Crowdsourcing vs. Laboratory Experiments - QoE Evaluation of Binaural Playback in a Teleconference Scenario

Thomas Volk[a], Christian Keimel[a], Michael Moosmeier[a], Klaus Diepold[a]

[a]*Institute for Data Processing, Technische Universität München, Arcisstr. 21, 80333 Munich*

Abstract

Experiments for the subjective evaluation of multimedia presentations and content are traditionally conducted in a laboratory environment. In this respect common procedures for the evaluation of teleconference systems are no different. The strictly controlled laboratory environment, however, often gives a rather poor representation of the actual use case. Therefore in this study we crowdsourced the evaluation of a teleconference system to perform the evaluation in a real-life environment. Moreover, we used the unique possibilities of crowdsourcing to employ two different demographics by hiring workers from Germany on the one hand and the US and Great Britain on the other hand. The goal of this experiment was to assess the perceived *Quality of Experience* (QoE) during a listening test and compare the results to results from a similar listening test conducted in the controlled laboratory environment. In doing so, we observed not only intriguing differences in the collected QoE ratings between the results of laboratory and crowdsourcing experiments, but also between the different worker demographics in terms of reliability, availability and efficiency.

Keywords: Crowdsourcing, QoE, Subjective evaluation, Binaural Audio, Teleconference

1. Introduction

Over the last decades teleconferencing has increasingly become an integral part of many people's everyday life. In the beginning the technology was limited to business environments using proprietary equipment, but especially due to the adoption of *voice over IP* (VoIP) in recent years conference calls have become an important element of private social interactions as well. In situations with multiple remote conferees in one session, however, the users' conference experience is often deteriorating when multiple speakers are simultaneously active and it becomes increasingly difficult to identify and understand a single speaker. To improve the users' experience in this conference situation we recently developed a teleconference application that allows for a spatially separated playback of the remote conference participants. The spatial playback was implemented using *binaural technology* in order to exploit the so-called *cocktail party effect* that allows humans to focus on particular sound sources in noisy environments [1].

Following the implementation and evaluation of several approaches to measure *head related transfer functions* that are used to create the virtual 3-D audio [2, 3], we implemented a real time

Email addresses: th.volk@tum.de (Thomas Volk), christian.keimel@tum.de (Christian Keimel), michael@moosmeier.org (Michael Moosmeier), kldi@tum.de (Klaus Diepold)

convolution engine allowing for spatially separated playback in a conference scenario with multiple remote conferees. The performance of this system was assessed in subjective evaluations to quantify the benefits of the spatially separated playback provided by the system for the user.

The standardized procedures for the subjective evaluation of teleconferencing applications require experiments to be conducted in a laboratory environment [4, 5, 6]. Even though a laboratory environment ensures strictly controlled surroundings regarding the listening conditions and the equipment –and thus usually provides highly repeatable and reliable results [7]– it can be argued that these fixed experimental conditions give a rather poor representation of the real world environment that most users find themselves when using a VoIP-based, but also traditional teleconference systems.

In this context crowdsourcing offers a promising alternative to avoid these shortcomings of traditional laboratory experiments. Crowdsourcing platforms such as Amazon's Meachnical Turk [8] or Microworkers [9] provide not only a large pool of potential test subjects, but also allow to perform the experiments in a more realistic environment, comparable to the environment used in everyday teleconferences e.g. at home in front of their computer using their own, usually non-professional equipment. In other words, crowdsourcing provides easy access to the real world use conditions of a teleconference system that are very difficult to recreate in a laboratory. There is, however, obviously much less control over the test conditions and there also other restrictions regarding the overall experimental design that have to be considered as already discussed by Hoßfeld et al. in [10]. In order to asses how crowdsourcing and its potential benefits could be incorporated and exploited for the evaluation of teleconference applications, especially applications utilising a virtual 3-D audio envrionemnt, we conducted a series of listening tests, performed both in the laboratory environment and using crowdsourcing, comparing the results gained with these two different set-ups.

In the remainder of this article we will give a review of related work and a brief introduction into binaural teleconference systems and their evaluation in Section 2, followed by a detailed description of the conducted experiments in Section 3. Then we will present the results of the experiments in Section 4 before concluding with discussing the results and some lessons learned regarding the subjective evaluation for teleconference scenarios via crowdsourcing.

2. Background

In this section we provide a review of current research on subjective evaluation via crowdsourcing, especially in the context of listening tests, and current methods for the subjective evaluation of binaural audio presentations. Furthermore, we will discus the use and potential benefits of binaural audio for teleconferencing and last but not least the concept of QoE.

2.1. Subjective Evaluation via Crowdsourcing

Crowdsourcing can be considered as the evolution of the outsourcing principle, where *tasks* are submitted to a huge crowd of usually anonymous *workers* by a *requester* in the form of an open call, instead of a designated employee or subcontractor that is assigned a specific job by the employer [11]. These tasks are often relatively short and therefore also called *micro-tasks* that can be done within a few minutes, but depending on the task, the granularity can differ [12]. As the goal of crowdsourced tasks is usually to delegate task that are simple for humans, but are extremely difficult or even impossible to be done using algorithms, such tasks are also often referred to as *Human Intelligence Tasks* or *HITs*.

In the context of subjective quality evaluation, the overall aim of crowd-based subjective evaluations is then to replace laboratory experiments with online, usually web-based experiments levering the huge pool of potential test subjects available using common crowdsourcing providers e.g. Amazon's MTurk [8] or Microworkers [9] that provide a mediation between the *requesters* and*workers*. Besides the easier access to test subjects, usually called *workers* in the context of crowdsourcing, this allows also for a more diverse test population [10], leading to a more realistic demographic. Moreover, depending on the location of the evaluation laboratory, the financial and logistical resources necessary for performing an evaluation can be significantly lowered using crowdsourcing, thus leading either to more subjects resulting in a statistical more representative population or allowing for more evaluations. In this context crowd-based subjective evaluation is often also referred to as *crowdtesting* [13].

