

# The Ventriloquist Illusion in Audiovisual Virtual Reality

Lukas Vollmer<sup>1</sup>, Janina Fels<sup>1</sup>

<sup>1</sup> *Institute for Hearing Technology and Acoustics, 52074 Aachen, Germany, Email: {lvo, jfe}@akustik.rwth-aachen.de*

## Introduction

Humans do not perceive their surroundings as distinct events in different sensory modalities. Instead, sensory information about a common source is integrated to a combined percept. Early demonstrations of these integrations are the Ventriloquist illusion [1], which describes the mislocalization of a sound towards a visual stimulus and the McGurk effect [3] that can be observed when the vocalization of a phoneme is presented with a mismatched mouth movement.

With the increasing demand for ecological validity of experimental environments, audiovisual integration plays an important role in various fields of research. However, setting up experiments in real environments can be demanding and costly while limiting experimental control at the same time. An elegant solution to these drawbacks is the use of virtual reality. Highly complex visual scenes can be created and reproduced in a head-mounted display (HMD) accompanied by spatial sound reproduction using binaural technology or loudspeaker-based methods. On the other hand, virtual reality technology imposes restrictions on the ecological validity by the HMD's limited screen resolution and field of view, and by shortcomings of spatial sound reproduction with respect to coloration and localization accuracy or spatial blur.

To address these shortcomings in terms of localization, the susceptibility of the ventriloquist illusion to increased source spread as a result of panning based sound reproduction was investigated by means of a comparison between real and virtual sound sources.

In the following sections the experimental design will be presented including short descriptions of the acoustic and visual reproductions as well as the experimental procedure. Afterwards the results are reported with special focus on the effects of reproduction methods and audiovisual offsets on localization errors. Finally, the results are discussed and conclusions for future work are drawn.

## Methods

### Experimental Design

A classical Ventriloquism paradigm [1] was setup involving three horizontal sound source positions  $\varphi = \{-40^\circ, 0^\circ, 18^\circ\}$  (left to right) and five audiovisual offsets  $\Delta\varphi = -20^\circ$  to  $20^\circ$  in  $10^\circ$  steps between the acoustic and visual stimulus. As previous studies on Ventriloquism typically used pure tones or clicks in contrast to pulsed noise which is commonly applied in localization tasks, both stimulus types are included in this study. Therefore, the acoustic stimuli consisted of three pulses of either pink noise or a 600 Hz sine tone of 300 ms duration and 10 ms on- and offset ramps. Silent intervals between

pulses had a duration of 150 ms. The visual stimulus consisted of a Gaussian blob ( $\sigma = 4.156$  cm) presented on a curved screen at a radius of 1.45 m in the virtual scene. The blinking pattern of the blob was equal to the temporal dynamics of the pulses and stimulus onsets were perceptually aligned and kept constant for all participants.

The experiment was conducted in the anechoic chamber of the Institute for Hearing Technology and Acoustics, RWTH Aachen University which is equipped with the surrounding 68-channel loudspeaker array SCaLAr [4]. As the aforementioned source positions are not covered by SCaLAr except for the  $0^\circ$  source, two additional loudspeakers at  $\varphi = \{-40^\circ, 18^\circ\}$  were mounted to the array, equalized and time-aligned.

### Acoustic Reproduction

Three panning-based spatial sound reproduction methods were compared to the reproduction using real loudspeakers.

Vector base amplitude panning (VBAP) [6] is the conceptual extension of Stereo panning to arbitrary sound source positions. The convex hull of a surrounding loudspeaker array is calculated and used to determine the active loudspeaker triangle by intersecting the sound incidence direction with the convex hull. Channel weights are found by calculating the linear combination of loudspeaker positions that results in the intersection point.

It is apparent that this approach results in a variable number of active loudspeakers which leads to a variable source width. Multiple direction amplitude panning (MDAP)[5] aims to reduce this variability by introducing additional spreading sources around the intended source direction. Ten sources were arranged in a circle with an "angular diameter" of  $\alpha_{\text{spread}} = 10^\circ$ .

