Standardized Methods for Subject Removal in Subjective Quality Assessment

Irene Viola, Lucjan Janowski
Problem Statement

- People are not precise in measuring
- Some people are especially bad in staying focused
- Decreasing variability in the data helps with obtaining precise conclusions
- One way to decrease variability is to remove random answers
- What is random?
Standards

- **BT.500**

\[ \beta_{2,jkr} = \frac{m_4}{(m_2)^2} \]

if \(2 \leq \beta_{2,jkr} \leq 4\), then:
  \[ \text{if } u_{ijkr} \geq \bar{u}_{jkr} + 2S_{jkr} \quad \text{then } P_i = P_i + 1 \]
  \[ \text{if } u_{ijkr} \leq \bar{u}_{jkr} - 2S_{jkr} \quad \text{then } Q_i = Q_i + 1 \]
else:
  \[ \text{if } u_{ijkr} \geq \bar{u}_{jkr} + \sqrt{20}S_{jkr} \quad \text{then } P_i = P_i + 1 \]
  \[ \text{if } u_{ijkr} \leq \bar{u}_{jkr} - \sqrt{20}S_{jkr} \quad \text{then } Q_i = Q_i + 1 \]

\[ \text{If } \frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05 \text{ and } \frac{|P_i - Q_i|}{P_i + Q_i} < 0.3 \quad \text{then reject observer } i \]

\[ r_1(x, y) = LPCC(x, y) \]

where in Equation (A-1):

- \(x_i\): MOS of all subjects per PVS
- \(y_i\): individual score of one subject for the corresponding PVS
- \(n\): total number of PVSs
- \(I\): PVS sequence number

- **P.913**

- **MLE - does not remove subjects**
Simulation setup

- Simulation parameters: 16 SRC, 2 distortion algorithms, 5 distortion points: **160 PVS**
- 25 subjects, 100 repetitions of the same test
- 2 scenarios:
  - Typical test: Bias: $\mu = 0; \sigma = 0.34$; PVS: $\sigma = 0.75; \alpha = 10$; [1]
  - Super-precise: Bias: $\mu = 0; \sigma = 0.01$; PVS: $\sigma = 0.3; \alpha = 100$;
- Subject removal parameters:
  - BT.500: $r1 = 0.05; r2 = 0.3$;
  - P.913: $r1 = 0.75$.
- Scramble test
  - For each repetition, 1 subject is randomly selected and scrambled 1000 times
  - Output: how many times scrambling is detected, how it affects other outliers

Results

Scenario 1: typical test
Results

Scenario 1: typical test

![Graphs showing data for BT.500 and P.913]
Results - Scrambling Test

Scenario 1: typical test

**BT.500**

**P.913**
Results

Scenario 2: super-precise test
Results

Scenario 2: super-precise test

BT.500

P.913
Results

Scenario 2: super-precise test
Conclusion

- We do not recommend the method described in Annex 1 to Part 1 “Analysis and presentation of results” of BT.500 for ACR tests
- P.913 method works fine. The threshold need further investigation
- Future work: different criteria for outliers beside scrambling