University of Huddersfield

Capturing and Rendering 360° VR Audio using Cardioid Microphones

Hyunkook Lee

h.lee@hud.ac.uk

Applied Psychoacoustics Lab (APL)

University of Huddersfield, UK

Motivation

University of HUDDERSFIELD

- Near-coincident mic arrays
  - ORTF, NOS, etc.
  - Arguably, preferred to pure coincident or pure spaced techniques by most professional recording engineers.
  - Rely on the trade-off between Time and Level differences.
  - Best of both worlds (Localisability & Spaciousness).

- Cardioid microphones
  - Most popular.
  - Most widely available.

- Record for VR using favourite cardioid mics arranged in a near-coincident fashion?

Contents

University of HUDDERSFIELD

- Research background
- Localisation test in loudspeaker reproduction
- Localisation test in binaural reproduction
- Discussion
- Summary

Research Background

Existing methods for VR audio capture

University of HUDDERSFIELD

- First Order Ambisonics (FOA)

Pros	Cons
• Very good "localisability" due to the coincident nature (But not necessarily good localisation "accuracy"). • Virtual microphones from flexible decoding. • Compact.	• High interchannel correlation. • Lack of spaciousness. • Comb-filtering and rapid change in image position even with a small head movement.

Existing methods for VR audio capture

University of HUDDERSFIELD

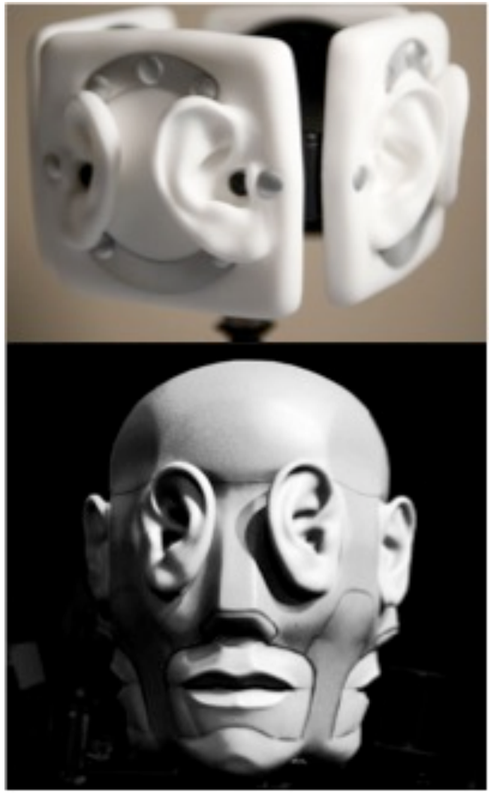
- Higher Order Ambisonics (HOA)

Pros	Cons
• Higher spatial resolution. • More accurate localisation.	• Requires a large number of channels for a proper decoding. $N = (M + 1)^2$ • Very expensive. • Tonal quality. • Spaciousness?

Existing methods for VR audio capture

University of HUDDERSFIELD

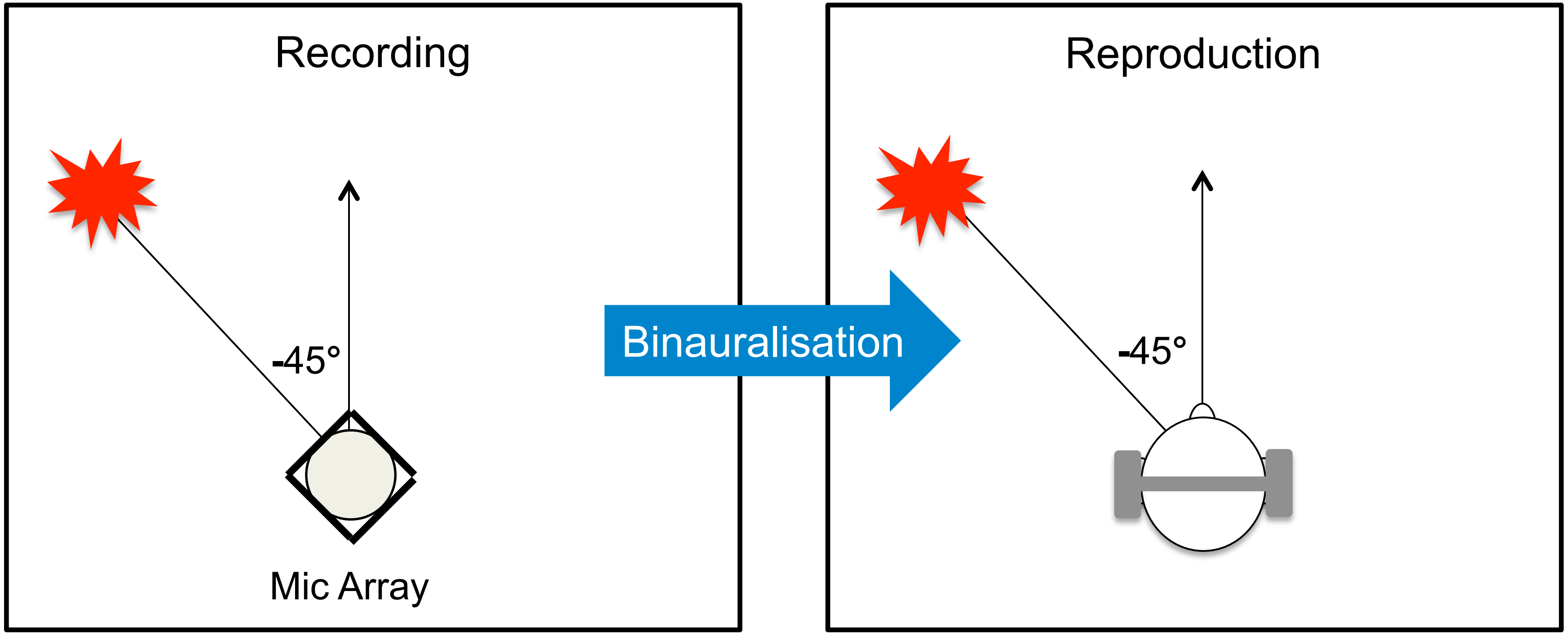
- Quad Binaural



Pros	Cons
• Direct pinnae filtering. • No need for extra binaural synthesis.	• Inaccurate localisation and comb-filtering due to crossfading between ear signals. • Not possible to use personal HRTFs. • Only for horizontal head rotation. • Expensive.

Psychoacoustic considerations for VR

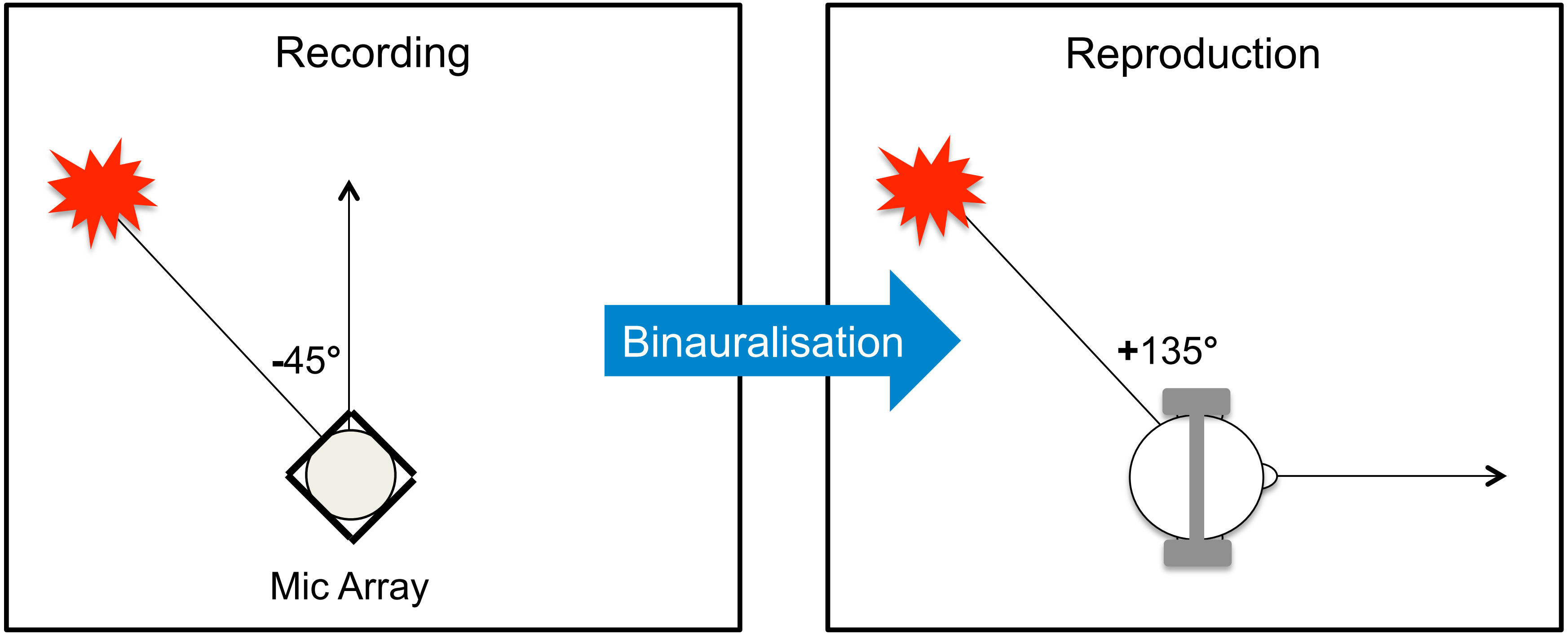
- In VR, it is important to match the actual and perceived source positions.



Psychoacoustic considerations for VR

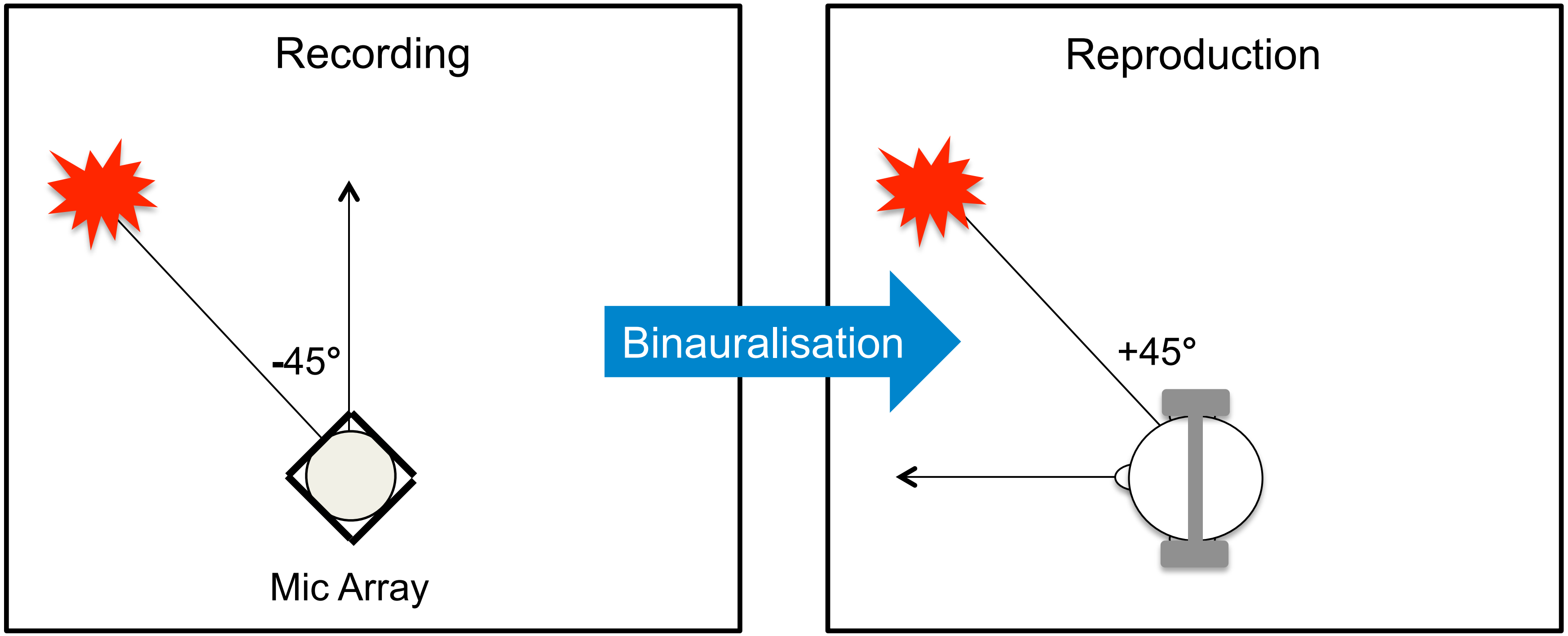
University of HUDDERSFIELD

- The perceived source position should stay the same as the head rotates.



