Naturalistic audio-visual volumetric sequences dataset of sounding actions for six degree-of-freedom interaction

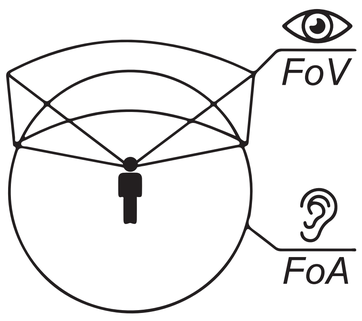
Hanne Stenzel, Davide Berghi, Marco Volino, and Philip J.B. Jackson

hanne.stenzel@iis.fraunhofer.de {d.berghi, m.volino, p.jackson}@surrey.ac.uk

Work supported by InnovateUK (105168) 'Polymersive: Immersive video production tools for studio and live events'.

Introduction

Sight and **hearing**, together, are primary means for experiencing real, virtual and mixed realities around us



Existing datasets do not reflect bimodal integration of the senses, as the data:

- address **specific** technical problems (e.g. only hand or body movements)
- are limited to **one modality** (i.e. audio only, or video only)
- **lack quality**, esp. of audio recordings

We present the **NAVVS** dataset:

- volumetric **3D-video** data high-quality **audio**
- **naturalistic** actions
- semantic and acoustic-feature **coverage**
- **freely-accessible** for research

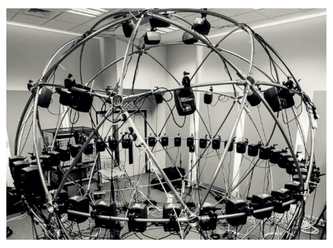
Use cases

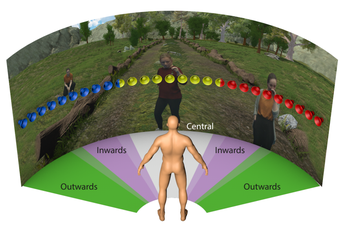
NAVVS dataset supports:

- technical and scientific studies VR/AR/XR
- 2D or 3D visual displays
- 3DoF or 6DoF interaction

For perceptual evaluations:

- transmission/rendering quality, localization, rendering techniques
- audio-visual congruence, synchronization, spatial alignment

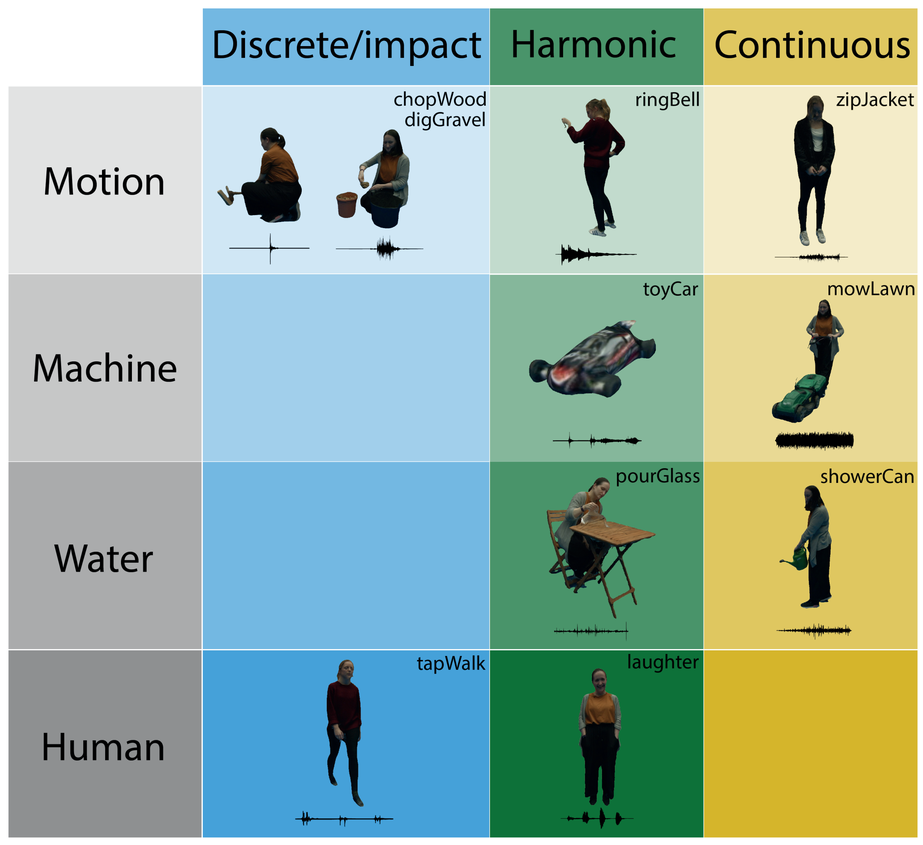




Design of dataset

- **Aim**: to enable subjective perceptual evaluation
- **Semantic categories** distinguish sound sources as linked to brain activity
- **Acoustic feature classes** describe sound character and relate to localisability

