Subjective and objective quality assessment

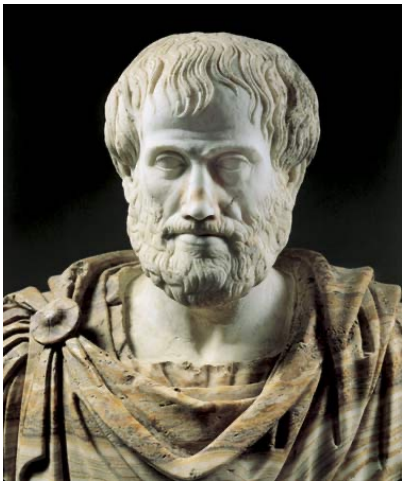
Prof. Dr. Touradj Ebrahimi
Touradj.Ebrahimi@epfl.ch

Quality – a simple yet difficult concept

- Like many human sensations quality is easy to understand but difficult to define
- According to Wikipedia:
  - A quality (from Latin - *qualitas*) is an attribute or a property.
  - Some philosophers assert that a quality cannot be defined.
  - In contemporary philosophy, the idea of qualities and especially how to distinguish certain kinds of qualities from one another remains controversial.

A fundamental yet largely under-investigated concept

- Aristotle classified every object of human apprehension into 10 *Categories*
  - Substance, Quantity, Quality, Relation, Place, Time, Position, State, Action, Affection

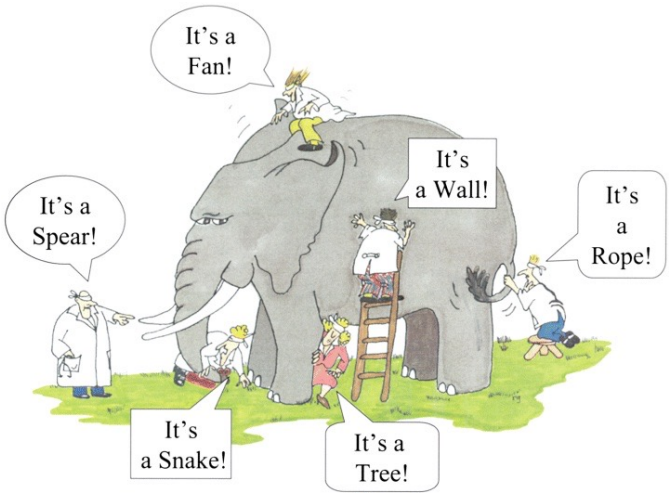


Aristotle 384 BC – 322 BC

Definition according to ISO 9000

- ISO 9000: a family of standards for quality management systems
- Quality of something can be determined by comparing a set of inherent characteristics with a set of requirements
  - High quality: if characteristics meet requirements
  - Low quality: if characteristics do not meet all requirements
- Quality is a relative concept
  - Degree of quality

