

Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications

Anna Watson and M. Angela Sasse
Dept. of CS
University College London, London, UK

Proceedings of ACM Multimedia
November 1998



Outline

- Introduction
- Measuring Perceived Quality
- What is Multimedia Quality?
- UCL Approach to Measuring Quality
- Summary



Motivation

- As power and connectivity of computers has increased
 - increase in Multimedia networking research
- Recognized that Multimedia has “special” constraints
 - Ex: delay, loss, jitter
 - Enter Network *Quality of Service (QoS)*
- QoS provides network guarantees on delay, loss, jitter, bwidth ...



Quality of Service

- Some say, QoS will be resolved through:
 - RSVP
 - Bandwidth increase
 - Consumers will want lower quality for low cost
- Need to know how QoS impacts the user to know what QoS to aim for!
 - Optimal conditions
 - Minimum QoS acceptable
 - + Ex: one-way delay less than 250ms
 - + Ex: need 3 frames per second
 - Maximum QoS beyond which does not make better
 - + Ex: one-way delay less than 100 ms
 - + Ex: 30 frames/second is max



User-Centric Performance

- Network QoS gives you objective measures to shoot for
- But the end-user is the one who finally matters
- Need a *subjective* assessment of quality
 - Called *Perceptual Quality (PQ)*
- Then, can tie an objective measure to PQ




Outline

- Introduction
- Measuring Perceived Quality
- What is Multimedia Quality?
- UCL Approach to Measuring Quality
- Summary



Measuring Perceived Quality

- Typically done by using standards
 - International Telecommunications Union (ITU)
- ITU for Traditional media
 - Speech quality (phone, etc)
 - Images (television, etc)
- ITU not suitable for computer based multimedia network communication
- Next up:
 - ITU recommended measures
 - Criticism




ITU on Measuring Speech Quality

<i>Quality of the speech/ connection</i>	<i>Score</i>
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

<i>Effort required to understand the meaning of sentences</i>	<i>Score</i>
Complete relaxation possible; no effort required	5
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1

- Based on 10 second test
- Quality and Effort
- Listening




ITU on Measuring Speech Quality

Did you or your partner have any difficulty in talking or hearing over the connection?

Yes	1
No	0


(c) Conversation difficulty scale

- Conversation



Criticism of ITU Speech Measure

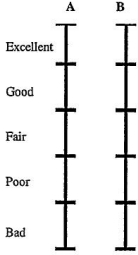
- Vocabulary-based poor
 - “Bad”, “Poor” and “Fair” difficult to define
 - Clusters at the low end
- Time-period is too short
 - Network conditions often unpredictable
 - Loss rates may be transient
- Effort scale is too simplistic
 - Again, network conditions change
 - Some effort for some of the talk but not all




ITU on Measuring Image Quality

<i>Image quality</i>	<i>Score</i>	
Excellent	5	Excellent
Good	4	Good
Fair	3	Fair
Poor	2	Poor
Bad	1	Bad

<i>Image Impairment</i>	<i>Score</i>	
Imperceptible	5	
Perceptible, but not annoying	4	
Slightly annoying	3	
Annoying	2	
Very annoying	1	




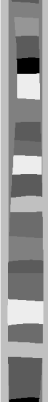
- Stimulus or Impairment scales



Criticism on ITU for Images



- Vocabulary not suitable
 - Same problems for “fair”, “poor” and “bad”
 - “Imperceptible” and “Perceptible” fine for television but not so good for lower-quality multimedia
- Time period too short
 - Same 10 second test not enough
- Artificiality of video test
 - Testing video without audio not good for multimedia
 - Unlikely would be watching video with no audio







International Interval Scale

- For an international measure, labels need to be translated equally
 - To compare research across countries
- Subjects given line:
 - “Worst Imaginable” at the bottom
 - “Best Possible” at the top
- Place the 5 labels on this line
 - Do we get 5 equal intervals in all languages?



ITU Scale in Different Languages

- In English “Poor” and “Bad” seen as the same
 - Points spaced to a 4 point, 3-interval scale
 - Not 5 points, as indicated
 - Users avoid the end (1 and 5), so 2 points
- In Italian,
 - no mid-point
 - “Ok” is equivalent to “Good”
- In Swedish,
 - “Poor” and “Bad” the same
 - “Fair” above mid-point
- In Dutch, also not equal
- In Japanese, all intervals equal!



Outline

- Introduction
- Measuring Perceived Quality
- What is Multimedia Quality?
- UCL Approach to Measuring Quality
- Summary



What is Multimedia Quality?

- Not one-dimensional
 - 1995 telecom identified at least 4 dimensions that affect quality
- Speech quality depends upon
 - Intelligibility, loudness, naturalness, listening effort, pleasantness of tone...
- Video quality depends upon
 - Color, brightness, background stability, speed in image reassembling...


What is Multimedia Quality?

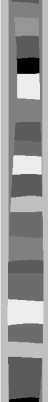
- A one-dimensional quality view
 - does not let us figure out where *bottleneck* is
 - leads to one-dimensional approach to fixing
- “Add more bandwidth to increase quality”
 - Probably many other ways to increase quality without increasing bandwidth

UCL Approach to Measuring Quality



- Identify suitable vocabulary to describe quality
- Identify key quality dimensions
- Employ knowledge in developing measure







Build Suitable Quality Vocabulary

- Don't supply words
 - Often too technical, may be lacking
 - Ex: "Does the picture have *jitter*?"
- Let users describe media in own terms
 - Ex: "choppy" or "buzzy" or "static"
- Build database of terms



Identify Dimensions

- Based on frequency of words associated with media quality
- For example, "choppiness" associated with:
 - Cut up
 - Irregular
 - Broken



Investigating New Scales

- Unlabeled scale
 - Subjects did not avoid endpoints
 - Consistent ratings across users
- Longer testing periods
 - But comparison across tests difficult
 - Cumulative affect on quality difficult
 - + Instead, get last impression
 - Users get bored, so tests less effective
- Combination of quality
 - Users will "forgive" bad video if followed by good
 - Good followed by bad is often bad
 - + Recency effect

Quality Assessment Sliders (QUASS)

- Unlabeled slider
- Records quality taken every second
 - Captures 'instantaneous' effects
- (Picture here?)

Summary

- We don't yet know how to measure MM quality
- Unlabeled scales look promising
- Worry about length of tested sample
 - Not too short, not too long
- Worry about order of samples
 - Avoid recency effect