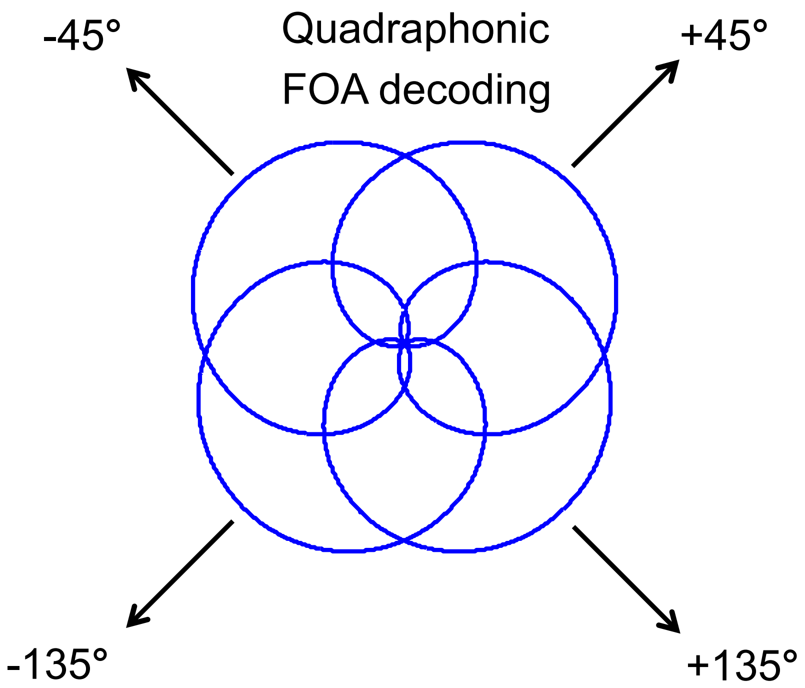
Psychoacoustic considerations for VR

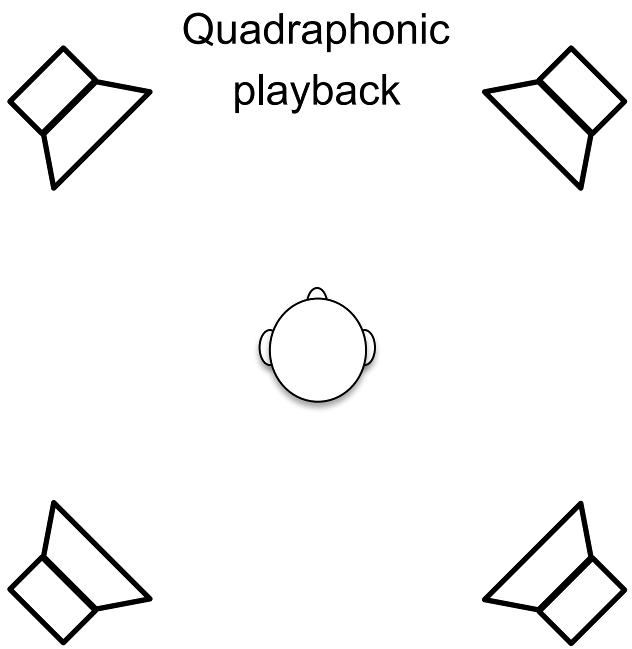
- The perceived source position should stay the same as the head rotates.



Psychoacoustic considerations for VR

- Limitation of FOA
  - Quadraphonic Cardioid decoding.

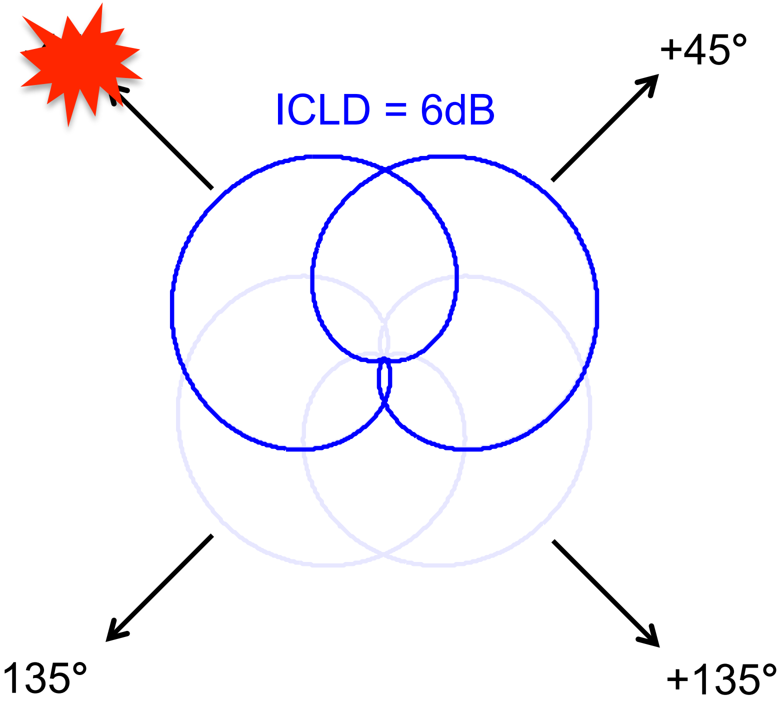


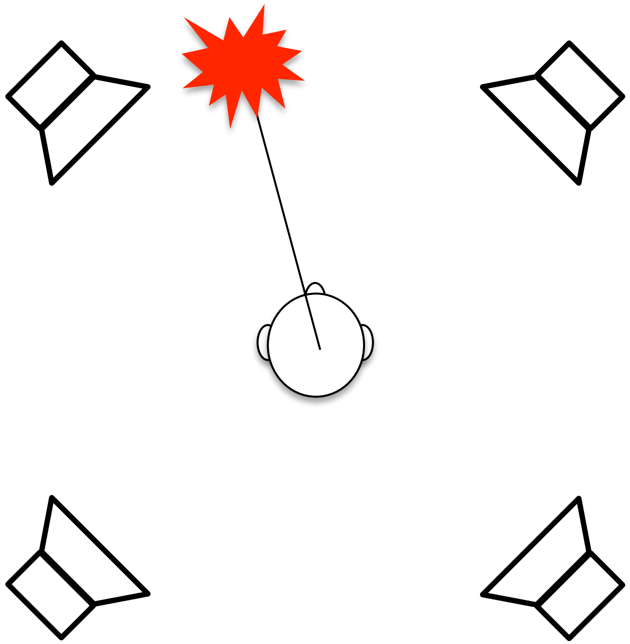


Psychoacoustic considerations for VR

- Limitation of FOA
  - Only 6dB ICLD (interchannel level difference) for the front pair for a source at 45°.

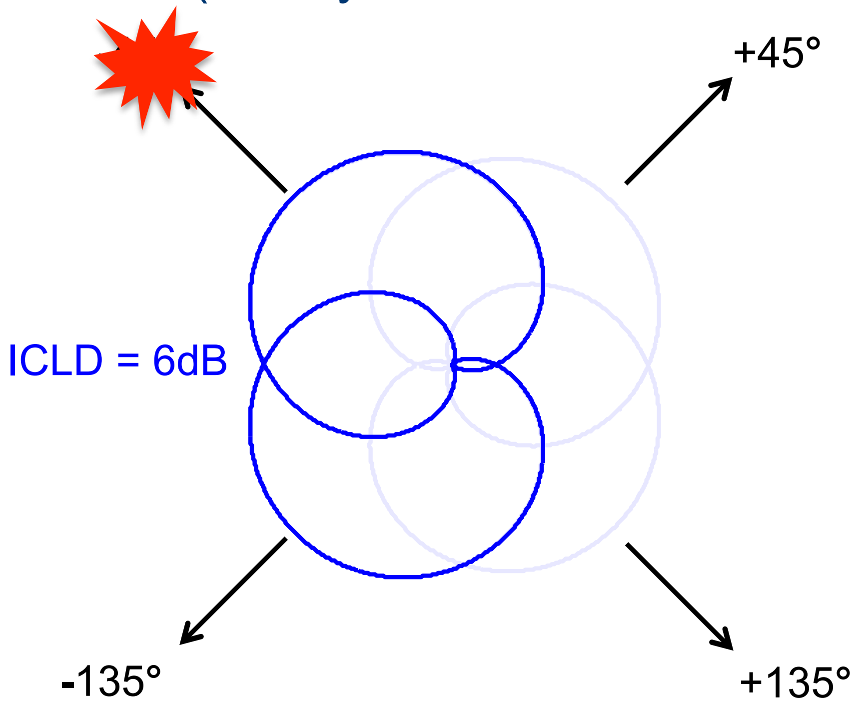
→ Not sufficient for a full phantom image shift to 45°.

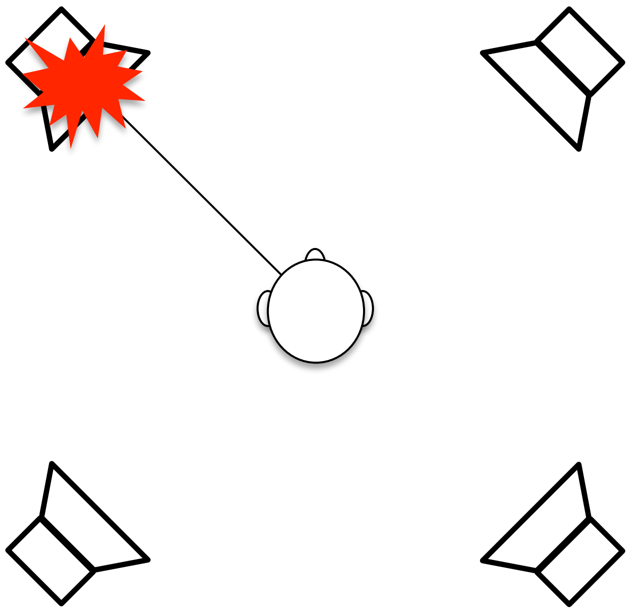




Psychoacoustic considerations for VR

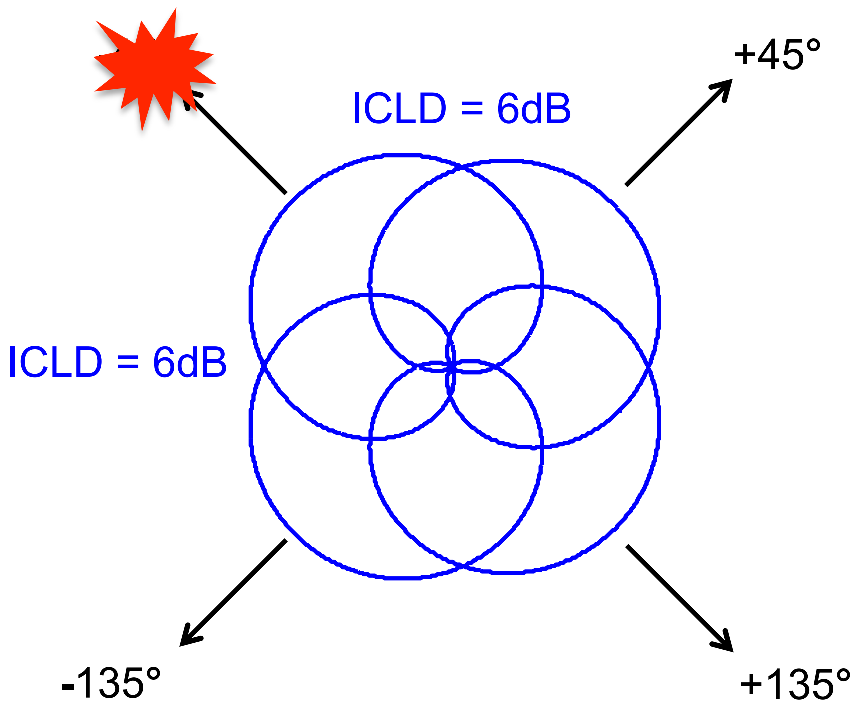
- Limitation of FOA
  - Another 6dB ICLD for the left pair.
  - The image is perceived almost at the front left speaker (mainly one ear → no effective interaural difference)

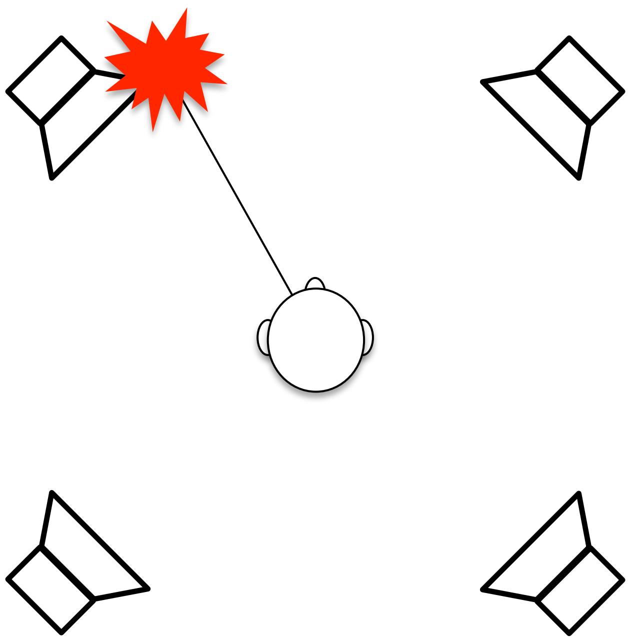




Psychoacoustic considerations for VR

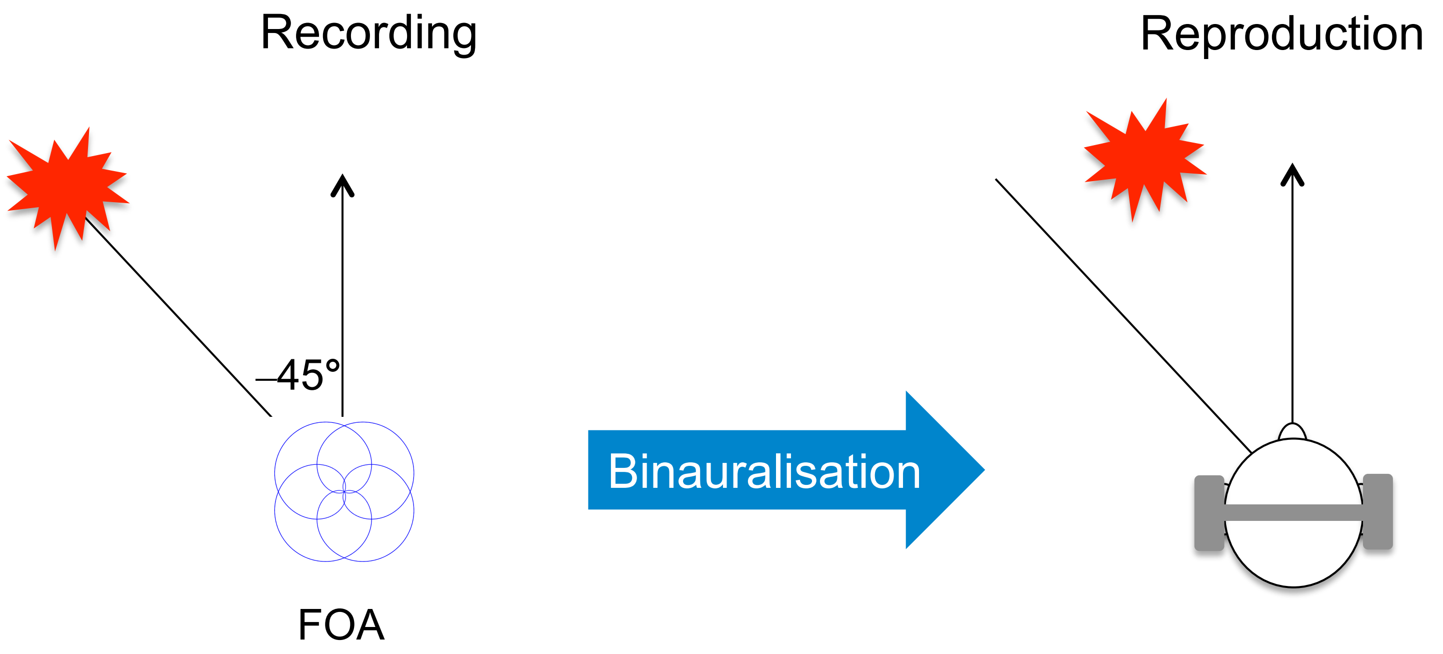
- Limitation of FOA
  - The resulting image position in the quadraphonic reproduction is still not fully shifted to 45°.