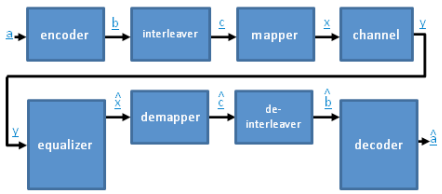
Quality – is in fact an elephant

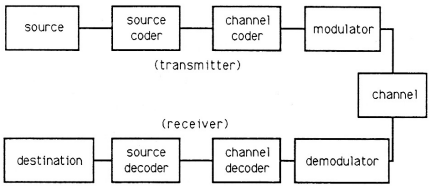


The blind men and the elephant: Poem by John Godfrey Saxe

Quality in telecommunications

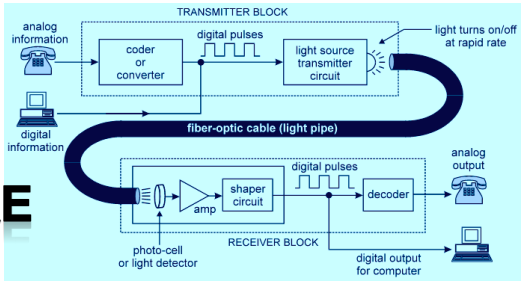
Quality in telecommunications





Objective: input = output

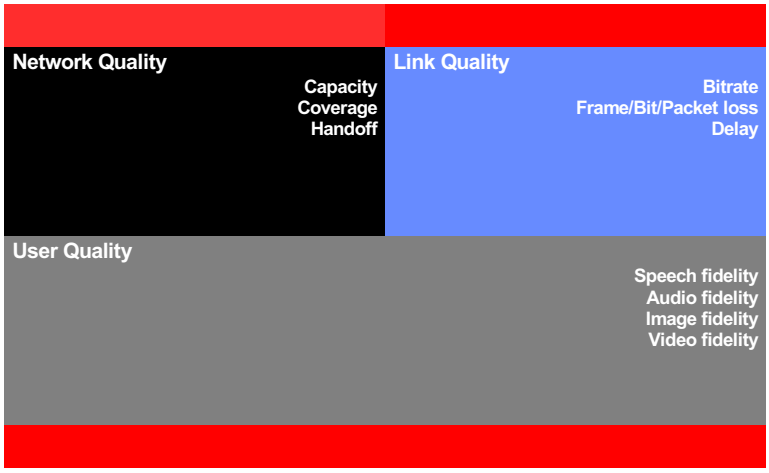
FIDELITY MEASURE



Quality of Service in computer networks and communications

- Quality of Service (QoS) refers to a collection of networking technologies and measurement tools that allow for the network to guarantee delivering predictable results

Quality in QoS framework



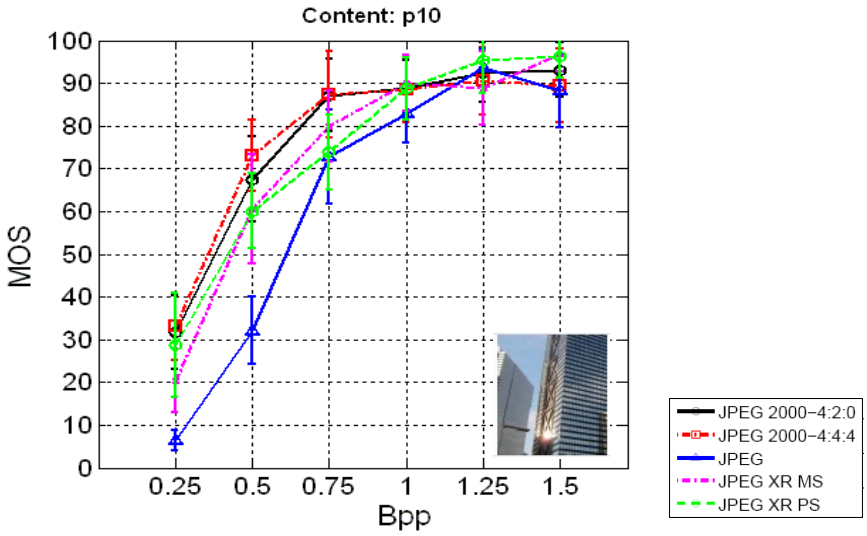
Quality of Service in computer networks and communications

- Quality of Service (QoS)
  - Resource reservation control mechanisms
  - Ability to provide different priority to different applications, users, or data flows
  - Guarantee a certain level of performance (quality) to a data flow
- (Service) Provide-centric concept
- Tightly related to the concept of Mean Opinion Score (MOS)

What is Mean Opinion Score (MOS)?

- Widely used in many fields:
  - Politics/Elections
  - Marketing/Advertisement
  - Food industry
  - Multimedia
  - …
- The likely level of satisfaction of a service or product as appreciated by an average user
- Should be performed such that it generates reliable and reproducible results
  - Subjective evaluation methodology
  - More complex and difficult that it *a priori* seems

What is behind a MOS?

