

A Crowdsourcable

QoE Evaluation Framework for Multimedia Content

Kuan-Ta Chen	Academia Sinica
Chen-Chi Wu	National Taiwan University
Yu-Chun Chang	National Taiwan University
Chin-Laung Lei	National Taiwan University



What is QoE?

Quality of Experience =

Users' satisfaction

about a service

(e.g., multimedia content)

Quality of Experience



Poor
(underexposed)



Good
(exposure OK)

Challenges

- How to quantify the QoE of multimedia content **efficiently** and **reliably**?



Q=?



Q=?



Q=?

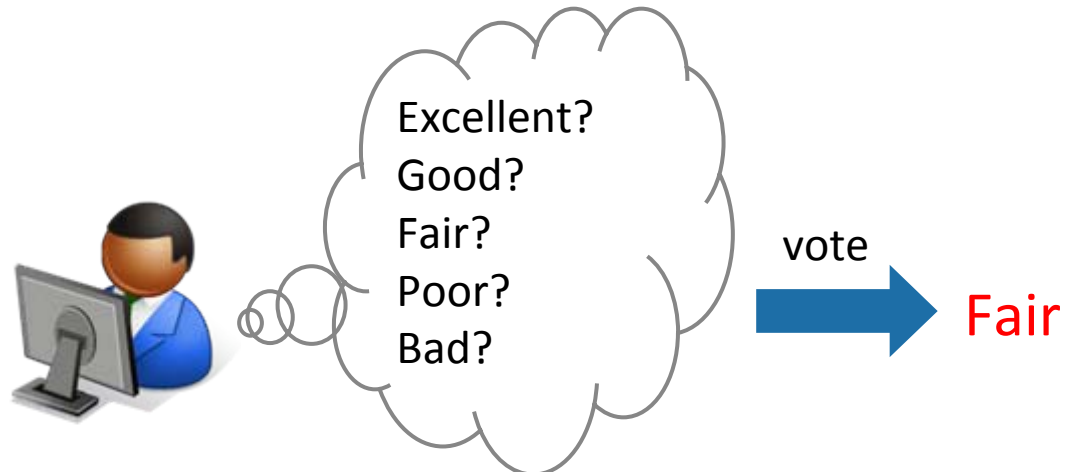


Q=?

Mean Opinion Score (MOS)

- Idea: Single Stimulus Method (SSM) + Absolute Categorical Rating (ACR)

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying



Drawbacks of MOS-based Evaluations

- ACR-based
 - Concepts of the scales cannot be concretely defined
 - Dissimilar interpretations of the scale among users
 - Only an ordinal scale, not an interval scale
 - Difficult to verify users' scores
- Subjective experiments in laboratory
 - Monetary cost (reward, transportation)
 - Labor cost (supervision)
 - Physical space/time/hardware constraints

Drawbacks of MOS-based Evaluations

■ ACR-based

- Concepts of the scales cannot be concretely defined
- Dissimilar interpretations of the scale among users
- Only an ordinal scale, not an interval scale
- Difficult to verify users' scores

