Auditory Immersion with Stereo Sound in a Mobile Robotic Telepresence System

Andrey Kiselev
Örebro University
70182 Örebro, Sweden
andrey.kiselev@oru.se

Mårten Scherlund
Giraff Technologies AB
Metallverksgatan 8, 721 30
Västerås, Sweden
marten.scherlund@giraff.org

Annica Kristoffersson
Örebro University
70182 Örebro, Sweden
annica.kristoffersson@oru.se

Natalia Efremova
Plekhanov University
Russia
natalia.efremova@gmail.com

Amy Loutfi
Örebro University
70182 Örebro, Sweden
amy.loutfi@oru.se

Abstract

Auditory immersion plays a significant role in generating a good feeling of presence for users driving a telepresence robot. In this paper, one of the key characteristics of auditory immersion - sound source localization (SSL) - is studied from the perspective of those who operate telepresence robots from remote locations. A prototype which is capable of delivering soundscape to the user through Interaural Time Difference (ITD) and Interaural Level Difference (ILD) using the ORTF stereo recording technique was developed. The prototype was evaluated in an experiment and the results suggest that the developed method is sufficient for sound source localization tasks.

Categories and Subject Descriptors

I.2.9 [**Robotics**]: Commercial robots and applications; H.5.2 [**User Interfaces**]: Auditory (non-speech) feedback

Keywords

Human-Robot Interaction; Mobile Robotic Telepresence; Teleoperation; Sound Source Localization; Auditory Immersion; User Interfaces; ORTF Stereo

1. INTRODUCTION

Mobile Robotic Telepresence (MRP) [4] is an emerging field in robotics which brings mobile robots equipped with telepresence and teleoperation capabilities into domestic environments. The main goal of these systems is to allow rich natural human-human interactions between distant locations.

Presence, an important aspect when people interact with virtual or distant environments, shows how effectively people can place themselves "in there" through some interface. Presence can be seen as one's psychological ability to be in a remote of virtual location, an ability which depends in large parts on the level of immersion [5]. Immersion, in turn, is a pure technical concept [6] which can be defined as *a technology's ability to provide reliable and consistent interpretation of remote or virtual environments.*

It has been previously shown [2] that immersion can have an impact on one's performance in remote or virtual environments. This means that by providing users with information allowing for a higher level of immersion, performance can also be improved. Immersion (and particularly spatial immersion) becomes more important to allow novice users to operate robots in a safe and efficient way.

Auditory immersion plays a significant role in creating a consistent and reliable scene for the user. Auditory immersion can be characterized by two components: perceived subjective feeling of immersion and objective SSL potential. Gustavino *et al.* [3] has shown that these two can conflict with each other and that the selection of a proper auditory immersion technique must be done with care.

In this paper, we present a study aiming at assessing the issue of sound source localisation from the perspective of telepresence robotics. A prototype capable of delivering the soundscape to the remote user has been developed and experimentally validated with 53 users. The prototype was implemented as an attachment to the Giraff MRP system. Particularly, two identical cardioid microphones were placed 17 cm from each other and at an angle of 110° implementing a standard ORTF stereo recording technique. This allowed to achieve a stereo effect using both ILD (difference in levels in two channels) and ITD (phase shift between channels).

2. EXPERIMENTAL VALIDATION

An experiment was conducted to evaluate how effective the newly developed stereo sound prototype is in helping users to localize sound sources in remote environments.

For this experiment, a test phrase was recorded from five different locations around the robot using the developed prototype (see Fig. 1). The phrase is sentence 2 from list 11 of Harvard sentences [1]: "Cats and dogs each hate the other." In all five recordings, the speaker is the same person. Both audio and video were recorded. Stereo sound was recorded

HRI'15, March 2–5, 2015, Portland, Oregon, USA.
Copyright © 2015 ACM 978-1-4503-3318-4/15/03.
http://dx.doi.org/10.1145/2701973.2702034.

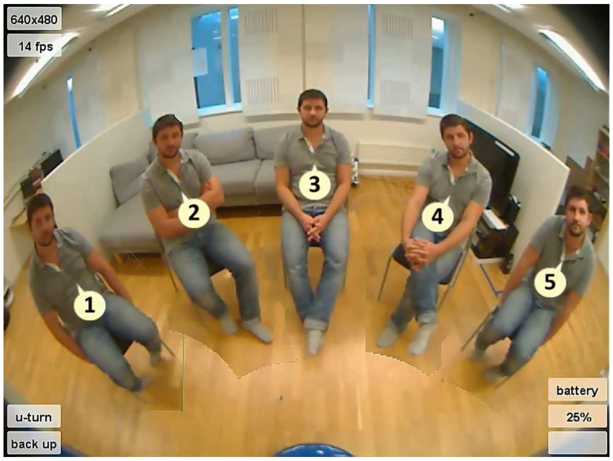

Figure 1: Screenshot from the videos shown to the subjects. The numbers denote the possible sound sources.

using the developed prototype with no further processing (no noise cancellation or stereo enhancement). The developed prototype provides both ILD and ITD stereo. Additionally, five samples with monophonic sound were made for control condition.