10 selected 2-s **scenes**
4 same-person **repetitions**
40 volumetric **sequences**



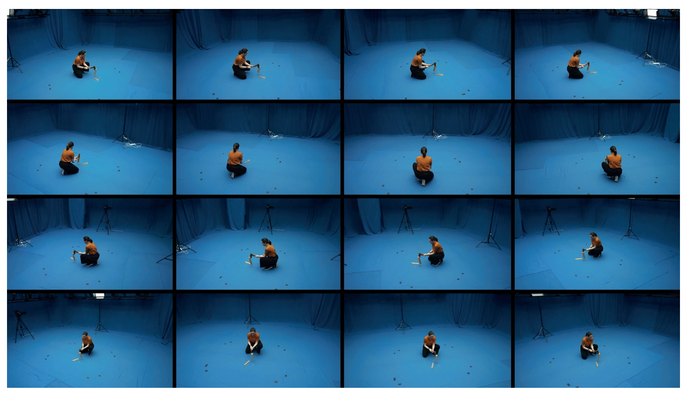
Data capture & processing

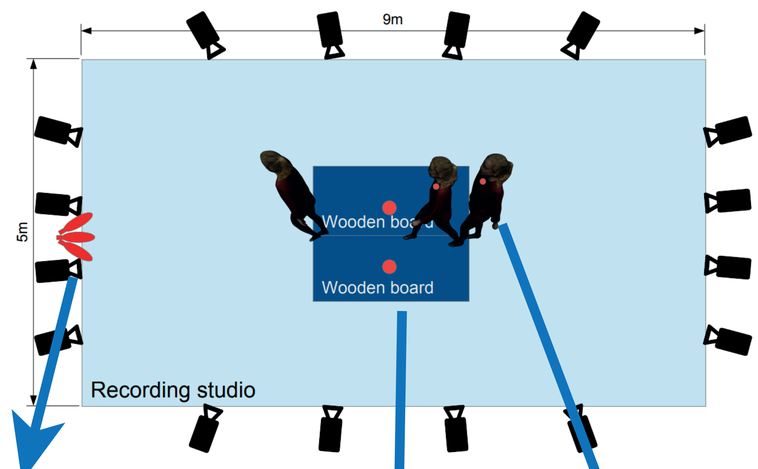
Scan to visit the website

Video

- 16 UHD video streams (30 fps)
- Background masks using chroma keying
- Foreground volumes recovered using shape-from-silhouette
- Volumetric graphs cut refines the visual hulls

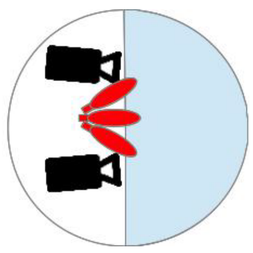
3D foreground mode per frame with UV texture atlasees





Audio

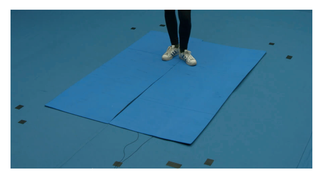
- Multimicrophone recording technique
- De-noise and de-crackle filtering(Izotope RX7)
- **Stereo-mix with left-right motion & distance cues**
- **Spot mic raw audio**



Shotgun microphones directed at actor

Lavalier clip-on mic

Wooden board with DPS contact mics
OPtional boom mic



Conclusion

- Audio-visual dataset needed for realistic tests & multimodal VR/AR/XR research
- We present **NAVVS** short volumetric action sequences, designed with semantic and acoustic diversity for technical evaluation and scientific perceptual studies
- Future work may compare threshold across scenes, add metadata & sequences
- For more info, visit **cvssp.org/data/navvs/**.

References

[1] A. Chatzitofis *et al*. HUMAN4D: A human-centric multimodal dataset for motions and immersive media. *IEEE Access*, 8:176241-176262, 2020. doi: 10.1109/access.2020.3026276.

[2] H. Stenzel & P.J.B. Jackson. Perceptual threshold of audio-visual spatial coherence for a variety of audio-visual objects. In *AES Int. Conf. on Audio for Virtual and Augmented Reality*. Redmond WA, USA, 2018.

[3] H. Joo *et al*. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Trans. PAMI*, 41(1):190-204, 2017.

[4] J. Starck & A. Hilton. Surface capture for performance-based animation. *IEEE. Comput. Graph*., 27(3):21-31, 5 2007, doi: 10.1109/MCG.2007.68.