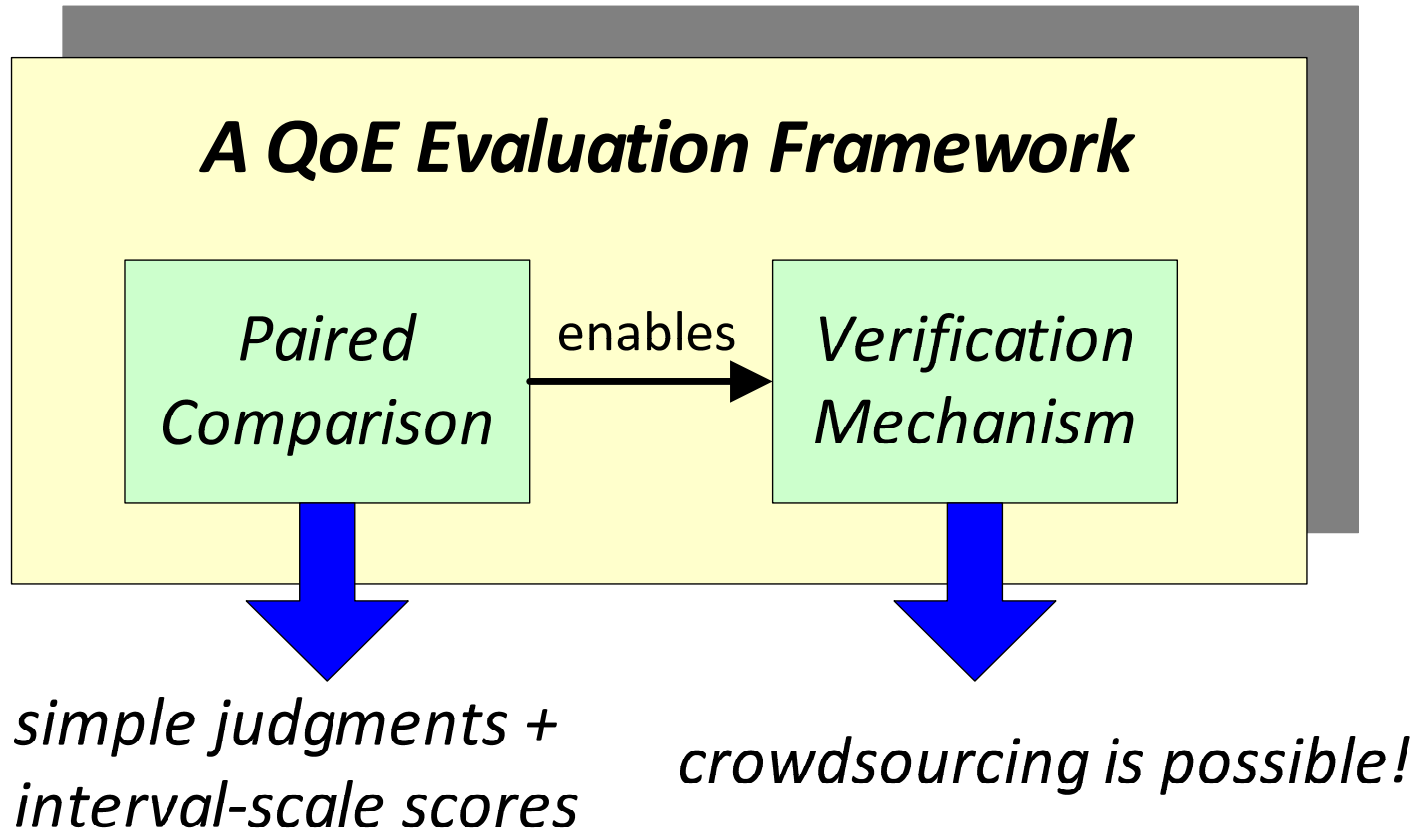
Paired
Comparison

■ Subjective experiments in laboratory

- Monetary cost (reward, transportation)
- Labor cost (supervision)
- Physical space/time/hardware constraints

Crowdsourcing

Contribution

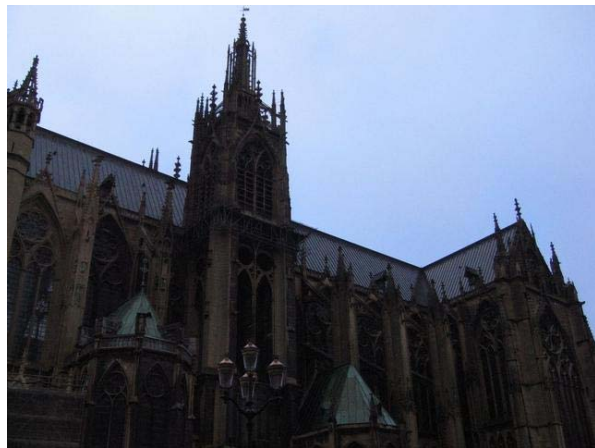


Talk Progress

- Overview
- Methodology
 - Paired Comparison
 - Crowdsourcing Support
 - Experiment Design
- Case Study & Evaluation
 - Acoustic QoE
 - Optical QoE
- Conclusion



Current Approach: MOS Rating



Excellent?
Good?
Fair?
Poor?
Bad?

