
link to paper

# Comparison of conditions for omnidirectional video with spatial audio in terms of subjective quality and impacts on objective metrics resolving power

Andréas Pastor[1], Pierre Lebreton[1,2], Toinon Vigier[1], Patrick Le Callet[1,3]

[1] Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

[2] Capacités SAS [3] Institut universitaire de France (IUF)




# Introduction

Motivations:

- not many work on 360° videos with Higher Order Ambisonic (HOA) spatial audio Quality of Experience
- reproduce [1] with naive assessors and compare data quality
- compare QoE perception between “reference” and “consumer-grade” audio setup

[1] R.F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, “Perceptual evaluation on audio-visual dataset of 360 content,” in 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2022, pp. 1–6.

# Introduction

Additional motivations:

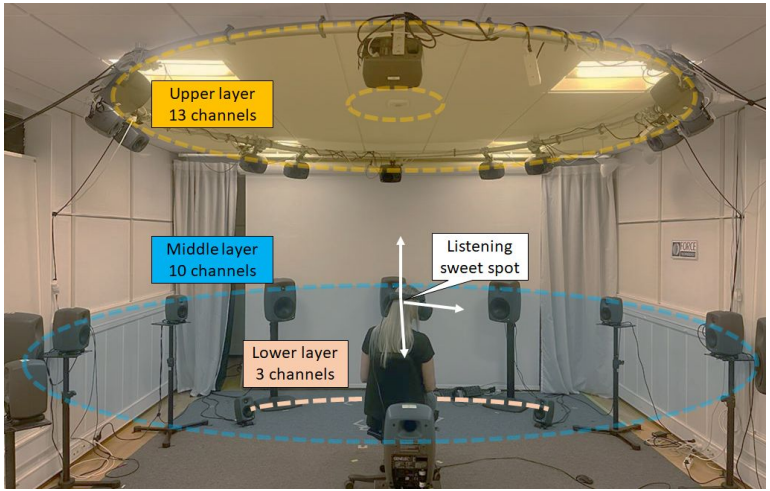
- Study of experimental effort cost and its relation with discriminability
- Study metric resolving power and its relation with discriminability
- propose a method to apply No-Reference parameter-based audiovisual quality estimation model from ITU-T P.1203 mode 0 to 360° AV content evaluation: toward greener metric

[1] R.F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, “Perceptual evaluation on audio-visual dataset of 360 content,” in 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2022, pp. 1–6.

# QOE conditions: reference versus consumer-grade setup

A reference playback setup from [1]:

- Odyssey+ HMD
- 26-channel 8040A Genelec loudspeaker setup



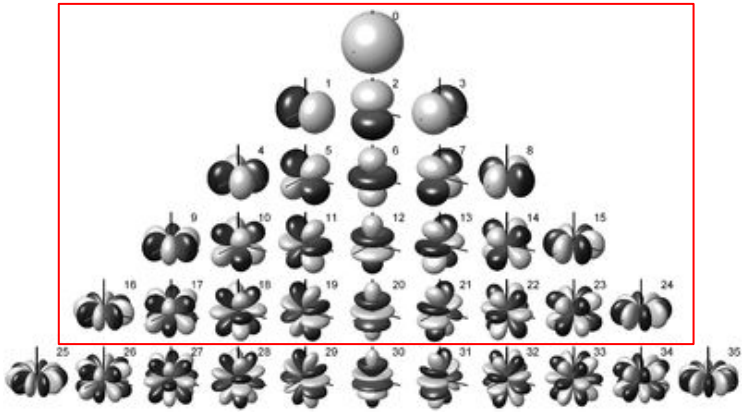

A consumer-grade setup:

- HTC Vive Pro eye
- HTC Vive Pro eye built-in headphones

[1] R.F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, “Perceptual evaluation on audio-visual dataset of 360 content,” in 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2022, pp. 1–6.

# 25 channel 4th Order Ambisonic audio playback

- 25 audio directions captured and compressed in an audio file
- Decoding with All-Round Ambisonic algorithm [4] into loudspeakers of reference setup
- Binaural rendering for build-in headphones in consumer-grade setup via Reaper, Sparta AmbiBIN plugin [5] and OSC messaging



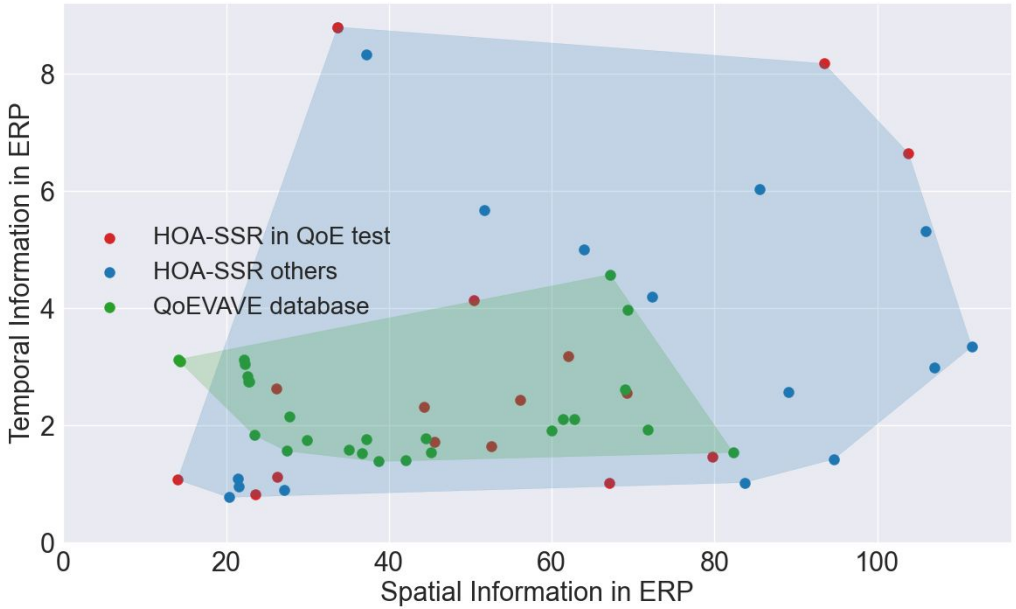

[4] F. Zotter and M. Frank, “All-round ambisonic panning and decoding,” Journal of the Audio Engineering Society, vol. 60, no. 10, pp. 807–820, 2012

[5] Leo McCormack and Archontis Politis, “SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods,” in 2019 AES International Conference on Immersive and Interactive Audio. Audio Engineering Society, 2019.

