Faculty of Computer Science

universität wien

Netzwerktechnologie für Multimedia Anwendungen (NTM)

Kapitel 5

Florian Metzger
florian.metzger@univie.ac.at

David Stezenbach
david.stezenbach@univie.ac.at

Bachelorstudium Informatik
WS 2014/2015

5. Quality of Experience

Mit Material von

- Markus Fiedler („Teletraffic Models for QoE Assessment", Euro-NF Summer School, Würzburg 2012)
- Touradj Ebrahimi („Quality of Multimedia Experience - Past, Present and Future", ACM Multimedia, Beijing 2009)
- Thomas Hoßfeld (u.a. „Internet Video Delivery in YouTube: From Traffic Measurements to Quality of Experience", 2013)
- Thomas Zinner (Vorlesung: "Neue Internet-Anwendungen: Technik und Modellierungsansätze - Medienübertragung über das Internet", Würzburg WS14)

Quality of Experience

- Kodier- und Netzwerkparameter definieren Qualität
  (z.B. Artefakte durch Kompression, Packet Loss)
- Wie empfinden **Menschen** diese Qualität?
  - Trade-off zwischen Videobitrate und Bildqualität?
  - Sehr gute Qualität erfordert möglicherweise eine enorme Bandbreite
- Quality of Experience (QoE)
  - Untersucht, wie menschliche Beobachter die **subjektive** (wahrgenommene) Qualität von Multimedia Inhalten bewerten
  - Abhängig von gewählten Kodier-Parametern, der (objektiven) QoS der Übertragung, den Umgebungsbedingungen, den Erwartungen der Benutzer und möglicherweise auch von weiteren Faktoren

Anwendungsbeispiel: Bitbudget

- Oft ist eine obere Grenze für die mögliche Bandbreite gegeben, z.B.
  - 90 Minuten auf einer BD
  - Begrenzung durch Access Network (DSL)
  - DVB-T Multiplex (~15 Mbit/s) für 4 Programme

- Gesucht ist ein Kompromiss:

  Ab welchen Kodier/Netzwerk-Parametern wird eine akzeptable **subjektive** Qualität erzielt?

  → Kann nur von menschlichen Beobachtern bewertet werden

Unterschiedliche QoE Definitionen

- White Paper [Nokia, 2005]:
  …*how a user perceives the usability of a service when in use – how satisfied he or she is with a service*
  - End-zu-Ende Netzwerk QoS
  - Netzabdeckung, Service- und Supportangebote, etc
  - Subjektive Faktoren wie Nutzererwartungen, -anforderungen und –erfahrungen
- Key Performance Indicators (KPI)
  - Zuverlässigkeit
  - Komfort/Zufriedenheit

Unterschiedliche QoE Definitionen

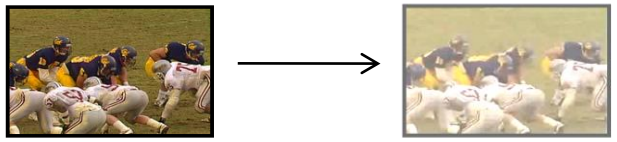
- QoE in ITU-T Rec. P.10/G.100 Am. 2 (2008):

The overall acceptability of an application or service, as perceived subjectively by the end user.

  - *NOTE 1 – Quality of experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.).*
  - *NOTE 2 – Overall acceptability may be influenced by user expectations and context.*

Quality of Experience

- Von QoS nach QoE
  - QoS: packet loss, delay, jitter, …
  - QoE: Subjektive Benutzerzufriedenheit eines Angebotes
- *Beispiel:* VoIP Nutzer ist an Sprachqualität interessiert
- QoE Messen/Bestimmen
  - Subjektiv: Testpersonen Ausgaben bewerten lasen und befragen
  - Objektiv: Modell nutzen
    - Full reference metric
    - Reduced reference metric
    - No reference metric