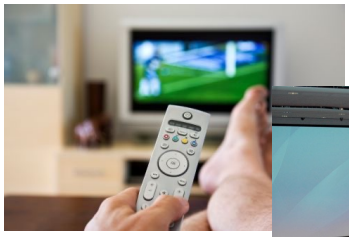


Subjective evaluation

- A subjective tests aiming at producing MOS is a delicate mixture of ingredients and choices:
  - **Test/lab environment**
  - **Test material**
  - **Test methodology**
  - **Analysis of the data**

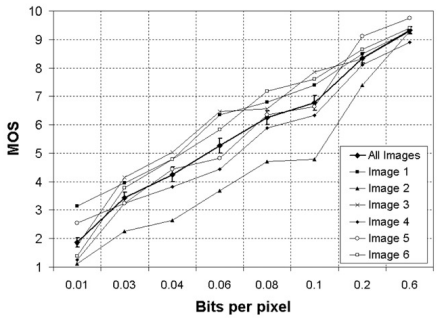
Test/lab environment

- Type of Monitors/Speakers and other test equipments
- Lighting /Acoustic conditions
- Laboratory architecture, background, …
- Viewing distance /Hearing position
- …



Test material

- Meaningful content for the envisaged scenario/application
  - Typical content
  - Worst case content
  - …



Test methodology

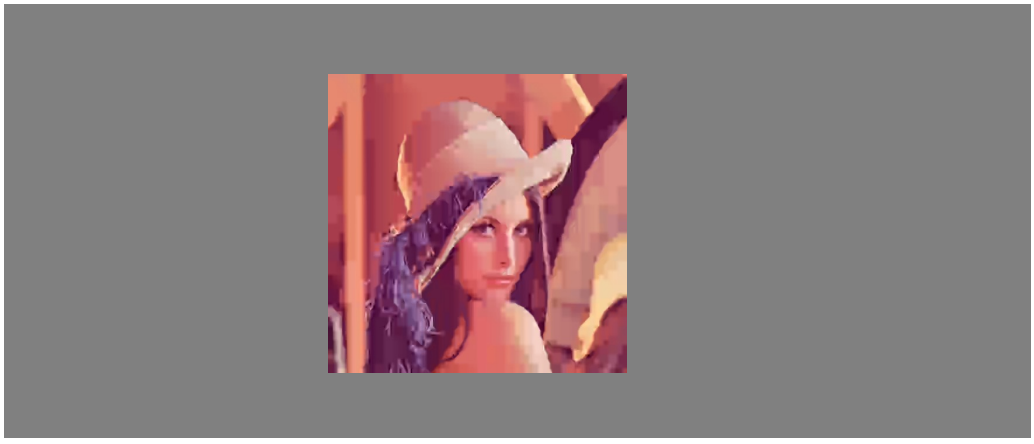
- Subjects
  - Naïve or Expert?
- Instructions
  - Which questions to ask subjects and how
  - Training
- Presentation
  - Single or double stimulus
  - Sequential or simultaneous
- Grading scale
  - Numerical
  - Categorical
- …

ITU Recommendations for test methodologies

- Test conditions and methodologies are specified in:
  - *Recommendation ITU-R BT. 500-11 "Methodology for the subjective assessment of the quality of television pictures" (1974-2002).*
  - *Recommendation ITU-T P. 910 "Subjective video quality assessment methods for multimedia applications" (1999).*
  - *Recommendation ITU-R BT. 1788 "Methodology for the subjective assessment of video quality in multimedia applications" (2007).*
- Based on television scenario!

Test methodology (I)

- Single Stimulus (SS)



Test methodology (I)

- Single Stimulus (SS)

- *Categorical numerical grading scale:*

"Rate from 1 to 10"

