Evaluating networked audio and video quality

Anna Watson

Department of Computer Science
University College London
A.Watson@cs.ucl.ac.uk
http://www.cs.ucl.ac.uk/Staff/A.Watson

Overview

- Characteristics of real-time interactive systems
- Evaluating end quality
  - searching for methods
    - controlled experiments
    - field trials
- Remaining issues
  - new vocabulary required
  - dynamic quality measurement

UCL

Evaluating interactive audiovisual systems

- Numerous variables
  - user(s)
  - environment(s)
  - task(s)
- Complex interactions

  e.g. perceived video quality can be raised by improving audio quality

Evaluating interactive audiovisual systems (2)

- Evaluating performance
  - which measures should be used?
  - Quality of Service: whose?

the *end user* will always determine the success of the system...

- Standard evaluation methods may not be applicable/available yet

UCL

Multimedia conferencing

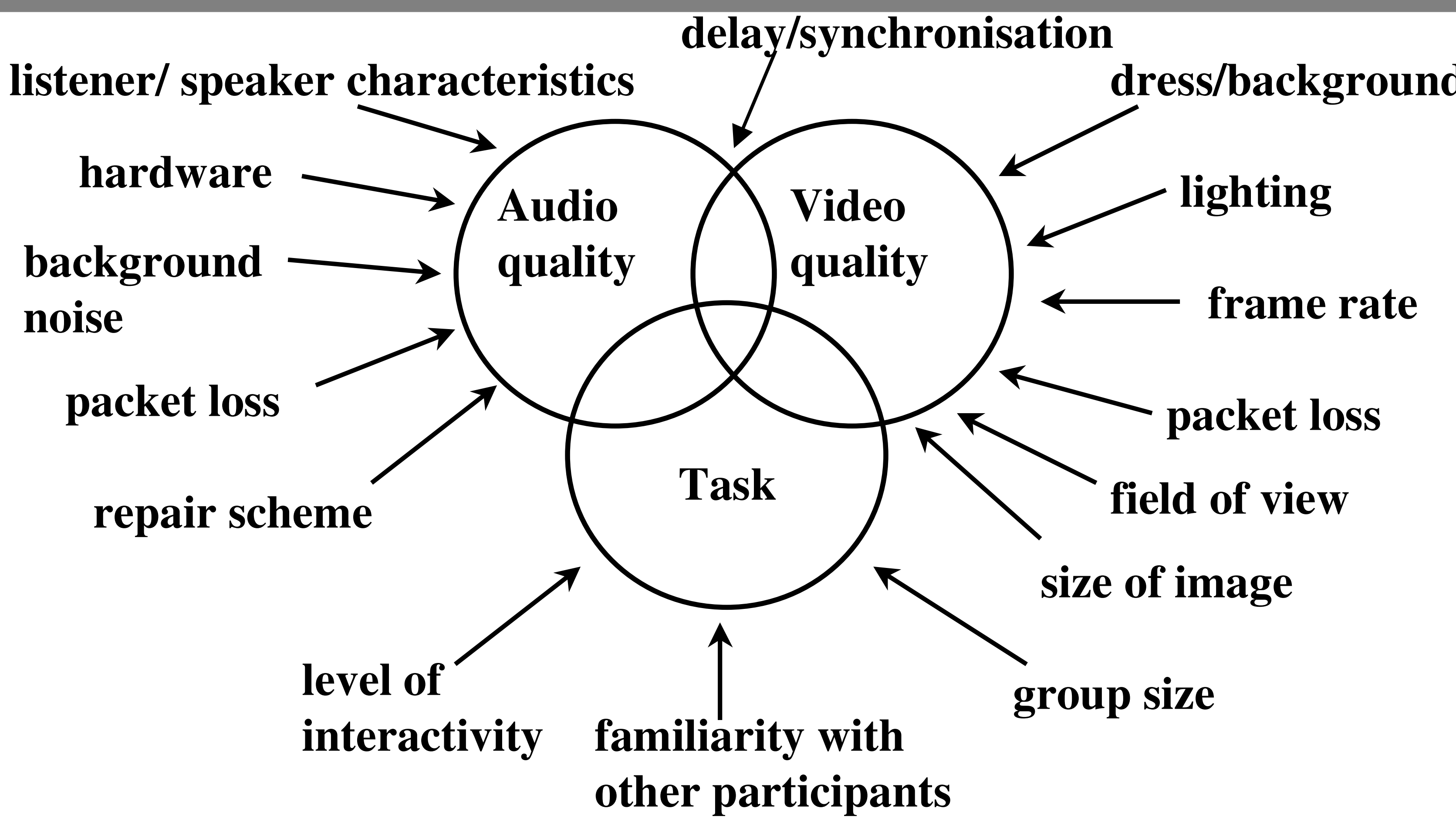
- Audio, video, shared workspace

- Multicast overlay of the internet
  - best effort, different routes

⇒ packet loss and delays are inevitable

