BNAM2012
Joint Baltic-Nordic Acoustics Meeting
June 18th - 20th 2012 Odense, Denmark

3D Sound in the Telepresence Project BEAMING

Søren Krarup Olesen
Aalborg University, Section of Acoustics, Fredrik Bajers Vej 7B, DK-9220, Aalborg Ø, Denmark, sko@es.aau.dk

Miloš Marković
Aalborg University, Section of Acoustics, Fredrik Bajers Vej 7B, DK-9220, Aalborg Ø, Denmark, mio@es.aau.dk

Esben Madsen
Aalborg University, Section of Acoustics, Fredrik Bajers Vej 7B, DK-9220, Aalborg Ø, Denmark, em@es.aau.dk

Pablo Faundez Hoffmann
Aalborg University, Section of Acoustics, Fredrik Bajers Vej 7B, DK-9220, Aalborg Ø, Denmark, pfh@es.aau.dk

Dorte Hammershøi
Aalborg University, Section of Acoustics, Fredrik Bajers Vej 7B, DK-9220, Aalborg Ø, Denmark, dh@es.aau.dk

The involvement of Aalborg University in the EU project BEAMING will be presented. BEAMING deals with telepresence including multiple modalities; vision, haptics and audio, of which the latter is of main interest here. The setup consists of two types of locations: The Destination, where the Locals reside, and the Transporter, where the Visitor resides. The Visitors will virtually visit the Locals through an Internet connection and both must feel the other party being physically present. All Locals wear close-up microphones and their positions are tracked. In order to support presence for the Visitor, 3D audio is provided through headphones. It is rendered based on the Locals' coordinates via a common Internet database including local positional tracking to ensure that information on the Visitor's head rotation has a minimum delay through the network. The BEAMING project currently addresses three applications: A general purpose theatrical scene, a teaching situation and a medical patient-visiting-doctor scenario. The March 2012 project review deals with the teaching situation. This involves a single microphone recording followed by signal processing that reconstructs the spatial content. The Visitor is represented as a robot with a loudspeaker.

1 Introduction

Immersive experience represents one of the primary goals of modern communication technologies. Communication with the feeling of being together and sharing the same environment is required [1]. The 4 year EU project BEAMING [2] which started in 2010 (Framework Programme 7) is addressing the issue of improving immersive communication interfaces through telepresence. BEAMING is a collaborate research project where the goal is to give people a real sense of physically being present in a remote location with other people.

The section of Acoustics at Aalborg University (AAU) is involved in this project as the only project partner who deals with the acoustical aspects. Close technical partners are University College London (UCL) and Technical University Munich (TUM), who deal with vision and robotics/haptics respectively. Scuola Superiore Sant'Anna (SSSA), Pisa, is jointly responsible for system integration and Starlab, Barcelona, mainly holds the project coordinating role.

BEAMING includes cross-modal research. This is accommodated across different work packages (WPs) in order to balance the needed basic research with the technical requirements associated to the capturing and rendering of the space where the users of the BEAMING system are physically located. For example, WP1 deals with capturing and rendering of the destination and WP2 deals with capturing and rendering of the visitor.

Since BEAMING investigates communication between different locations, all modality data plus various extra controlling data must be transferred through a network. The network can be local (LAN) or the Internet, which in the

latter case poses problems about synchronization, delays and even lost data. The project has currently not reached the state where cross-modal synchronization has been thoroughly investigated, hence AAU/audio (and others) has mainly focused on absolute delays and data loss.

1.1 Roles

The user of the BEAMING telepresence systems can take one of three different roles, being either 1) Visitor, 2) Local or 3) Spectator. There can be one or more of the above, although normally there will be several Locals, typically forming a conference where they (the Locals) meet in person. The standard scenario is that the Locals (e.g. at a meeting) are *visited* by someone not physically present, i.e. the Visitor, and *observed* by someone not physically present, i.e. the Spectator.

The Visitor must be able to interact with the Locals, so they can see his gestures and facial expressions and hear his voice. For this to happen the Visitor must have either a physical/mechanical representation or a holographic representation of himself at the location of the Locals (the Destination). Such a representation (the Avatar) can be (simplest) a computer screen or (more advanced) a robot, depending on the application (see 1.2). In the physical case loudspeakers are mounted on the devices in order to support audio (see 2 for details). Presenting the Visitor to the Locals is handled by WP2.

Also the Locals must be presented to the Visitor in a convincing way. Since the Visitors are not physically present at the Destination, they will view the Locals through a head-mounted display (HMD) or use a CAVE. They will wear gloves to provide haptic stimuli and headphones and microphone for playback and capturing of sound. The concept is that the Locals primarily are situated in a “normal” room wearing as little equipment as possible, while the Visitors will be in less natural surroundings (the Transporter) wearing more complex and sofar heavy equipment. Presenting the Locals to the Visitor is handled by WP1.