- *Categorical adjectival grading scale:*

5 Excellent
4 Good
3 Fair
2 Poor
1 Bad

5 Imperceptible
4 Perceptible but not annoying
3 Slightly annoying
2 Annoying
1 Very annoying

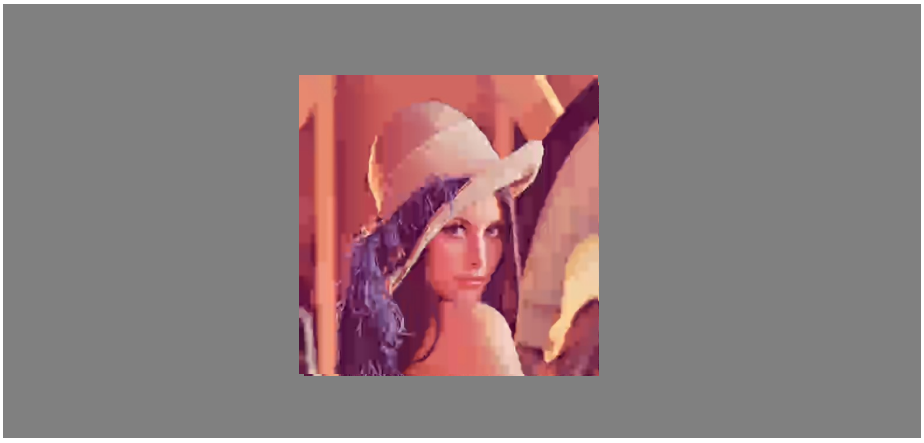
- *Non-categorical adjectival or numerical grading scale*

Test methodology (II)

- Double Stimulus (DS) Sequential

Test methodology (II)

- Double Stimulus (DS) Sequential



Test methodology (II)

- Double Stimulus (DS) Side by Side Hidden Reference

Test methodology (II)

- Double Stimulus (DS) Side by Side

Test methodology (II)

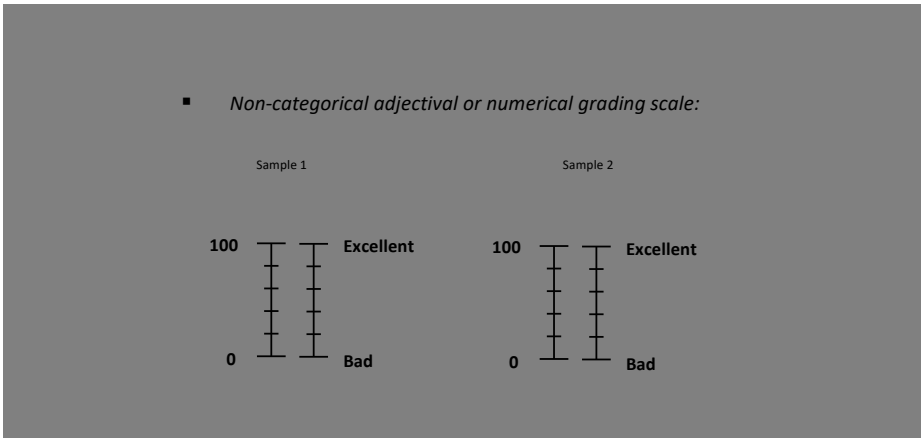
- Double Stimulus Impairment Scale (DSIS)

- *Categorical impairment grading scale:*

5 Imperceptible
4 Perceptible but not annoying
3 Slightly annoying
2 Annoying
1 Very annoying

Test methodology (III)

- Double Stimulus Continuous Quality Scale (DSCQS) Hidden Reference



Test methodology (IV)

- Stimulus Comparison (SC)

Test methodology (IV)

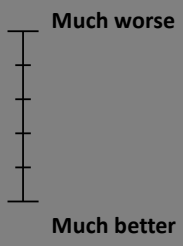
- Stimulus Comparison (SC)

▪ *Categorical adjectival comparison scale:*

"same or different"

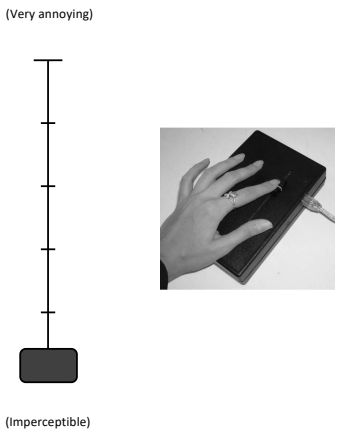
much worse
worse
slightly worse
the same
slightly better
better
much better