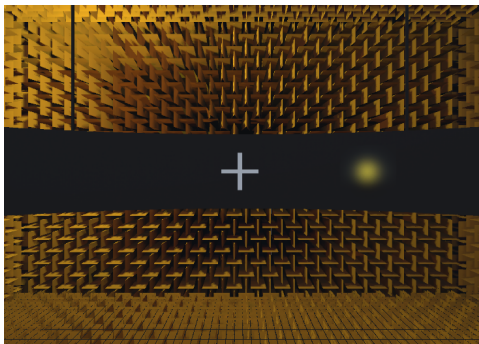
The third method is higher-order Ambisonics (HOA) [8]. Here, the soundfield of a source is encoded in spherical harmonics and decoded into channel weights by a decoder matrix designed specifically for the surrounding loudspeaker array. The implementation of the decoder matrix is in accordance with the all-round Ambisonics decoder [7] ( $N = 7$ ) using a 15-design with 120 sampling points [2].

The hardware setup for the acoustic reproduction was equal to that described in [4] and the auralization software received scene updates by TCP/IP connection from an additional computer which was responsible for visual reproduction.

### Visual Scene

A room model was designed to replicate the anechoic room in which the experiment was conducted. To avoid

visual cues by SCaLAR's loudspeakers the array was not included in the model. Instead, a curved screen (angular width: 100°, height: 30 cm) was added to the scene for plausible presentation of the visual stimuli (c.f. Fig.1).



**Figure 1:** Example rendering of the visual scene including the fixation cross (light gray) and the visual stimulus (yellow).

The visual scene was reproduced using Unity™ (Unity Technologies, 2019) on the additional computer (Intel® Core™ i7-7700, 16 GB RAM, NVIDIA® GeForce GTX™ 1080) driving an HTC VIVE Pro Eye™ (HTC Corporation) HMD which provides eye tracking capabilities. Due to the HMD's limited field of view, left-sided offsets of the visual stimulus to the acoustic source position at  $\varphi = -40^\circ$  were mirrored to the right side, limiting the analysis of the results to absolute angular offsets.

### Motion Tracking

Instead of using the HTC VIVE's built-in tracking system, external tracking by an OptiTrack (NaturalPoint Inc.) optical tracking system was used which allows for the same calibrated tracking area for all participants. Reflective markers were attached to the controller and HMD, which required additional alignment of the tracked positions in the VIVE's and OptiTrack's coordinate spaces.

### Experimental Procedure

In the beginning of the experiment participants were seated inside SCaLAR and aligned to the array's center with the aid of a two-axes laser. Participants were explicitly instructed to localize the acoustic source in the horizontal plane and ignore the visual stimulus. Furthermore, they were informed that audiovisual stimuli could, but do not have to be spatially aligned. Head movements were not restricted.

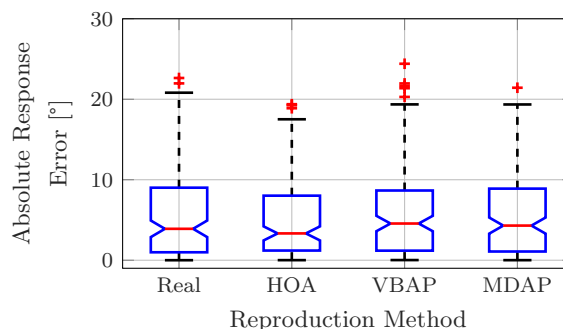
Before each trial, participants oriented their gaze towards a fixation marker. If the gaze remained on the marker for two seconds the trial started. Participants were instructed to keep looking at the marker during stimulus presentation. Responses were given with the aid of a visual pointer that was deactivated during stimulation. Response positions were calculated by projecting the controller position and direction of the visual pointer onto the horizontal plane and finding the nearest point where the visual pointer intersects the circle with radius 1.45 m around the coordinate space's origin.

## Results

In the following, the pilot-study results of five normal-hearing participants (three male, mean age 29.8 yrs within 23 yrs - 38 yrs) are reported. Due to the small sample size, no statistical analysis is performed. Instead, the feasibility of the hardware setup and experimental design are of interest.

