

Subjective Analysis and Objective Characterization of Adaptive Bitrate Videos

Jacob Søgaard, Samira Tavakoli, Kjell Brunnström,
Narciso Garcia



POLITÉCNICA



Mittuniversitetet
MID SWEDEN UNIVERSITY



Motivation: Evaluating HAS QoE

- Increasing number of services based on HTTP Adaptive Streaming (HAS)

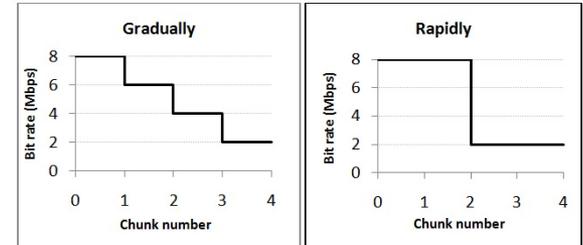


- High interest on identifying the factors influencing on Quality of Experience (QoE) of HAS

Challenges

- **Multiple parameters can influence**

- Switching frequency
- Switching amplitude
- Content
 - Genre
 - Objective characterization



- **Goal: Evaluation (improvement) of whole HAS session**

- Analyzing the effect of each parameters
 - Testing certain patterns of quality switching
 - NOT the video quality in each time instant

Challenges

- **Adaptation event can last up to several seconds**

- **Lack of appropriate testing method**

- Mostly for assessment of short sequences (~10 sec)

- Absolute Category Rating (ACR)



- Recency, hysteresis effect and decreasing user attention

- Single Stimulus Continuous Quality Evaluation (SSCQE)



- **Hard to design experiments on event level**

- Test session should not last longer than 1.5 hrs (*Rec. ITU-T P.913*)

- Avoiding user fatigue and boredom

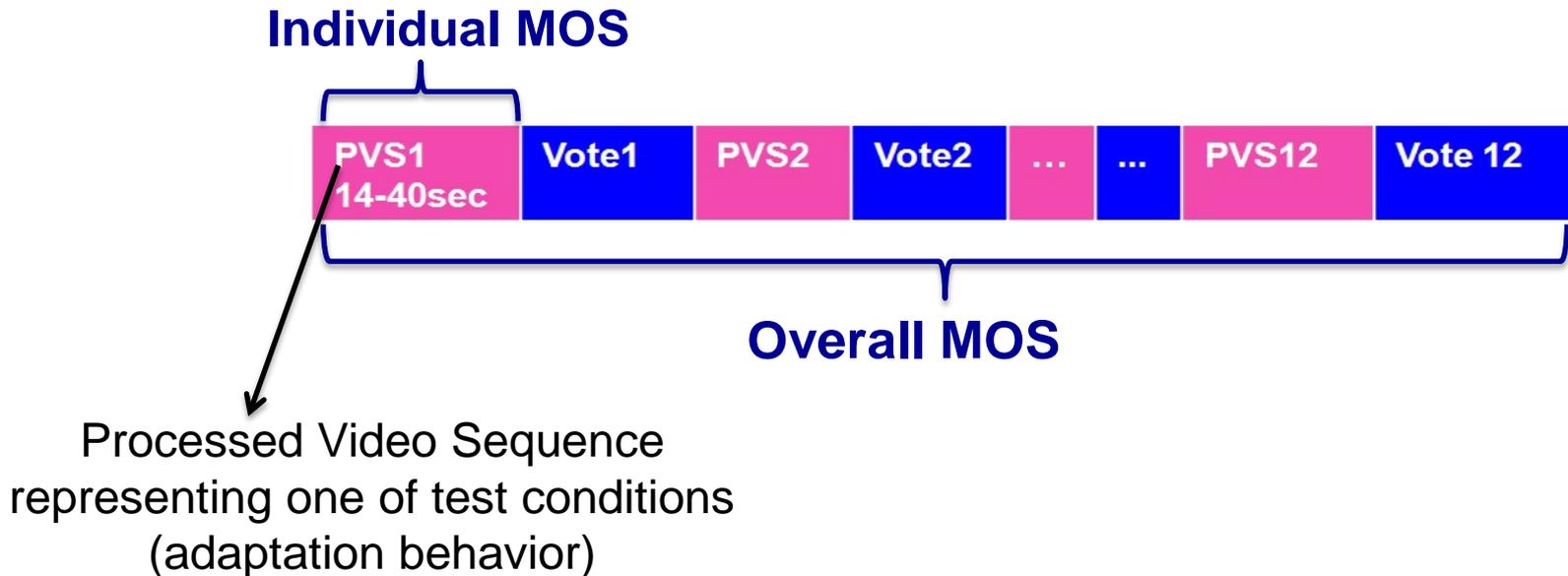
Objectives of this study

1. Toward improving **subjective testing methodology** for adaptive streaming
 - Subjective Experiment
2. **Objective characterization of adapted videos** for understanding **subjective ratings**
 - Objective characterization

