Computer Networks 000 (2015) 1–11

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Crowdsourcing vs. laboratory experiments – QoE evaluation of binaural playback in a teleconference scenario

Thomas Volk*, Christian Keimel, Michael Moosmeier, Klaus Diepold

Institute for Data Processing, Technische Universität München, Arcisstr. 21, Munich, 80333, Germany

ARTICLE INFO

Article history:
Received 15 December 2014
Revised 10 April 2015
Accepted 15 May 2015
Available online xxx

Keywords:
Crowdsourcing
QoE
Subjective evaluation
Binaural audio
Teleconference

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.

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 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 standardised procedures for the subjective evaluation of teleconferencing applications require experiments to be conducted in a laboratory environment [4–6]. 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

* Corresponding author. Tel.: +498928923629.
E-mail addresses: th.volk@tum.de (T. Volk), christian.keimel@tum.de (C. Keimel), michael@moosmeier.org (M. Moosmeier), kldi@tum.de (K. Diepold).

http://dx.doi.org/10.1016/j.comnet.2015.05.021
1389-1286/

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 us 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 assess 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 environment, 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 discuss 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 tasks 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 evaluation [14–19]. 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

Please cite this article as: T. Volk et al., Crowdsourcing vs. laboratory experiments – QoE evaluation of binaural playback in a teleconference scenario, Computer Networks (2015), http://dx.doi.org/10.1016/j.comnet.2015.05.021