Audio-Visual Experiments on a Life-Size Videoconferencing System Combining Spatialized Audio and 2D Video Projection

Werner P.J. de Bruijn[1], Marinus M. Boone[2]

[1] *Philips Research, Eindhoven, The Netherlands, Email: werner.de.bruijn@philips.com*
[2] *Delft University of Technology, Delft, The Netherlands, Email: rinus@akst.tn.tudelft.nl*

Introduction

Spatialized reproduction of audio in an audio-visual system, such as a life-size videoconferencing system, can in general greatly enhance the feeling of 'presence'. Additionally, by reproducing speech signals in a spatialized way, the intelligibility of the speech can be improved and the identification of individual speakers when multiple speakers are present can be facilitated for the listeners.

Wave Field Synthesis

A very realistic spatialized audio reproduction, including a realistic reproduction of acoustic source distance, can be obtained with the technique of *Wave Field Synthesis* (WFS), which uses arrays of closely spaced loudspeakers to generate a desired sound field within an extensive listening area [1]. Because of its excellent capabilities for spatial reproduction of sound sources, this technique seems ideal for implementation in a high-end videoconferencing system of which the goal is to achieve a sound reproduction that is as realistic as possible.

Combining 3D Audio with 2D Video

Ideally, a high-quality spatial audio reproduction as the one achieved with WFS would be combined with three-dimensional video projection. Unfortunately, such systems are not yet available for implementation in a real-time videoconferencing system, so in practice conventional two-dimensional video projection will be used. However, the combination of sound reproduction that includes depth with conventional 2D video projection has effects on the resulting audio-visual experience of observers. The exact nature of these effects is not easy to predict, because of the audio-visual interactions that are involved. Also, these effects are not necessarily all beneficial for the overall perceived audio-visual quality. Specifically, a mismatch between perceived auditory and visual source directions may occur for observers that are not in the correct, unique, viewpoint of the 2D video projection. This is illustrated in Figure 1. While the sound source will be localized at the true source position for all listening positions, the perceived position of the corresponding visual source depends on the position from which the 2D projection is viewed.

This paper presents a series of subjective experiments that have been carried out to investigate these effects in the context of a life-size videoconferencing system. Issues that were investigated were the subjective correspondence of auditory and visual source positions, source identification performance and overall realism of the audio-visual scene. More details about all experiments described in this paper can be found in [2].

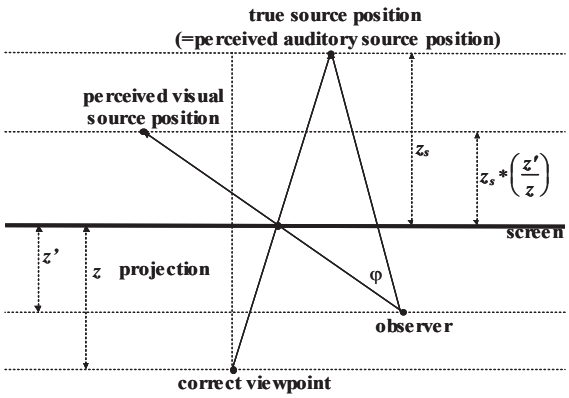

Figure 1: Illustration of the mismatch between perceived auditory and visual source positions for an observer who is not located at the viewpoint of the 2D video projection.

Correspondence of Auditory and Visual Source Positions

In this first series of experiments, we wanted to investigate if perceptible discrepancies of auditory and visual source positions for non-viewpoint observers, as described above, actually occur in practical situations.

Set-Up and Source Material

The set-up for all experiments described in this paper consisted of an acoustically transparent screen on which a life-sized perspective still image of three people standing at different positions in a room was projected. The (virtual) visual source positions relative to the center of the screen are shown in Figure 2. The viewpoint of the projection was located at (0, 2.74) m. A horizontal loudspeaker array was positioned behind the screen, consisting of 32 small loudspeakers with a spacing of 12.7 cm.

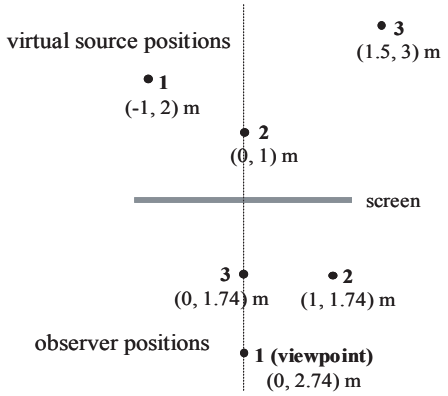

Figure 2: Source- and observer positions used in the experiments.

Given a desired source position and input signal, a DSP system generated the individual loudspeaker signals according to WFS theory, such that the input signal was reproduced by the array as a virtual point source at the desired position. The audio input signal was a mono male speech signal.