Psychoacoustic considerations for VR

- Problems of B-format (FOA) binauralisation for VR
  - Inaccurate localisation due to insufficient ICLD.
  - The image follows you when you rotate the head.

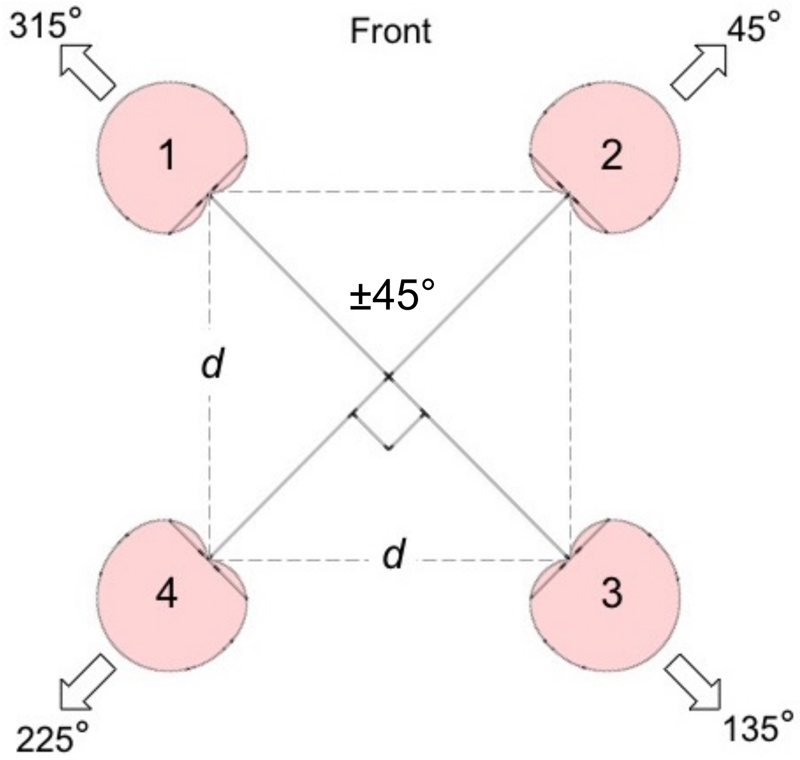


Proposed Technique

Design philosophy

University of HUDDERSFIELD

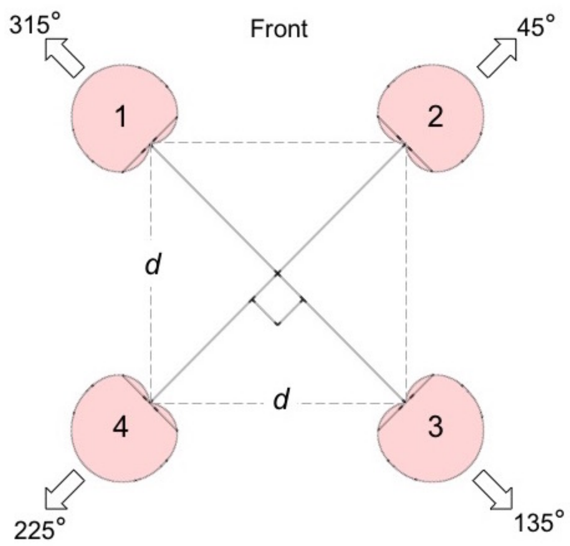
- Equal Segment Microphone Array (ESMA)
  - A design concept proposed by Williams (1991), but for 360 multichannel reproduction.
- Requirements
  1. Equal subtended angle for all stereo segments (±45°).
  2. The stereophonic recording angle (SRA) of each segment should match the subtended angle of the segment. (±45°)



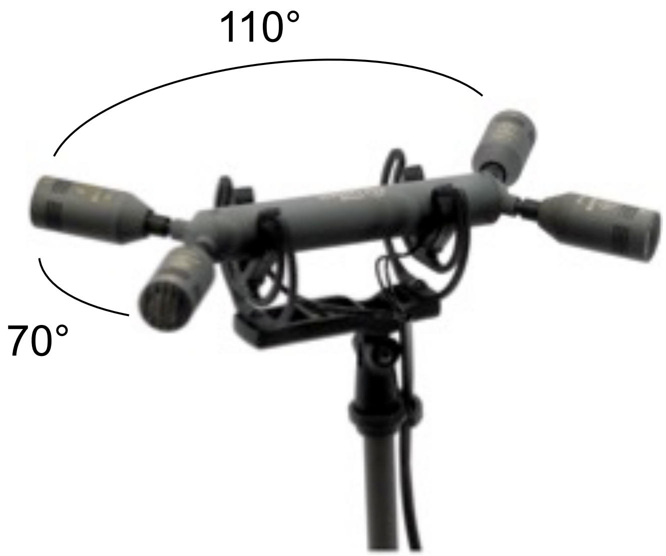
Design philosophy

University of HUDDERSFIELD

- IRT-Cross by Theile
  - Originally designed for ambience capture.
  - *d* = 20 to 25cm.



- ORTF-Surround (or 3D)
  - SRA not consistent for every segment.
  - Not suitable for ESMA.



Design philosophy

University of HUDDERSFIELD

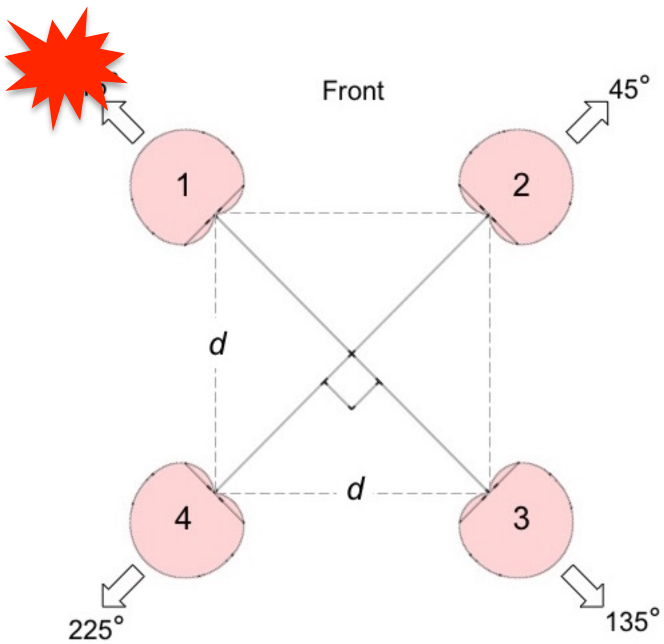
- BBC Proms using ORTF 3D

Design philosophy

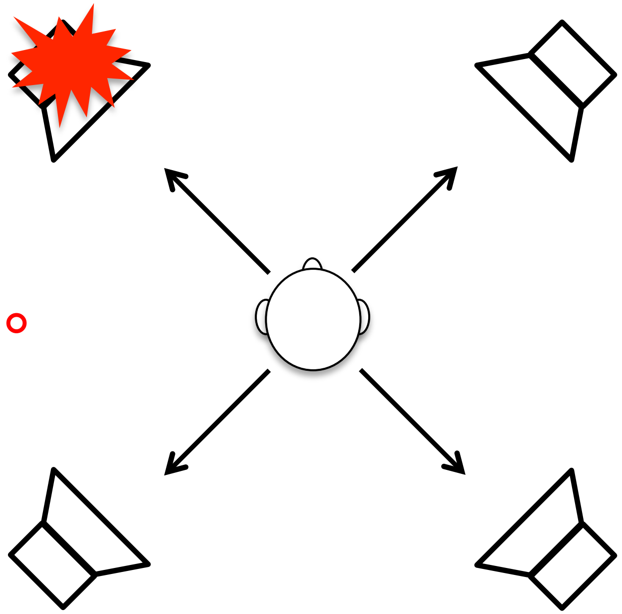
University of HUDDERSFIELD

- The SRA of ±45° for quadraphonic ESMA

→ A source at ±45° in recording should be localised at ±45° in reproduction.



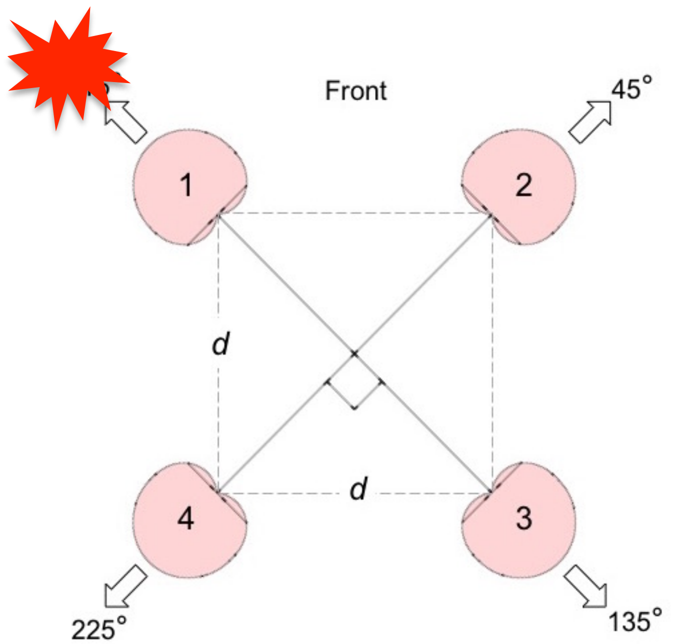
SRA = ±45°

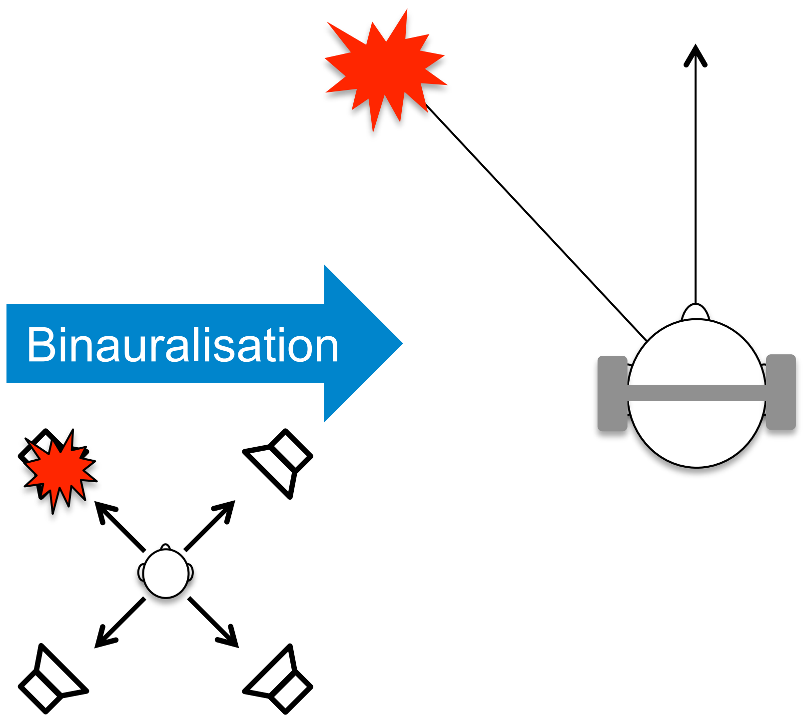


Design philosophy

- The SRA of ±45° for quadraphonic ESMA

→ A source at ±45° in recording should be localised at ±45° in reproduction.

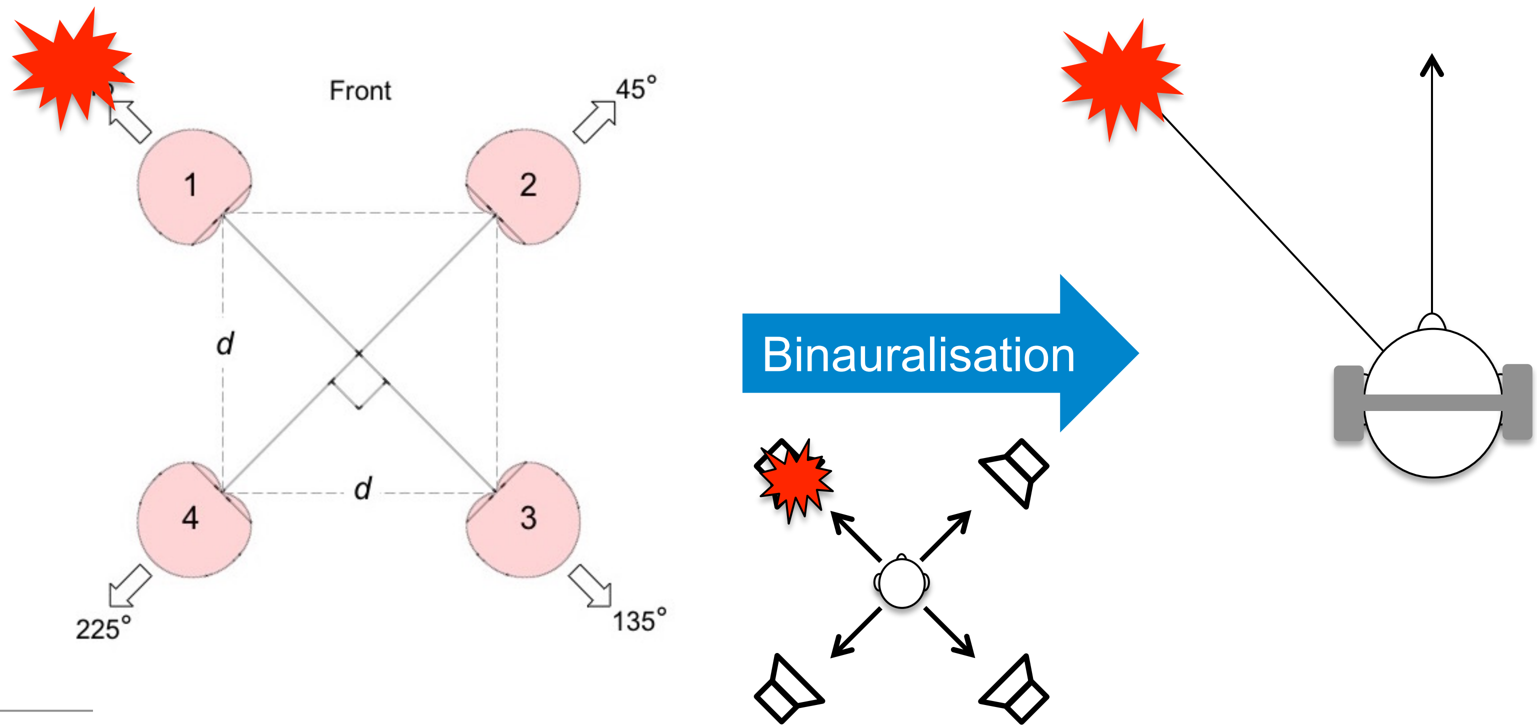




Design philosophy

University of HUDDERSFIELD

- Suitable for VR applications with head-tracking.