- Different conferencing tasks will have different requirements

What affects usability in MMC?



Where to start?

- Absence of guidelines - new technologies...
- Twofold approach
  - Controlled experiments
  - Field trials
    - (informal) observations
    - questionnaires
    - interviews (group and individual)

UCL

Existing measurement methods

- Established telecommunications methods
  - intelligibility tests (for speech)
  - subjective rating scales (for speech and image quality)
- The problems
  - level of material in intelligibility tests
  - vocabulary of labels on scales
  - unidimensionality of scales
  - novel and rapidly changing conditions over the Internet are not modelled well in a controlled environment

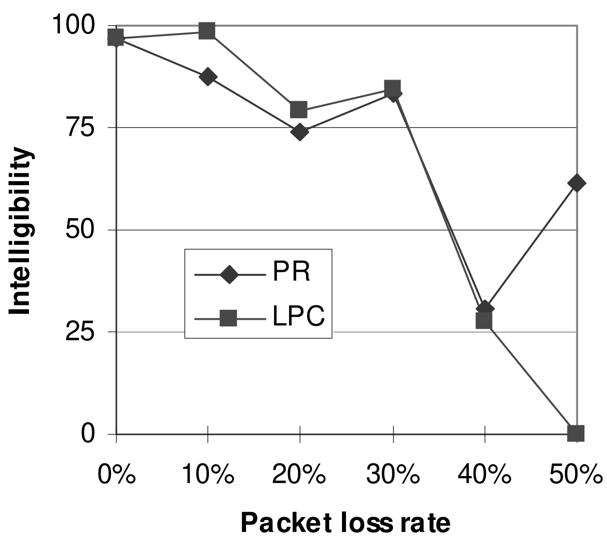
UCL

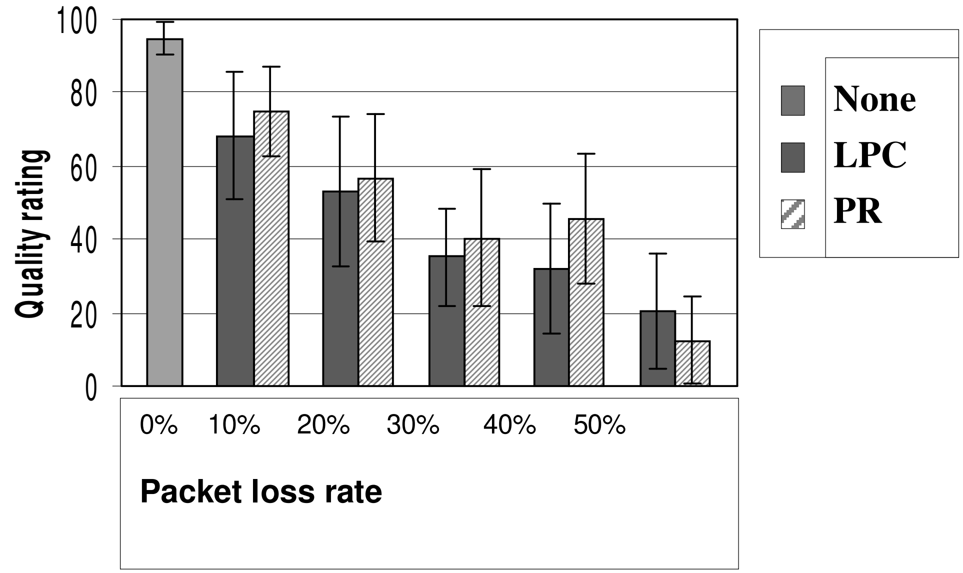
Controlled experiments

- Need to control and investigate key factors
  - Audio: packet loss and repair scheme
  - Video: frame rate
- Measures of performance
  - speech intelligibility and perceived quality
  - perceived synchronisation
  - task performance