Subjective Experiment

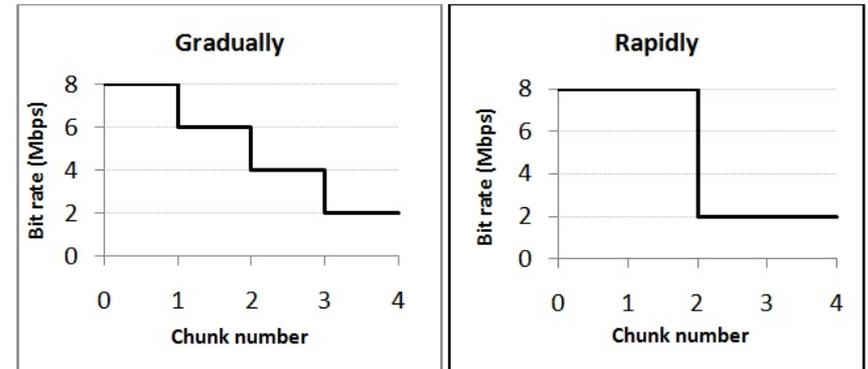
Toward improving subjective testing methodology:

- Impact of evaluation methodology on test subjects evaluation
- Relationship between **MOS of individual adaptation events** and **overall MOS of whole sequence**



Adaptation study parameters

- Switching behavior
 - Period (chunk length): 2 sec and 10 sec
 - Amplitude (difference between consecutive quality levels)
gradual vs. rapid switching
 - Decreasing and increasing



- Adaptation dimension
 - Video quality (QP)
 - H.264/AVC
 - 4 streams: 5Mbps, 3Mbps, 1Mbps, 600kbps
 - 1280×720/ 25fps
- MOS based on > 60 test subjects
 - 132 adaptation events with variable length video

Test video sequences

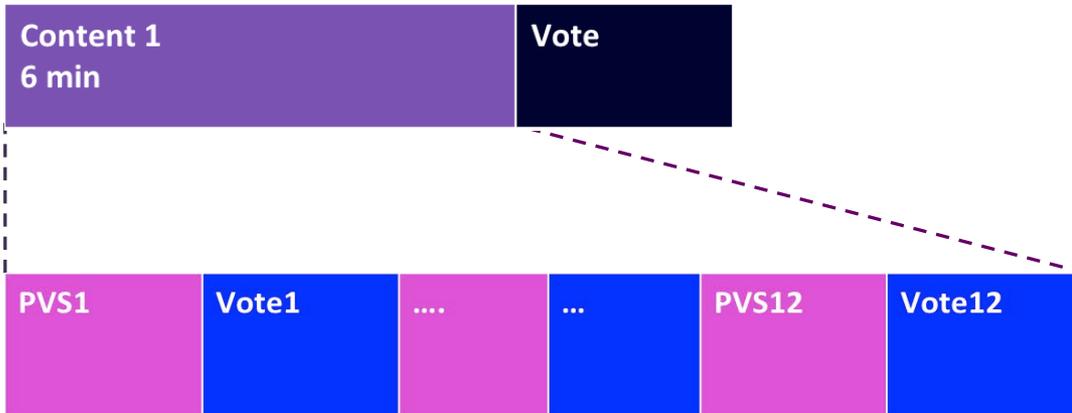


- 7 commercial content
 - Movie, Documentary, Sport, News, Music
- 6 min; originally 1080p Blu-ray video, 24/50 fps
- Different spatial and temporal characteristics

Evaluation methodology

Cross-lab experiment (same PVS and rating question in both labs)

- ① **UPM:** Continuous presentation of 6min video including subsequent PVS
- Evaluation of individual PVS and overall quality of each video
 - Two experiments: **with and without audio**



Rate the following properties of the video sequence

Overall Quality

Excellent

Good

Fair

Poor

Bad

Did you see a change?

Decreased Quality

Not visible

Increased Quality

OK

- ② **Acree:** Evaluation of overall quality of 6 min video



Experimental Setup

- Environment and equipment in both labs
 - Satisfying Rec. ITU-R BT. 500-11
 - 46" Hyundai S465D display
 - Controlled lightning system
 - Viewing distance: 4H
- Observers' number (after screening)
 - UPM-Audio: 21 (6 female & 15 male)
 - UPM-NonAudio: 22 (5 female & 17 male)
 - Acreo: 30 (10 female & 20 male)
- Test presentation
 - One subject per session
 - Training