### Reproduction

Fig. 2 depicts the absolute angular response error split into the four sound reproduction conditions. Judg-

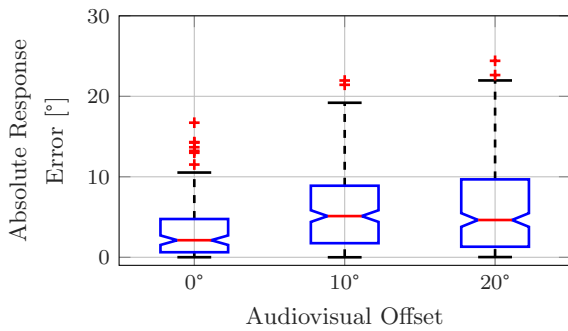


**Figure 2:** Boxplot of absolute response errors for each reproduction method. Two outliers of the *Real* condition at  $\approx 60^\circ$  are not shown. A clear tendency that panning-based sound reproduction increases the localization error cannot be observed.

ing from the median values (red lines) and interquartile ranges (blue boxes) there is no increasing tendency in absolute angular errors for the panning-based reproductions compared to real sources. While this result seems unexpected, considering SCaLAR's high loudspeaker number offers an explanation based on the rather dense placement of the loudspeakers, thus limiting the source spread of VBAP. For HOA however, the result indicates a sufficient localizability of virtual sound sources reproduced with seventh-order Ambisonics even if a conflicting visual stimulus is introduced.

### Offset

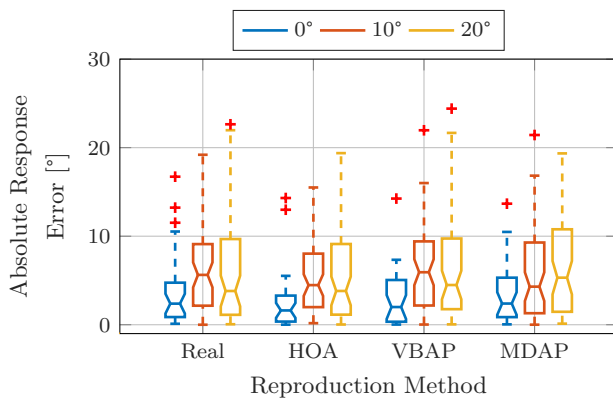
While different reproduction methods did not result in increased angular response errors, the influence of the visual stimulus on sound localization can be analyzed. Shown in Fig. 3 are the response error boxplots grouped by the absolute angular offset between acoustic and visual stimulus position. An increased median response error can be observed for both offset conditions, however there is a slight decrease of the median error of the 20° condition compared to 10° offset. Furthermore, the interquartile range shows an increase between both offset conditions. As expected, these results indicate the occurrence of Ventriloquist illusions irrespective of the reproduction method. Both the decreased median error and increased interquartile ranges indicate an exceeding of the spatial integration window which was reported to be  $\approx 15^\circ$  [1]. By exceeding the integration window the perceived sound source location is less likely to be affected by the visual stimulus. However, if the illusion occurs responses will be further off and thus increase the interquartile range.



**Figure 3:** Absolute angular response error grouped by absolute offset between acoustic and visual stimulus. Two outliers, one for each offset condition, are not shown. While the median error increases for 10° offset and slightly decreases afterwards, the interquartile range increases with the offset.