Psychoacoustic basis

University of HUDDERSFIELD

- The appropriate spacing between microphones to produce the ±45° SRA?
  - Depends on what psychoacoustic time-level trade-off model is used for calculating the SRA.

Model	Microphone spacing	Source to mic array distance
Williams	23.8cm	unknown
Sengpiel	25cm	unknown
Wittek + Theile	24cm	2m
Lee + Theile	30cm	2m
Lee	50cm	2m

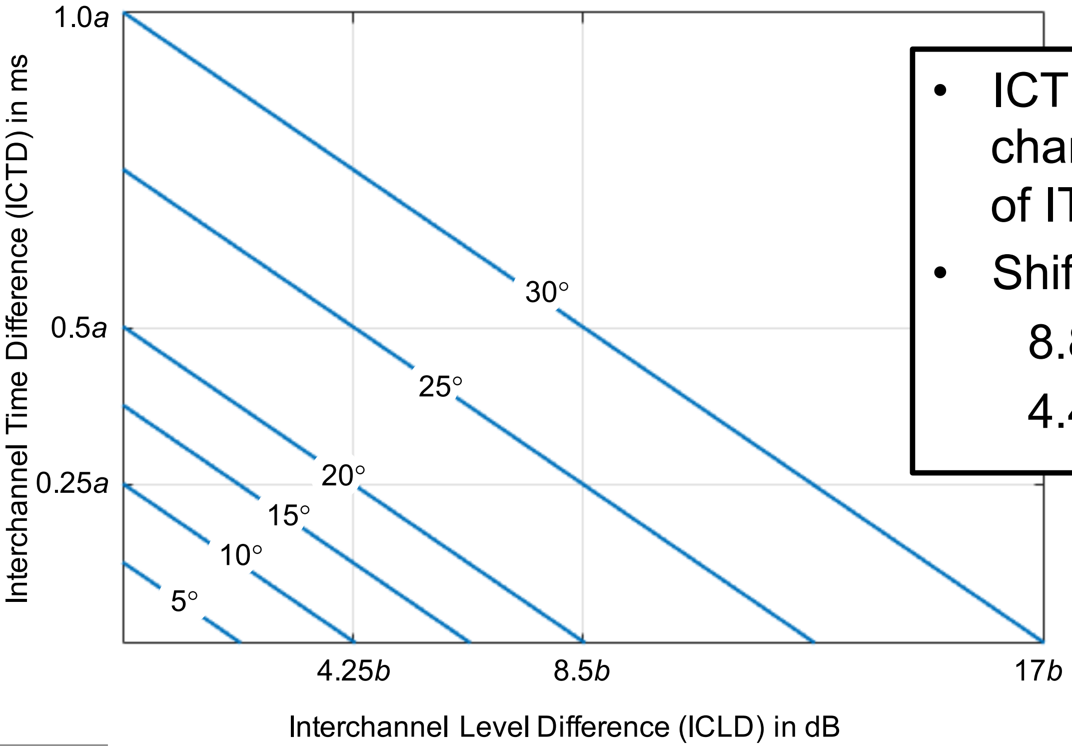
Based on ICTD and ICLD data obtained using ±30 setup°

Optimised for ±45 setup°

Designing a near-coincident VR mic array

University of HUDDERSFIELD

- Linear time-level trade-off functions (Lee 2016)
  - Shift region dependent.
  - Loudspeaker base angle dependent.



- ICTD and ICLD image shift factors change in proportion to the change of ITD and ILD.
- Shift factors for ±45° base angle.

  8.8%/0.1ms; 6%/dB (< 30°)

  4.4%/0.1ms; 3%/dB (30°- 45°).

Experiments

Aim

- To evaluate the localisation accuracies of the quadraphonic FOA and ESMA.
  - If the SRA of ±45° can be achieved.
  - Loudspeaker and headphone reproduction tests in simulated head rotation scenarios.

- Microphone spacing tested:
  - 0cm (FOA)
  - 24cm (Wittek + Theile)
  - 30cm (Lee + Theile)
  - 50cm (Lee)

Stimuli creation

- Recording setup

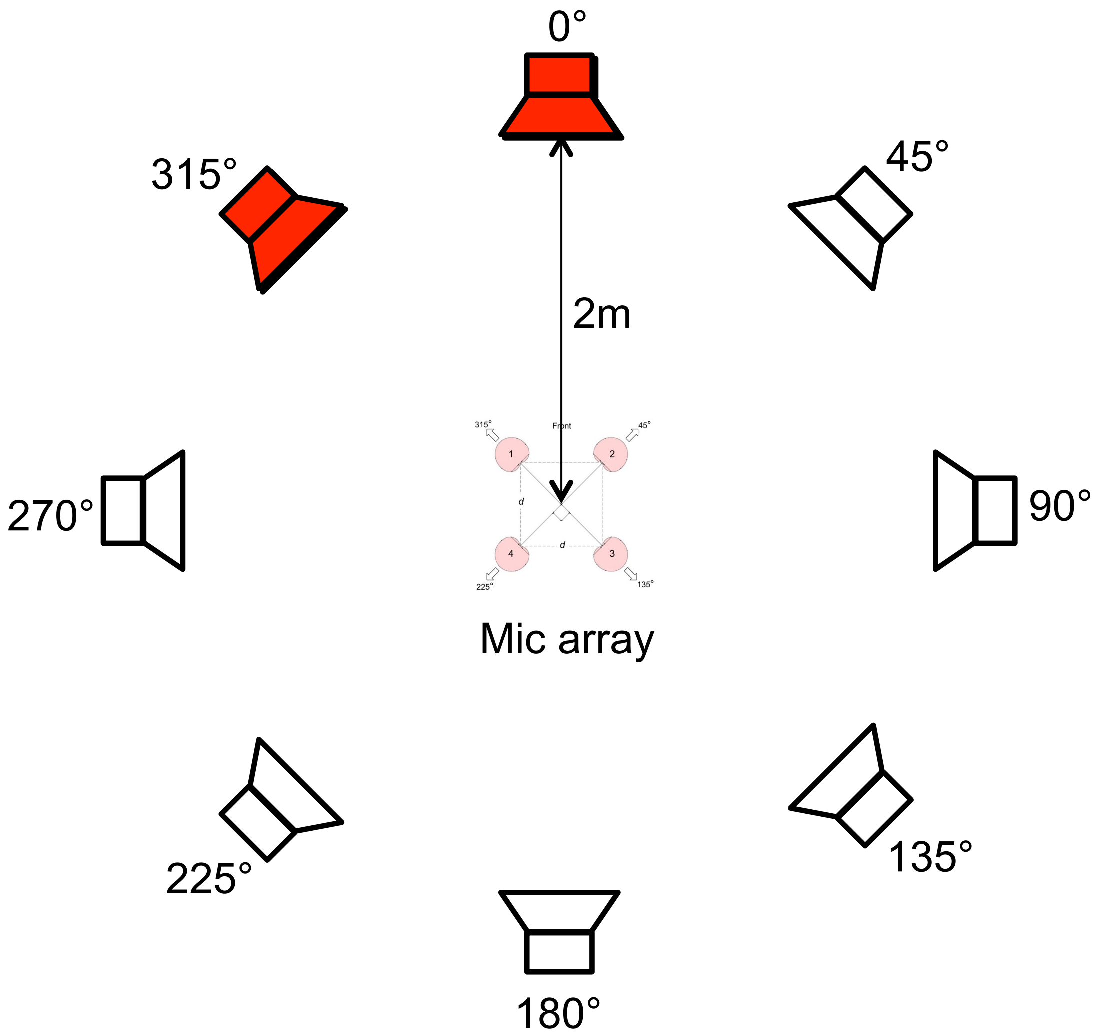
- ITU-R BS.1116 standard room.

- 8 Genelec 8040As arranged in an octagonal layout.

- Room impulse responses (RIRs) captured for 0° and 45°.

- Soundfield SPS 422b for FOA.

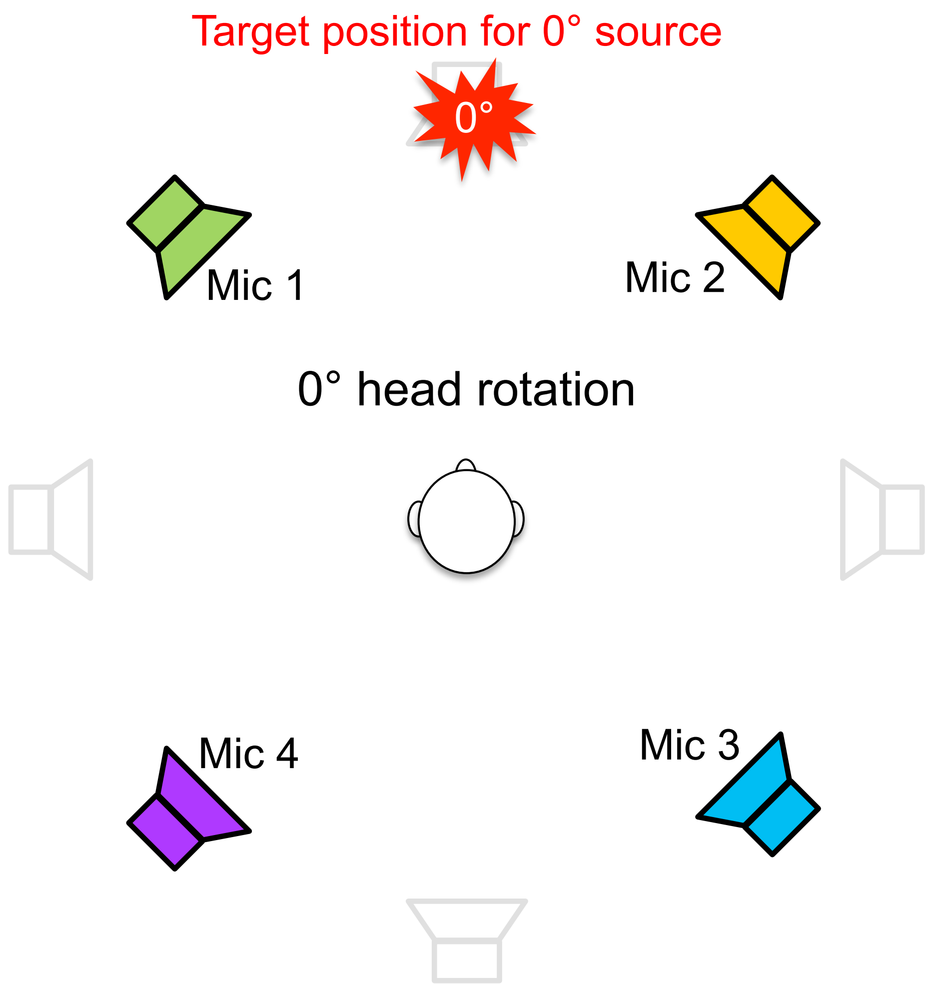
- Neumann KM184 for ESMA.



Stimuli creation

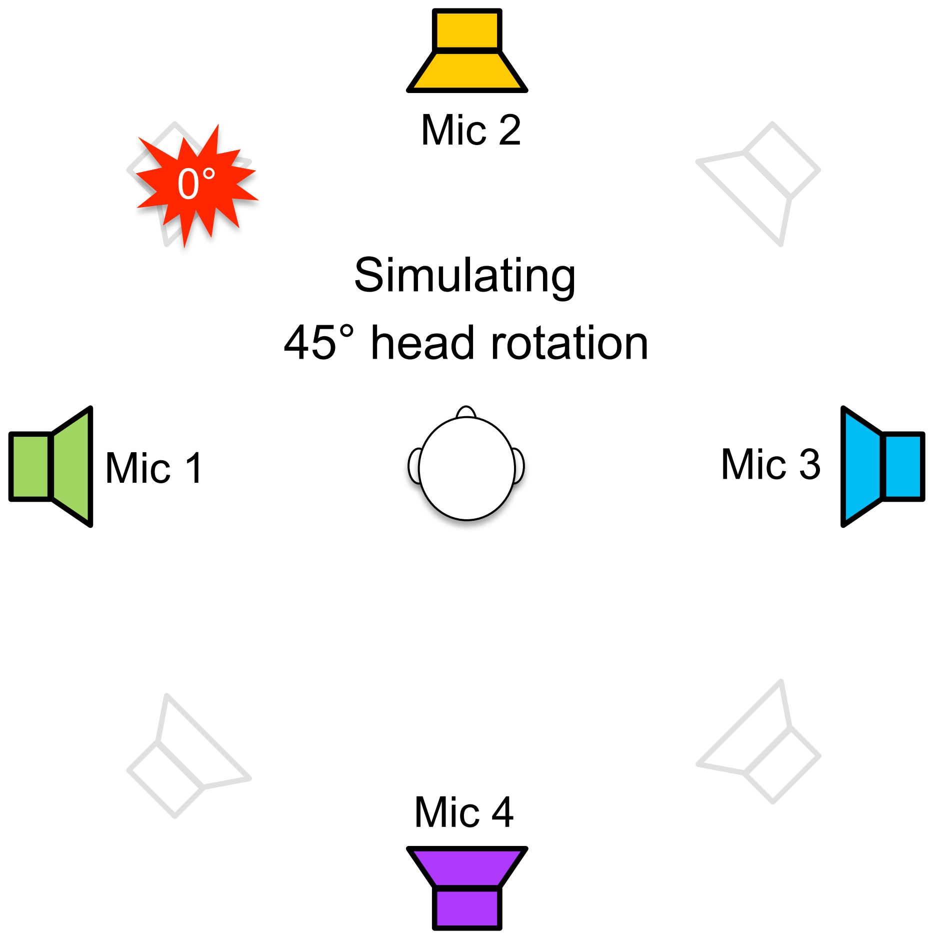
University of HUDDERSFIELD

- Stimuli for Experiment 1 (Loudspeaker playback)
  - An anechoic speech signal was convolved with the direct sounds of the RIRs (reflections removed).
  - Head rotations simulated for 0°, ±45°, ±90°, ±135° and ±180° (Soundfield rotation).