Instead of implementing a separate testing application for each experiment, a number of different frameworks have been proposed that provide an out-of-the-box web-based online test environment, requiring only little or no programming skills to configure the evaluating [14, 15, 16, 17, 18, 19]. Two frameworks often utilised are the *Quadrant of Euphoria* by Chen et al. [14, 20, 21] and the *QualityCrowd* framework by Keimel et al. [16] that is also used in this contribution. For a detailed discussion about web-based crowdsourcing frameworks for subjective quality assessment we refer to the survey in [22]

The QualityCrowd framework was chosen in this contribution on the one hand due to its availability as open source, but on the other hand also because it provides a multitude of different options for the test design, allowing for any number of questions, and more importantly in the context of this contribution, it also supports different stimuli e.g. videos, sounds or images or any combination. In addition, it allows the use of different testing methodologies, e.g., single stimulus or double stimulus, and different scales, e.g., discrete or continuous quality or impairment scales, enabling us to tailor the test setup to our specific requirements.

2.2. Listening Tests via Crowdsourcing

So far there have been relatively few studies investigating whether crowdsourcing is an appropriate tool for subjective listening tests.

In [14], Chen et al. presented the results of two listening tests that were conducted using a newly implemented crowdsourcing framework. The first experiment dealt with the perceived QoE resulting from different MP3 compression levels of music files, whereas the second experiment investigated the effect of packet loss on the QoE of a VoIP application. Both experiments were conducted using the crowdsourcing platform MTurk, recruiting workers without any selection according to demography or geographic location. For comparison, the same setup was replicated in a laboratory environment with local test subjects. Although the study showed differences regarding the reliability of the workers depending on their origin, the overall test results were found to be reasonably consistent.

In [15], Ribeiro et al. suggested a framework called crowdMOS for subjective crowdtesting and presented a case study that compared the perceived naturalness of different speech synthesis algorithms. The study consisted of two crowdsourcing experiments: one in which the participants used headphones and one in which the participants used loudspeakers as playback device. The results of these experiments were compared to a benchmark experiment that was also carried out with paid online, but not crowdsourced participants. The study found a high correlation for all listening tests with headphones, whereas the correlation for the test with loudspeakers was considerably lower, but the study did not include results from a comparable laboratory experiment as a ground truth benchmark.

These results show that crowdsourcing is a promising approach for conducting listening tests. However, there are clearly some restrictions. For example, it is mentioned in [14] and [15] that the experiments carried out via crowdsourcing are not able to provide any expert training to the participants. Furthermore any experiment that requires a very specific equipment is not suitable for crowdtesting as also discussed in [13].

2.3. Binaural Audio and Subjective Evaluation

Most current sound systems that are designed for home entertainment purposes follow the concept of stereo playback. This approach permits the positioning of virtual sound sources along a straight line between two loudspeakers (2-D audio) as opposed to the simple monophonic playback, which would feed the same input signal to both speakers and make any virtual sound source appear to be located at a fixed position between the speakers (1-D audio). Most approaches to facilitate the reproduction of spatial or 3-D audio use arrays of more than two loudspeakers, such as the very common 5.1 speaker setup or the more elaborate concept of wavefield synthesis [23], but are clearly more expensive due to the increased number of loudspeakers.

Moving beyond loudspeakers, another approach to reproducing spatial audio using simple stereo headphones is the *binaural* or *immersive audio* approach. The binaural method uses head related transfer functions (HRTFs) to produce virtual 3-D audio scenes via headphone playback [1]. When measured in an anechoic environment, HRTFs contain the spectral characteristics of the acoustic reflections caused by the pinnae and the torso that allow humans to determine the position of a sound source. Filtering an audio signal with the corresponding HRTFs for the left and right ear can therefore make the sound source virtually appear at any position where HRTFs were measured. An in depth review of the state of the art methods for HRTF measurement as well as binaural synthesis and playback can be found in [24].

To achieve the most authentic binaural audio experience possible, the synthesized signal should be indistinguishable from the natural sound field produced by the actual sound source in a real environment. Therefore, the compensation of spectral distortions caused by the electroacoustic transducers involved is an important issue [25]. Most of the distortion is usually induced by the loudspeakers and microphones used during the HRTF-measurement process and by the headphones used for playback. The spectral characteristics of the loudspeakers and microphones are usually compensated during or immediately after the measurement. In most cases a recording of the measurement stimulus from a microphone placed at the position of the center of the interaural axis is used to create an inverse filter to equalize the recorded HRTFs. This approach was also used for the HRTFs that are contained in the database that was measured at our institute [26]. A detailed description of the recording method is given in [3]

Headphone equalization on the other hand is influenced by a multitude of factors. There are for example several reports that even different positioning of the headphones on the listeners head can lead to different *headphone transfer functions* (HpTFs) [25, 27]. Another important question is the choice between individual (measured from the listener) or general (measured from a dummy head) headphone compensation. The studies presented in the literature [28, 27] do not come to an unanimous answer to this question. They agree, however, on the fact that in general headphone compensation leads to a more realistic binaural playback and should be used whenever possible.

But apart from the headphone compensation the choice of the headphone itself has a significant influence as well. A study by Schonstein et al. [29] showed that different headphones can lead to significantly different results in a localization experiment. Surprisingly, headphone compensation did not always improve the localization performance in this study. This is, however, not an

explicit contradiction to the results of the studies discussed in the previous paragraph, since these studies investigated perceptual differences to the natural sound field and not the localization performance. These results illustrate, that optimal binaural synthesis is a highly sensitive process with numerous influencing variables, which is one of the main reasons why at this point there are only few real-life general applications that make use of this technology despite its immense potential.