# Source content characterisation: video part

SI TI computed on ERP:

- 16 sequences of HOA–SSR database QoE subjectively evaluated in [1]
- other HOA–SSR database sequences (not evaluated)
- QoEVAVE database [2]

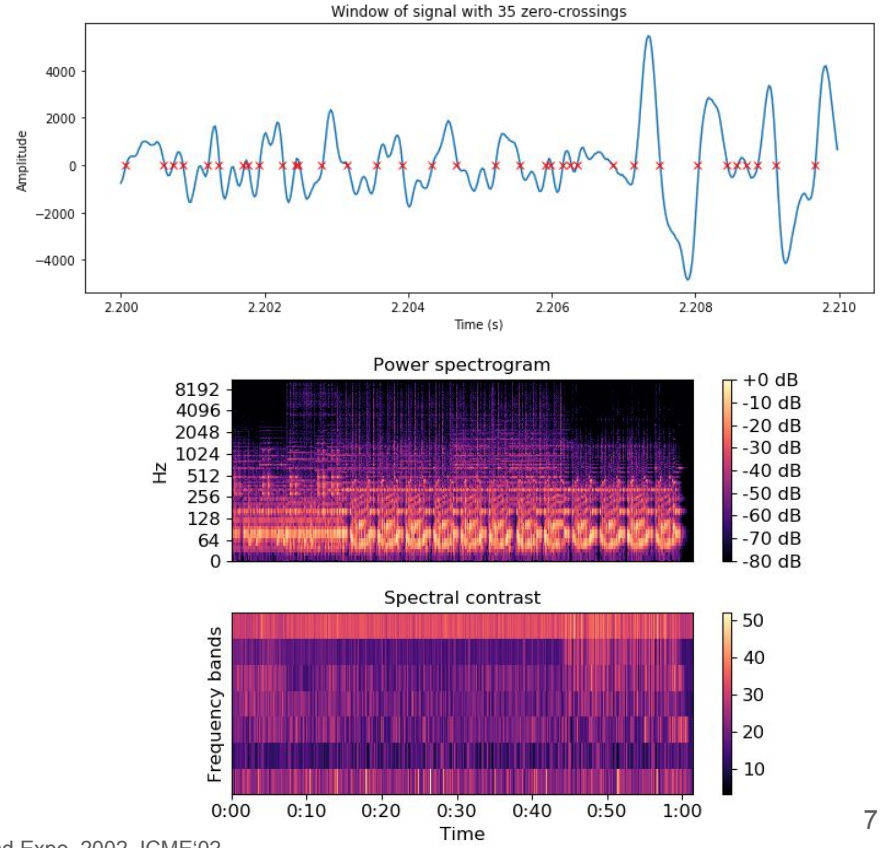



[1] R.F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, “Perceptual evaluation on audio-visual dataset of 360 content,” in 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2022, pp. 1–6.
[2] T. Robotham, A. Singla, O. S. Rummukainen, A. Raake, and E. A. P.Habets, “Audiovisual database with 360° video and higher-order ambisonics audio for perception, cognition, behavior, and qoe evaluation research,” in 14th International Conference on Quality of Multimedia Experience, Lippstadt, Germany, 2022, pp. 1–6.

# Source content characterisation: audio part

Zero Crossing Rate (ZCR) and Spectral Contrast (SC) [6] computed on W-ambisonic channel (0th order):

- ZRC express rate at which signal is changing sign: translating the smoothness of a signal
- SC measures energy differences between spectral peak and valley in each frequency subband, see [6]. SC translate the clarity of a signal



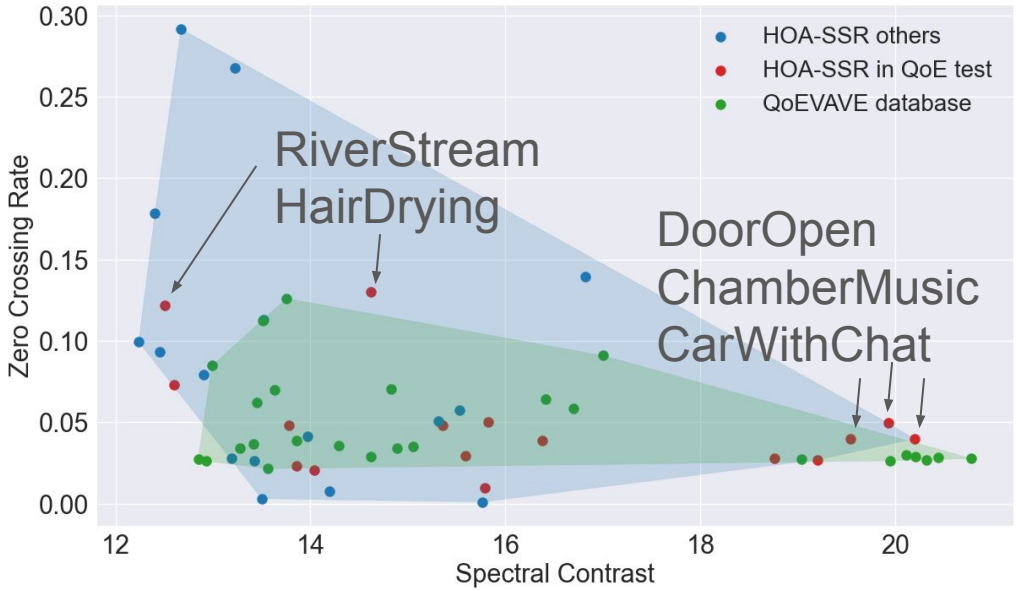



[6] Jiang, Dan-Ning, and al. Cai. “Music type classification by spectral contrast feature.” In Multimedia and Expo, 2002. ICME‘02.

