, the shown results are room dependent, since each room exhibits different position depended acoustics.

References

[1] M. Kleiner and J. Tichy, *Acoustics of small rooms*. Boca Raton: Tayler & Francis Group, 2014.

[2] J. C. G. Carvajal, S. Santurette, J. Cubick, and T. Dau, "Spatial hearing with incongruent visual or auditory room cues," *Scientific Reports*, vol. 6, no. 37342, 2016.

[3] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, Lissabon, Portugal, 2016.

[4] ISO, *ISO 3382-1:2009(en) Acoustics — Measurement of room acoustic parameters — Part 1: Performance spaces*. International Organization for Standardization, 2009.

[5] H. P. Seraphim, "Untersuchungen über die unterschiedsschwelle exponentiellen abklingens von rauschbandimpulsen," *Acta Acustica united with Acustica*, vol. 8, no. 1, pp. 280–284, 1958.

[6] M. G. Blevins, A. Buck, Z. E. Peng, and L. Wang, "Quantifying the just noticeable difference of reverberation time with band-limited noise centered around 1000 hz using a transformed up-down adaptive method," in *Proceedings of the International Symposium on Room Acoustics*, Toronto, Canada, 06 2013.

[7] I. Frissen, B. F. G. Katz, and C. Guastavino, "Effect of sound source stimuli on the perception of reverberation in large volumes," in *Prcoeedings of International Conference on Auditory Display*. Springer Berlin Heidelberg, 01 2009, pp. 358–376.

[8] F. Klein, S. Werner, G. Götz, and K. Brandenburg, "Auditory adaptation in real and virtual rooms," in *Proceedings of the International Symposium on Auditory and Audiological Research*, vol. 7, Nyborg, Dänemark, 2020, pp. 341–348. [Online]. Available: https://proceedings.isaar.eu/index.php/isaarproc/article/view/2019-39

[9] S. Amengual Garí, H. G. Hassager, F. Klein, J. M. Arend, and P. Robinson, "Towards determining thresholds for the room divergence effect: A pilot study on perceived externalization," in *to be published in Proceedings of International Conference on Immersive and 3D Audio*, Italy, September 2021.

[10] S. Amengual Garí, J. M. Arend, P. T. Calamia, and P. W. Robinson, "Optimizations of the spatial decomposition method for binaural reproduction," *Journal of the Audio Engineering Society*, vol. 68, no. 12, pp. 959–976, december 2021.

[11] A. Neidhardt, F. Klein, N. Knoop, and T. Köllmer, "Flexible python tool for dynamic binaural synthesis applications," in *142nd AES Convention*, Berlin, 5 2017. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18721

[12] Z. Meng, F. Zhao, and M. He, "The just noticeable difference of noise length and reverberation perception," 10 2006, pp. 418 – 421.