The Spectator can only view and hear the scene but is not able to interact. His equipment could be a smartphone or a PC.

1.2 Applications

BEAMING currently addresses three applications. The applications are meant for testing and are not demonstrators which must all work at the end of the project. They are, 1) A general purpose theatrical scene, where a remote actor (Visitor) can practise a play together with another actor and a director who are both Locals, 2) A medical scenario where the patient visits the doctor. For the sake of the patient, he does not wear an HMD. Instead focus has been on haptic interaction and the doctor is simply displayed on a normal computer screen, 3) A teaching situation, exemplified during the March 2012 project review in Munich, as a teacher (the Visitor) training a novice (a Local) in playing the xylophone. The teacher's Avatar is a robot.

The following paragraphs will concentrate on application 3 because it entails specific acoustical challenges beyond those met in 1 and 2.

2 Audio in BEAMING

The perceived auditory realism in a telepresence system is important. It was decided to utilize binaural techniques [3,4] to provide spatial information to those users who need it. The binaural processing includes usage of digital filters known as Head Related Transfer Functions (HRTFs) which are sets of transfer functions that represent the change of sound from a given direction to the two human ears in a free field condition. Ideally the set of HRTFs should be measured for each individual who uses the system, but a database based on measurements on a dummy head is also applicable [5,6]. The HRTF database employed in this system is based on 2 degree angular resolution measurements of the “Valdemar” dummy head developed at AAU [7].

The BEAMING users who need synthesized spatial audio (3D sound) are the Visitors and the Spectators. The Locals get spatial sound automatically, either provided directly by other Locals or by the Avatar(s) as they physically move about at the Destination.

With only few modifications an audio setup can be created that covers all three BEAMING applications (see 1.2 and Figure 1) [8].

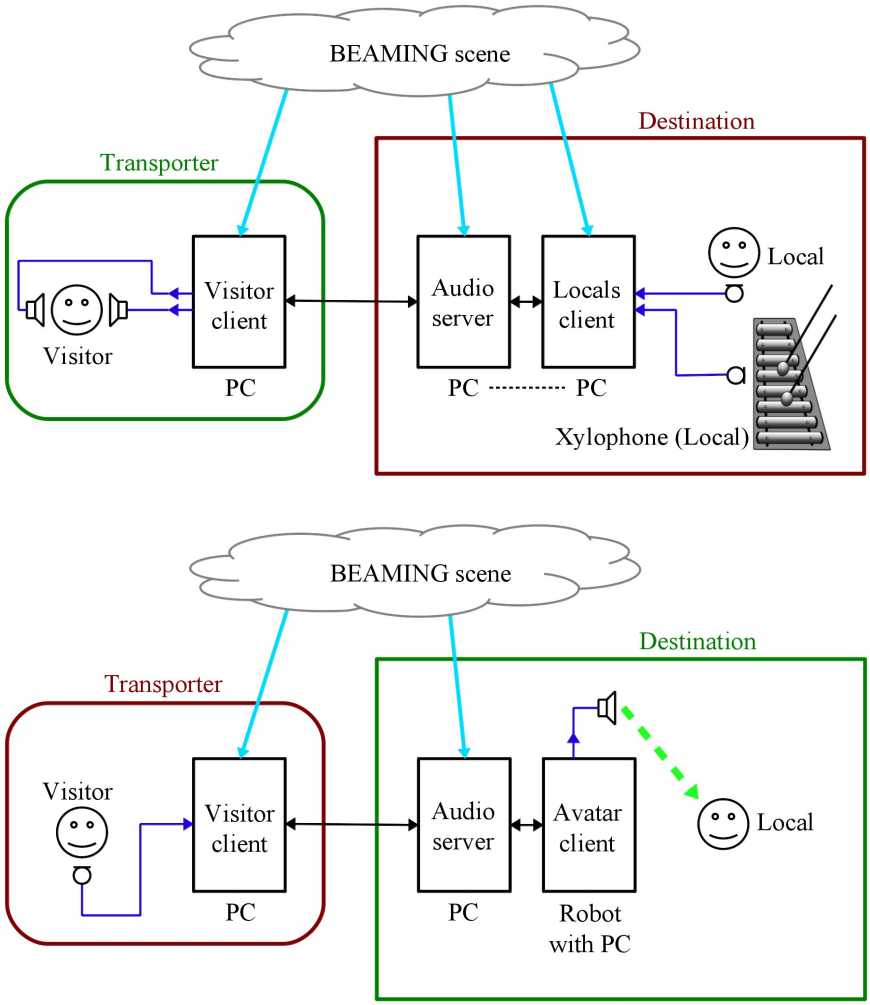

Figure 1: Top: Capturing and rendering of the Destination, WP1. Bottom: Capturing and rendering of the Visitor, WP2.