Experiments

First an experiment was conducted to determine, for each of the three observation positions shown in Figure 2, the range of lateral audio source positions that corresponds subjectively to a certain (virtual) visual source position. The procedure was as follows: the subject was seated at one of the three observer positions and was instructed which of the three visual sources was the target source. The audio source signal was reproduced as a virtual source at a random initial lateral position at the distance of the target source. The subject was then asked to position the sound source (using a graphical computer interface) such that it corresponded best to the visual scene. The results are summarized in Figure 3.

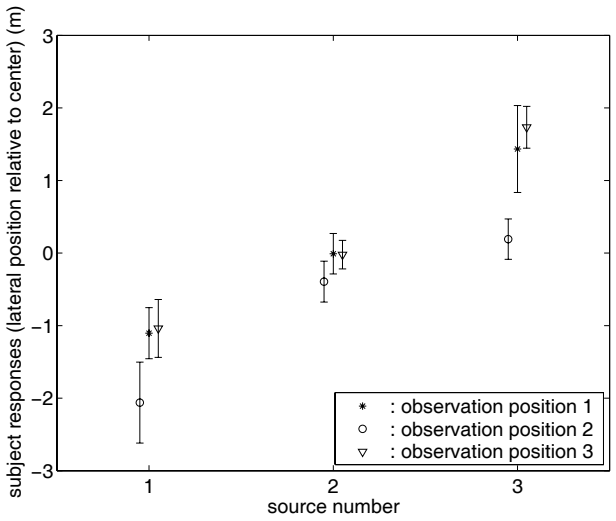

Figure 3: Summary of the results of the lateral source positioning experiment for the 3 observer positions. Horizontal axis: source number (see Figure 2). Vertical axis: lateral position (m).

From Figure 3 we see that for source numbers 1 and 3 the range of the preferred audio source position for the off-axis observer position 2 has no overlap with the corresponding preferred range for the other two observation positions. This means that it is not possible to position sound sources 1 and 3 at such positions that no observer perceives a discrepancy with the position of the corresponding visual sources.

To investigate the subjective degree of annoyance when a discrepancy *is* perceived, a next experiment was carried out in which subjects graded how annoying the discrepancy was that they perceived when observing an audio-visual scene with the audio source at the correct, original, source position as shown in Figure 2. A 5-point scale was used, extending from *no discrepancy* to *very annoying discrepancy*. The main result of this experiment was that for the laterally displaced observer position 2, a serious degradation of the subjective spatial A/V correspondence was found, compared to the results for observer positions 1 and 3.

Multiple-Source Experiments

We also wanted to investigate the subjective performance of the videoconferencing system when several speakers are active simultaneously, in terms of speaker identification and overall realism of the A/V scene. The set-up and visual source material for the experiments described in this section were identical to those from the previous section. The audio source material consisted of three mono signals: one male speech signal and two female speech signals.

Experiments

The three speech signals were distributed randomly over the three visual sources and reproduced by the WFS array as virtual point sources from the correct positions, as shown in Figure 2. The subjects' task was to identify the visual source that corresponded to the male speech signal. Results showed that identification performance was significantly degraded for observer position 2, dropping from 98% correct identification at the correct viewpoint to 71%, which confirms the conclusion of the single-source experiments.

In a final experiment, using the same set-up as in the previous experiment, the subjects' task was to grade the realism of the A/V scene in terms of overall spatial correspondence between what they heard and what they saw. Results again confirmed the conclusions from the previous experiments.

Auditory Depth Compression

A simple way to avoid or reduce the discrepancies that were found in the experiments is to compress the perspective of the audio reproduction to some extent, by pulling all sound sources somewhat closer to the screen in the direction of the viewpoint. A compromise must be found between avoidance of discrepancies and retaining the benefits of audio reproduction that includes depth, like improved speech intelligibility. For details the reader is referred to [3].

Conclusion

In A/V systems combining 2D video and audio that includes correct reproduction of distance, discrepancies occur between perceived source directions in the two modalities for observers that are laterally displaced from the viewpoint of the video image. This discrepancy can be avoided or reduced by proper compression of the audio perspective.

References

[1] A.J. Berkhout, D. de Vries and P. Vogel, *Acoustic control by Wave Field Synthesis*, J. Acoust. Soc. Am. **93**, pp. 2764–2778, 1993.

[2] W.P.J. de Bruijn and M.M. Boone: *Subjective Experiments on the Effects of Combining Spatialized Audio and 2D Video Projection in Audio-visual Systems*, Proc. 112th AES Convention (paper 5582), Munich, 2002.

[3] W.P.J. de Bruijn and M.M. Boone: *Application of Wave Field Synthesis in Life-size Videoconferencing*, Proc. 114th AES Convention (paper 5801), Amsterdam, 2003.