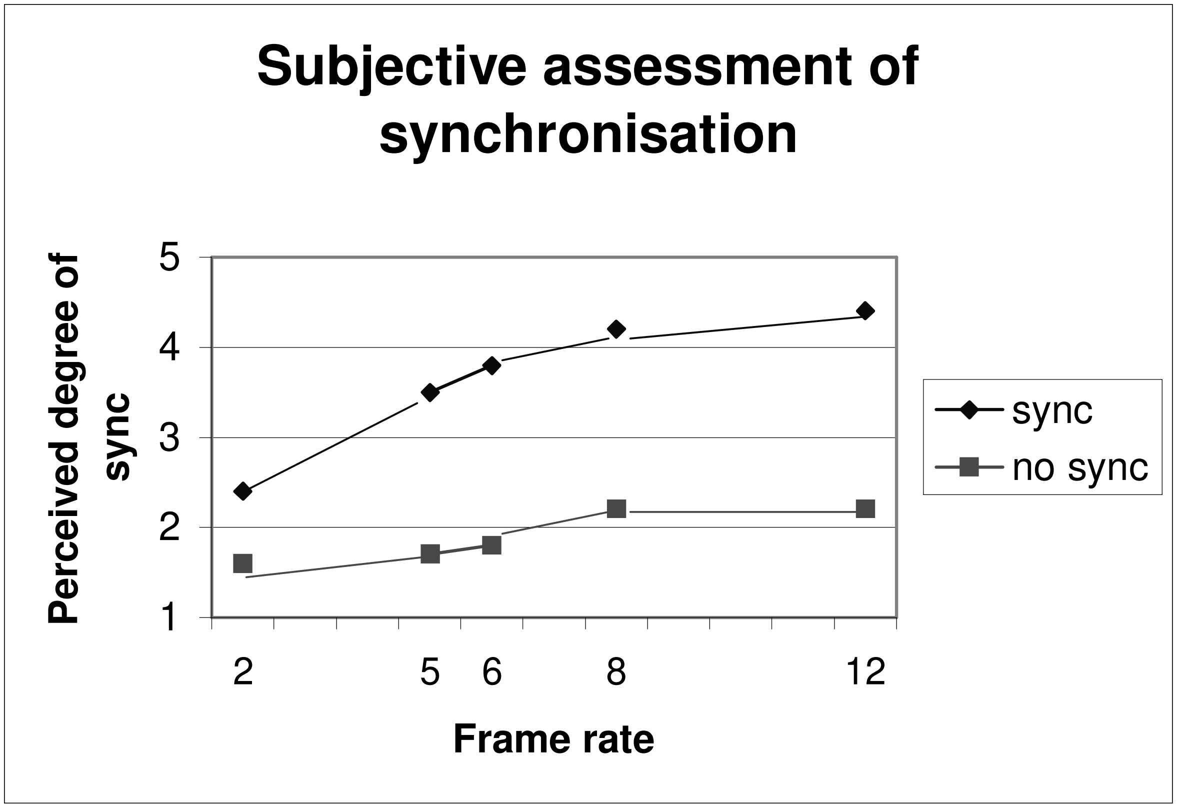
Some experimental results

- **Intelligibility is a fundamental part of speech quality, but the two are not the same…**





Experimental results (2)



Comments on experiments

- Speech intelligibility and speech quality are not the same thing
- Current rating scales treat speech quality as uni-dimensional
- Tasks can be artificial and can confound results
- Tendency to focus on one media only

→ We need to look at applications in use in the real world...