▪ *judgement:Non-categorical*



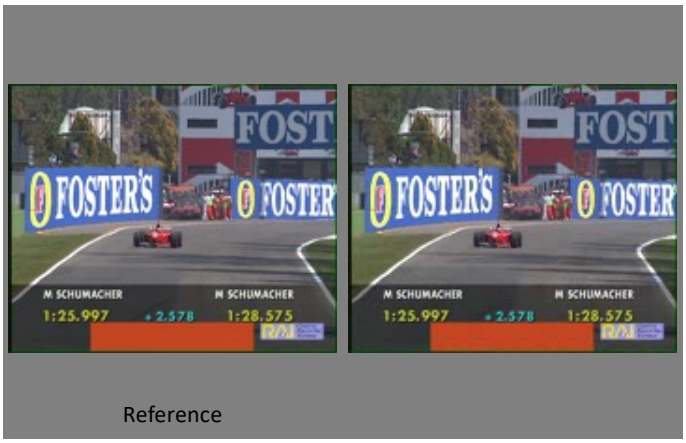
Test methodology (V)

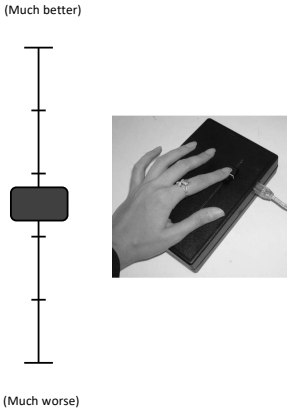
- Single Stimulus Continuous Quality Evaluation (SSCQE)



Test methodology (VI)

- Simultaneous Double Stimulus for Continuous Evaluation (SDSCE)





Analysis of the data

- Scores distributions across subjects is assumed to be close to normal distribution
- Outlier detection and removal
- Mean Opinion Scores (MOS) and 95% confidence intervals

$$MOS_j = \frac{\sum_{i=1}^{N} m_{ij}}{N}$$

$$CI_j = t(1-\alpha/2, N) \cdot \frac{\sigma_j}{\sqrt{N}}$$

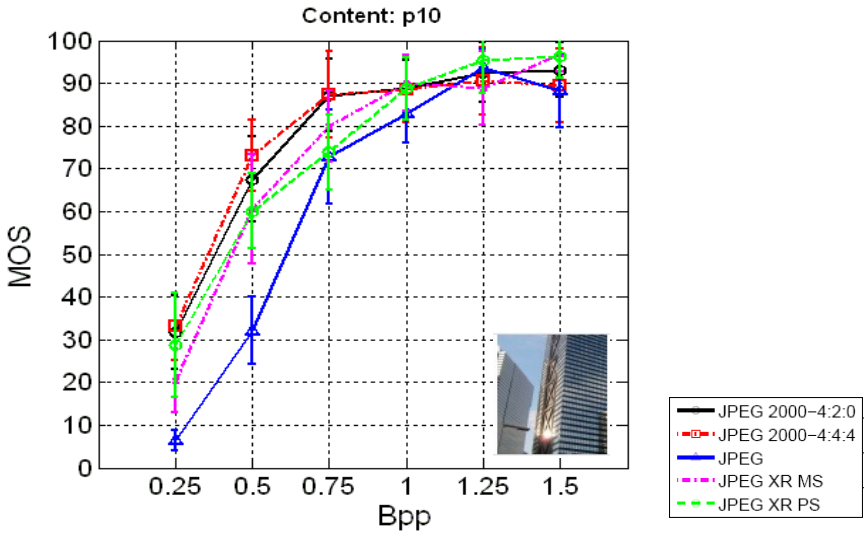
m_{ij} = score by subject *i* for test condition *j*.

N = number of subjects after outliers removal.

$t(1-\alpha/2,N)$ = t-value corresponding to a two-tailed t-Student distribution with N-1 degrees of freedom and a desired significance level α (α=0.05 in our case).

σ_j = standard deviation of the scores distribution across subjects for test condition *j*.

What is behind a MOS?



What is wrong with MOS?