Subjective evaluation of binaural audio in general also faces some challenges. Since HRTFs are usually measured for discrete positions on a spherical surface surrounding the listener, the most common approach for the subjective evaluation of binaural audio presentations are localization experiments. There is a multitude of studies to be found describing such experiments and their results, e.g. in [30] or [31]. It is questionable, however, whether localization accuracy is an adequate criterion to assess the *quality* of binaural audio presentations. It can be argued that for many applications of binaural audio, e.g. in video games or in VoIP applications, the fact how precisely the position of a virtual sound source can be determined is not as important to the listener as an overall satisfactory impression of spatial audio. Also comparing the virtual 3-D audio to a real auditory scene does not always appropriately reflect the benefits of binaural audio, especially if approaching the issued from an end user's perspective e.g. if we consider the cocktail party effect. An alternative to the use of localization accuracy or perceived realism as criteria to evaluate immersive audio presentations generated with HRTFs is therefore the concept of QoE.

2.4. Binaural Audio in Teleconference Applications

As mentioned in [24] there are numerous applications for binaural synthesis such as surround sound by headphones, 3-D auditory displays or virtual reality. Another interesting area of application, where users can benefit from binaural playback, are communication systems. The *cocktail party effect* [1] is used in this connection to help the listener to distinguish and distribute his attention between multiple active communication channels at the same time. The main objective in such applications is to improve intelligibility and speaker identification for the user. While a very realistic 3-D audio impression is certainly preferable for communication systems as well as for any other application of binaural audio, it is not necessarily the top priority in this application context, as the main goal of using binaural synthesis is the separation of the different participants in a teleconference. This is also one of the reasons why we decided to investigate the use of crowdsourcing for the perceptual evaluation of the system despite its obvious shortcomings regarding the listening equipment as discussed in section 2.3. An example for such a communication system is the teleconferencing system that was developed at the Institute for Data Processing at TUM [3]. Among other features it enables spatially separated playback of remote conference participants in real time. To do so, the audio signals from the individual speakers are filtered with HRTFs from our HRTF database [26] on the receiving end of the system and the listener can place the remote participants along a full circle at an elevation angle of 0° with an azimuth resolution of 1°.

To evaluate the benefits provided by the system for the user we also performed a series of listening tests investigating the users *QoE* and *cognitive load* while using the binaural playback compared to simple monophonic playback [3]. The test results showed that in conversations with four active speakers the listeners could remember more of the conversation's content and that the perceived effort to identify and follow the individual speakers decreased significantly with binaural playback. Furthermore, in terms of QoE the test participants clearly preferred the binaural over the monophonic playback. Note, that no headphone compensation was used during

any of these evaluation experiments. As mentioned in the previous section, headphone compensation is indeed crucial to achieve the most realistic spatial audio impression. For VoIP based teleconferencing, however, individual measurements are out of the question and the flexibility of a teleconference system is significantly improved if we do not require a specific type of headphone for all potential users. Furthermore, the results of the experiments in [3] show clearly that binaural playback induces considerable benefits for the user in terms of *QoE* and *cognitive load* even without headphone compensation.

2.5. Quality of Experience (QoE)

Quality of experience *QoE* is a comparably new concept in the realm of multimedia quality assessment. Unlike more established concepts, in particular quality of service (QoS), QoE focuses on the user's perception and his expectations towards an application or presentation rather than objective measures such as ubiquitous *signal to noise ratio (SNR)* or similar objective measurements. Furthermore, QoE sets itself apart from many studies concerned with audio quality assessment, by considering the user's overall impression rather than evaluating several different attributes of sound quality as suggested by Letowski in [32]. Concepts and examples of such experiments can be found in [33] and [34].

To this date there is no commonly accepted definition of the QoE concept, even though the ITU has incorporated the notion of quality of service experienced as "A statement expressing the level of quality that customers/users believe they have experienced" in ITU-T E.800 [35]. The most recent definition of QoE that extends the previous work is the definition in the "Qualinet White Paper on Definitions of Quality of Experience" [36]:

> **Quality of Experience (QoE)** is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state.

Compared to previous definitions, it also takes into account the user's current context and provides a more holistic view of the users. In this contribution, we therefore use this latest QoE definition.

3. Experiments

The listening tests were conducted as part of the subjective evaluation of the aforementioned immersive teleconference system. The system allows for a spatially separated playback of remote conference participants using the binaural technology explained in the previous section. In contrast to conventional teleconference systems that offer merely simple monophonic playback of remote participants, the possibility to spatially separate multiple talkers enhances the user's ability to identify the single speakers, thus increasing their intelligibility and consequently improving the overall teleconferencing experience. Therefore the focus of the evaluation was to examine how users rate the system's binaural playback in terms of QoE. During the listening tests, the subjects were requested to take on the role of a teleconference participant and quantify their subjective overall impression and satisfaction regarding the playback of the remote participants. To collect the QoE ratings, we employed the continuous five point scale suggested in ITU-R BS.1534-1 [37], also known as MUSHRA as depicted in figure 2.

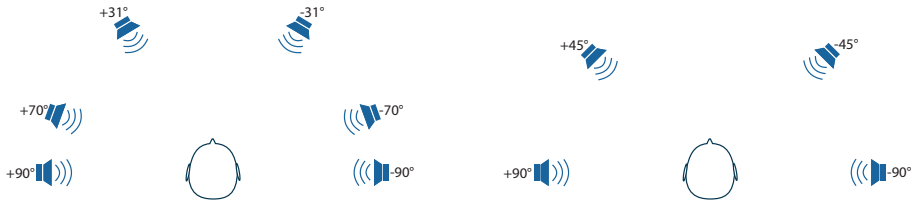

Figure 1: Virtual speaker alignment for (a) six and (b) four participants

3.1. Audio Material

In our experiments we used snippets from two different conference/meeting corpora. The German dialogue used in the laboratory experiment and in the first crowdsourcing experiment (CS1) was taken from a corpus designed especially for the evaluation of teleconference systems [38]. The English dialogue used in the second crowdsourcing experiment (CS2) was taken from the AMI meeting corpus [39]. Since the recordings from [38] were edited to provide a more fluent listening experience we decided to edit the AMI recordings in a similar manner. Furthermore we removed annoying "pop" noises which occur when a conference participant is talking too close to a microphone. During the editing process great care was taken not to impair the natural character of the speech recordings while providing an equal standard for both, the German and the English recordings as well. In total we collected six dialogue sequences from each corpus, each lasting for about one minute. The German dialogue included six different speakers while the AMI recordings included four. Therefore we chose two different virtual speaker alignments reported in [40] for the binaural rendering. The virtual speaker alignments are depicted in Figure 1.