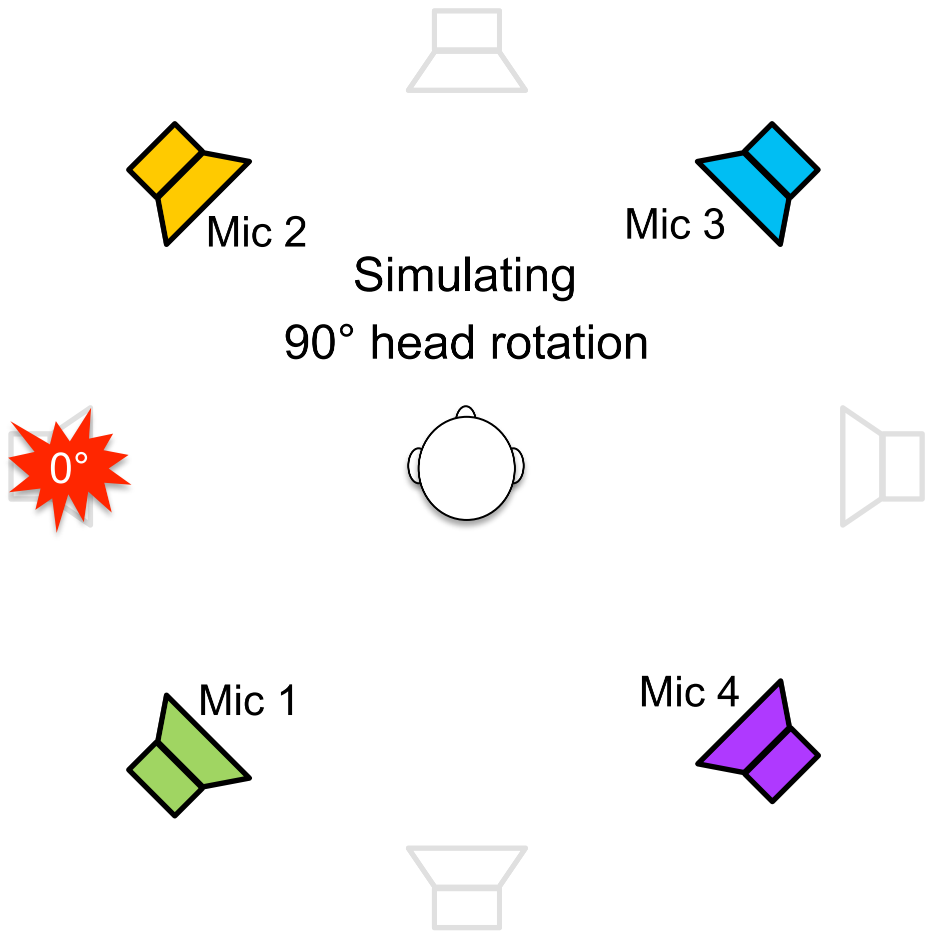
Stimuli creation

- Stimuli for Experiment 1 (Loudspeaker playback)
  - An anechoic speech signal was convolved with the direct sounds of the RIRs (reflections removed).
  - Head rotations simulated for 0°, ±45°, ±90°, ±135° and ±180° (Soundfield rotation).



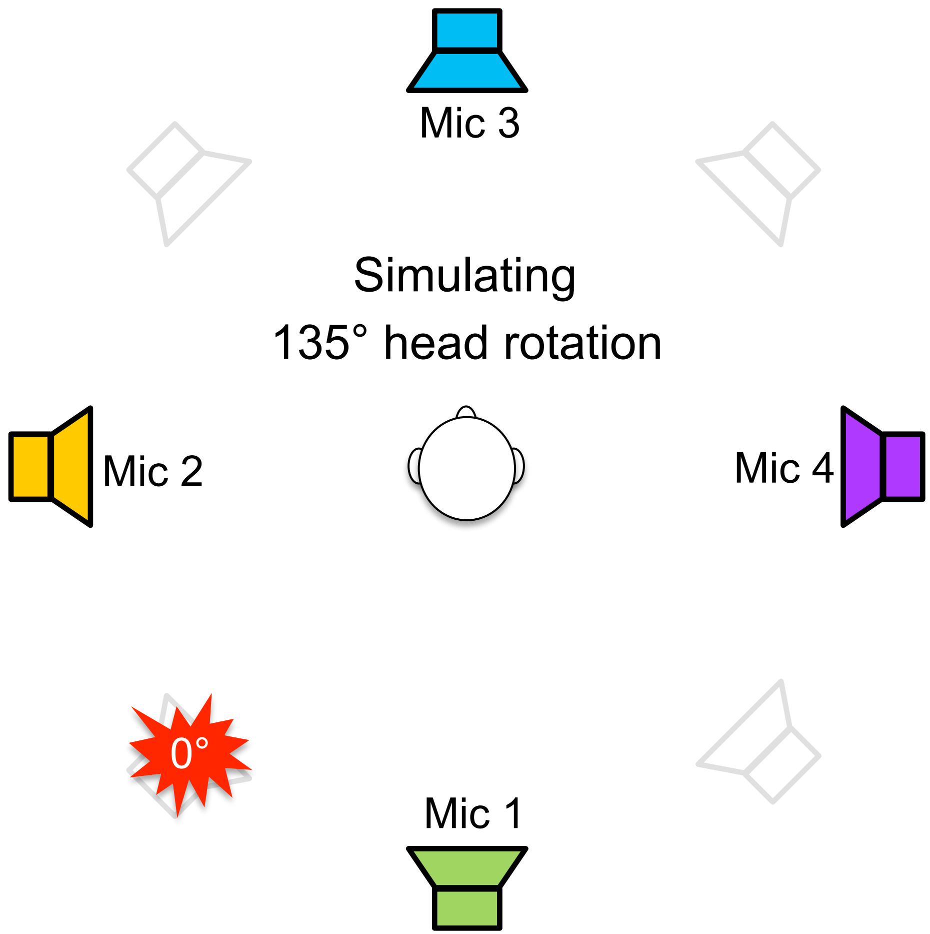
Stimuli creation

- Stimuli for Experiment 1 (Loudspeaker playback)
  - An anechoic speech signal was convolved with the direct sounds of the RIRs (reflections removed).
  - Head rotations simulated for 0°, ±45°, ±90°, ±135° and ±180° (Soundfield rotation).



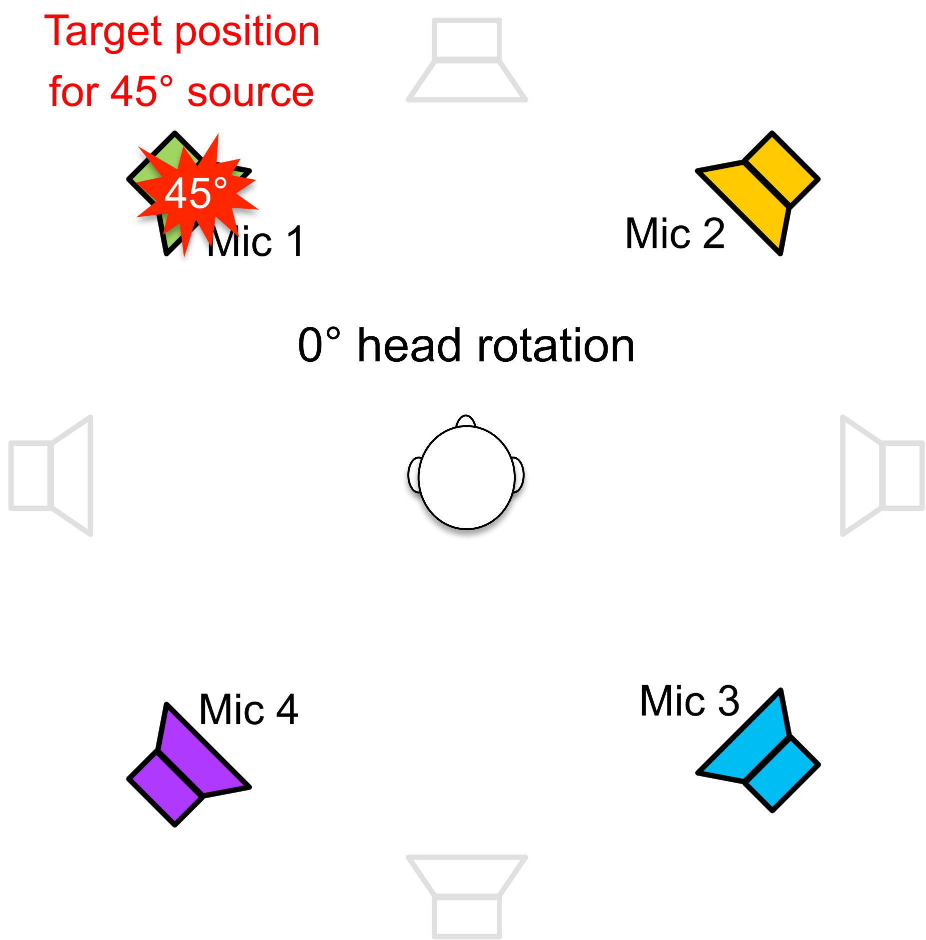
Stimuli creation

- Stimuli for Experiment 1 (Loudspeaker playback)
  - An anechoic speech signal was convolved with the direct sounds of the RIRs (reflections removed).
  - Head rotations simulated for 0°, ±45°, ±90°, ±135° and ±180° (Soundfield rotation).



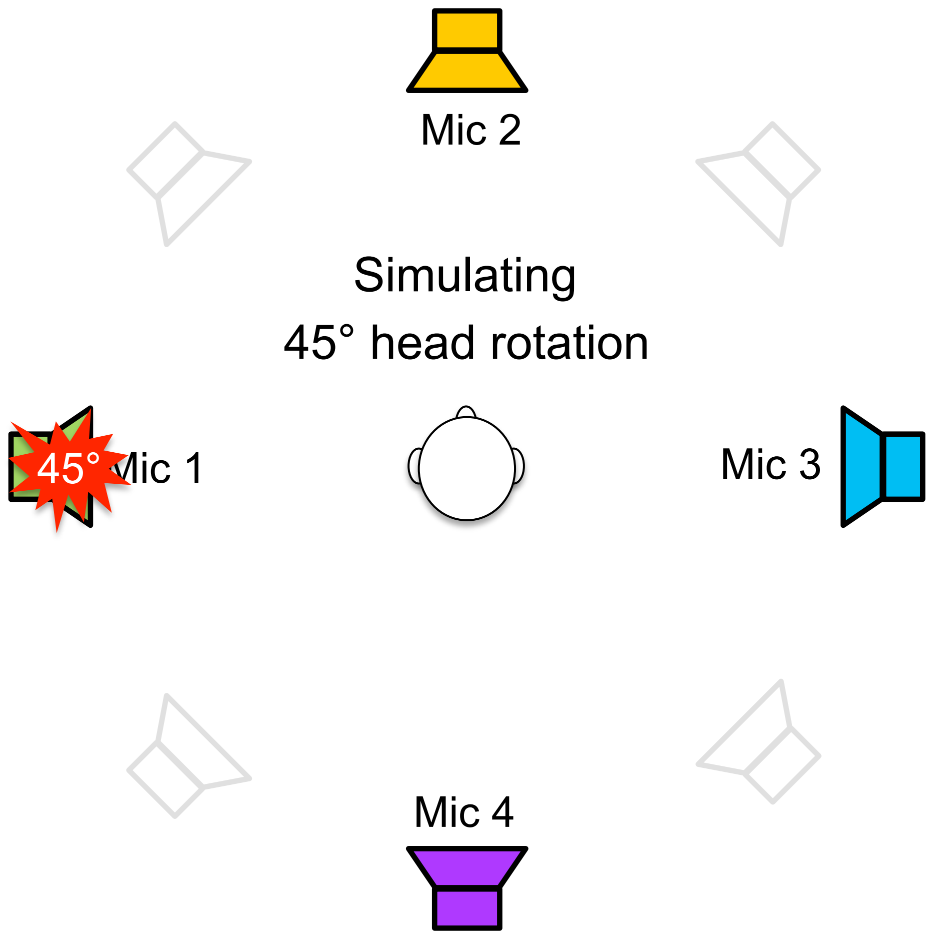
Stimuli creation

- Stimuli for Experiment 1 (Loudspeaker playback)
  - An anechoic speech signal was convolved with the direct sounds of the RIRs (reflections removed).
  - Head rotations simulated for 0°, ±45°, ±90°, ±135° and ±180° (Soundfield rotation).