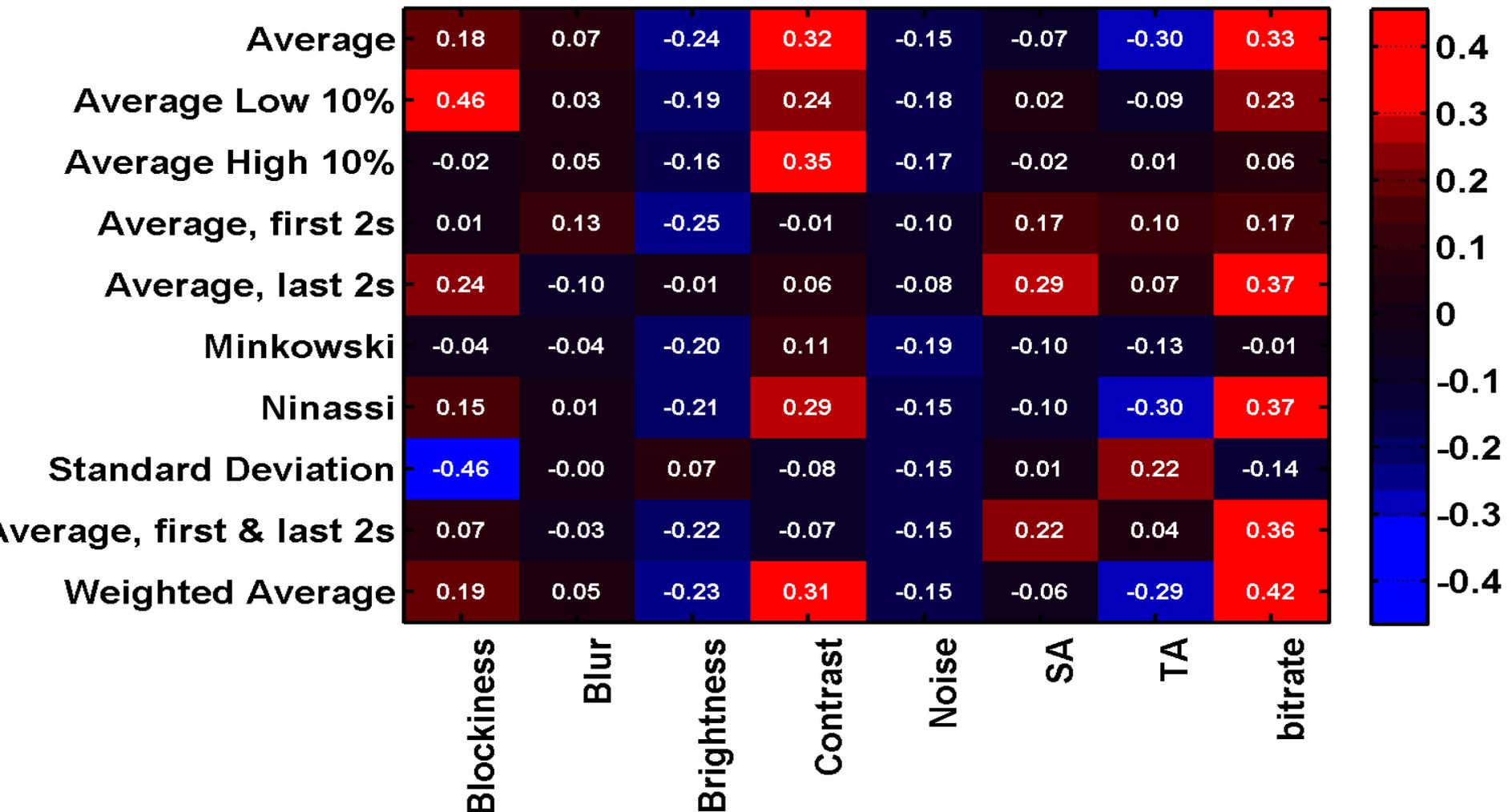
Objective Characteriation

Aim: Identify the PVS characteristics influential on QoE

Approach:

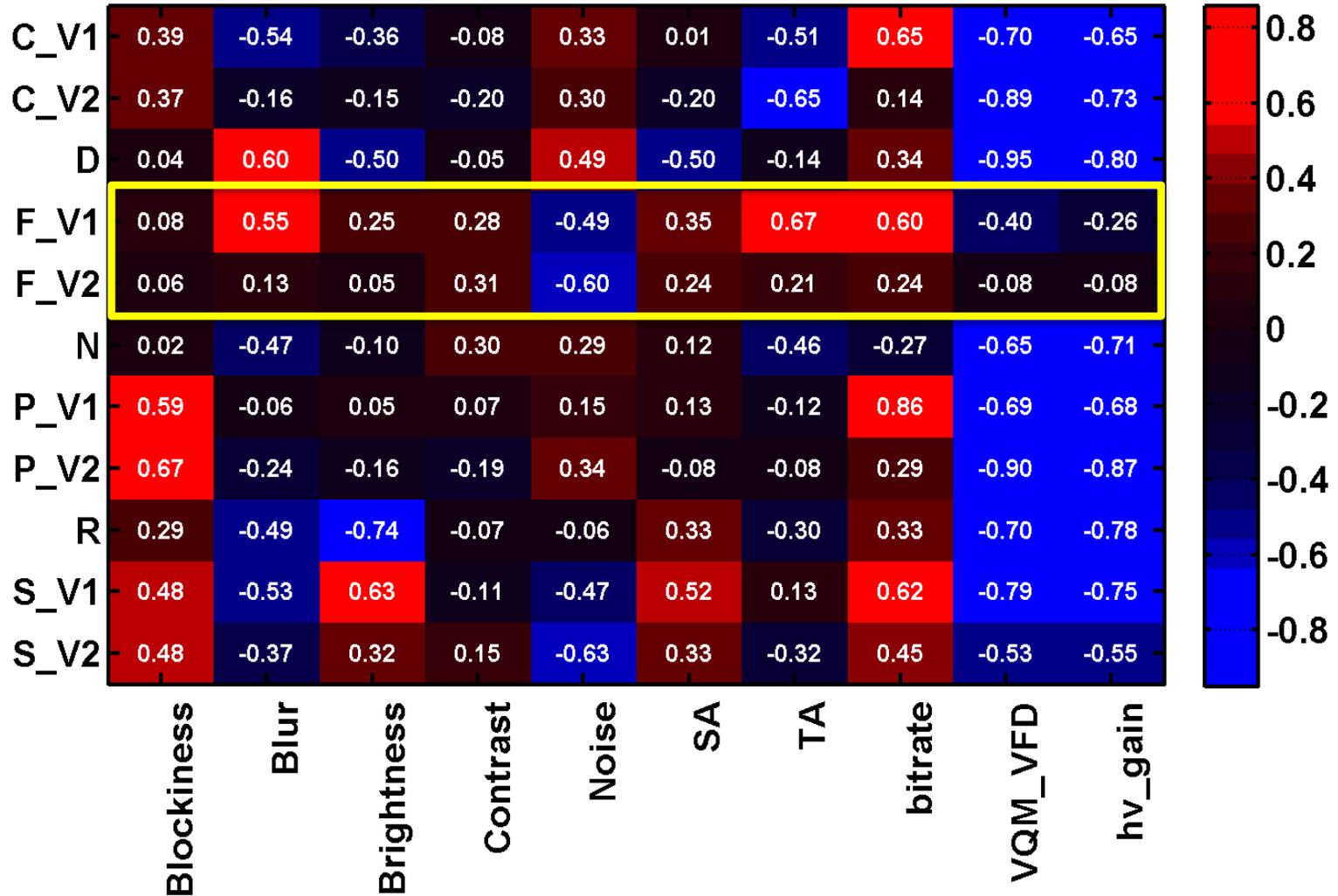
- No-Reference objective characterization tools
 - Blockiness
 - Blur
 - Brightness
 - Noise
 - Bitrate
 - Contrast
 - Spatial Activity
 - Temporal Activity
- Full-Reference tool
 - Video Quality Model with Variable Frame Delays (VQM-VFD)
- Temporal pooling technique
 - Avrage
 - Av. low 10%
 - Av. High 10%
 - Av. first 2 sec
 - Av. first 2 sec
 - Av. first & last 2 sec
 - Weighted Average
 - Standard Deviation
 - Minkowski
 - Ninassi