Vote



?

Our Proposal: Paired Comparison



A



B



Vote








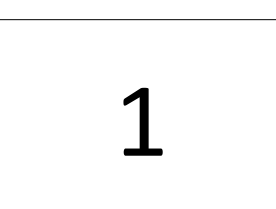

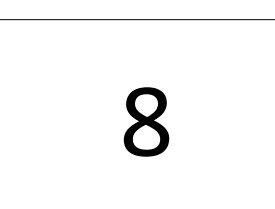

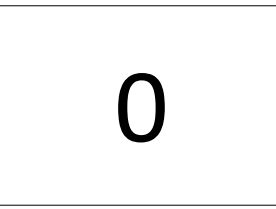
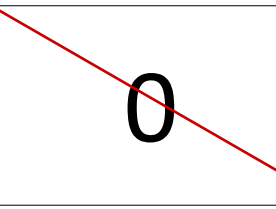
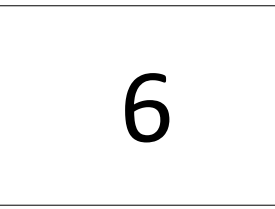


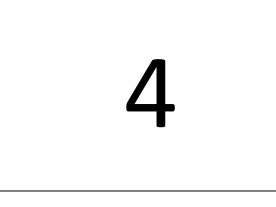
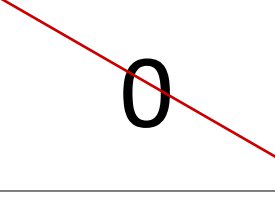
B

Properties of Paired Comparison

- *Generalizable* across different content types and applications
- *Simple* comparative judgment
 - dichotomous decision easier than 5-category rating
- *Interval-scale* QoE scores can be inferred
- The users' inputs can be *verified*

Choice Frequency Matrix

10 experiments, each containing $C(4,2)=6$ paired comparisons

	A	B	C	D
A	 0	 9	 10	 9
B	 1	 0	 7	 8
C	 0	 3	 0	 6
D	 1	 2	 4	 0

Inference of QoE Scores

- Bradley-Terry-Luce (BTL) model
 - input: choice frequency matrix
 - output: an interval-scale score for each content (based on maximum likelihood estimation)