The binaural playback was achieved by convolution of the edited single speaker channels with HRTFs, which were measured from a KEMAR dummy head in a semi anechoic chamber at the audio laboratory of the Institute for Data Processing at Technische Universität München using the measurement approach described in [2]. Further information on the HRTF data is provided in [26]. During the editing and binaural rendering we used 48 kHz WAV-files which were then encoded to 320 kBit/s MP3-files to make them suitable for crowdtesting. To obtain comparable results from laboratory and crowdsourcing experiments we decided to use the 320 kBits/s MP3 codec in all our experiments.

3.2. Laboratory Experiments

The laboratory experiments were conducted at the Institute for Data Processing at Technische Universität München and sixteen subjects took part in our study. All of them were students between the age of 21 and 25 and none of them reported any known hearing damages. In order to emulate the recruitment of subjects using crowdsourcing, no test for confirming normal hearing abilities were performed as this can not be done reliably in a crowdtesting environment. Each subject received a reimbursement of € 5 for participating in the test that lasted approximately 20 minutes. The participants performed the evaluation task using the same QualityCrowd instance and user interface that was also used in the subsequent CS1 and CS2 tests. The equipment used in the laboratory experiment is listed in Table 1.

Upon entering the laboratory the interface was shortly introduced to the participants. They then received a written introduction followed by a training consisting of one audio example for

PC (OS)	Lenovo Thinkpad T520 (Ubuntu Studio 13.04 64 Bit)
Browser	Firefox 27.0.1
Audio Interface	Roland Cakewalk UA-25 EX
Headphones	Beyerdynamic DT 990 Pro

Table 1: Equipment used in the laboratory experiment

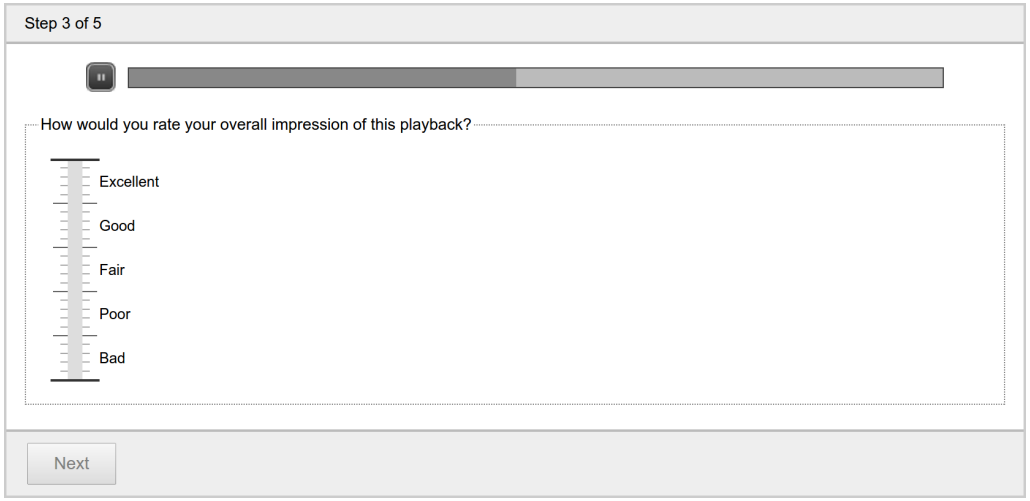

Figure 2: Playback and rating screen from the QualityCrowd user interface

the monophonic and one for the binaural playback method, both presented within the interface of the QualityCrowd framework. After completing the introduction and the training, they were given the possibility to ask questions in case they were insecure about their understanding of the evaluation task or procedure.

The stimulus set consisted of the six German dialogue sequences discussed in Section 3.1. Each sequence was presented once using monophonic playback and once using binaural playback. To ensure a similar structure of the experiment in both the laboratory and using crowdsourcing we decided to combine the monophonic and binaural treatments of the same stimulus in a stimulus pair resulting in six stimulus pairs overall. Three of these pairs started with the monophonic treatment, while the others started with the binaural treatment to avoid sequencing bias [7]. Nevertheless, each of the stimulus treatments was assessed separately. Also the same control questions used in the crowdsourcing experiments CS1 and CS2 were presented to the subjects in the laboratory experiment in order to maintain an experimental design as close as possible to the crowdsourcing experiments. The questions were always presented to the participants after they assessed the last of the two treatments in a pair. Finally, to control the sequencing bias between the stimulus pairs we used a balanced Latin square design [7] to determine the presentation order. With 16 subjects each rating six stimulus pairs, we therefore collected a total of 192 single user ratings.