# Source content characterisation: audio part

ZCR and SC computed on W-ambisonic channel (0th order):

- 16 sequences of HOA–SSR database QoE subjectively evaluated in [1]
- other HOA–SSR database sequences (not evaluated)
- QoEVAVE database [2]



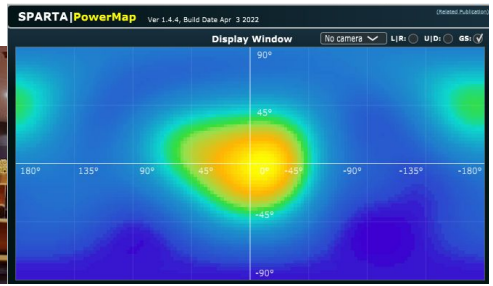

[1] R.F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, “Perceptual evaluation on audio-visual dataset of 360 content,” in 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2022, pp. 1–6.
[2] T. Robotham, A. Singla, O. S. Rummukainen, A. Raake, and E. A. P.Habets, “Audiovisual database with 360° video and higher-order ambisonics audio for perception, cognition, behavior, and qoe evaluation research,” in 14th International Conference on Quality of Multimedia Experience, Lippstadt, Germany, 2022, pp. 1–6.

# ERPs of 16 evaluated sequences in [1]: diversity of scenes

[1] R.F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, “Perceptual evaluation on audio-visual dataset of 360 content,” in 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2022, pp. 1–6.

# Selected test sequences for our test reproduction with naives observers

A sub-group of 7 sequences + 1 sequences for calibration selected based on SI–TI and audio characteristics: source nature, clarity, location and movement




Sequence used for participants calibration and associated 4HOA power map of localized sound with reverb

# About the tested Hypothetical Reference Circuits (HRCs)

HRCs covers video and audio degradations

- Video degradations:
  - SRCs ERP in 6K and encoded QP0 (6KQP0 – Hidden Reference)
  - Downscaling of ERP degradation to 4K or 2K
  - Encoding of ERP with libx265 at fix QPs (0, 22, 28, 34)
- Audio degradations:
  - SRCs audio 1152 kbps/channel “PCM”
  - Audio encoding with AAC–LC at (64, 32, 16 kbps/channel)

→ 48 HRCs all evaluated in [1]

[1] R.F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, “Perceptual evaluation on audio-visual dataset of 360 content,” in 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2022, pp. 1–6.

# Methods tested in “consumer-grade” setup for audiovisual quality assessment

# Summary of the experiments: methodologies

ACR-HR methodology: ITU-T Rec. P.910

- 5-grade ACR scale
- no repetition
- calibration over 3 PVS: low, mid, high quality
- conversion of raw OS to DMOS
- more than 25 assessors

| Scores | Quality items |
|---|---|
| 5 | Excellent |
| 4 | Good |
| 3 | Fair |
| 2 | Poor |
| 1 | Bad |

$$DMOS_o^j = 5 - (MOS_o^{ref} - MOS_o^j)$$

$$DMOS^j = \frac{1}{N}\sum_{o=1}^{N} DMOS_o^j$$

# Summary of the experiments: methodologies

DCR methodology: ITU-T Rec. P.910

- 5-grade DCR scale
- no repetition
- calibration over 3 PVS: low, mid, high impairment
- more than 25 assessors

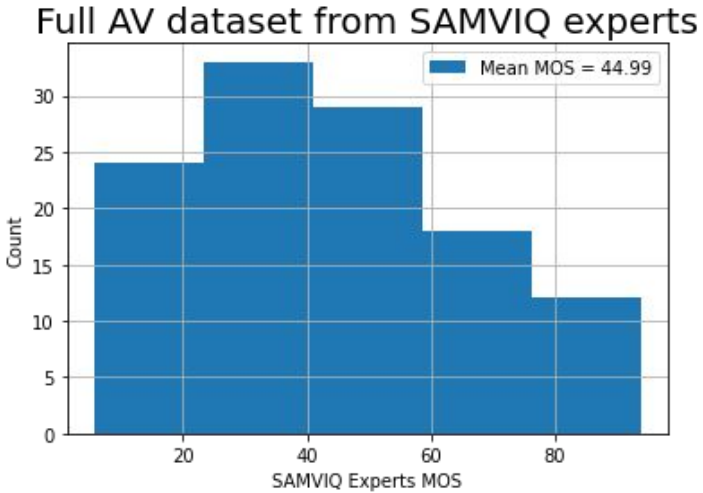
| Scores | Impairment items |
|---|---|
| 5 | Imperceptible |
| 4 | Perceptible but not annoying |
| 3 | Slightly annoying |
| 2 | Annoying |
| 1 | Very annoying |

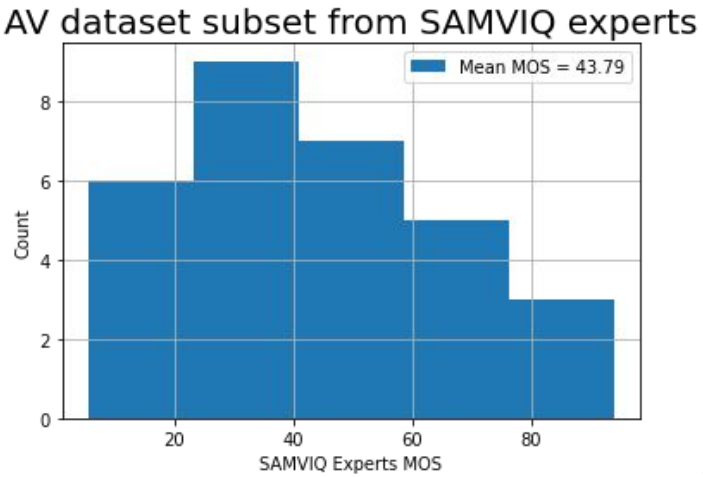
$$| Source | 5\,sec\ Gray\ \ Screen | Impaired\ \ PVS_A |$$

# Tested HRCs for ACR–HR and DCR conditions

48 possible HRCs

Subset of 6x4 HRCs: video (6KQP0, 6KQP22, 4KQP0, 4KQP28, 2KQP0, 2KQP34) and audio degradation (PCM, 64, 32, 16 kbps/channel) encoded with AAC.