Field trials

- Remote language teaching (ReLaTe) between UCL and Exeter
  - first systematic field trial
  - demanding application
  - small group tutoring
- Networking course between UCL and Westminster
  - larger groups (5 students)
- PIPVIC :-)

UCL

Early field trial results

- Observations
  - audio is the critical component
    - push-to-talk should be avoided
  - people use video more than they think
    - psychologically important
  - shared workspace is a new domain
  - people adapt, both quickly and over time
  - control groups are hard to find, and subjective measurement is difficult..!

Field trial results (2)

- Interviews
  - group interviews/focus groups are valuable

- Questionnaires
  - rating scales - problem labels, cumulative rating
  - descriptive terms are limited
    - what do ‘choppy’ and ‘blocky’ mean?

UCL

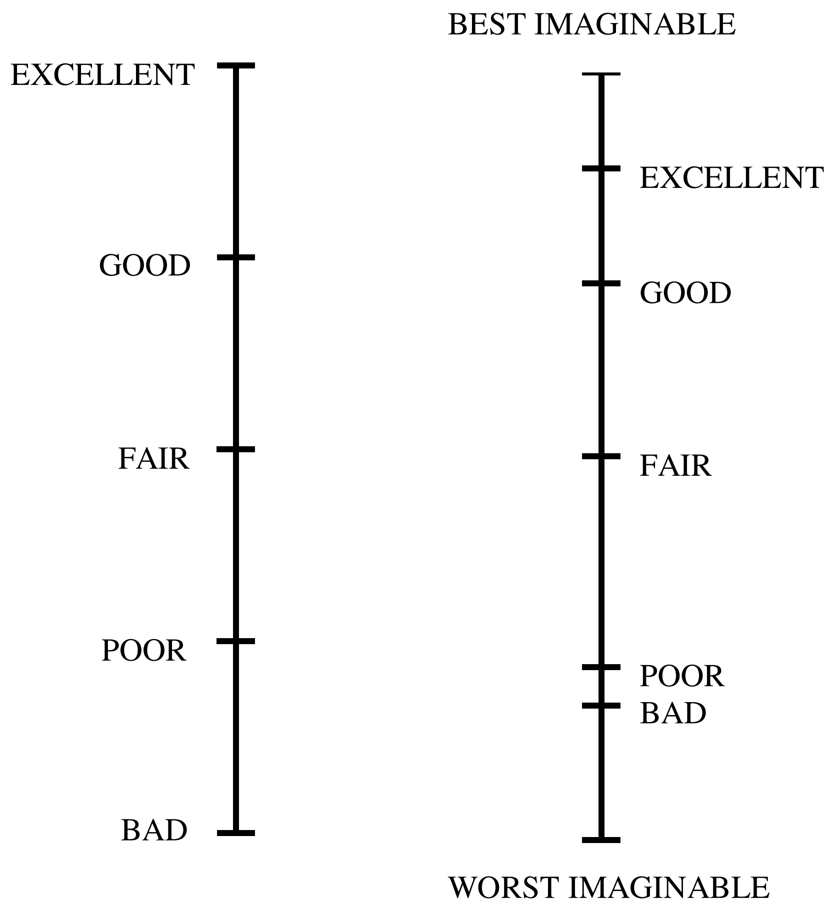
Rating scales revisited

Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

- **Hour long sessions, one word quality descriptions?**
  - **cumulative rating, primacy/recency effects**
- **Vocabulary is unsuitable**
- **Multidimensionality of quality is denied**
- **The scale is not interval...**

UCL

The quality rating scale



Field trials & controlled expts compared

Field trials

+

- **High in ecological validity**
- **Many applications**

-

- **Control groups hard to find**
- **Cumulative rating**

Experiments

+

- **Confounding variables limited**
- **Control groups**

-

- **Can't generalise**
- **What happens when it's *multi*media?**

Plus the methods for gathering subjective quality measurements are not ideal...

UCL

New approaches

- Common descriptive vocabulary is being collected
- Unlabelled scale avoids cultural and vocabulary difficulties, but the cumulative rating factor remains

→ *dynamic* slider being developed

  - rating of different dimensions, in experimental set-up to begin with
  - control to the user

- Ultimate aim:

Guidelines on AV quality requirements for different real-time multimedia tasks

UCL