- Quantifizierung für QoE benötigt um Dienste für bestimmte QoS-Szenarien bewerten zu

QoE-Messansätze

- Umfragen unter Benutzern
  - Genaue Messung möglich
  - Lange Dauer um “objektive” Bewertungen zu erhalten (Einschätzungen immer subjektiv, persönliche Vorlieben)
  - Schwierigkeit menschliche Reaktionen mathematisch korrekt festzuhalten und weiter zu verarbeiten
- Quantitative QoE Metriken
  - Mathematisch beschreibbar
  - Nicht unbedingt akkurate Wiedergabe der Wahrnehmung
- Inhärent subjektiv, mehr als eine reine Umsetzung von QoS
- Offene Forschungsaufgabe: Finden von mathematischen Modellen der menschlichen Wahrnehmung für bestimmte Anwendungen.

Subjektive QoE Messungen

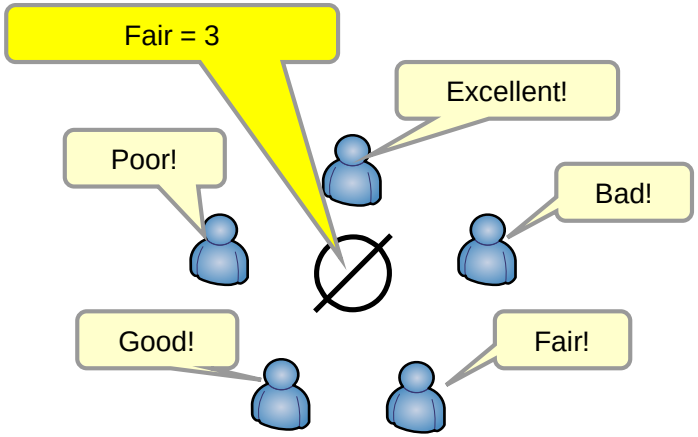
- Korrekte Wahl von Werkzeugen, Umgebung und Inhalten
- Direkte Befragung oder Beobachtung
  - Fragebögen
  - Benotung
  - Anmerkungen und Kommentare
  - Non-/Verbale Methoden
  - Etc.
- Mögliche (indirekte/ungewollte) Einflußfaktoren
  - Aufsichts-/Begleitpersonen
  - Inhalt und Art der Präsentation der zu testenden Anwendung
  - Test- und Auswertungsmethodik
  - Testumgebung (Labor, ...)

Mean Opinion Score (MOS)

- Bewertungseinheit für subjektive Meinungsumfragen
- Numerische Darstellung der wahrgenommenen Qualität von dargestellten Medien nach deren Komprimierung und/oder Übertragung
- Durchschnittlicher Grad der Zufriedenheit für einen durchschnittlichen Benutzer
- Kann leicht falsch angewendet werden („Meaningless Opinion Score“)
  - Durchschnitt nicht unbedingt aussagekräftig, z.B. durch Ausreißer
  - Meinungsumfragen nicht immer sinnvoll anwendbar

Mean Opinion Score (MOS)

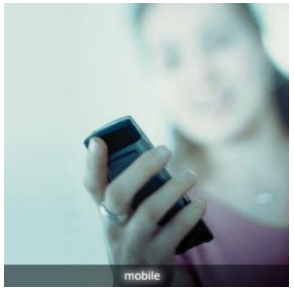
- Durchschnitt aus einzelnen Bewertungen (OS)
- Typischerweise kontinuierliche oder diskrete Einheit mit **ungerader** Stufenzahl
  - 1, 2, 3, 4, 5(,.., 7(,.., 9))
  - 0, 1, 2, …, 10
  - 0…100
  - Alternativ auch binäre ja/nein Entscheidung