- Ground truth
  - It is assumed to be the optimal solution (ground truth)
- Defies user preference
  - It is assumed to be independent of specific subjects
- Ignores key parameters
  - It is often assumed to be independent of content and context
- Misuse as a comparison approach
  - It is often used to compare performance between alternative tools for a same task/operation
- Impractical or expensive
- Statistical assumptions not always realistic

Relationship between estimated mean values

- Hypothesis test to find out whether the difference between two MOS values are statistically significant
  - Two-sided t-test:

$$\begin{cases} H_0 : MOS_A = MOS_B \\ H_a : MOS_A \neq MOS_B \end{cases}$$

- t statistic:

$$t_{obs} = \frac{MOS_A - MOS_B}{\sqrt{\frac{\sigma_A^2}{N} + \frac{\sigma_B^2}{N}}}$$

- Decision rule to reject H_0:

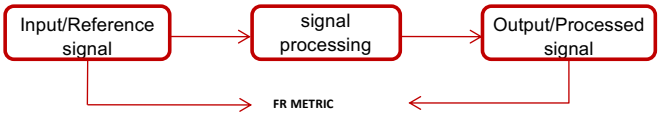
$$t_{obs} < t(\alpha/2, N) \quad OR \quad t_{obs} > t(1-\alpha/2, N)$$

Objective quality metrics

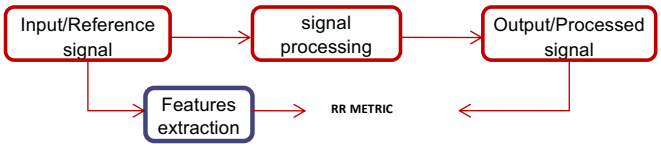
- Subjective tests are time consuming, expensive, and difficult to design
- Objective algorithms, i.e. metrics, estimating subjective MOS with high level of correlation are desired
  - Full reference metrics
  - No reference metrics
  - Reduced reference metrics

FR, RR and NR scenarios

- Full Reference approach:



- Reduced Reference approach:



- No Reference approach:

MOS predictors based on fidelity measures

- Full Reference scenario
- Metrics which look at the **fidelity** of the signal when compared to an explicit reference:

processed signal

=

perfect quality reference signal

+

error signal

MOS predictors based on fidelity metrics

- Examples
  - Mean Square Error (MSE)
  - Peak Signal to Noise Ratio (PSNR)
  - Maximum Pixel Deviation (L_{inf})
  - Weighted PSNR
  - Masked PSNR
  - Structural SIMilarity (SSIM)
  - Multiscale Structural Similarity (MSSIM)
  - Visual Information Fidelity (VIF)
  - etc…

Peak Signal to Noise Ratio

$$PSNR = 10\log_{10}\frac{(2^B-1)^2}{MSE}$$

where:

$$MSE = \frac{1}{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}[Im_a(x,y) - Im_b(x,y)]^2$$

M, N = image dimensions

Im_a , Im_b = pictures to compare

B= bit depth

- Widely used because of its simplicity and ease in formalizing optimization problems!
- For image and video data (Y component), a correlation of circa 80% reported when compared to subjective MOS evaluation

PSNR for color images/video (I)

- Several alternatives to compute PSNR for color images/video:

$$WPSNR = w_1 PSNR_1 + w_2 PSNR_2 + w_3 PSNR_3$$

$$WPSNR\_MSE = 10\log_{10} \frac{(2^B - 1)^2}{(w_1 MSE_1 + w_2 MSE_2 + w_3 MSE_3)}$$

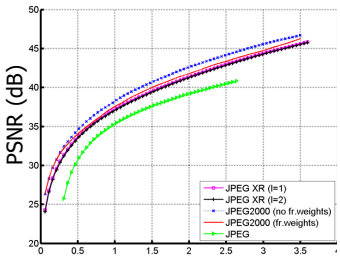
$$WPSNR\_PIX = 10\log_{10} \frac{(2^B - 1)^2}{\frac{1}{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}\left[\left(w_1 Im_{a1}(x,y) + w_2 Im_{a2}(x,y) + w_3 Im_{a3}(x,y)\right) - \left(w_1 Im_{b1}(x,y) + w_2 Im_{b2}(x,y) + w_3 Im_{b3}(x,y)\right)\right]^2}$$

PSNR for color images/video (II)

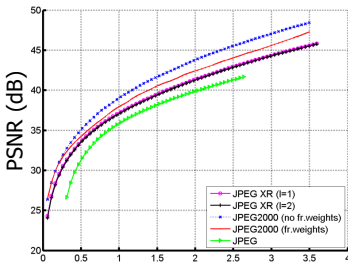
- Which color space to use?
  - RGB
  - Y'CbCr
  - other?
- Which weights to use?
  - w1=w2=w3=1/3
  - w1=0.8, w2=w3=0.1
  - other?