n content: T_1, \dots, T_n

P_{ij} : the probability of choosing T_i over T_j

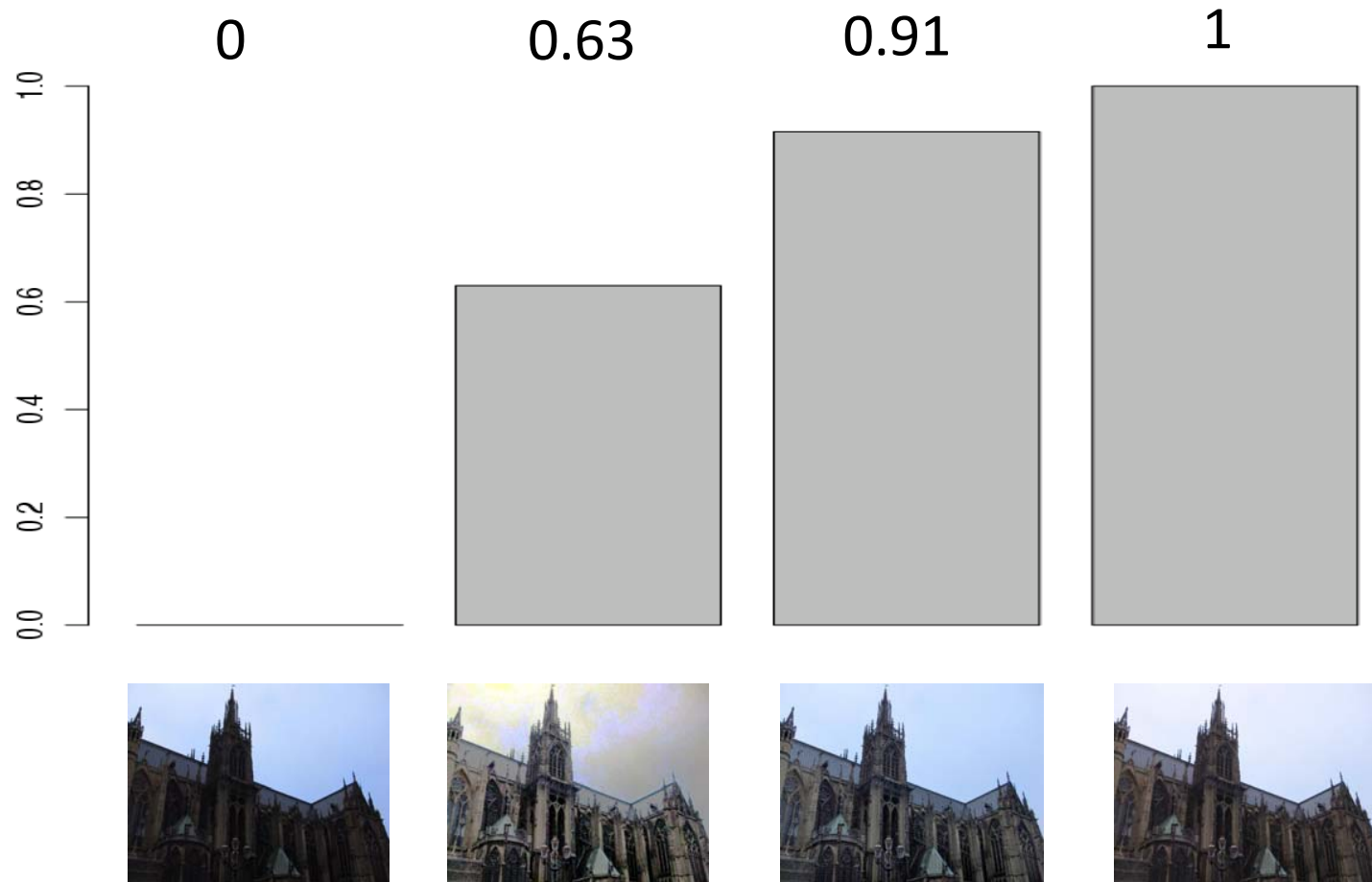
$u(T_i)$ is the estimated QoE score of the quality level T_i

$$P_{ij} = \frac{\pi(T_i)}{\pi(T_i) + \pi(T_j)} = \frac{e^{u(T_i) - u(T_j)}}{1 + e^{u(T_i) - u(T_j)}}$$

- Basic Idea

- $P_{12} = P_{23} \rightarrow u(T_1) - u(T_2) = u(T_2) - u(T_3)$

Inferred QoE Scores



Talk Progress

- Overview
- Methodology
 - Paired Comparison
 - ✂ ■ Crowdsourcing Support
 - Experiment Design
- Case Study & Evaluation
 - Acoustic QoE
 - Optical QoE
- Conclusion

Crowdsourcing

= Crowd + Outsourcing

“soliciting solutions via open calls to large-scale communities”



Image Understanding

- Reward: 0.04 USD

main theme?
key objects?
unique attributes?



Instructions: Provide information about the following image(s) by accurately answering the following questions.

Guidelines:

- Specific terms are preferred (Disneyland vs amusement park)
- Correct spelling is required (Hint: Use Firefox for spellchecker functionality)
- Don't repeat terms

Main Theme
What is the MAIN theme of the image?

Key Objects
What are the key objects in the image?

Unique Attributes
What are some unique attributes about this image? (actions, emotions, colors)



Linguistic Annotations

- Word similarity (Snow et al. 2008)

Word pair similarity

Below is a list of pairs of words. For each pair, please assign a numerical similarity score between 0 and 10 (0 = words are totally unrelated, 10 = words are VERY closely related). By definition, the similarity of the word to itself should be 10. You may assign fractional scores (for example, 7.5).

boy lad

Similarity (0-10)

coast shore

Similarity (0-10)

Submit HIT

USD 0.2 for labeling 30 word pairs

More Examples

- Document relevance evaluation
 - Alonso et al. (2008)
- Document rating collection
 - Kittur et al. (2008)
- Noun compound paraphrasing
 - Nakov (2008)
- Person name resolution
 - Su et al. (2007)

The Risk



Not every Internet user is trustworthy!