All samples were presented in a pseudo-random order. Two opposite sequences of trials were made for counterbalancing. Stereo samples were presented in two sets, thus each sample was presented twice. Subjects were presented with the samples in monophonic mode to build a control condition prior to the stereo set. Thus, each subject had five trials in mono session and ten trials in stereo session. Subjects were allowed repeat samples several times, but not allowed to go back and modify their answers. Two additional trials with all mono or all stereo samples blended together were used before mono and stereo sessions respectively to pre-expose subjects to complete soundscapes and verify audio settings.

We hypothesized that for the samples presented in stereo mode, subjects would be able to identify the sound source with a statistically significant distinction whereas no statistical difference among samples would be found for the samples presented in mono mode.

53 subjects participated in this study, 31 and 22 for each pseudo-random sequence respectively. The age varied from 20 to 66 ($= 32.74, \sigma = 9.45$). 36 of them were men, 17 of them were women. The subjects used their own computers and their own headphones or speakers to implement natural variety of devices in this experiment.

2.1 Data analysis

All results are shown graphically in Fig. 2. The horizontal axis represents the actual sound source, whereas the vertical axis shows the subjects' responses. In each group from 1 to 5 horizontally, the left (white) bar shows the response in *stereo* mode and the right (gray) bar shows the response in *mono* mode. On each bar, the bold line shows the $\widetilde{\chi}$ value, the lower and the upper bounds of the bar are $\mu - \sigma$ and $\mu + \sigma$ respectively.

A repeated measure ANOVA was used on both mono and stereo sequences. No statistically significant distinction between five speakers identified by subjects can be found in mono mode ($F(4, 52) = 1.774, \rho = 0.01$), whereas for stereo mode there is a statistically significant difference between speakers ($F(4, 105) = 122.223, \rho = 0.01$). This result supports the research hypothesis that the implemented prototype allows users to identify the location of the source of sound with a reasonable reliability.

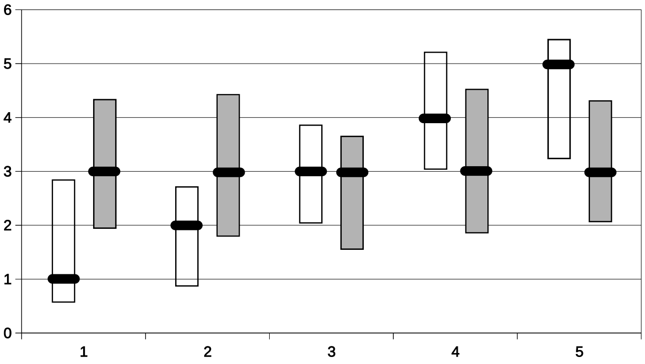

Figure 2: Experiment results. The horizontal axis represents the sound sources, the vertical axis shows the $\widetilde{\chi}$ (bold horizontal lines), μ, and σ of results in stereo (white bars) and mono (gray bars) modes.

In the experiment, a misalignment between audio and video stimuli was observed. The actual angle between speakers 1 and 5 is approximately 130°. Indeed, the audio samples were recorded at this wide angle. At the same time, the image from the wide angle lens camera (horizontal field of view is approximately 170°) is significantly compressed to around $20° - 30°$ when subjects use standard computer screens. As a result, the sound sources *1* and *5* are heard much further to the left and to the right respectively than they appear on the video. Surely, the impact of this inconsistency depends on each subject's individual perception of presence.

3. ACKNOWLEDGEMENTS

This work is financed by Västmanlands County Administrative Board, Sweden (303-3149-13) and Russian Foundation for Basic Research, Russia (14-07-00603).

4. REFERENCES

[1] IEEE Recommended Practices for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, 17:227–246, 1969.

[2] K. Gruchalla. Immersive well-path editing: investigating the added value of immersion. In *IEEE Virtual Reality 2004*, pages 157–164. IEEE, 2004.

[3] C. Guastavino, V. Larcher, G. Catusseau, and P. Boussard. Spatial audio quality evaluation: Comparing transaural, ambisonics and stereo. In *Proc. of ICAD'07*, pages 53–58, 2007.

[4] A. Kristoffersson, S. Coradeschi, and A. Loutfi. A Review of Mobile Robotic Telepresence. *Advances in Human-Computer Interaction*, 2013:1–17, 2013.

[5] M. Slater, M. Usoh, and A. Steed. Taking steps: the influence of a walking technique on presence in virtual reality, 1995.

[6] C. M. Usoh and M. Slater. Presence: Experiments in the Psychology of Virtual Environments. *135th AES Convention*, 2013.