# Tested HRCs for ACR–HR and DCR conditions

DCR test:

- 5 SRCs, 6PVS per SRC from above HRCs: 30 PVS
- 25 naives assessors for a 30min in-lab experiment (24 min for annotation)

ACR-HR test:

- 7 (5+2) SRCs, 6 same PVS per SRC + Hidden Reference: 49 PVS
- HR is HRC 6K_QP0_PCM
- 25 naive assessors for a 30min in-lab experiment (19 min for annotation)

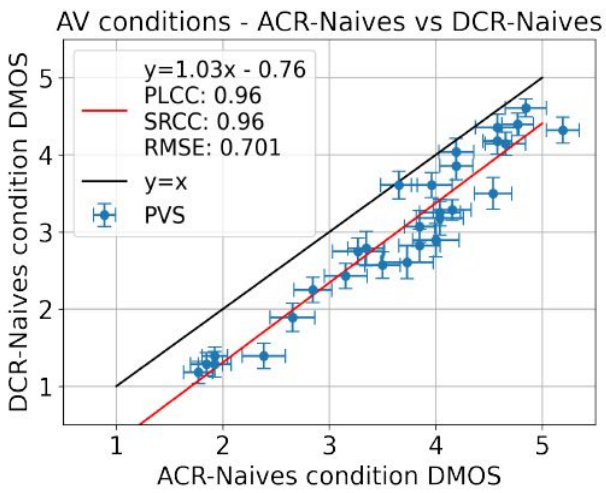
# QoE conditions subjectively evaluated recap

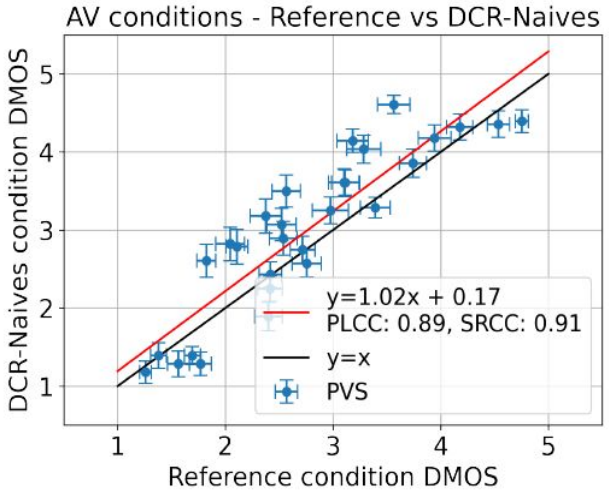
| Condition name | Sound system | Assessors | Rating scale |
|---|---|---|---|
| “Reference” from [1] | 26 channel loudspeaker | Trained following procedure in [2] | SAMVIQ |
| “DCR–Naives” | HMD headphone (binaural rendering) | naive | DCR |
| “ACR–Naives” | HMD headphone (binaural rendering) | naive | ACR–HR |

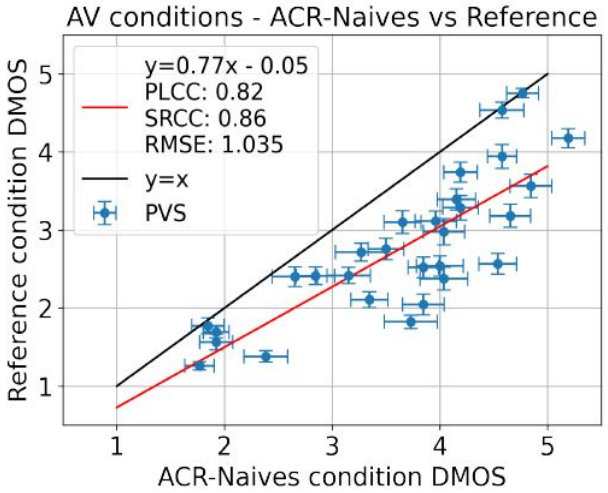
[1] R.F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, “Perceptual evaluation on audio-visual dataset of 360 content,” in 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2022, pp. 1–6.
[2] Randy Frans Fela, Nick Zacharov, and Søren Forchhammer, “Assessor selection process for perceptual quality evaluation of 360 audiovisual content,” Journal of the Audio Engineering Society, vol. 70, no. 10, pp. 824–842, 2022.

# Results and analysis

# Usage of the scale across conditions




red and black line close = similar overall usage of the scale range

Strong agreement between “DCR–naives” and “ACR–naives” experiments

Agreement lower between “Reference” and “DCR–naives”, and low between “Reference” and “ACR–naives”

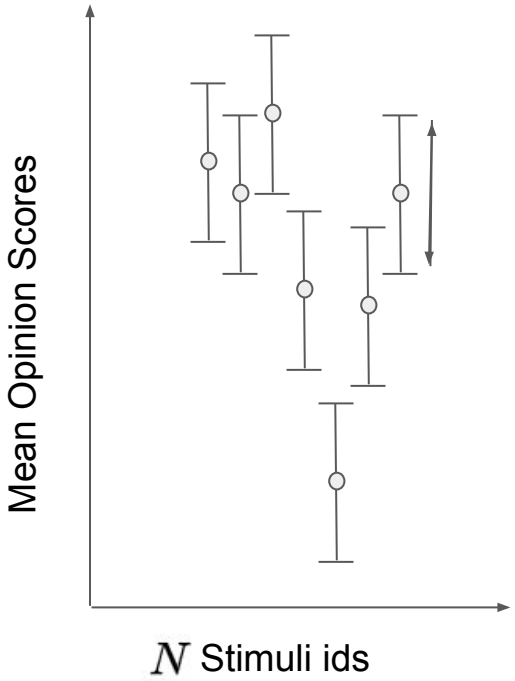
# Going deeper into MOS analysis:

## Subjective data precision and subjective methodologies efficiency

# Data precision: mean MOS Confidence Interval size

Mean CI: average over all the estimated MOS Confidence Intervals - smaller is better