- Users may give erroneous feedback perfunctorily, carelessly, or dishonestly
- Dishonest users have more incentives to perform tasks

Need to have an **ONLINE algorithm** to detect problematic inputs!

Verification of Users' Inputs (1)

- **Transitivity property**

- If $A > B$ and $B > C \rightarrow A$ should be $> C$

- **Transitivity Satisfaction Rate (TSR)**

$$\frac{\text{\# of triples satisfy the transitivity rule}}{\text{\# of triples the transitivity rule may apply to}}$$

Verification of Users' Inputs (2)

- Detect inconsistent judgments from problematic users
 - $TSR = 1 \rightarrow$ perfect consistency
 - $TSR \geq 0.8 \rightarrow$ generally consistent
 - $TSR < 0.8 \rightarrow$ judgments are inconsistent



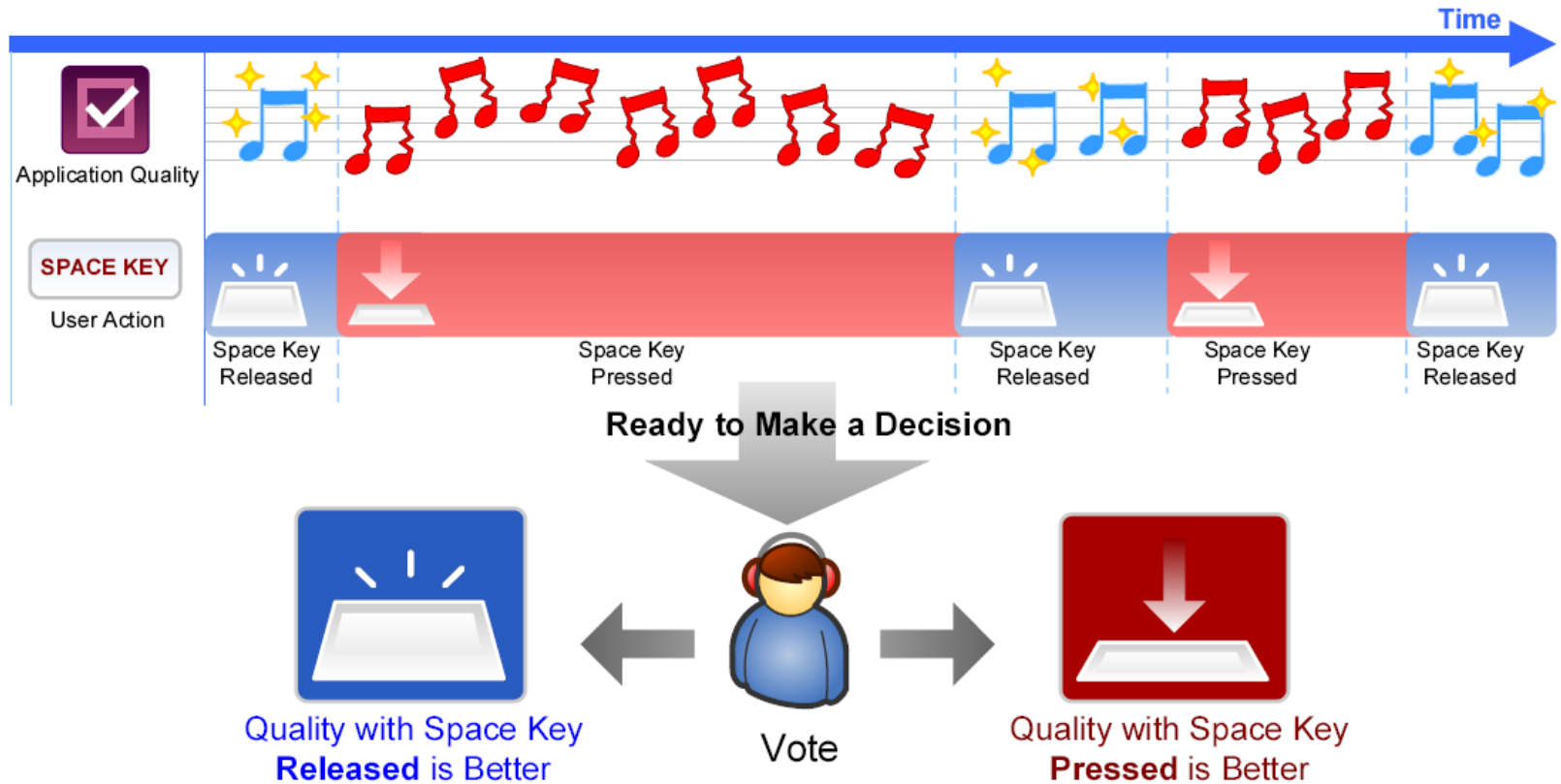
TSR-based reward / punishment
(e.g., only pay a reward if $TSR > 0.8$)