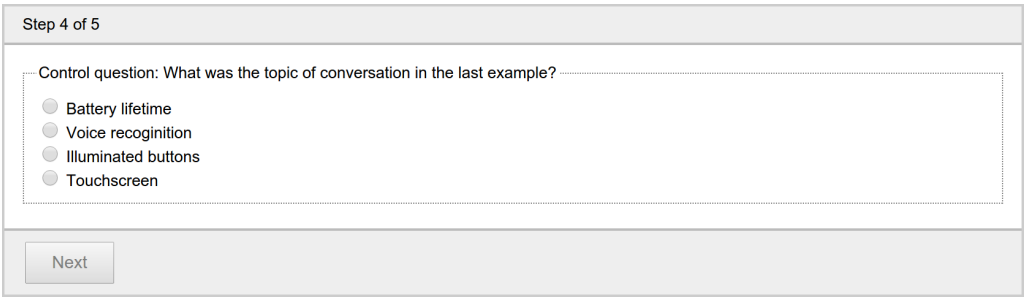

Figure 3: Control question screen from the QualityCrowd user interface

3.3. Crowdsourcing Experiments

The crowdsourcing experiments were conducted using Microworkers [9] as a crowd provider. For each completed HIT the workers received €0.47. Since the total duration of the laboratory experiment, approximately 20 minutes, exceeds the recommended time for a HIT as suggested in [13], we therefore split the entire procedure into six HITs, each consisting of one stimulus pair and an additional control question to monitor the workers' attention as recommended in [41]. The control questions required simple multiple choice answers about the content of the conference dialogue. We decided not to make correct answers to the control questions a requirement for receiving the payment to keep the workers' focus on the evaluation task. When a worker accepted the first HIT from our experiment they were given an introduction and training that were the same as the introduction and training in the laboratory experiment, with the exception that no question could be asked by the workers to clarify any questions regarding the test setup. Only the workers that successfully completed the training were able to participate in the test itself. This represents a two stage design as recommended in [13]. The workers performed the evaluation task and received a token to invoke the payment via Microworkers. After completing their first HIT, workers could accept further HITs for the same payment without having to get through the introduction and perform training again. Controlling the presentation order of the stimulus pairs in a similar way as in the laboratory was not possible since different workers completed different amounts of HITs. However, since they were able to choose the order in which they completed the HITs for themselves at least a certain amount of randomization was provided. The aim of the campaigns CS1 and CS2 was to complete 192 HITs in each campaign which equals a total of 384 single user ratings.

3.3.1. Crowdsourcing Experiment One (CS1)

In the first crowdsourcing campaign *(CS1)* we used the same (German) conference recordings as in the laboratory experiment. Therefore, the worker audience from Microworkers was limited to Germany. Otherwise the workers would not have been able to answer the control questions correctly. However, a study by Hirth et. al. [42] suggests that European workers only play a minor role in Microwrkers' worker demographic. They make up 8% of the potentially available workers. Poland and Romania provide 3% and 5%, respectively, indicating that the remaining European countries, including Germany, offer only a very small proportion of the worker population. As expected, the CS1 campaign required more time than most crowdsourcing campaigns reported in the literature: it took a total of 13 weeks and 48 workers to collect the desired 192 HITs.

3.3.2. Crowdsourcing Experiment Two (CS2

The second crowdsourcing campaign *(CS2)* used the English conference recordings from the AMI meeting corpus and the worker audience was limited to the US and Great Britain. We added this second campaign to the study since we were not aware of any previous studies using an exclusively German worker demographic. Due to the demographic structure of the crowdsourcing platform as discussed for the CS1 experiment, it was unclear whether such an experiment would be feasible at all. But considering the results by Hirth et. al. [42] that workers from the US make up 11% of the population on Microwrkers, it seemed reasonable to assume that a campaign using this different demographic could be completed in a sensible amount of time. Moreover, the second campaign was intended to be a validation experiment for CS1. The CS2 campaign was completed over a relatively short period of time. It took less than 7 hours and 41 workers to collect the desired amount of 192 HITs.

4. Results

In this section we provide a review of the experiments' results. Besides the outcome of the QoE evaluation task we also took a closer look at the differences between laboratory and crowdsourcing experiments regarding the reliability of the test subjects, the financial efforts that were necessary to collect the QoE ratings and especially the effects of the diverse environments and equipment conditions in the crowdsourcing experiments as opposed to the controlled set up in the laboratory.

4.1. Reliability

Getting reliable results from crowdworkers is one of the most crucial issues in crowdtesting. Therefore we excluded all results from any worker who answered one or more of the control questions wrong. The same practice was applied in similar studies comparing a laboratory experiment to a crowdsourced experiment e.g. in [43]. As discussed previously, we also included the control questions in the laboratory experiment to ensure a preferably similar evaluation procedure on one hand, but also to provide a benchmark when comparing laboratory and crowdsourcing experiments to each other in terms of reliability.

As expected, the participants in the laboratory experiment were very reliable and answered all of the control questions correctly. The workers in CS1 and CS2, however, showed surprisingly different behaviour. The German participants in CS1 answered 97% of the questions correctly and 92% of them were classified as reliable. The workers in CS2 on the other hand answered only 76% of the questions correctly and only 56% of them qualified as reliable.

Another noticeable observation we made was that all of the five workers from CS1 that were classified as unreliable only answered one of the control questions incorrectly whereas the majority of the unreliable participants in CS2 answered two or more questions incorrectly resulting in a percentage of 21% (CS1) and 45% (CS2) of false answers among the unreliable test subjects. This suggests that the unreliable workers in CS1 were more likely to be inattentive when they gave wrong answers while the unreliable workers in CS2 seem to tend more towards deliberate cheating. Figure 4 shows the distribution of incorrect answers per worker for CS1 and CS2.

4.2. Costs of Crowdsourcing vs. Laboratory

Since one of the main assets of crowdtesting compared to conventional laboratory experiments is the reduction of the costs, we examine briefly the financial efforts that were made to obtain the

	workers	reliable	%	questions	correct	%
LAB	16	16	100	96	96	100
CS1	48	44	92	192	187	97
CS2	41	23	56	192	145	76

Table 2: Overview about the reliability of the workers and the correct answers to the control questions

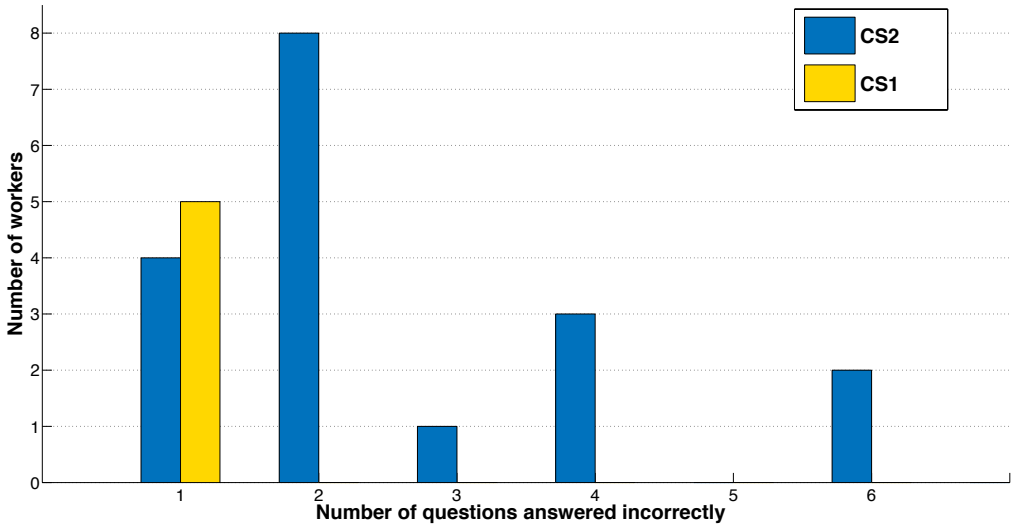

Figure 4: Number of questions answered incorrectly among the workers classified unreliable in CS1 and CS2