$$Mean_{CI} = \frac{1}{N} \sum_{n=1}^{N} CI_{MOS}^{n}$$



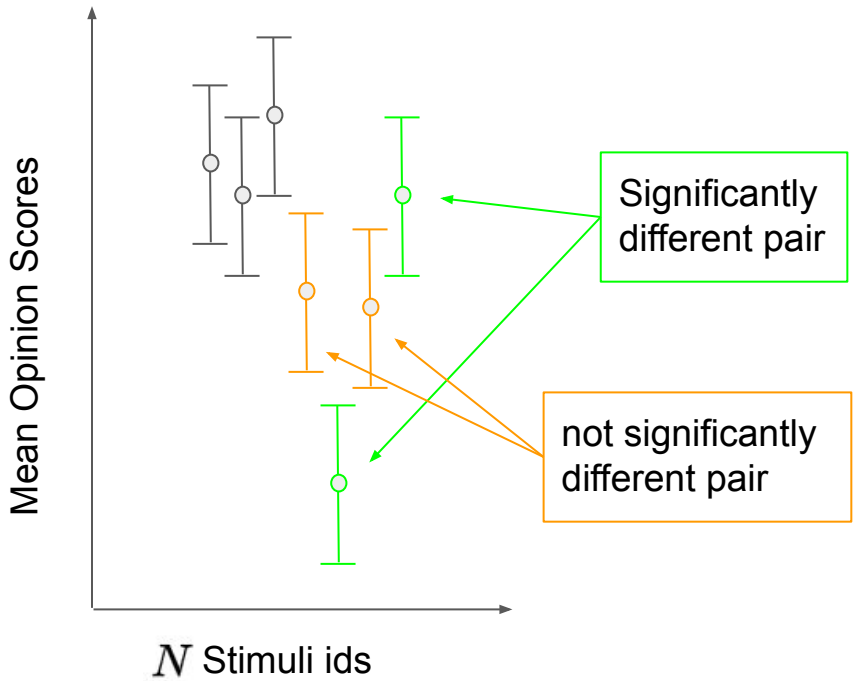
# Data precision: discriminability ratio on MOS

T-test analysis on pairs of stimuli MOS

Discriminability ratio: number of significantly different pairs among all the possible ones - higher is better

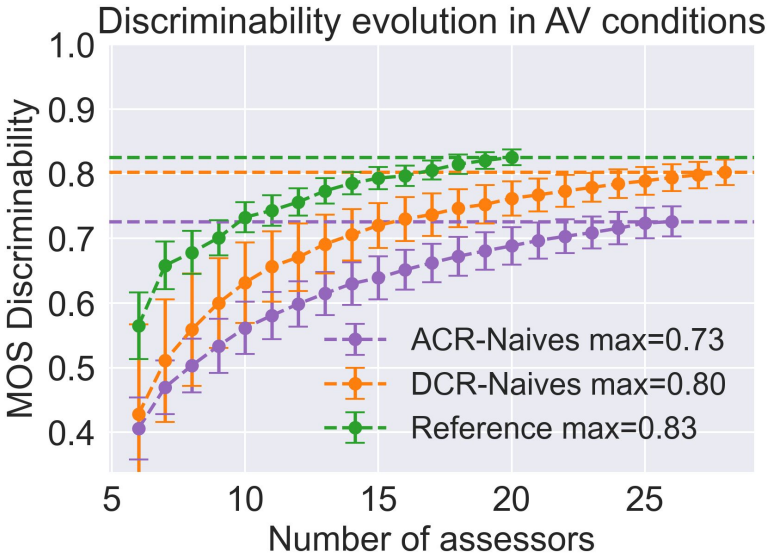
$$D_{ratio} = \frac{1}{M} \sum_{m=1}^{M} Sig_m$$

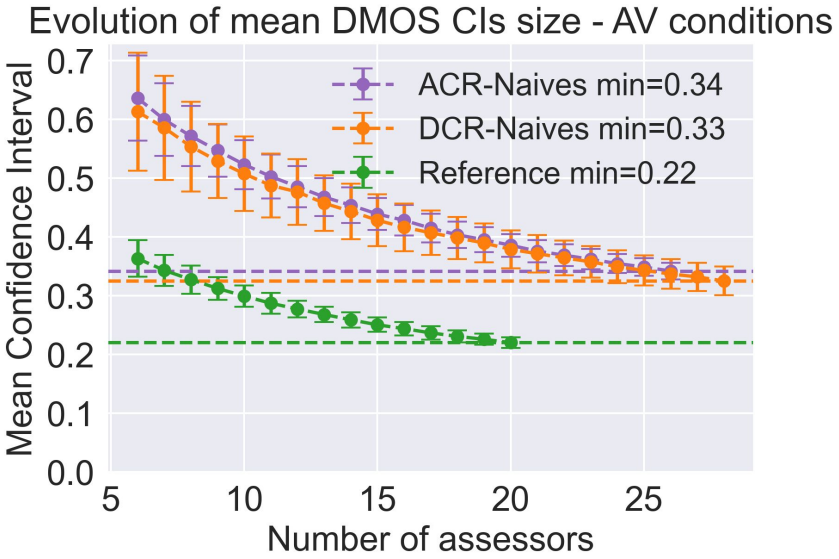
$$M = \frac{N * (N-1)}{2}$$




[1] Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué, “Comparison of subjective methods for quality assessment of 3D graphics in virtual reality,” ACM Transactions on Applied Perception (TAP), vol. 18, no. 1, pp. 1–23, 2020.