Experiment Design

- For n algorithms (e.g., speech encoding)
 1. a source content as the evaluation target
 2. apply the n algorithms to generate n content w/ different Q
 3. ask a user to perform $\binom{n}{2}$ paired comparisons
 4. compute TSR after an experiment

reward a user **ONLY** if his inputs are self-consistent
(i.e., TSR is higher than a certain threshold)

Concept Flow in Each Round



Audio QoE Evaluation

- Which one is better?



(SPACE key released)



(SPACE key pressed)

Video QoE evaluation

- Which one is better?



(SPACE key released)



(SPACE key pressed)

Talk Progress

- Overview
- Methodology
 - Paired Comparison
 - Crowdsourcing Support
 - Experiment Design



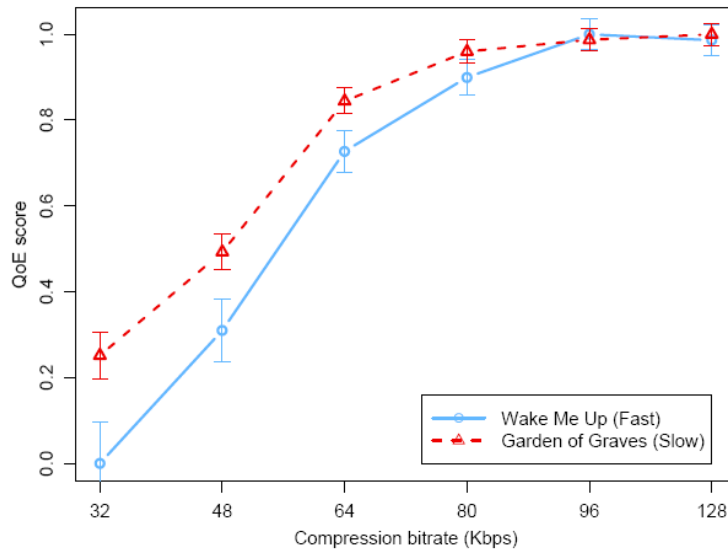
- **Case Study & Evaluation**
 - Acoustic QoE
 - Optical QoE
- Conclusion

Audio QoE Evaluation

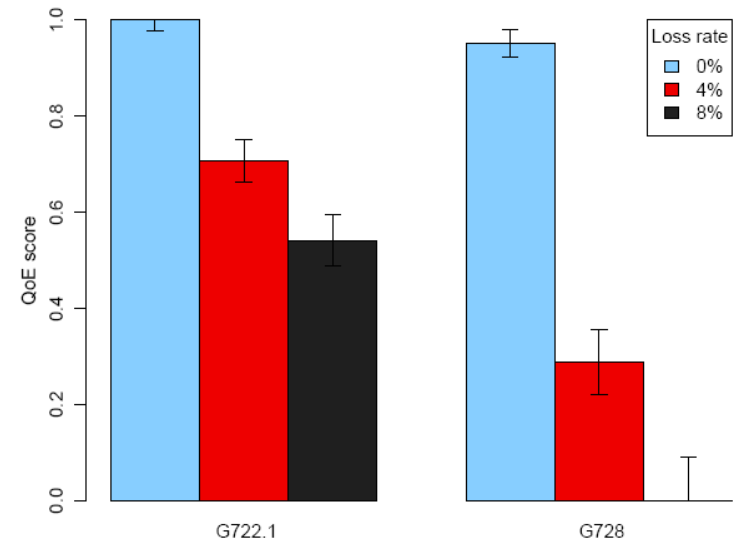
- MP3 compression level
 - Source clips: one fast-paced and one slow-paced song
 - MP3 CBR format with 6 bit rate levels: 32, 48, 64, 80, 96, and 128 Kbps
 - 127 participants and 3,660 paired comparisons
- Effect of packet loss rate on VoIP
 - Two speech codecs: G722.1 and G728
 - Packet loss rate: 0%, 4%, and 8%
 - 62 participants and 1,545 paired comparisons