Figure 1 shows the setup demonstrated during the project review in Munich. It should be noted that the two figures must be viewed as a merger; for example the Visitor will wear both headphones and a microphone at the same time. For simplicity we shall mainly discuss the elements related to audio.

The central entity is the Audio server. It is responsible for providing an Internet IP to which the Visitor client can connect. Also a Locals client can connect and an Avatar client, which happens locally if the Audio server is physically placed at the Destination. There is one Visitor client per Visitor but several Locals must share a common Locals client because a multi-channel soundcard is installed on this PC and in order to keep the number of computers to a minimum.

The particular setup of Figure 1 is an example of the 3rd application where the second Local is the xylophone and not a person. The Visitor is the teacher who visits the student (Local) and give instructions in playing. The Local wears a close-up microphone and the xylophone is captured with a studio microphone on a stand. Thus the teacher can hear both the student and the xylophone as 3D sound. The 3D sound is generated by the Visitor client where the HRTF database and convolution software is placed.

During the review the teacher (Visitor) was represented at the Destination as a Kali-type robot. It is capable of showing various facial expressions and move its arms so that it can physically play the xylophone. A loudspeaker is mounted below the robot's head which makes the Visitor able to speak via the robot, that is the teacher can both show how to play and give instructions to the student. The loudspeaker chosen is a spherical closed-box with a 2” unit connected to the soundscard on the robot. The Visitor wears a close-up microphone.

Tracking is needed both for the Visitor and for the Locals. The tracking information which contains position and rotation of everyone involved is continuously stored in a common Internet database known as The Beaming Scene. In the case of audio it is mainly used by the Visitor client. By combining positions of Locals (incl. the xylophone), the

position/rotation of the Avatar (robot) and the rotation of the Visitor himself, it is possible to select the correct HRTF and produce corresponding 3D sound to the Visitor. This means that currently no 3D/binaural sound needs to be passed through the network. All filtering by HRTFs is done "locally" at the Transporter side.

3 Discussion

A system for 3D sound in the telepresence project BEAMING has been developed. Its current state is a prototype which mostly needs finish, e.g. with the graphical user interface, the implementation of a Spectator role (see 1.1) and bug tracking. With two years left of the BEAMING project, the software, and the whole audio concept used requires testing in terms of timing and delay, synchronization with other modalities and potential data loss through the network connections, as well as experiments with human subjects.

Optimization was carried out in transferring sound from the xylophone (see 2). Since a xylophone is a distributed sound source it cannot acoustically be represented by the signal filtered with a single HRTF (as currently is the case of the Locals). Instead the xylophone was captured with one studio microphone. This signal was sent via the Audio server to the Visitor client, where the spatial and distributed nature of the xylophone would be recreated and rendered employing several HRTFs. The advantage is that only a single audio channel must be passed through the network. The algorithm is reported by Marković et al [9].

References

[1] Y. Huang, J. Chen and J. Benesty, Immersive audio schemes, *Signal Processing Magazine, IEEE*, 28(1), 2011, 20-32.

[2] BEAMING Project, the official website http://beaming-eu.org/

[3] H. Møller, Fundamentals of Binaural Technology, *Applied Acoustics*, 36(3/4), 1992, 171-218.

[4] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, The MIT Press, 1983.

[5] H. Møller, M. F. Sørensen, C. B. Jensen and D. Hammershøi, Binaural technique: Do we need individual recordings?, *Journal of Audio Engineering Society*, 44(6), 1996, 451-469.

[6] D. R. Begault, E. M. Wenzel, A. S. Lee and M. R. Anderson, Direct comparison of the impact of head tracking, reverberation, and individualized Head-related Transfer Functions on the spatial perception of a virtual speech source, *Proceeding of 108th Audio Engineering Society Convention*, Paris, 2000, 1-19.

[7] B. P. Bovbjerg, F. Christensen, P. Minnaar and X. Chen, Measuring the head-related transfer functions of an artificial head with a high directional resolution, *Proceedings of 109th Audio Engineering Society Convention*, Los Angeles, 2000, 1-17.

[8] E. Madsen, S. K. Olesen, M. Marković, P. F. Hoffmann and D. Hammershøi, Setup for demonstrating interactive binaural synthesis for telepresence applications, *Proceedings of Forum Acusticum 2011*, Aalborg, 1281-1286.

[9] M. Marković, E. Madsen, S. K. Olesen, P. F. Hoffmann and D. Hammershøi, BEAMING teaching application: recording techniques for spatial xylophone sound rendering, To be presented at *Acoustics 2012*, Hongkong.