# Data precision and discriminability at same assessors number

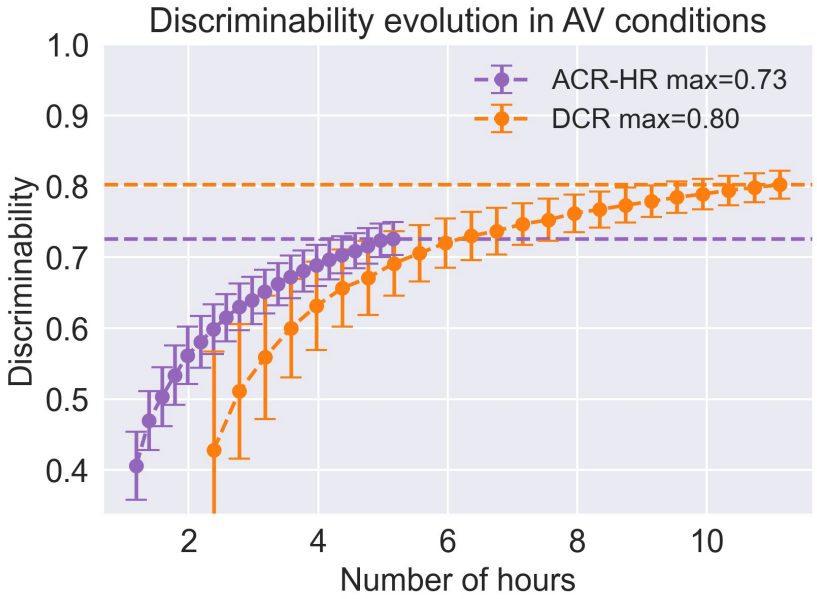


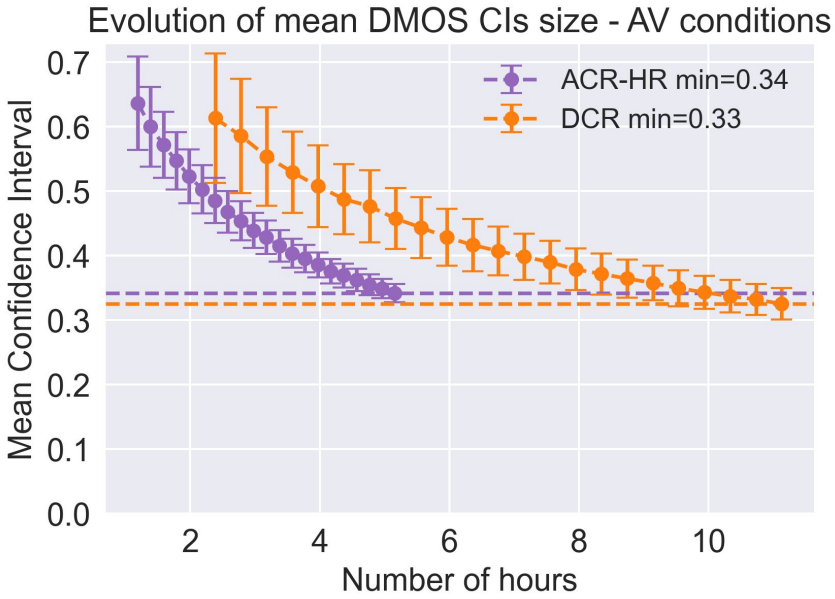




Reference condition with trained assessors gives highest discriminability and smallest average CI size

ACR–Naives and DCR–Naives have similar average CI size but the discriminability is higher in the DCR–Naives conditions

# Data precision and discriminability at same experimental effort



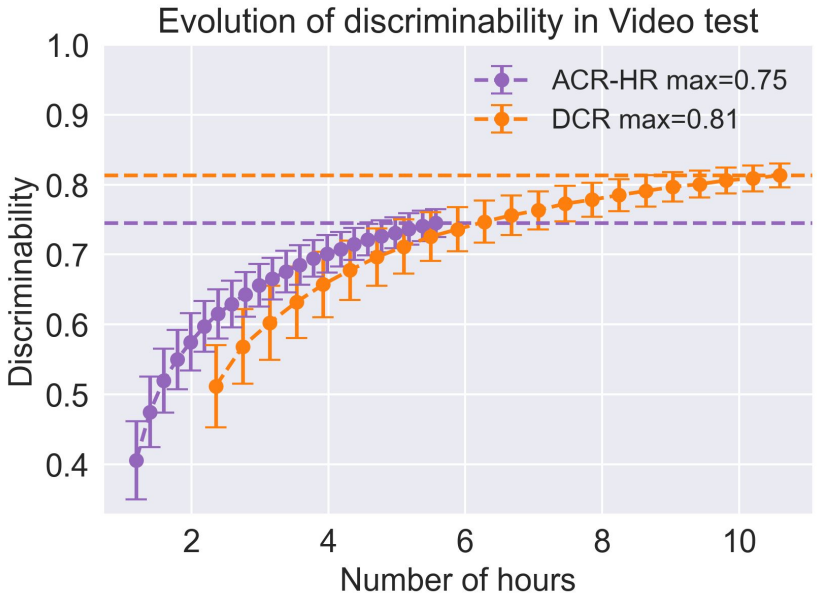



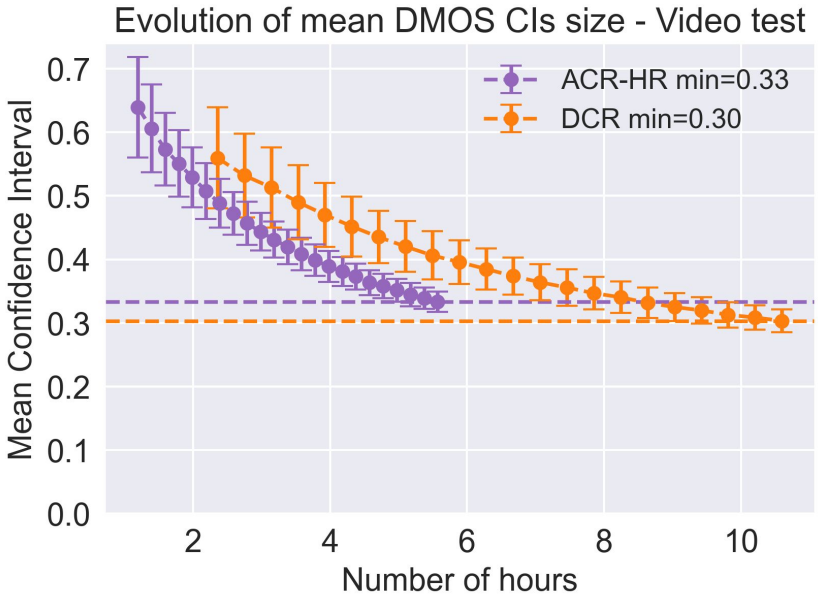

It is difficult to estimate the hours spent for training phase of Reference condition with trained assessors

ACR–Naives has benefit at first to collect accurate estimate but with increasing experimental effort DCR remains more efficient and require a smaller population

*Additional slide for video-only experiment (no audio playback), assessors judge only video quality

# Data precision and discriminability at same experimental effort




ACR–Naives has benefit at first to collect accurate estimate but with increasing experimental effort DCR remains more efficient and require a smaller population (eventually ACR–Naives could become less efficient method)

Discriminability more important than Confidence Interval mean as discriminability comes with small CI.

# Objective metrics and resolving power

# Audiovisual objective quality metrics

Goals:

- currently, no existing audiovisual quality metrics benchmark on HOA
- Green coding and efficiency of two modalities evaluation

We proposed to explore ITU bitstream and parameter based methods, as well as accurate and proven traditional Full Reference metrics VMAF [1] and ViSQOL [2]

[1] Netflix, “VMAF v0.6.1 Model,” https://github.com/Netflix/vmaf.
[2] Andrew Hines, Eoin Gillen, Damien Kelly, Jan Skoglund, Anil Kokaram, and Naomi Harte, “ViSQOLAudio: An objective audio quality metric for low bitrate codecs,” The Journal of the Acoustical Society of America, vol. 137, no. 6, pp. EL449–EL455, 2015.