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

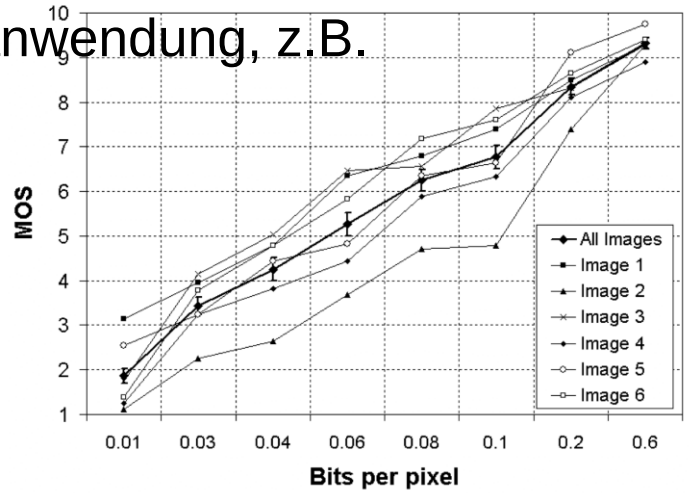
Testumgebung

- Messergebnisse müssen zuverlässig und reproduzierbar sein
  - Subjektive Evaluationsmethodiken
  - Hohe Voraussetzungen an Testumgebung
- Viele Einflussfaktoren, z.B.
  - Art und Qualität von Monitor, Lautsprechern und anderen Testgeräten
  - Licht und Akustikbedingungen
  - Laborarchitektur und Grundriss
  - Betrachtungsdistanz und –winkel



Testmaterial

- Auch Medieninhalte beeinflussen Bewertung
  - Wahl des richtigen Inhalts für die Zielanwendung, z.B.
    - Typischer Inhalt
    - Worst-case Inhalt




P01


p06


p10


bike


cafe


woman

Testmethodik

- Wahl der Teilnehmer (Laien oder Experten)
- Instruktionen (z.B. Welche Fragen wie stellen?)
- Präsentation (Single oder double stimulus, gleichzeitig oder nacheinander)
- Benotungsschema (Numerisch oder kategorisch)
- ITU Recommendations
  - ITU-R BT. 500-11 “Methodology for the subjective assessment of the quality of television pictures” (1974-2002).
  - ITU-T P. 910 “Subjective video quality assessment methods for multimedia applications” (1999).
  - ITU-R BT. 1788 “Methodology for the subjective assessment of video quality in multimedia applications” (2007).

Testmethodik

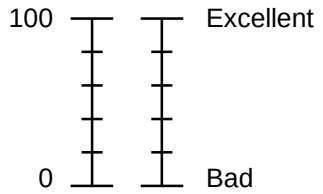
Single Stimulus (Einzelbild)

Kategorische Attributsskala

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

Kategorische und numerische Skala:
“Benote zwischen 1 und 11”

Nicht-Kategorische Attributsskala oder numerische Skala



Testmethodik

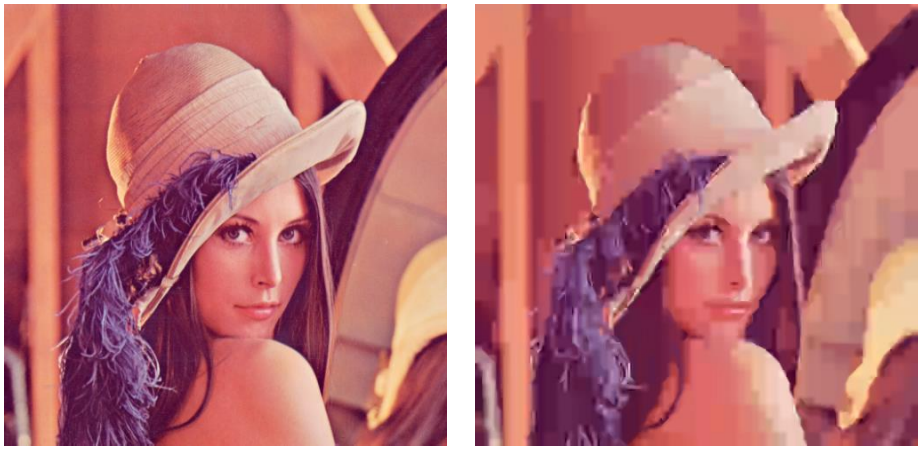
Double Stimulus Impairment Scale
(Quelle und Resultat nacheinander)