SROCC Result: NR measurement and MOS



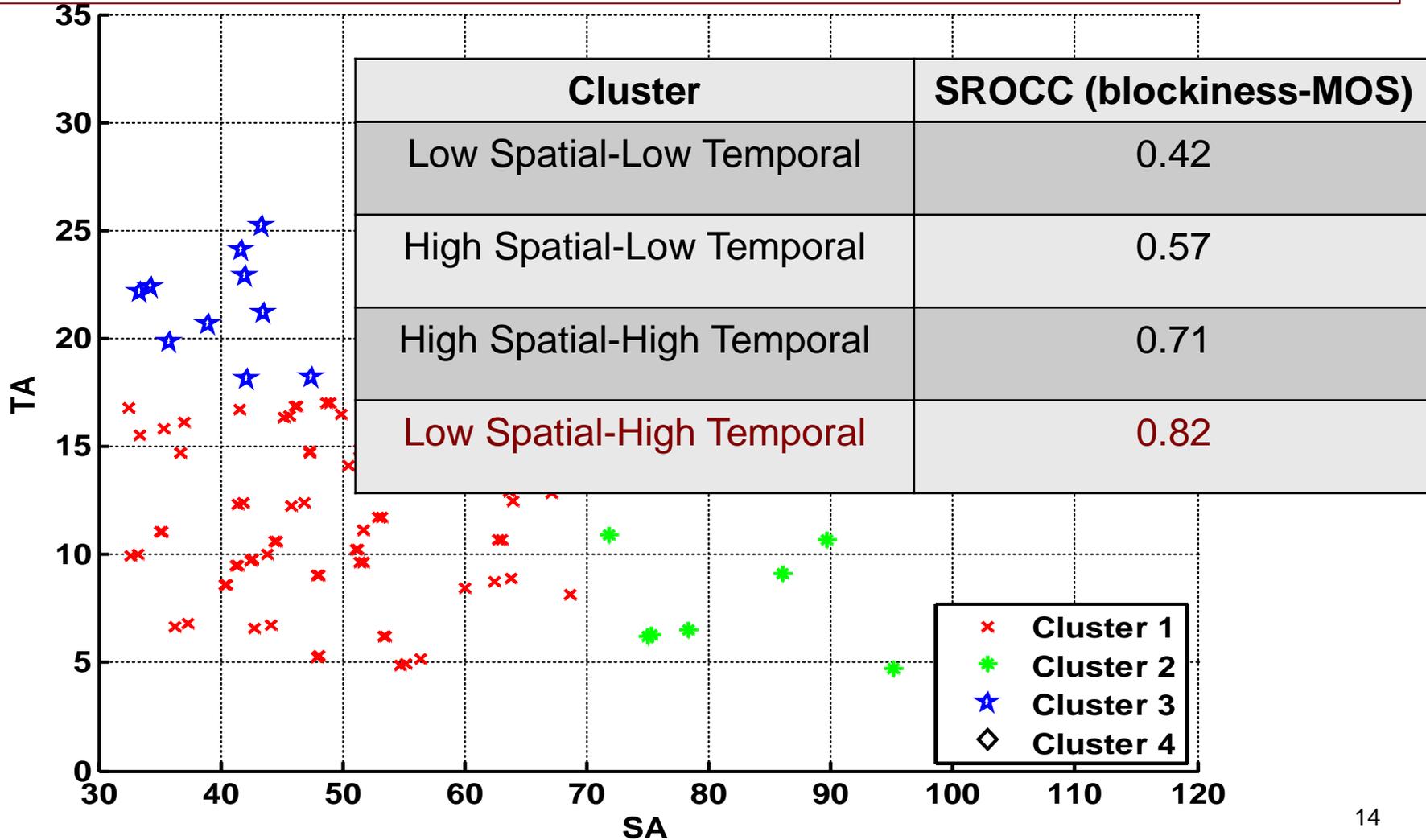
SROCC Result: NR and FR measurement for each content