	raliable ratings	costs	costs / reliable HIT
LAB	192	€ 100	€ 0.52
CS1	322	€ 96	€ 0.30
CS2	194	€ 96	€ 0.50

Table 3: Review of the costs for the LAB, CS1 and CS2 experiments and the price for a single reliable QoE rating

	Background noise			Price range of headphones			
	quiet	moderate	loud	< 20 €	20 - 49 €	50 -100 €	> 100 €
LAB	16	0	0	0	0	0	16
CS1	38	6	0	11	21	11	1
CS2	20	2	1	5	12	6	0

Table 4: Overview about the reliable crowdworkers' statements regarding background noise and the price range of their headphones

QoE ratings from the laboratory participants and the workers. Table 3 gives an overview about the expenses that were necessary to complete the three experiments and especially the costs to get one single reliable rating in each experiment and it can be seen that in CS1 we had to spend only about 60% compared to the laboratory experiments and CS2.

Putting these expenses into perspective, however, one still has to consider the costs for staff, laboratory maintenance, administrative effort etc. that are necessary to conduct laboratory experiments.

4.3. Survey on environment and headphones

During the introduction to the crowdsourcing experiments we asked the workers to describe the level of background noise in their environment by assigning it to one of three classes: *quiet (e.g. alone in a room)*, *moderate (e.g. conversation in the background)*, and *loud (e.g public space)*. Furthermore, we asked them about the price range of the headphones they were using as an indication of the headphones audio quality i.e. undistorted sound reproduction. The given categories were: *< 20 €*, *20 - 49 €*, *50 -100 €* and *> 100 €*. The results of the survey as shown in Table 4 revealed that most of the crowdworkers (87%) performed the evaluation task in a quiet environment as did all the participants in the laboratory experiment. Regarding the question about the headphones, however, we discovered that 73% of the crowdworkers used headphones below a price of € 50, whereas only one person specified the price range that also applies for the equipment used in the laboratory. And while we are aware that the price of the headphones is not an accurate indicator for their suitability for binaural playback this observation nevertheless reflects a considerable discrepancy in the equipment conditions between the laboratory and the crowdsourcing experiments.

There is only little research available in general on the topic of the influence of different headphones on binaural playback. The most extensive study was done by Schonstein et al. [29] comparing eight different headphones in a localization experiment. They came to the conclusion that the localization accuracy varied widely between different headphone types while headphone equalization of the HRTFs, which is obviously not feasible for crowdsourcing, had very little effect in comparison.

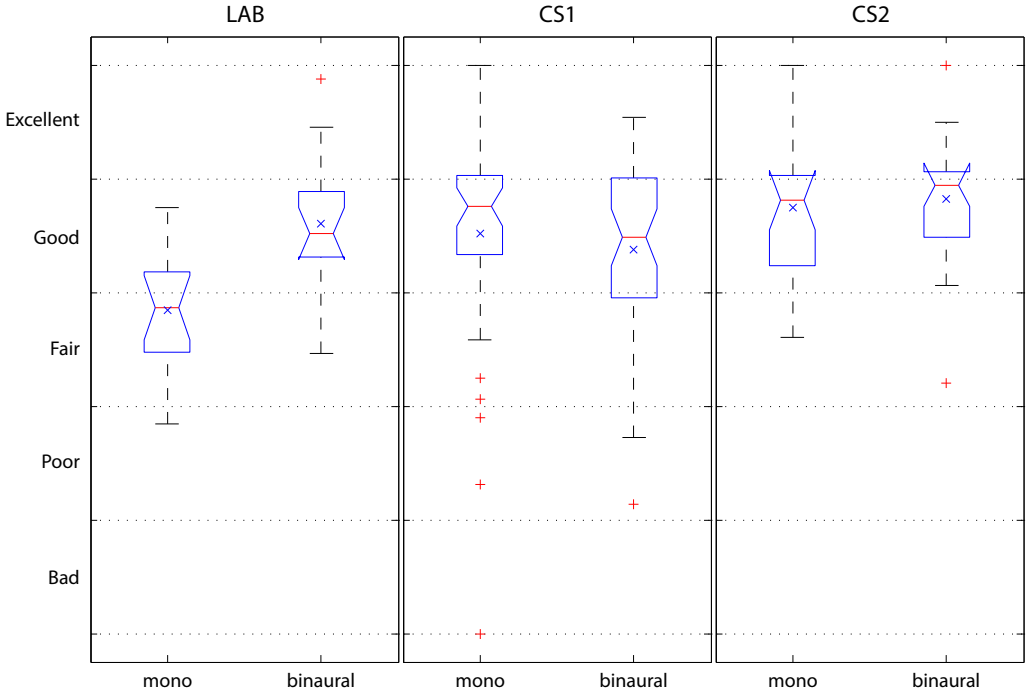

Figure 5: Boxplots of the QoE evaluation ratings (the red line marks the median and "x" the arithmetic mean value, the box includes the upper and lower quartiles and the whiskers are 1.5 times the length of the respective quartile, the narrow areas of the box show the 95% confidence intervals, red crosses mark outliers)