### Reproduction and Offset

While no global difference between reproduction methods is observed, Fig. 4 offers a more detailed view on the interaction of reproduction and offset. For Real, HOA

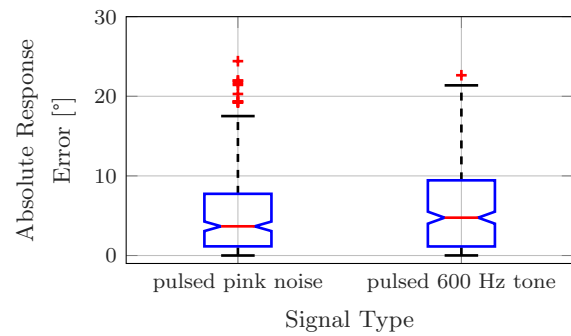


**Figure 4:** The angular response error is now grouped by reproduction method and offset. While *Real*, *HOA* and *VBAP* show the median error decrease observed above, *MDAP* shows an increase of the median error.

and *VBAP* reproduction the same trend that is depicted in Fig. 3 can be observed. *MDAP*, however, shows a further increase of the median response error for 20° offset compared to 10°. This result can be explained by the design of *MDAP* aiming at uniform source spread for all sound source positions at the expense of generally increased source widths. However, more data for a statistical analysis is needed, which would allow statements if this result suggests an increased susceptibility to Ventriloquism for wider source widths, or is an artifact of the small sample size. On the other hand, *HOA* and *VBAP* appear to be suitable candidates for the realistic reproduction of audiovisual scenes, at least from a localization point of view.

### Signal Type

Finally, the response errors depending on the signal type are shown in Fig. 5. The slight increase in the pure tone's median response error and interquartile range indicate better localizability for the noise stimulus.



**Figure 5:** Effect of signal type on the response error. Again, two extreme outliers are not shown. Both median error and interquartile range are increased for the pure tone compared to the noise signal.

## Discussion and Conclusion

Firstly, the presented two computer solution to conduct audiovisual experiments with computationally demanding acoustic reproduction methods such as HOA was successfully implemented and tested using a classical Ventriloquism paradigm.

Secondly, the results presented in the previous section suggest, that the Ventriloquist illusion can be replicated in audiovisual virtual reality by similarly simple and abstract means as has been demonstrated in previous research [1]. However, a tendency that the panning-based reproduction methods influence the Ventriloquist illusion could not be found. Although an increasing trend in response errors with wider audiovisual offsets was observed for *MDAP*, further data is needed to prove the presence of this trend in a statistical manner. One contributing factor to the lack of influence is the high loudspeaker placement density of *SCaLAR*, which effectively limits the source spread of *VBAP* and thus the hypothesized susceptibility of the source localization to simultaneously but spatially offset presented visual stimuli. Furthermore, the similarly limited influence of visual stimuli on the localization of seventh-order-Ambisonics-reproduced stimuli suggests high spatial fidelity of the reproduction.

On the other hand, these results were achieved with a low number of participants and should be treated with care. Further experiments will be necessary to solidify and validate the result and conclusions. Future work should also focus on precise synchronization of audiovisual stimuli.

### Acknowledgement

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 368482240/GRK2416. Special thanks goes to Chalotorn Möhlmann for improving the room model of the anechoic chamber and to Rolf Kaldenbach and Uwe Schlömer for extensive technical support.

### References

- [1] Chen, L., and Vroomen, J.: Intersensory binding across space and time: A tutorial review. *Attention, Perception, & Psychophysics* 75, 5 (2013), 790-811

- [2] Hardin, R. H., and Sloane, N. J. A.: McLaren's improved snub cube and other new spherical designs in three dimensions. *Discrete & Computational Geometry*, 15, 4 (1996) 429-441
- [3] McGurk, H., and MacDonald, J.: Hearing lips and seeing voices. *Nature* 264, 5588 (1976), 746-748
- [4] Pausch, F., Behler, G., and Fels, J.: SCaLAr - a surrounding spherical cap loudspeaker array for flexible generation and evaluation of virtual acoustic environments. *Acta Acustica* 4, 5 (2020)
- [5] Pulkki, V.: Uniform spreading of amplitude panned virtual sources. *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (1999), 187-190
- [6] Pulkki, V.: Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society* 45, 6 (1997), 456-466
- [7] Zotter, F., and Frank, M.: All-round ambisonic panning and decoding. *Journal of the Audio Engineering Society* 60, 10 (2012), 807-820
- [8] Zotter, F., and Frank, M.: *Ambisonics*. Springer International Publishing, 2019