Pooling approach for NR measurement: Average

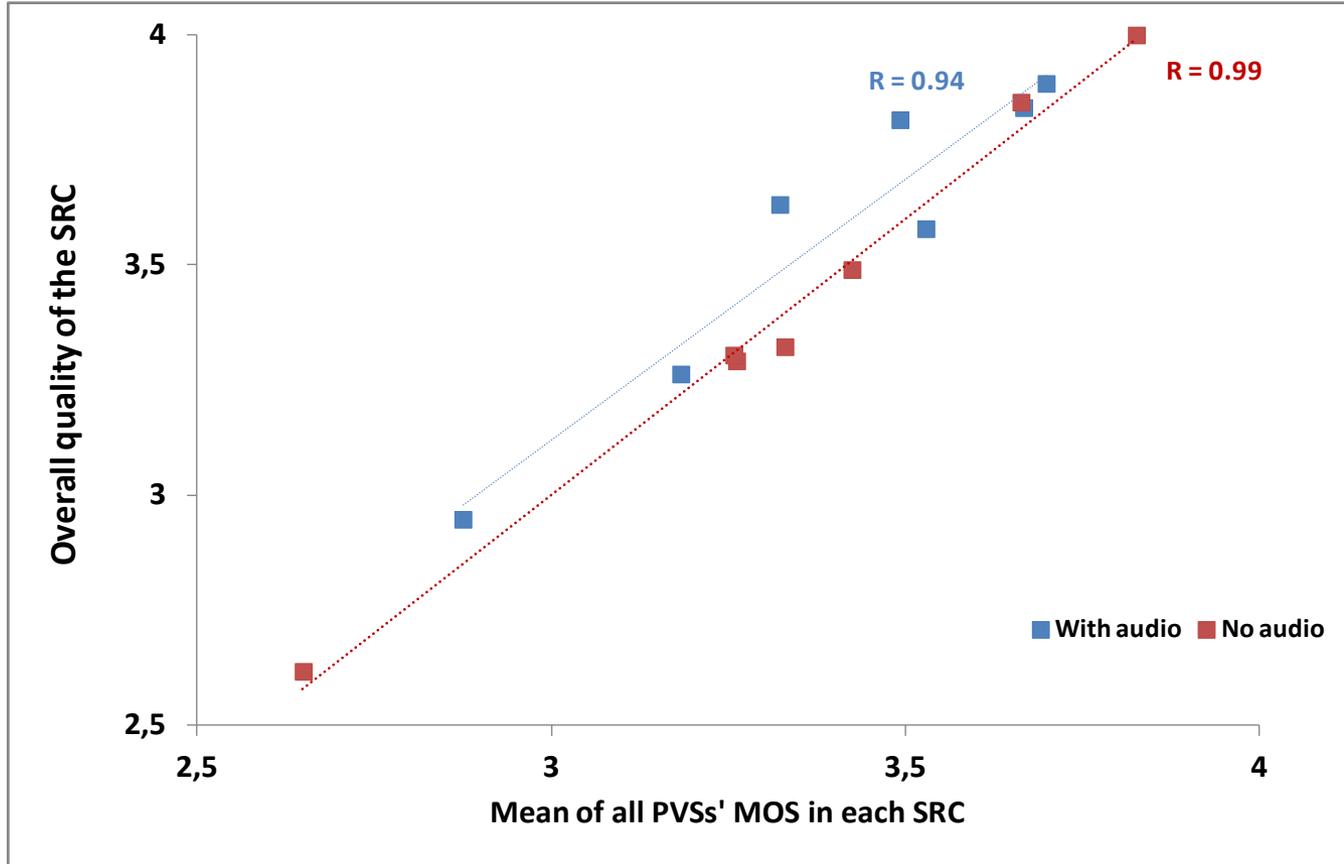


Clustering based on Spatial and Temporal Activity

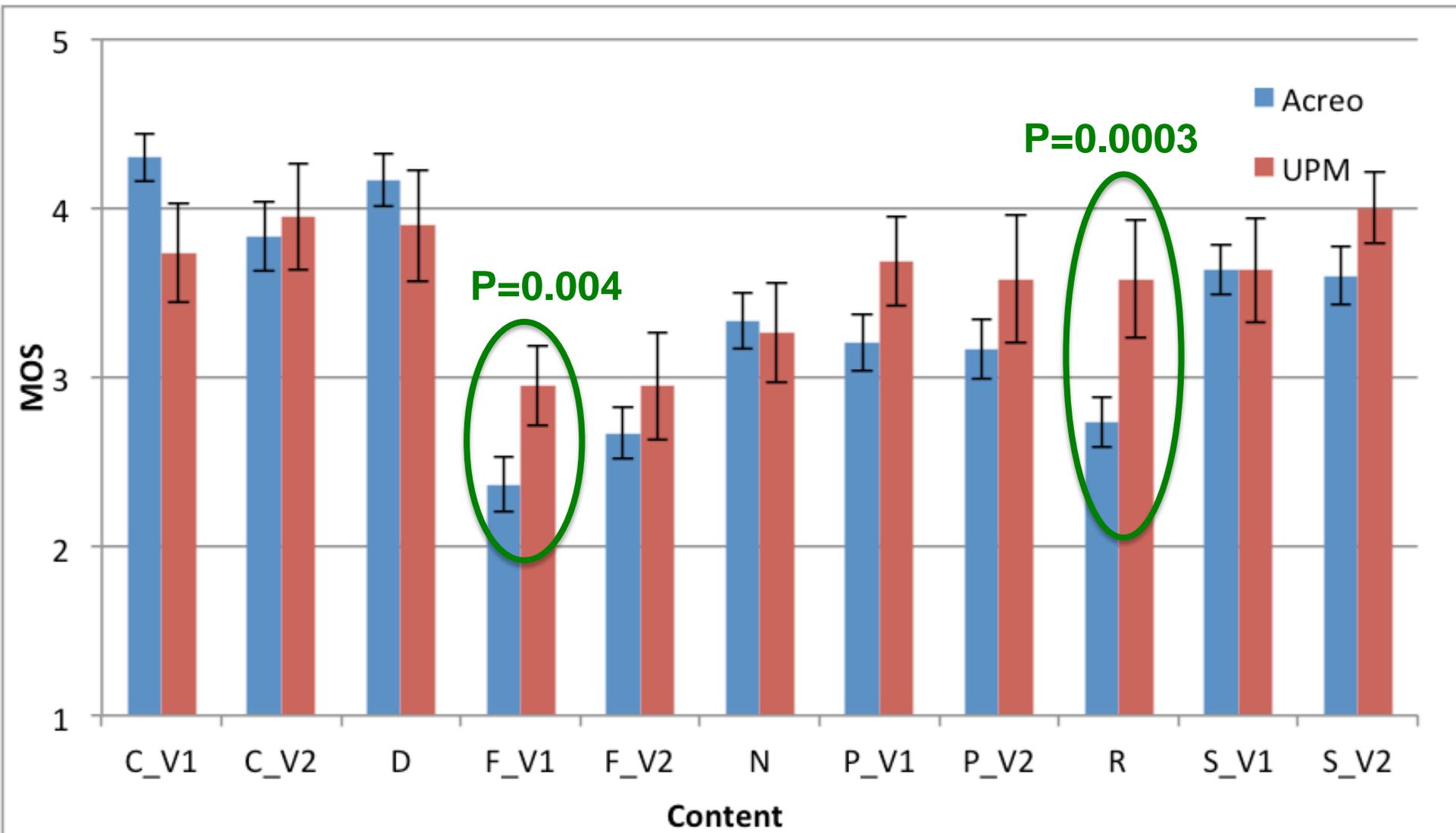
Pooling: Average of 10% of the lowest NR blockiness values