4.4. QoE Results

The boxplots of the QoE ratings from the laboratory and crowdsourcing experiments are shown in Figure 5. Graphic analysis of the data indicates that the participants of the laboratory experiment perceived a clear difference in terms of QoE between the monophonic and binaural playback whereas in the CS1 and CS2 experiments this tendency appears far less distinct. The overlapping 95% confidence intervals suggest that the difference between the mean values in the CS1 and CS2 data might statistically not be significant. After establishing the normal distribution of the data, we therefore applied the t-test to the QoE data to investigate whether the differences between the means were rather based on random or systematic effects. Judging from histograms and boxplots of the differences between the compared samples the data met the required precondition of normal distribution. The results of the t-test are shown in Table 5: it confirmed our assumption that the crowdworkers in CS1 and CS2 perceived no significant difference in terms of QoE. With p-values of 39% (CS1) and 57% (CS2), respectively, we accepted the null hypothesis whereas we rejected it with a p-value of 0.04% for the data from the laboratory experiment. The threshold level α was set to 5% while changing it to other common values such as 1% or 10% did not affect the outcome of the analysis.

4.5. Discussion

As expected, comparing the crowdsourcing experiments to the experiments conducted in a laboratory environment, we found that crowdworkers are less reliable than laboratory participants

	p	Null-hypothesis
LAB	$0.04\% \ll \alpha$	rejected
CS1	$39\% > \alpha$	accepted
CS2	$57\% > \alpha$	accepted

Table 5: Results of the t-test to compare the means, $\alpha = 5\%$

but more cost efficient.

Comparing CS1 and CS2, however, we made an interesting observation. Even though crowdsourcing is less popular in Germany than in the US and Great Britain, which is also reflected in the vast difference regarding the amount of time needed to collect the desired amount of QoE ratings, German workers were a lot more reliable and cost efficient. We assume that their motivation to participate in crowdsourcing is more likely to be out of personal interest whereas workers especially from the US, where crowdsourcing is a lot more popular, are more likely to primarily be motivated by the financial reward and thus tend more towards cheating.

On the topic of QoE evaluation of the two different playback methods, we arrived at diverging observations between laboratory and crowdsourcing experiments as well. While the participants of the laboratory experiment clearly preferred the binaural playback option, the crowdworkers rated both, the monophonic and the binaural playback options, equally. Judging from the additional information we collected from the crowdsourcing participants using self-reports the main hypothesis explaining the reason for the differing results is the use of different headphones by the crowdworkers. And while we are not aware of any previous studies investigating the effect of different headphones on the QoE of binaural audio presentations there is evidence in literature to support the assumption that such varying playback conditions have a considerable effect on the human perception as for example discussed by by Schonstein et al. in [29]. Besides that, leaving out the headphone equalization could also have an influence on the binaural playback. In the laboratory, however, this did not have the same effect. Another hypothesis regarding the diverging results is that the crowdworkers did not pay sufficient attention to the introduction and therefore did not fully understand their task. The presence of an instructor in the laboratory is commonly known to increase the participants' motivation.

Therefore we conclude that the particular type of experiment discussed in this contribution is not necessarily suitable to be repeated via crowdsourcing as it does not yield similar results to the laboratory. The data collected in the crowdsourcing experiments can, however, still be considered very valuable, since for a teleconferencing application, such as the one examined in this study, the environment in which the crowdworkers performed the evaluation task is very a much more accurate representation of the actual real-world use case than an experiment conducted in a laboratory.

But on the other hand the results from the comparison between laboratory and crowdsourcing experiments in this study illustrate the need for further research on QoE-influencing factors such as e.g. the choice of headphones for binaural audio presentations in order to achieve a better understanding of the possibilities and limitations of this technology from a user's perspective.

5. Conclusion

The main contribution gained in the experiments described in this paper is that for the particular test case of binaural playback in a teleconferencing application the results from a laboratory

experiment can not be repeated via crowdsourcing. There are several possible explanations for this observation. Firstly, there is the discrepancy regarding the headphones that were used by the participants in the laboratory and crowdsourcing experiments. Another difference between the two scenarios is that the test subjects in the laboratory experiment had the opportunity to ask questions during the introduction in case they were not exactly sure about the purpose of their task. Moreover, the presence of an instructor at the beginning of the experiment may also have increased their motivation. The crowdtesting participants on the other hand only received the written introduction and the same listening examples, but there was no room for further explanations to prevent possible misunderstandings. To achieve a better understanding of the influence that these factors have on the QoE results, however, further research is needed.

During the crowdsourcing experiments we also gained information especially about the behaviour of the two different demographics we addressed (Germany and US/Great Britain): while the German crowdworkers were more reliable and thus cost-efficient, the experiment with workers from the US and Great Britain took significantly less time to complete and was thus more time-efficient. If this difference in the workers' performance can be generalised or is specific to the used crowdsourcing platform and/or demographics needs further studies.

Eventually, this study emphasizes the question whether crowdsourcing experiments can and/or should always have the aim of replacing conventional laboratory experiments. In our case they were much more of an insightful addition to the laboratory study that raised important questions about the performance of the evaluated system under real world conditions than a replacement of the laboratory experiment itself.

6. References

[1] J. Blauert, Spatial hearing (1997).

[2] M. Rothbucher, K. Veprek, P. Paukner, T. Habigt, K. Diepold, Comparison of head-related impulse response measurement approaches, The Journal of the Acoustical Society of America 134 (2) (2013) EL223–EL229.

[3] M. Rothbucher, Development and evaluation of an immersive audio conferencing system, Ph.D. thesis, München, Technische Universität München, Diss., 2014 (2014).

[4] ITU-T, ITU-T P.800 Methods for subjective determination of transmission quality.

[5] ITU-T, ITU-T P.1301 Subjective quality evaluation of audio and audiovisual multiparty telemeetings.

[6] ITU-T, ITU-R BS.1284-1 General Methods for the Subjective Assessement of Sound Quality.

[7] S. Bech, N. Zacharov, Perceptual audio evaluation-Theory, method and application, John Wiley & Sons, 2007.