# Five audiovisual quality metrics

**VMAF_ViSQUOL**: VMAF as video quality predictor from ERP is combined with ViSQOL for audio using linear combination defined for audiovisual quality integration module of ITU-T Rec. P.1203.3 [1]

**P1204m3_ViSQUOL**: bitstream based video model from ITU-T P.1204 [2] combined with ViSQOL

Two parameter-based models from ITU-T Rec. P.1203: parameters are video bitrate, resolution, framerate and audio bitrate and adapted to HTC Vive viewport size by scaling factor on video bitrate.

Scaling factor is ratio between HMD viewport size and ERP size

$$Ratio = \frac{360 \times 180}{110 \times 110} \approx 5.35$$

[1] ITU-T Rec. P.1203.3, “Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport quality integration module,” 2019.
[2] ITU-T Rec. P.1204.3, “Video quality assessment of streaming services over reliable transport for resolutions up to 4k with access to full bitstream information,” 2020.

# Five audiovisual quality metrics

**VMAF_ViSQUOL**: VMAF as video quality predictor from ERP is combined with ViSQOL for audio using linear combination defined for audiovisual quality integration module of ITU-T Rec. P.1203.3 [1]

**P1204m3_ViSQUOL**: bitstream based video model from ITU-T P.1204 [2] combined with ViSQOL

Two parameter-based models from ITU-T Rec. P.1203: parameters are video bitrate, resolution, framerate and audio bitrate and adapted to HTC Vive viewport size by scaling factor on video bitrate.

- **P1203m0-O35v**: O35 beeing quality score after audiovisual and temporal pooling
- **P1203m0-O46v**: O46 final prediction of ITU-T Rec. P.1203 mode 0 (O35 + RF model)

**P1203m0-022v_ViSQUOL**: to replace parameter based audio quality estimation by ViSQUOL

[1] ITU-T Rec. P.1203.3, “Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport quality integration module,” 2019.
[2] ITU-T Rec. P.1204.3, “Video quality assessment of streaming services over reliable transport for resolutions up to 4k with access to full bitstream information,” 2020.

# Performance of the selected audiovisual metrics

SRCC evaluation of metrics on the three conditions dataset

Parametric based models from P1203 adapted to viewport size are performing the best across the 3 conditions

VMAF_ViSQUOL with linear combination from ITU-T Rec. P.1203.3 the worst (*could be improve with better linear combination parameters)

**Table 17**: Objective metrics SRCC performance on data of the three audiovisual conditions: bold for best and italic for worst.

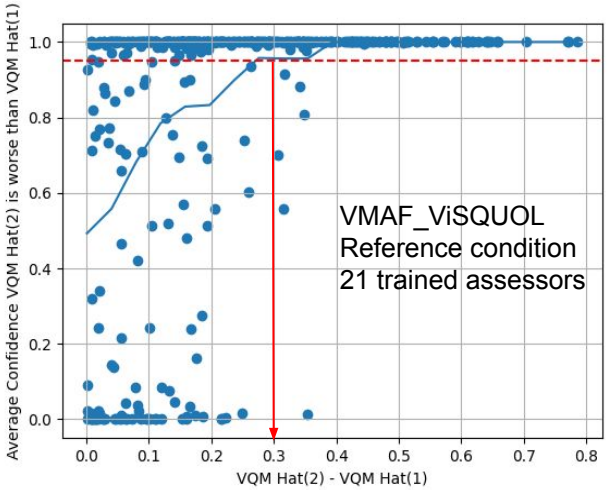
| Metrics | Reference | ACR–Naives | DCR–Naives |
|---|---|---|---|
| VMAF_VISQUOL | *0.846* | 0.855 | 0.870 |
| P1203m0-O22v_VISQUOL | **0.882** | 0.857 | 0.879 |
| P1204m3_VISQUOL | 0.877 | *0.689* | *0.731* |
| P1203m0-O35v | 0.872 | 0.882 | **0.920** |
| P1203m0-O46v | 0.870 | **0.892** | 0.895 |

# Resolving power: ITU-R Rec BT.1676 [1]

The resolving power of a VQM can be defined as the difference between two VQM values for which the corresponding subjective-score distributions have means that are statistically different from each other (typically at the 95% significance level).

- form pairs of stimuli (blue dots)
- z-test performed on pairs, significance on y-axis
- deltaVQM per pair on x-axis

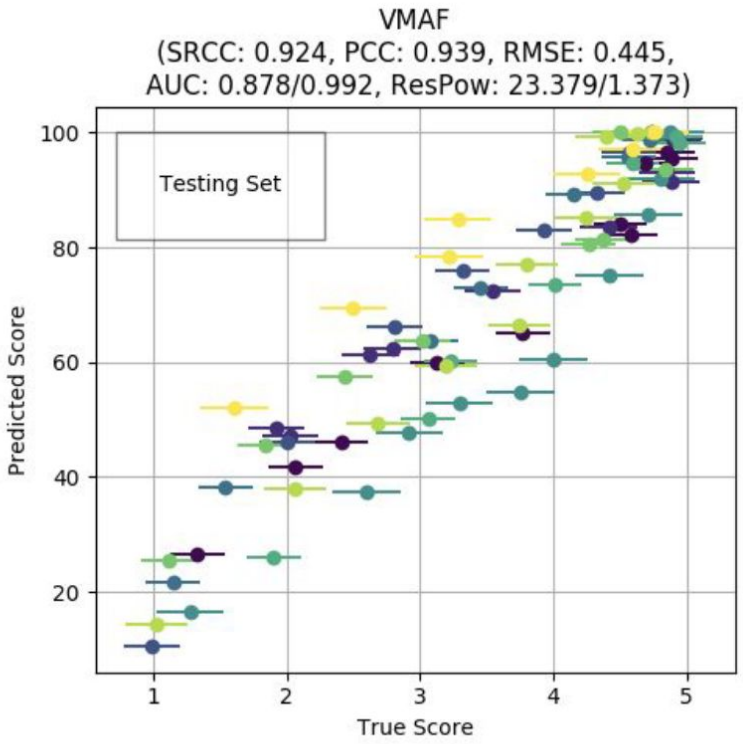
⇒ RP@95% of VMAF_ViSQUOL is 0.3 (scale 0–1)




[1] ITU-R Rec. BT.1676, “Methodological framework for specifying accuracy and cross-calibration of video quality metrics,” 2004.