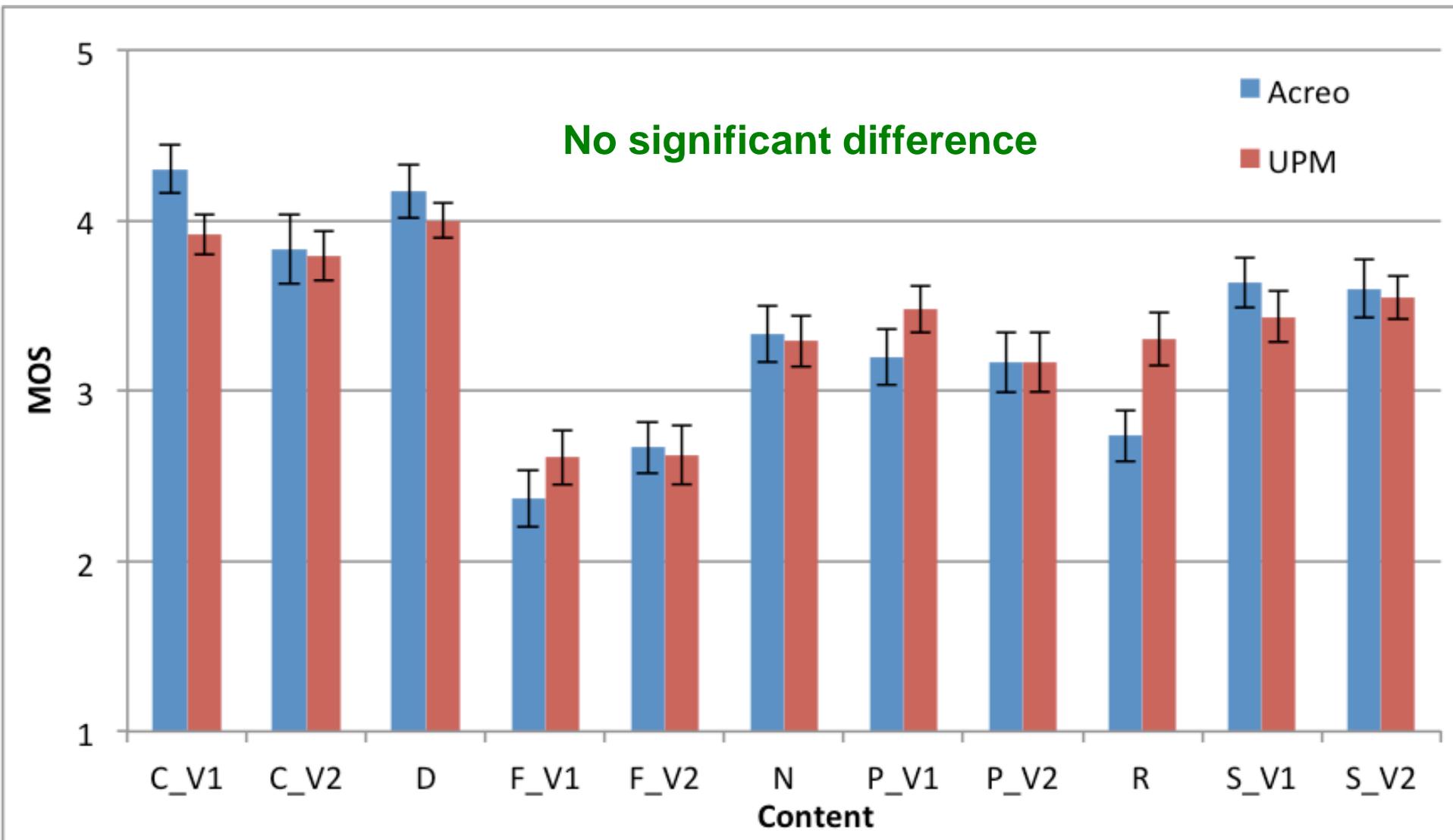
Earlier Subjective results



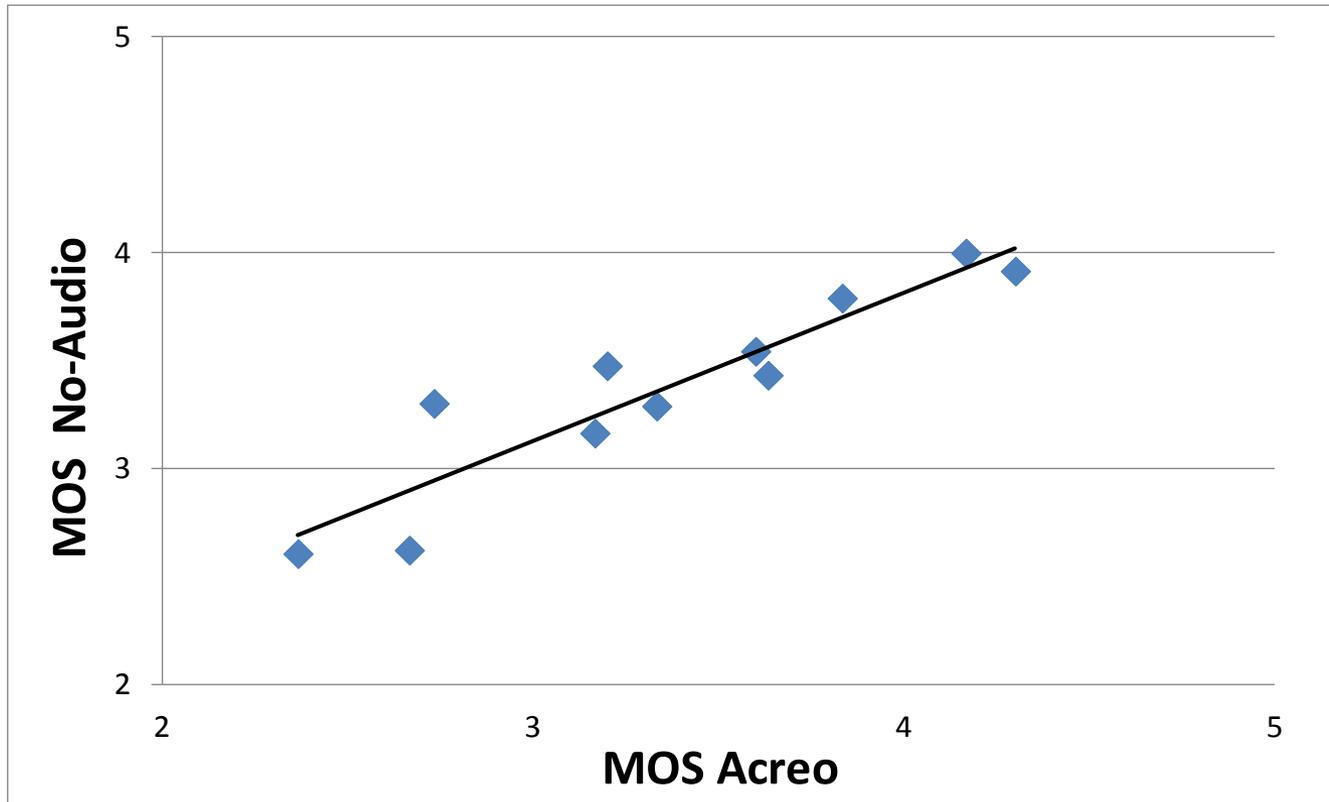
Impact of evaluation methodology: Acreo vs. UPM-Audio