[8] Amazon Mechanical Turk (Feb. 2013). [link].
URL `http://mturk.com`

[9] Microworkers (Feb. 2013). [link].
URL `http://microworkers.com`

[10] T. Hoßfeld, C. Keimel, Crowdsourcing in QoE Evaluation, in: Quality of Experience, Springer, 2014, pp. 315–327.

[11] E. Estellés-Arolas, F. González-Ladrón-de Guevara, Towards an integrated crowdsourcing definition, Journal of Information science 38 (2) (2012) 189–200.

[12] T. Hoßfeld, M. Hirth, P. Tran-Gia, Modeling of Crowdsourcing Platforms and Granularity of Work Organization in Future Internet, in: International Teletraffic Congress (ITC), San Francisco, USA, 2011.

[13] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, P. Tran-Gia, Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing, Multimedia, IEEE Transactions on 16 (2) (2014) 541–558. doi:10.1109/TMM.2013.2291663.

[14] K.-T. Chen, C.-C. Wu, Y.-C. Chang, C.-L. Lei, A crowdsourceable qoe evaluation framework for multimedia content, in: Proceedings of the 17th ACM international conference on Multimedia, ACM, 2009, pp. 491–500.

[15] F. Ribeiro, D. Florêncio, C. Zhang, M. Seltzer, Crowdmos: An approach for crowdsourcing mean opinion score studies, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 2416–2419.

[16] C. Keimel, J. Habigt, C. Horch, K. Diepold, Qualitycrowd—a framework for crowd-based quality evaluation, in: Picture Coding Symposium (PCS), 2012, IEEE, 2012, pp. 245–248.

[17] B. Rainer, M. Waltl, C. Timmerer, A web based subjective evaluation platform, in: Workshop on Quality of Multimedia Experience, Klagenfurth, AT, 2013.
[18] S. Kraft, U. Zölzer, BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality, in: Linux Audio Conference, Karlsruhe, DE, 2014.
[19] B. Gardlo, S. Egger, M. Seufert, R. Schatz, Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing, in: International Conference on Communications, Sydney, AU, 2014.
[20] K.-T. Chen, C.-J. Chang, C.-C. Wu, Y.-C. Chang, C.-L. Lei, Quadrant of euphoria: A crowdsourcing platform for QoE assessment, Network 24 (2).
[21] C. Wu, K. Chen, Y. Chang, C. Lei, Crowdsourcing multimedia QoE evaluation: A trusted framework, Transactions on Multimedia 15 (99).
[22] T. Hoßfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, C. Timmerer, Survey of web-based crowdsourcing frameworks for subjective quality assessment, in: IEEE 16th International Workshop on Multimedia Signal Processing (MMSP), 2014.
[23] A. J. Berkhout, D. de Vries, P. Vogel, Acoustic control by wave field synthesis, The Journal of the Acoustical Society of America 93 (5) (1993) 2764–2778.
[24] D. Hammershøi, H. Møller, Binaural technique—basic methods for recording, synthesis, and reproduction, in: Communication Acoustics, Springer, 2005, pp. 223–254.
[25] Z. Schärer, A. Lindau, Evaluation of equalization methods for binaural signals, in: Audio Engineering Society Convention 126, Audio Engineering Society, 2009.
[26] M. Rothbucher, P. Paukner, M. Stimpfl, K. Diepold, The tum-ldv hrtf database.
[27] B. Masiero, J. Fels, Perceptually robust headphone equalization for binaural reproduction, in: Audio Engineering Society Convention 130, Audio Engineering Society, 2011.
[28] A. Lindau, F. Brinkmann, Perceptual evaluation of individual headphone compensation in binaural synthesis based on non-individual recordings, in: 3rd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, 2010, pp. 137–142.
[29] D. Schonstein, L. Ferré, B. F. Katz, Comparison of headphones and equalization for virtual auditory source localization, Journal of the Acoustical Society of America 123 (5) (2008) 3724.
[30] D. R. Begault, et al., 3-D sound for virtual reality and multimedia, Vol. 955, AP professional Boston etc, 1994.
[31] E. M. Wenzel, M. Arruda, D. J. Kistler, F. L. Wightman, Localization using nonindividualized head-related transfer functions, The Journal of the Acoustical Society of America 94 (1) (1993) 111–123.
[32] T. Letowski, Sound quality assessment: concepts and criteria, in: Audio Engineering Society Convention 87, Audio Engineering Society, 1989.
[33] F. Rumsey, Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm, Journal of the Audio Engineering Society 50 (9) (2002) 651–666.
[34] J. Berg, F. Rumsey, Systematic evaluation of perceived spatial quality, in: Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality, Audio Engineering Society, 2003.
[35] ITU-T, ITU-T E.800 Definitions of terms related to quality of service.
[36] P. Le Callet, S. Möller, A. Perkis, et al., Qualinet white paper on definitions of quality of experience, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003).
[37] ITU-R, ITU-R BS.1534-1, method for the subjective assessment of intermediate quality levels of coding systems (MUSHRA), International Telecommunication Union.
[38] J. Skowronek, A. Raake, K. Hoeldtke, M. Geier, Speech recordings for systematic assessment of multi-party conferencing, in: Proceedings of Forum Acusticum, 2011, pp. 111–116.
[39] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al., The ami meeting corpus, in: Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, Vol. 88, 2005.
[40] D. S. Brungart, B. D. Simpson, Improving multitalker speech communication with advanced audio displays, Tech. rep., DTIC Document (2005).
[41] T. Hossfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, C. Keimel, Best practices and recommendations for crowdsourced qoe - lessons learned from the qualinet task force crowdsourcing, Tech. Rep. 1.0, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003 Qualinet) (October 2014).
[42] M. Hirth, T. Hoßfeld, P. Tran-Gia, Anatomy of a crowdsourcing platform-using the example of microworkers. com, in: Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference On, IEEE, 2011, pp. 322–329.
[43] J. A. Redi, T. Hoßfeld, P. Korshunov, F. Mazza, I. Povoa, C. Keimel, Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal, in: Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia, ACM, 2013, pp. 29–34.