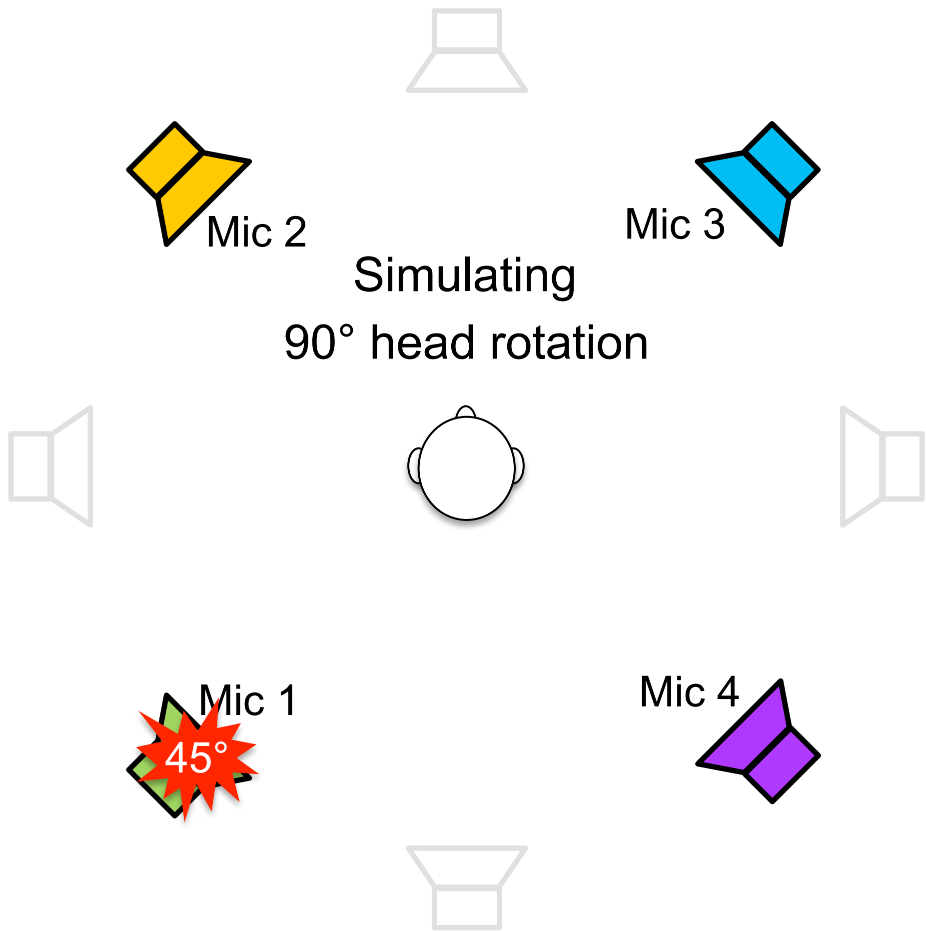
Stimuli creation

- Stimuli for Experiment 1 (Loudspeaker playback)
  - An anechoic speech signal was convolved with the direct sounds of the RIRs (reflections removed).
  - Head rotations simulated for 0°, ±45°, ±90°, ±135° and ±180° (Soundfield rotation).



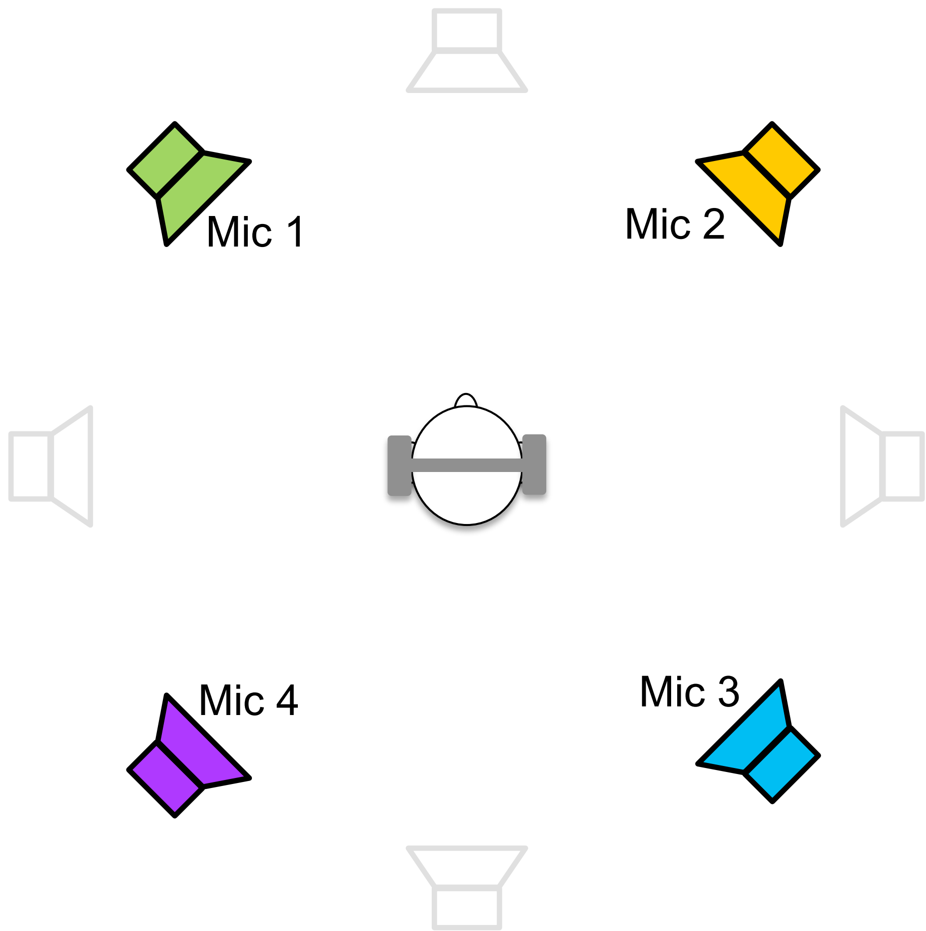
Stimuli creation

- Stimuli for Experiment 1 (Loudspeaker playback)
  - An anechoic speech signal was convolved with the direct sounds of the RIRs (reflections removed).
  - Head rotations simulated for 0°, ±45°, ±90°, ±135° and ±180° (Soundfield rotation).



Stimuli creation

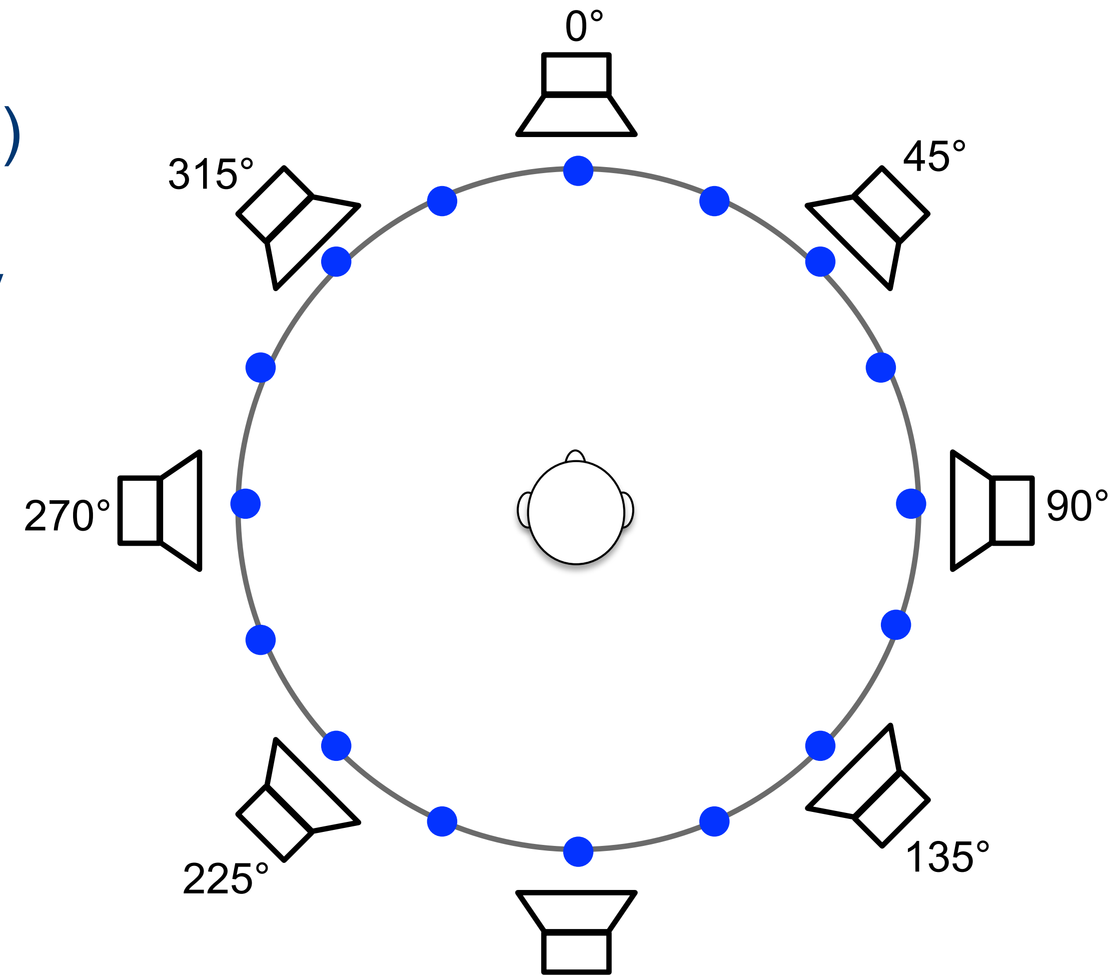
- Stimuli for Experiment 2 (Binaural playback)
  - Same conditions as Experiment 1, but with the full RIRs (reflections included).
  - The multichannel stimuli were binauralised with dry KU100 dummy head HRIRs from the ‘SADIE’ database (Kearney 2015).



Listening tests

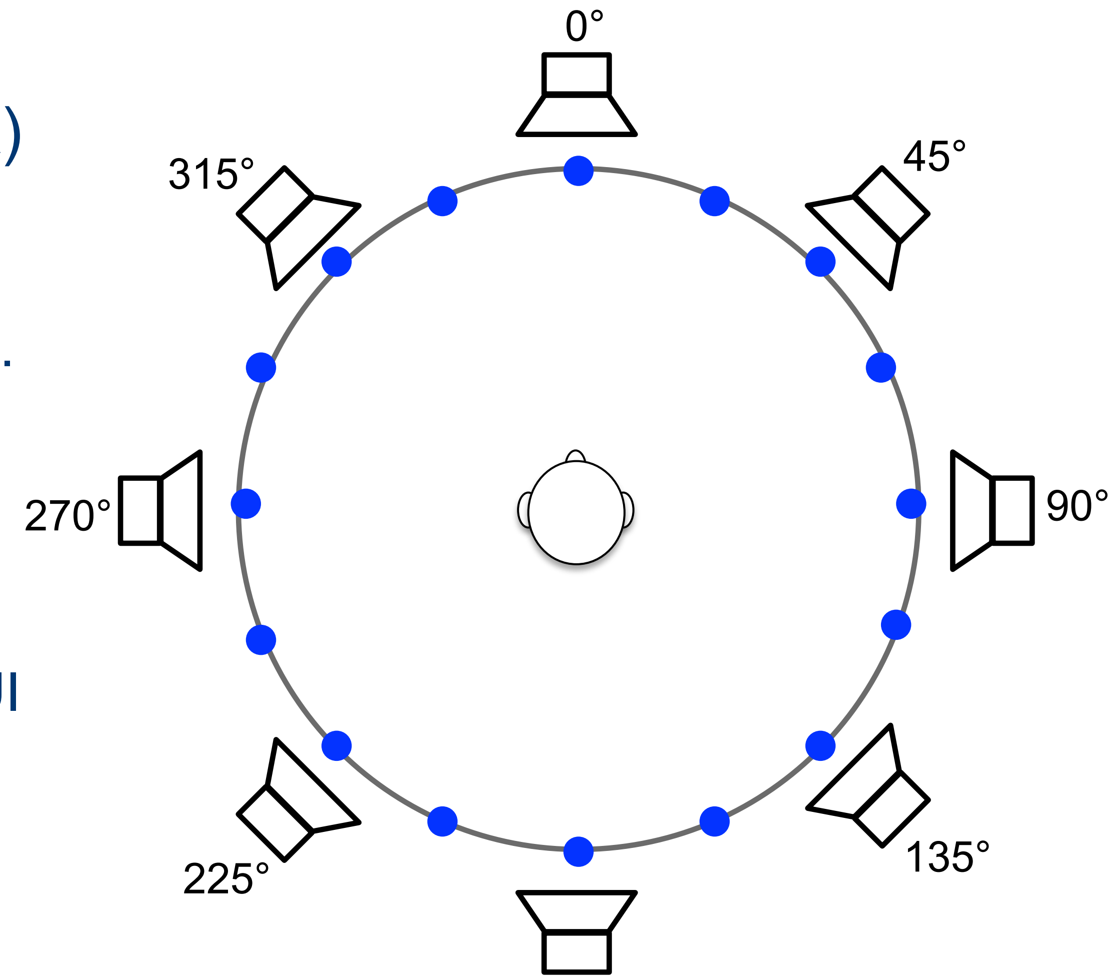
University of HUDDERSFIELD

- Experiment 1 (Loudspeaker playback)
  - Loudspeakers hidden by acoustically transparent curtains.
  - Small markers were placed on the curtain from 0° with 22.5° intervals.
  - 70dBA playback level.