Inferred QoE Scores

MP3 Compression Level



VoIP Packet Loss Rate

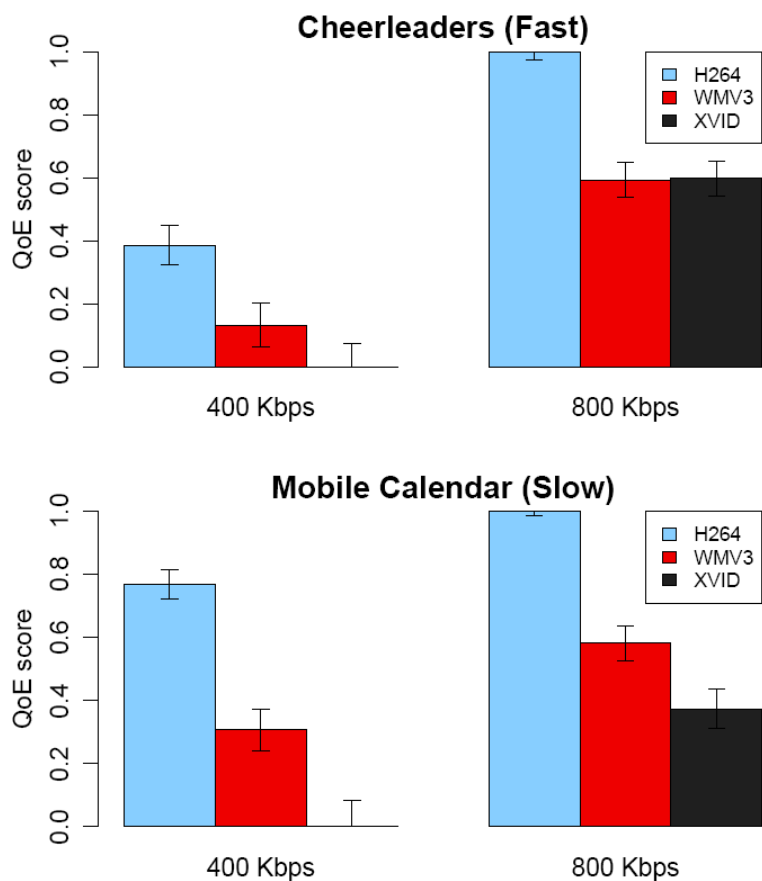


Video QoE Evaluation

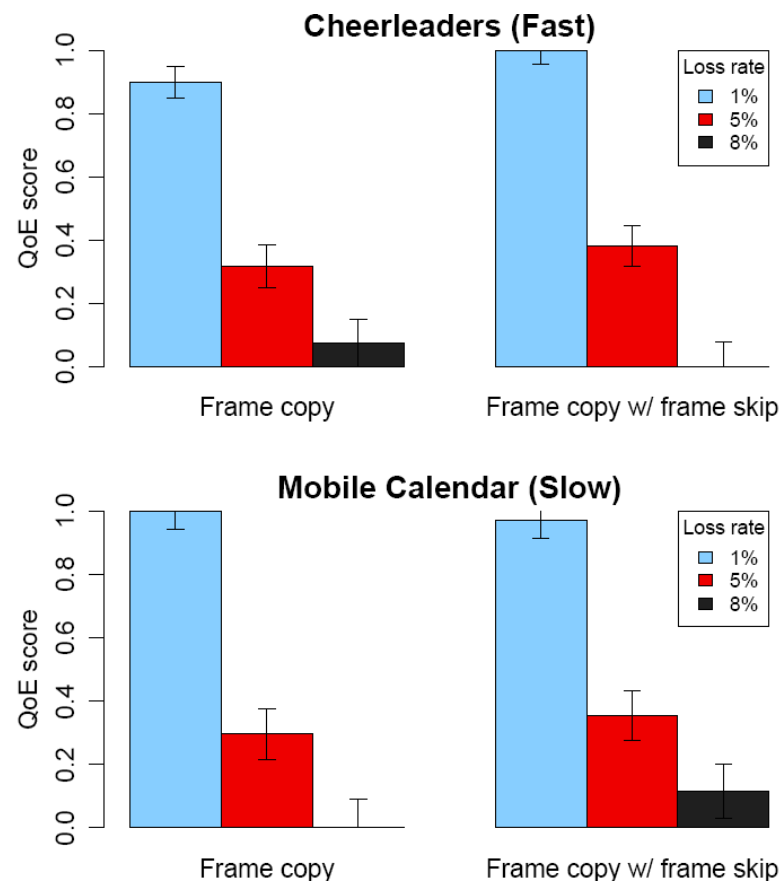
- Video codec
 - Source clips: one fast-paced and one slow-paced video clip
 - Three codecs: H.264, WMV3, and XVID
 - Two bit rates: 400 and 800 Kbps
 - 121 participants and 3,345 paired comparisons
- Loss concealment scheme
 - Source clips: one fast-paced and one slow-paced video clip
 - Two schemes: Frame copy (FC) and FC with frame skip (FCFS)
 - Packet loss rate: 1%, 5%, and 8%
 - 91 participants and 2,745 paired comparisons

Inferred QoE Scores

Video Codec



Concealment Scheme



Participant Source

- Laboratory
 - Recruit part-time workers at **an hourly rate of 8 USD**
- MTurk
 - Post experiments on the Mechanical Turk web site
 - Pay the participant **0.15 USD** for each qualified experiment
- Community
 - Seek participants on the website of an Internet community with 1.5 million members
 - Pay the participant an amount of virtual currency that was equivalent to **one US cent** for each qualified experiment

Participant Source Evaluation

- With crowdsourcing...
 - lower monetary cost
 - wider participant diversity
 - maintaining the evaluation results' quality



Case Study	Experimenter Source	Total Cost (dollar)	# Rounds	# Person	Qualified Rate	Cost / Round (cent)	Time / Round (sec)	Avg. TSR