PSNR for color images/video (III)

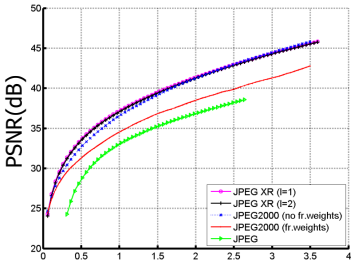
on R component:



on G component:

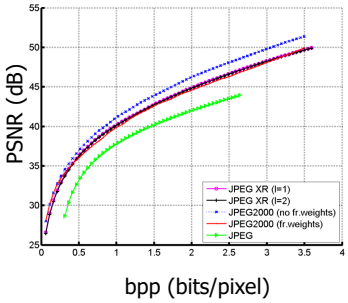


on B component:

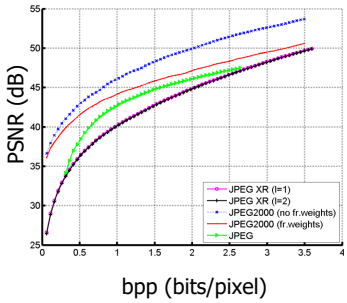


PSNR for color images/video (IV)

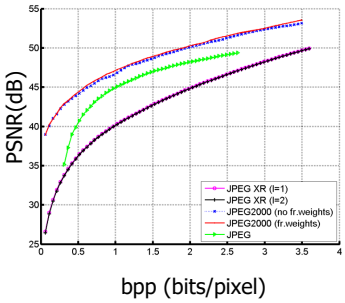
on Y' component:



on Cb component:

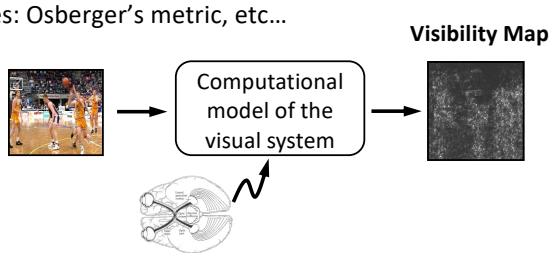


on Cr component:



MOS predictors based on fidelity metrics

- Human Visual System (HVS) based metrics
  - simulating properties of the early stages of the HVS
  - Examples: PSNR-HVS-M, etc…
  - simulating high level features of the HVS
  - Examples: Osberger's metric, etc…



☺ Better correlation with human perception.

☹ High complexity.

PSNR-HVS-M

$$PSNR-HVS-M = 10\log_{10}\frac{(2^B-1)^2}{MSE_H}$$

$$MSE_H = K\sum_{i=1}^{M-7}\sum_{j=1}^{N-7}\sum_{m=1}^{8}\sum_{n=1}^{8}\left[X(m,n)_{\Delta ij}\, T_c(m,n)\right]^2$$

where:

M, N = image dimensions

K = constant

$X(m,n)_{\Delta ij}$ = visible difference between DCT coefficients of the original and distorted based on a contrast masking

T_c = matrix of correcting factors based on standard visually optimized JPEG quantization tables

B= bit depth

MOS predictors based on fidelity measures

- Metrics based on the hypothesis that the **HVS is highly adapted for extraction of structural information from the content** of a still image or video.
  - degradation of still images or video = perceived structural information variation
  - Structural Similarity by *Wang et al.*

Structural Similarity (SSIM)

$$SSIM(\mathrm{Im}_1,\mathrm{Im}_2) = \left[l(\mathrm{Im}_1,\mathrm{Im}_2)\right]^{\alpha}\left[c(\mathrm{Im}_1,\mathrm{Im}_2)\right]^{\beta}\left[s(\mathrm{Im}_1,\mathrm{Im}_2)\right]^{\gamma} \quad (\alpha>0,\beta>0,\gamma>0)$$