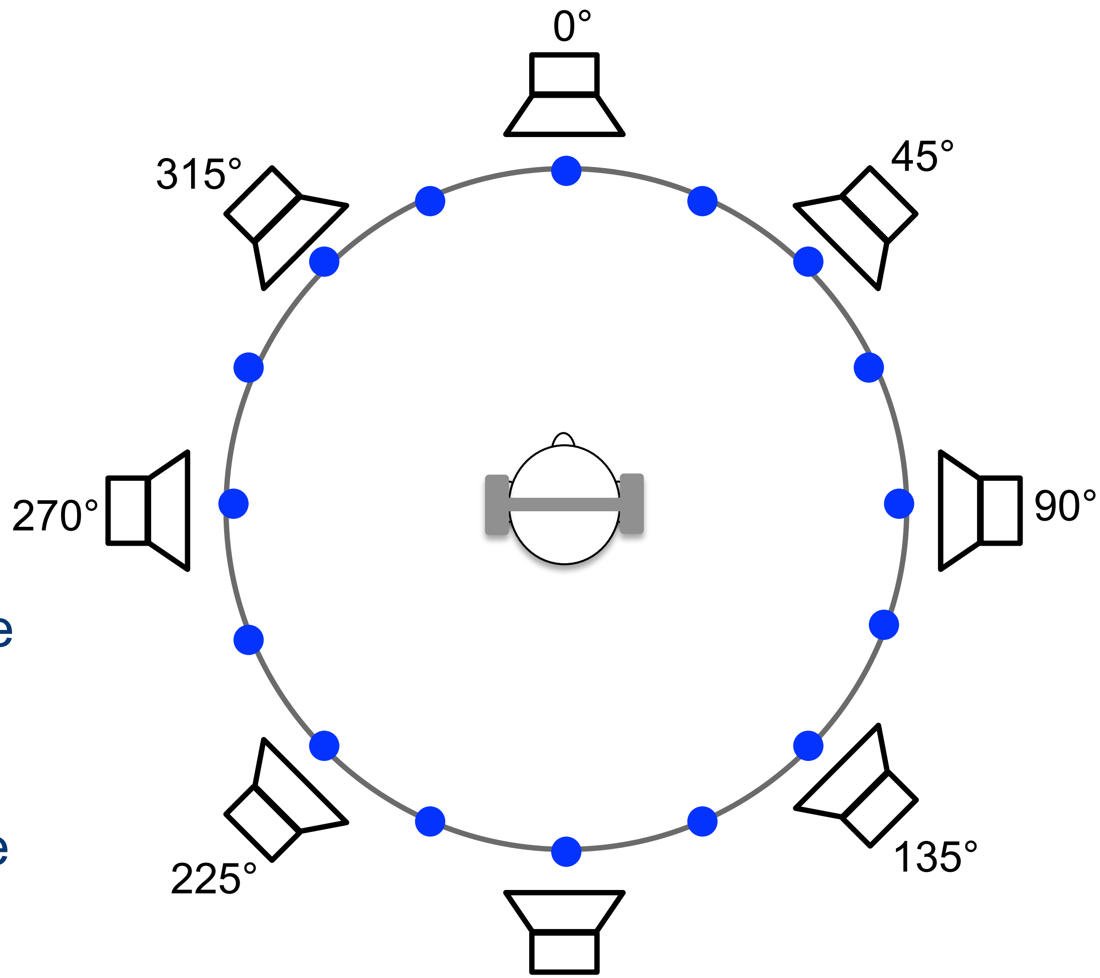
Listening tests

- Experiment 1 (Loudspeaker playback)
  - 9 experienced subjects repeated each test twice.
  - The task was to mark down the perceived image position on a horizontal circle on a GUI with markers indicated with 22.5° intervals.

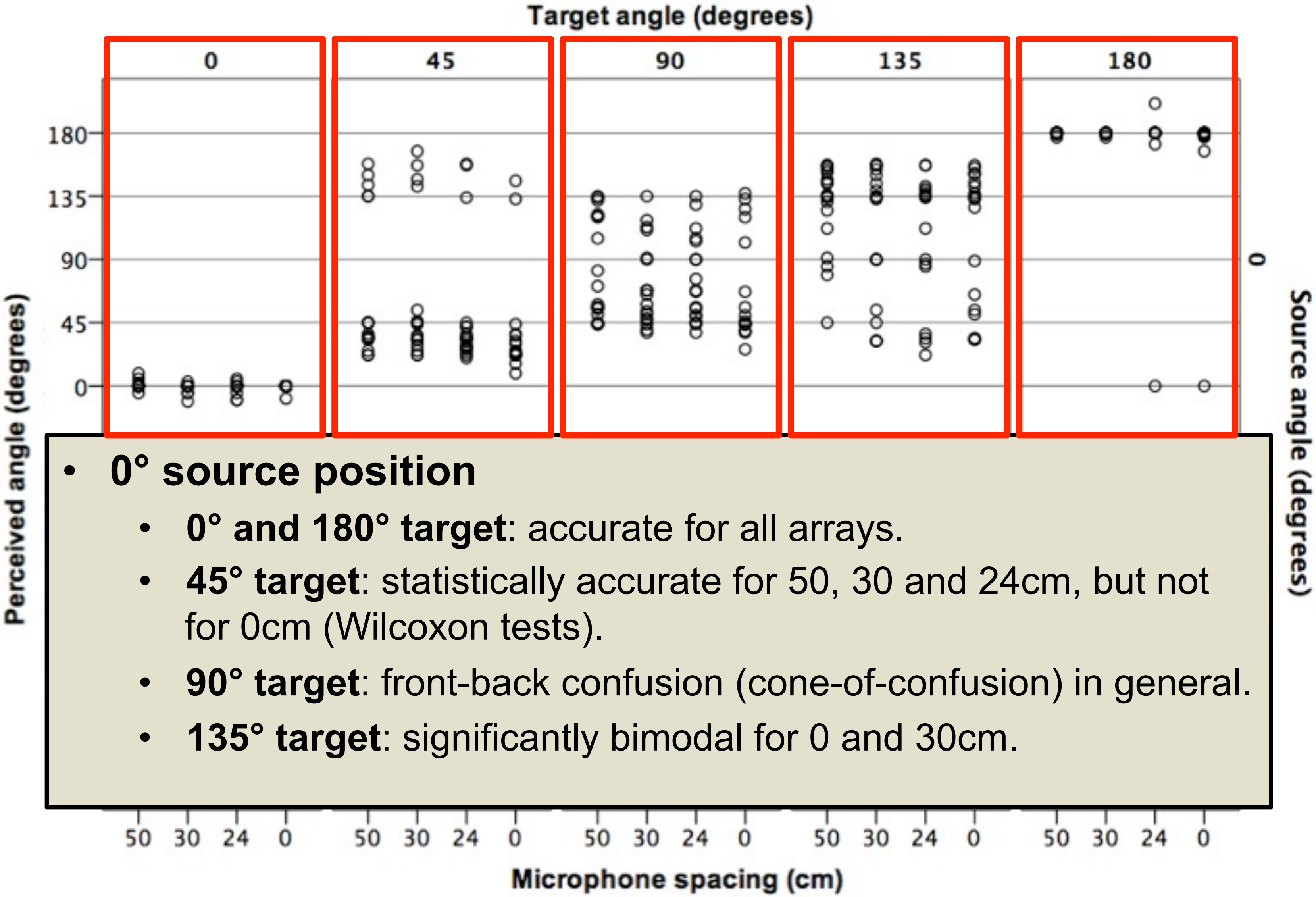


Listening tests

- Experiment 2 (Binaural playback)
  - The same room, subjects, task and method as Experiment 1.
  - Equalised Sennheiser HD650 headphones were used.
  - Loudness matched to the playback levels of multichannel stimuli.



Results – Loudspeaker experiment

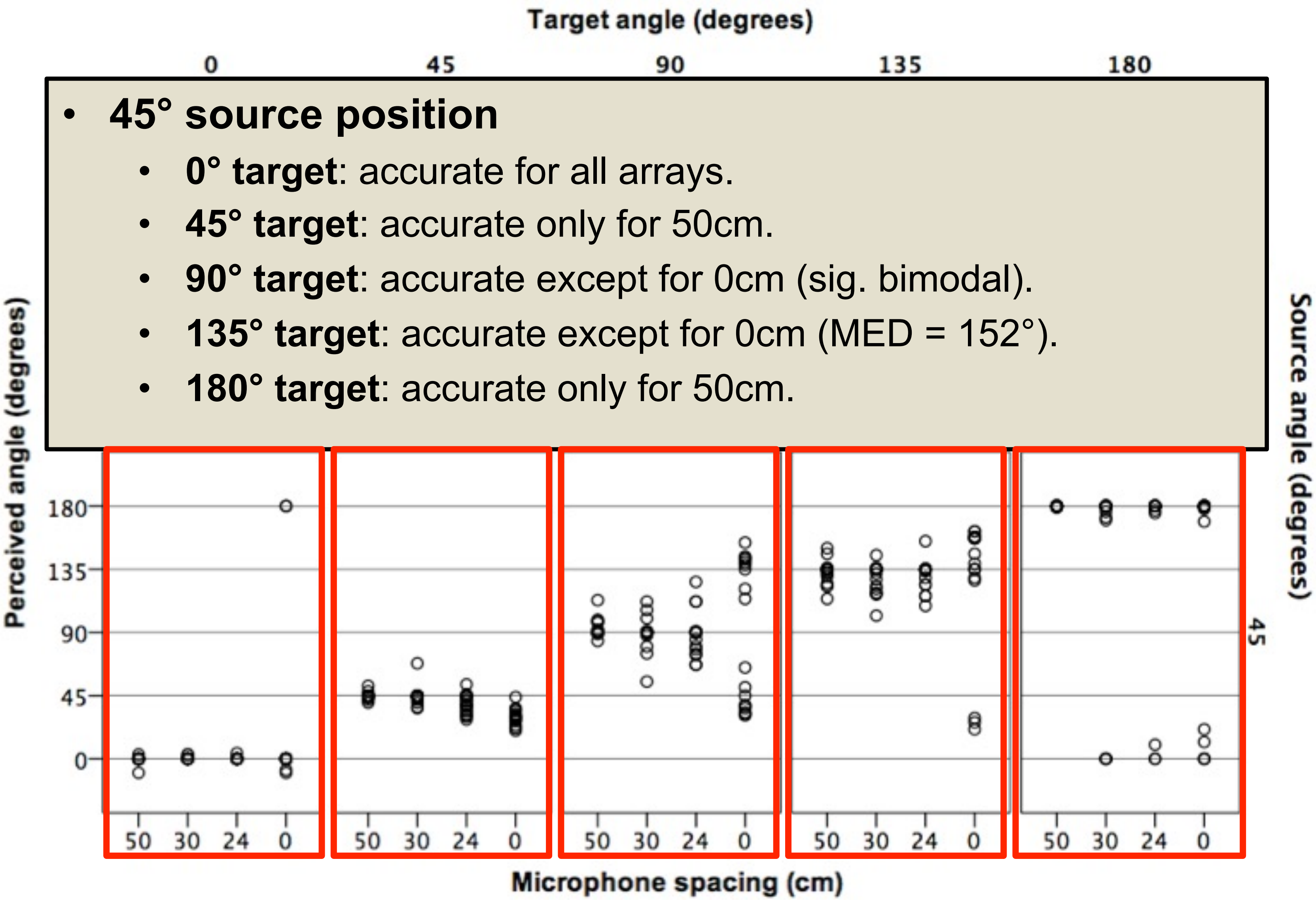


- **0° source position**
  - **0° and 180° target**: accurate for all arrays.
  - **45° target**: statistically accurate for 50, 30 and 24cm, but not for 0cm (Wilcoxon tests).
  - **90° target**: front-back confusion (cone-of-confusion) in general.
  - **135° target**: significantly bimodal for 0 and 30cm.

Results – Loudspeaker experiment

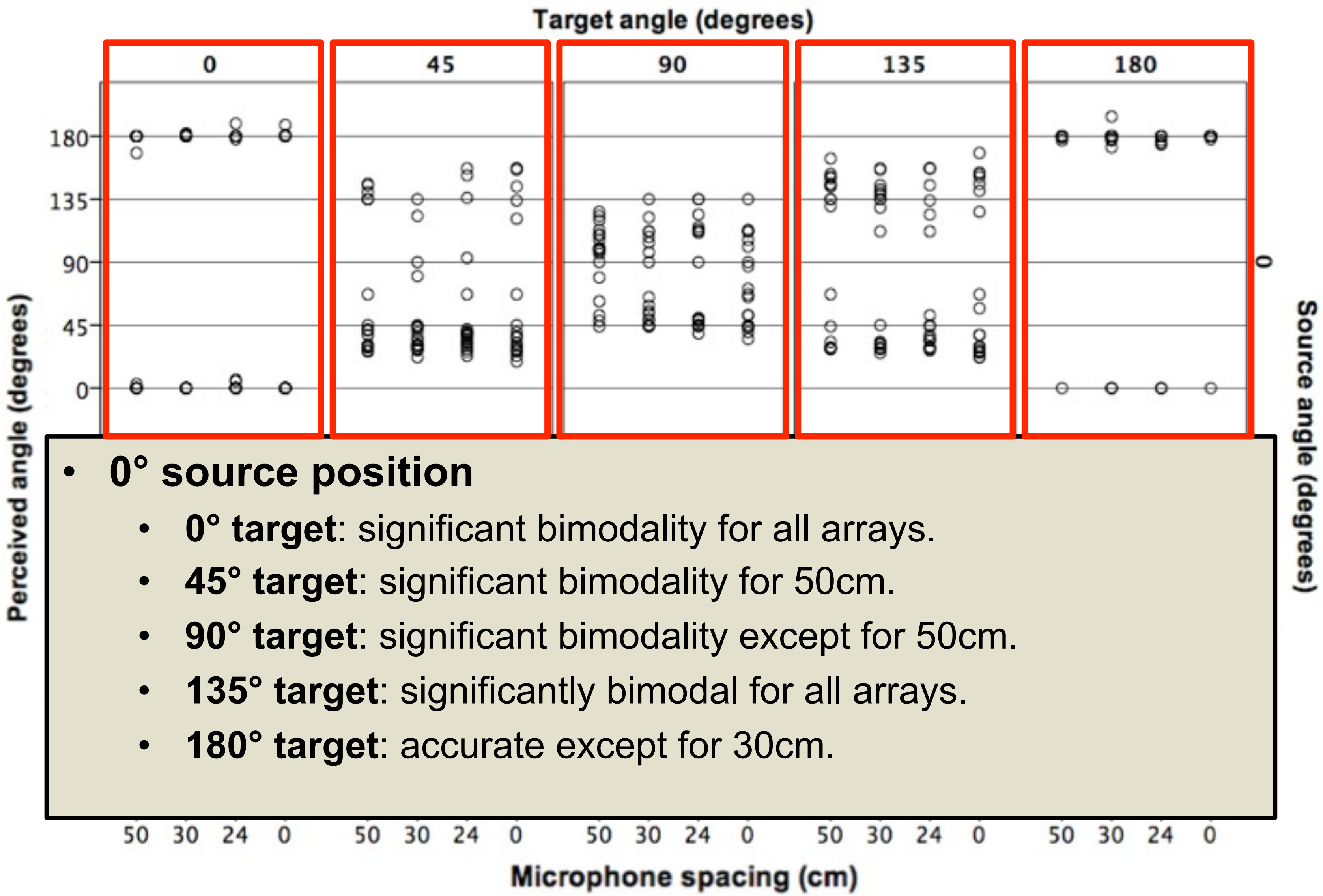
University of Huddersfield

- **45° source position**
  - **0° target**: accurate for all arrays.
  - **45° target**: accurate only for 50cm.
  - **90° target**: accurate except for 0cm (sig. bimodal).
  - **135° target**: accurate except for 0cm (MED = 152°).
  - **180° target**: accurate only for 50cm.