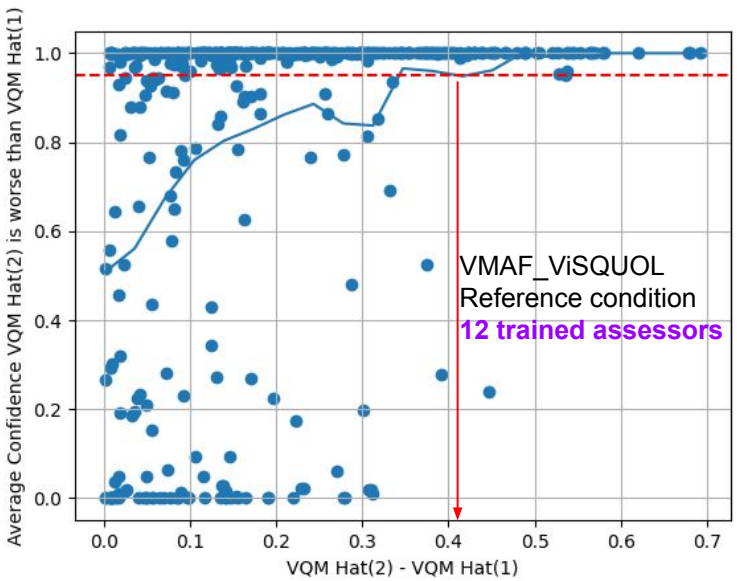
# From VQEG_SAM_2018_111_AnalysisToolsInVMAF

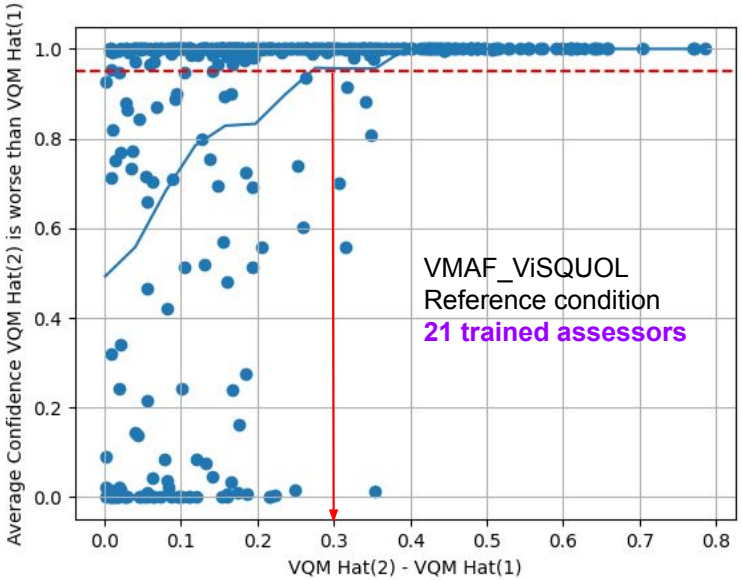
RP of 0.23 (0–1 scale)




- ResPow
  - 23.379 - resolv. power in VMAF score scale (0 - 100)
  - 1.373 - resolv. power in subjective scale (1 - 5)
- AUC
  - 0.878 - different/similar (DS) AUC analysis
  - 0.992 - better/worse (BW) AUC analysis

# Comparison of a metric resolving power on different data accuracy




Adding more assessors, retrieving more accurate data, helps to improve the estimated resolving power of metrics

# Resolving Power threshold as a function of discriminability in datasets (scale 0–1)

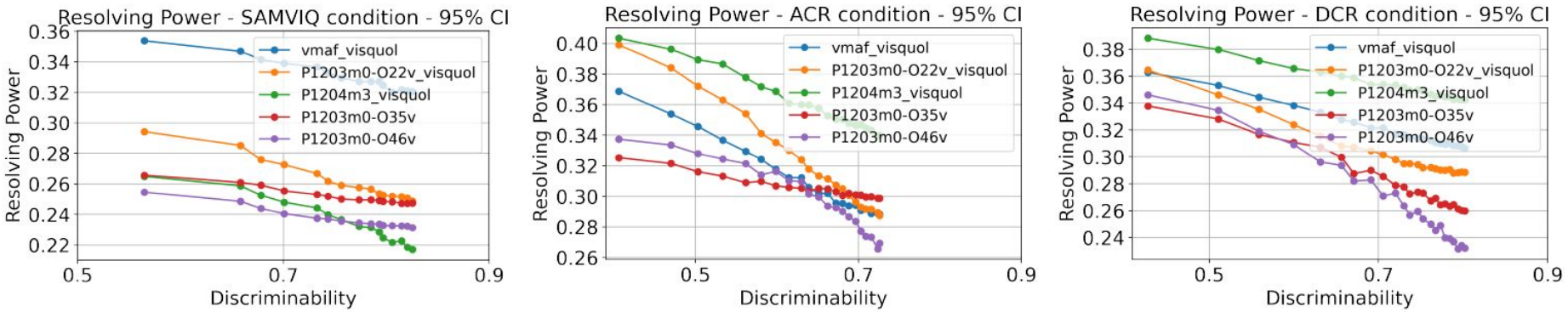



**Fig. 3**: Audiovisual objective quality metrics Resolving Power [14] for the 3 conditions in the function of the discriminability

Higher discriminability in the subjective data helps to improve the resolving power evaluation of objective quality metrics

P1203 mode 0 models have best Resolving Power score at high discriminability

link to paper

# Conclusions

- comparison on efficiency of subjective methodologies for quality assessments of 360° videos with HOA audio across different setups
- High discriminability obtained by trained assessors can be replicated with naive assessors (DCR)
- Explore relation between discriminability and resolving power of quality metrics
- proposal of a new parameter-based quality estimation model adapted to viewport resolution

# Futurs works

- Other collected data to analyse:
  - on video only quality scores:
    - video only versus video + audio: explore the impact of HOA degradation in QoE
    - comparison with/without audio impact in both “consumer-grade” and “reference” setup
  - from eye tracking data:
    - impact of audio on content exploration
    - impact of methodologies for subjective evaluation on content exploration