- **Luminance comparison function:** $l(\mathrm{Im}_1,\mathrm{Im}_2)=\frac{2\mu_1\mu_2+C_1}{\mu_1^2+\mu_2^2+C_1}$ (C_1=constant)
- **Contrast comparison function:** $c(\mathrm{Im}_1,\mathrm{Im}_2)=\frac{2\sigma_1\sigma_2+C_2}{\sigma_1^2+\sigma_2^2+C_2}$ (C_2=constant)
- **Structure comparison function:** $s(\mathrm{Im}_1,\mathrm{Im}_2)=\frac{\sigma_{1,2}+C_3}{\sigma_1\sigma_2+C_3}$ (C_3=constant)

Mean SSIM

- MSSIM vs PSNR

PSNR = 24.9 dB

for all the images

Specific distortion metrics

- At times one is interested in specific types of distortions that occur in multimedia systems
- Examples
  - Bluriness
  - Blockiness
  - Ringing/Mosquito noise
  - Jerkiness
  - etc…

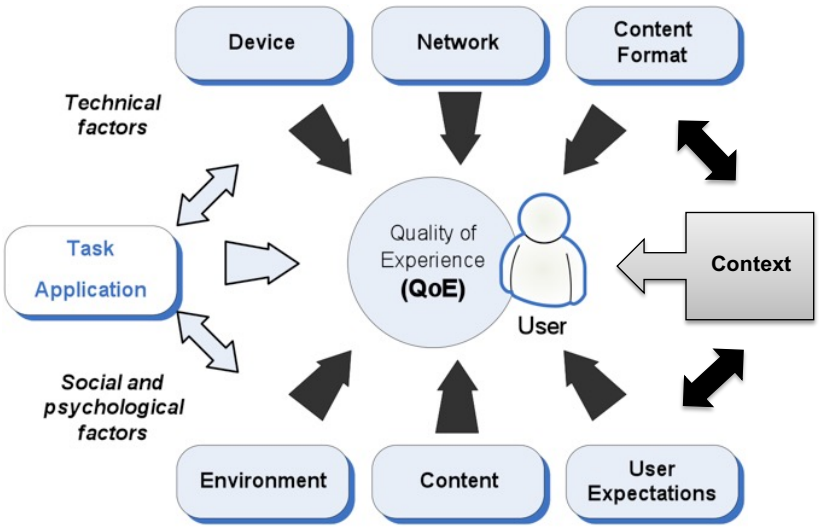
Moving to Quality of Experience

- **Quality of Service:** Value of the average user's service richness estimated by a service/product/content provider
- **Quality of Experience**: Value (estimated or actually measured) of a specific user's experience richness

Quality of Experience is the dual (and extended) view of Quality of Service !

QoS=provider-centric
QoE=user-centric

Factors impacting Quality of Experience

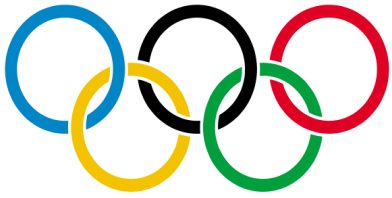


Experiences are Individual !

- Applications and Services may have to be designed to provide individual experiences ...
- This involves capabilities allowing the user to gain control, e.g. interaction, personalization, recommendation, etc.
- In fact, the user contributes to build is own experience ... if the system/service allows ...

QoE Modeling

- **System** factors
  - technical properties (as in QoS)
- **Human/User** factors
  - individual properties
  - sensorial properties
  - perceptual properties
  - emotional properties
- **Context** factors
  - environmental/physical properties
  - temporal properties
  - service properties
  - economic properties
  - social properties
- **Content** factors
- **…**



QoE modeling may consider more or less influence factors depending on the service/application, each with a different weight on the overall assessment.

QoE is multi-dimensional, multi-modal and multi-sensorial.

User centered influence factors are expected to be dominating.