MP3 Bit Rate	Laboratory	50.97	1440	10	67%	3.54	16	0.96
	MTurk	7.50	750	24	47%	1.00	9	0.96
	Community	1.03	1,470	93	54%	0.07	25	0.96

Crowdsourcing seems a good strategy for multimedia QoE assessment!



Quadrant of Euphoria



Researchers



Image

Register

Login



Audio

Register

Login



Video

Register

Login

- 💡 If you are a researcher who interested in Quadrant of Euphoria and want to try it out, we provide a demo profile for each type of experiment for you.
- 💡 Please login the demo profile by using name: **demo**, password: **qoedemo**



Experiment Participants

Type	Exp	Description	Reward	Link
	jpg2000	JPEG 2000 Quality Study.	\$1.0	go
	new_jpg	We want to test our new compression method.	N/A	go
	compression	Audio VBR compression level.	\$1.5	go
	mp3_lossless	Verify the loss-less MP3 codec.	\$1.5	go
	h264_test	Test if the new codec have significant quality boost.	\$1.0	go

[More...](#)

<http://mmnet.iis.sinica.edu.tw/link/qoe>

Conclusion

- Crowdsourcing is not without limitations
 - physical contact
 - environment control
 - media
- With paired comparison and user input verification,
 - less monetary cost
 - wider participant diversity
 - shorter experiment cycle
 - evaluation quality maintained



***MAY THE CROWD FORCE
WITH YOU!***

Thank You!



Thanks to LucasArts

Kuan-Ta Chen
Academia Sinica