Results – Binaural experiment

University of HUDDERSFIELD

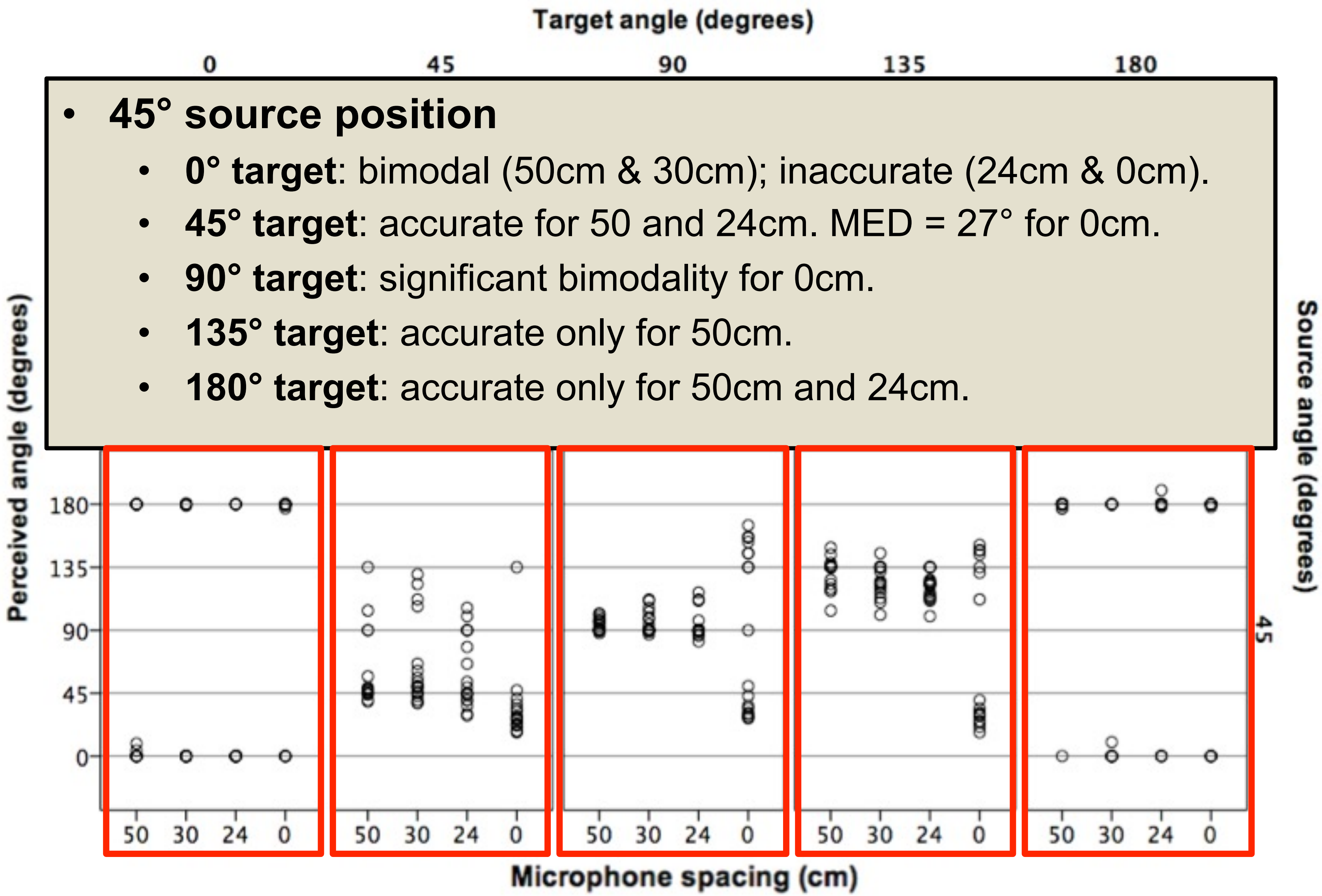


- **0° source position**
  - **0° target**: significant bimodality for all arrays.
  - **45° target**: significant bimodality for 50cm.
  - **90° target**: significant bimodality except for 50cm.
  - **135° target**: significantly bimodal for all arrays.
  - **180° target**: accurate except for 30cm.

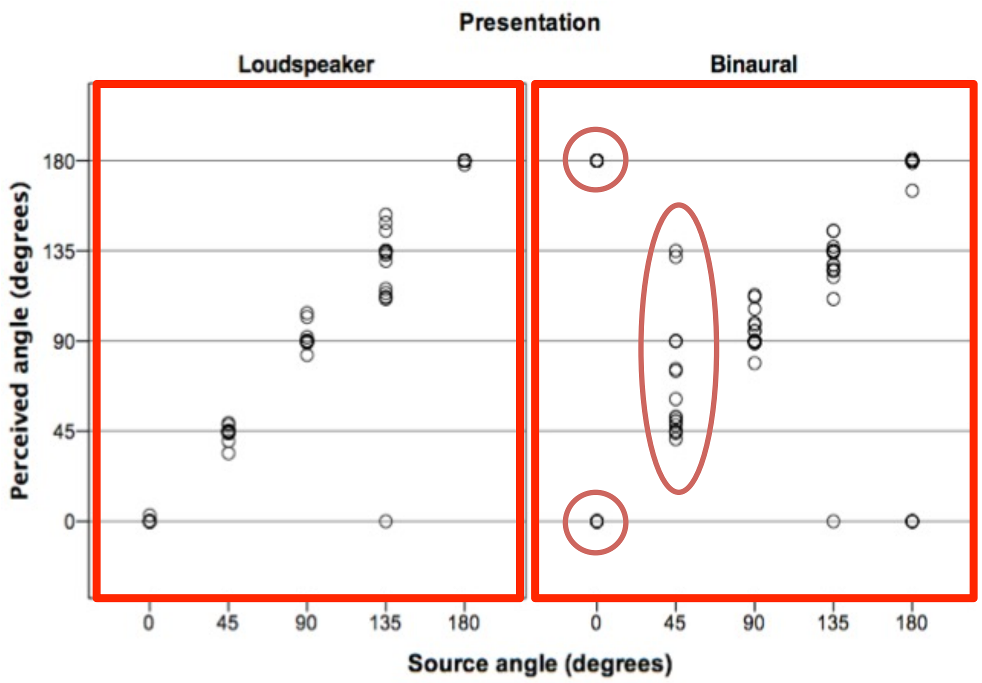
Results – Binaural experiment

University of HUDDERSFIELD

- **45° source position**
  - **0° target**: bimodal (50cm & 30cm); inaccurate (24cm & 0cm).
  - **45° target**: accurate for 50 and 24cm. MED = 27° for 0cm.
  - **90° target**: significant bimodality for 0cm.
  - **135° target**: accurate only for 50cm.
  - **180° target**: accurate only for 50cm and 24cm.



Results – Real source



- **Loudspeaker**
  - Loudspeaker: accurate for all source angles.
- **Binaural**
  - Binaural responses are generally more spread than loudspeaker ones.
  - 0°: significantly bimodal.
  - 45°: inaccurate, MED = 52°.
  - 90°, 135°: accurate.
  - 180°: inaccurate, bimodal.

Discussion

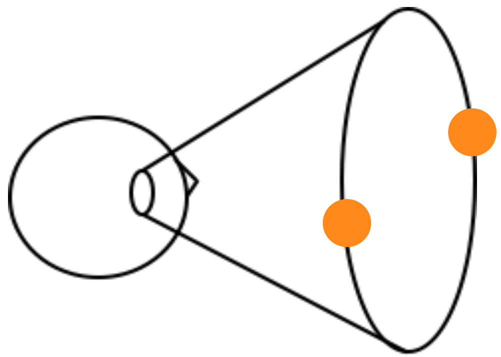
University of HUDDERSFIELD

- Microphone spacing effect
  - 0cm had the worst localisation performance overall.
    - Significant bimodal distributions for many target angle conditions.
    - Perceived to be significantly narrower for the 45° source in both loudspeaker (MED = 30°) and binarual (MED = 27°).
  - 50cm was the only spacing that achieved the SRA of ±45°.
    - Seems to validate the new psychoacoustic model.
  - 50cm had slightly better consistency and accuracy than the other configurations overall.
    - But a smaller size might be more beneficial in practical situations.
    - Practical importance of localisation accuracy in VR?

Discussion

University of HUDDERSFIELD

- Source angle effect
  - The 0° source produced larger response spreads and more bimodal distributions than the 45°.
    - Front-back confusion (Cone of confusion), especially for the 90° target angle.
    - Lateral phantom image localisation is highly unstable (Theile and Plenge 1977, Martin et al 1999).



Discussion

University of HUDDERSFIELD

- Loudspeaker vs. Binaural

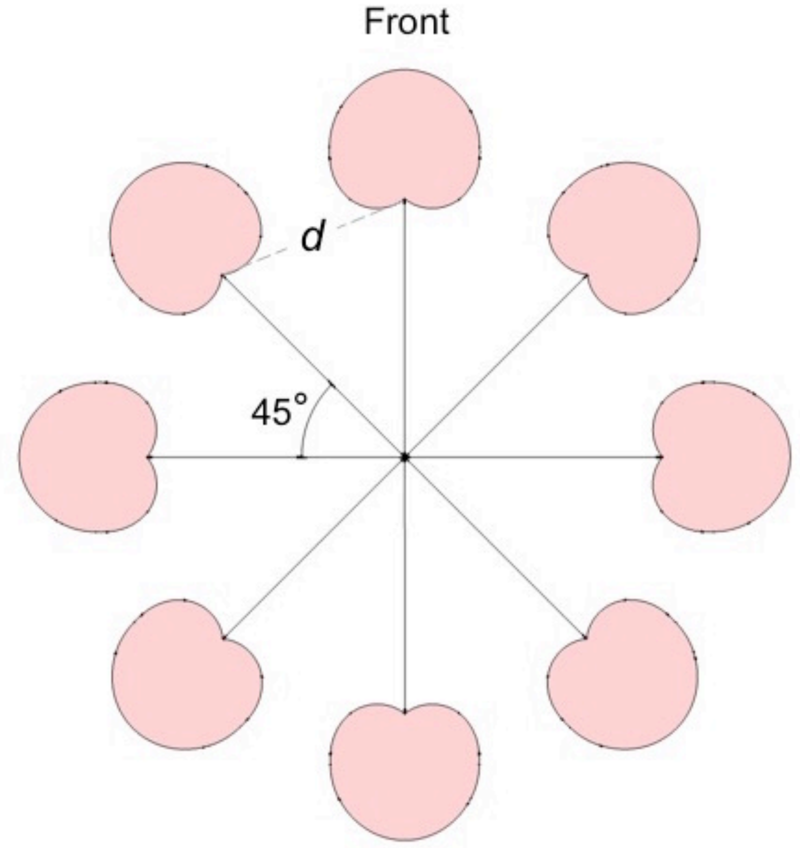
  - **Front-back confusion** was more frequently observed in the binaural presentation, but not in the loudspeaker one.
  - The binaural presentation had more spread responses.
  - Real source results also show similar tendencies for the 0° and 45°.

  - Might be due to the use of **non-individualised HRTF**, rather than the microphone arrays.
  - But more about the **lack of head movement**?
    - FB confusion can occur even with individualised HRTF when head rotation is not allowed (Wightman and Kistler 1999).
  - The FB confusion problem might be largely resolved in practical VR applications with head tracking.

Discussion

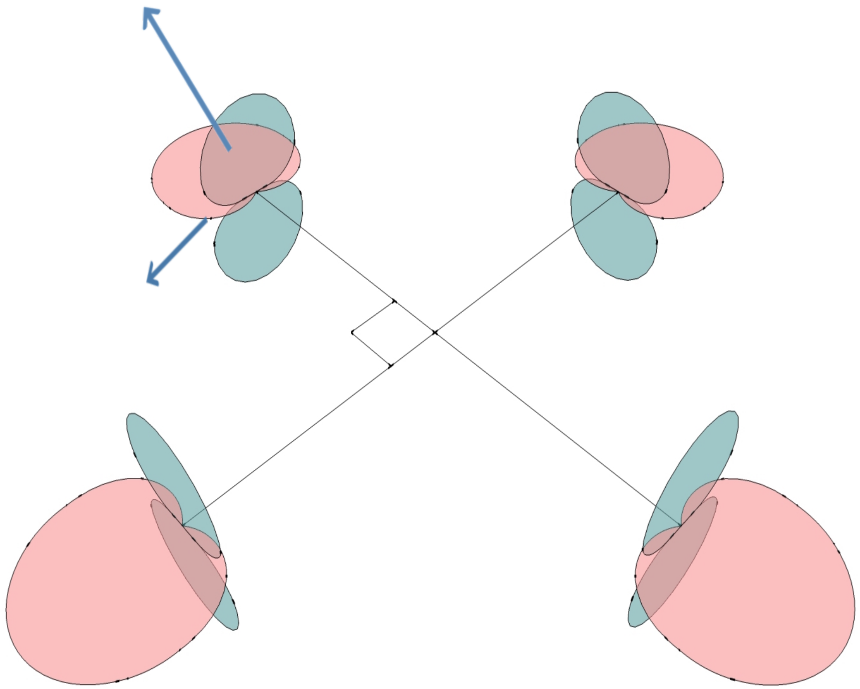
University of HUDDERSFIELD

- Higher Order ESMA
  - For an octagonal setup, each segment should have the SRA of 45° (±22.5°).
  - Can potentially solve the problem of unstable side image localisation.
  - Mic spacing *d*
    - *Williams: 82cm*
    - *Lee: 55cm*



Discussion

- Adding vertical dimension to ESMA
  - Cardioid + Figure-of-eight in a vertically coincident fashion.
    - Vertical Mid-Side decoding.
    - Vertical microphone spacing has little effect on LEV (Lee and Gribben JAES 2014).
    - Vertical level panning can provide source imaging with a limited resolution (Barbour 2003, Mironovs and Lee 2016).
    - Vertical time panning is highly unstable (Wallis and Lee JAES 2015).



Conclusions

- ESMAs had a better localisation accuracy than FOA.
- 50cm spacing had the best localisation accuracy, but 30cm or 24cm might still be acceptable.
- Front-Back confusion in binaural reproduction without head rotation.
- Ongoing works
  – Investigations on different attributes.
  – Externalisation, tonal quality, spaciousness, naturalness, etc.
  – Practical evaluations with head tracking.

Thank you for listening.

Hyunkook Lee
h.lee@hud.ac.uk