Impact of evaluation methodology: Acreo vs. UPM-NoAudio

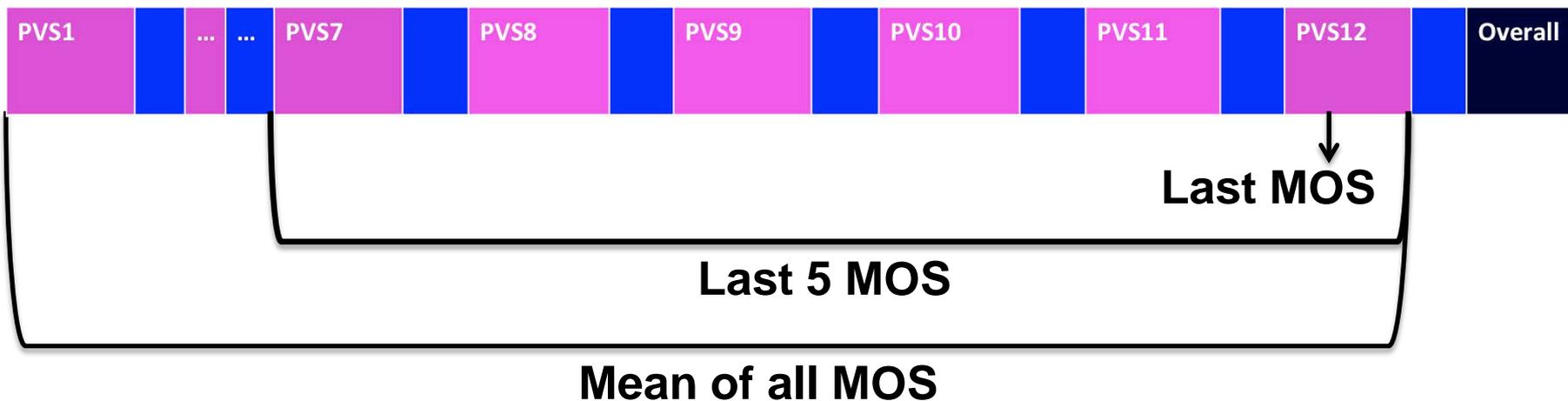


Estimating overall QoE from individual MOS



Estimating overall QoE from individual MOS?

Pearson Correlation	Overall	Last 5 MOS	Last MOS	Mean of All MOS
Acreo-Audio	0.79	0.66	0.54	0.81
Acreo-NoAudio	0.93	0.70	0.71	0.90



Discussion

On subjective evaluation methodology

- Effect of rating individual events on given overall scores (UPM experiments)
- Influence of language and culture
 - Big difference in Spanish language videos (Acreo vs. UPM)
 - High similarity of Acreo and No-Audio
 - More focus on visual quality
- Mean MOS of individual event: a good correlation with overall quality
 - Higher correlation in Audio-NoAudio

Discussion

On objective characterization of adaptive streaming videos

- A selection of tools can be used
- Depends on the content the performance of the tools can be different
 - Football content due to the characteristics of the original video

Conclusions

- Mean of the MOS of individual events: a good predictor of the overall MOS for the full length 6 minute video
- Low performance of objective characterization (tools) in our dataset
- Clustering videos based on Spatial and Temporal characteristics: increasing the performance
- Best predictor: Full-Reference blockiness

Future works

- Repeating experiment with different content (English language)
- Incorporating other characteristics on perceived quality
 - e.g. scene changes

Thank you

- UPM's work was partially funded by Ministerio de Economía y Competitividad of the Spanish Government.
- Acreo's work was funded by Vinnova (Sweden's innovation agency) and EIT ICT Labs, which is hereby gratefully acknowledged.



Adaptation scenarios

Status	Possible Client Behavior		Code
Increasing quality	Gradually (600-1-3-5)	2 sec chunk	IGR2
		10 sec chunk	IGR10
	Rapidly (600-5)	2 sec chunk	IRP2
		10 sec chunk	IRP10
Decreasing quality	Gradually (5-3-1-600)	2 sec chunk	DGR2
		10 sec chunk	DGR10
	Rapidly (5-600)	2 sec chunk	DRP2
		10 sec chunk	DRP10
Constant quality	No degradation- The whole segment at 5Mbps		N5
	No degradation- The whole segment at 3Mbps		N3
	No degradation- The whole segment at 1Mbps		N1
	No degradation- The whole segment at 600kbps		N600

SROCC Results between NR and FR measurements for each content

Pooling approach: Average of 10% of the lowest values