Kategorische Beeinträchtigungsskala

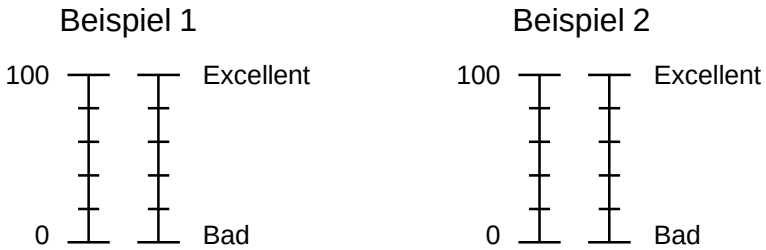
5 Excellent	5 Imperceptible
4 Good	4 Perceptible but not annoying
3 Fair	3 Slightly annoying
2 Poor	2 Annoying
1 Bad	1 Very annoying

Testmethodik

Double Stimulus Continuous Quality Scale (DSCQS)



Nicht-Kategorische Attributsskala oder numerische Skala



Testmethodik

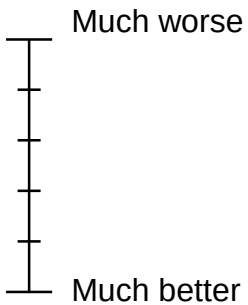
Stimulus Comparison (SC, gleichzeitig gezeigt)


Relative Kategorische Attributsskala

“gleich oder unterschiedlich”

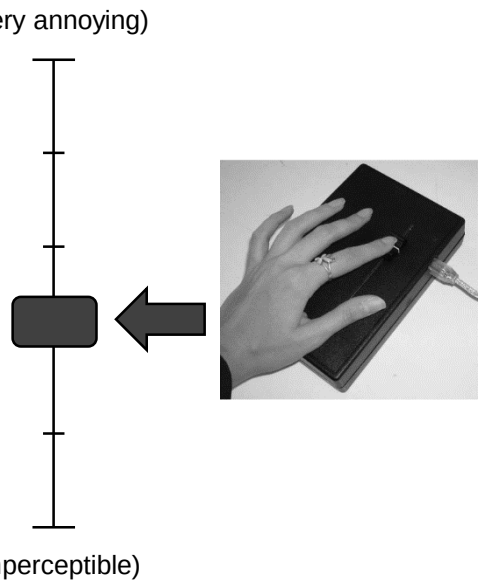
much worse
worse
slightly worse
the same
slightly better
better
much better

Relative Nicht-kategorische Bewertung



Testmethodik

Single Stimulus Continuous Quality Evaluation (für Video, SSCQE)

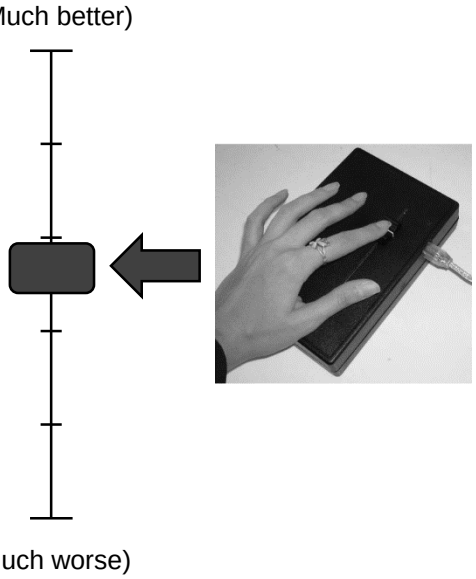


Testmethodik

- Simultaneous Double Stimulus for Continuous Evaluation (für Video, SDSCE)

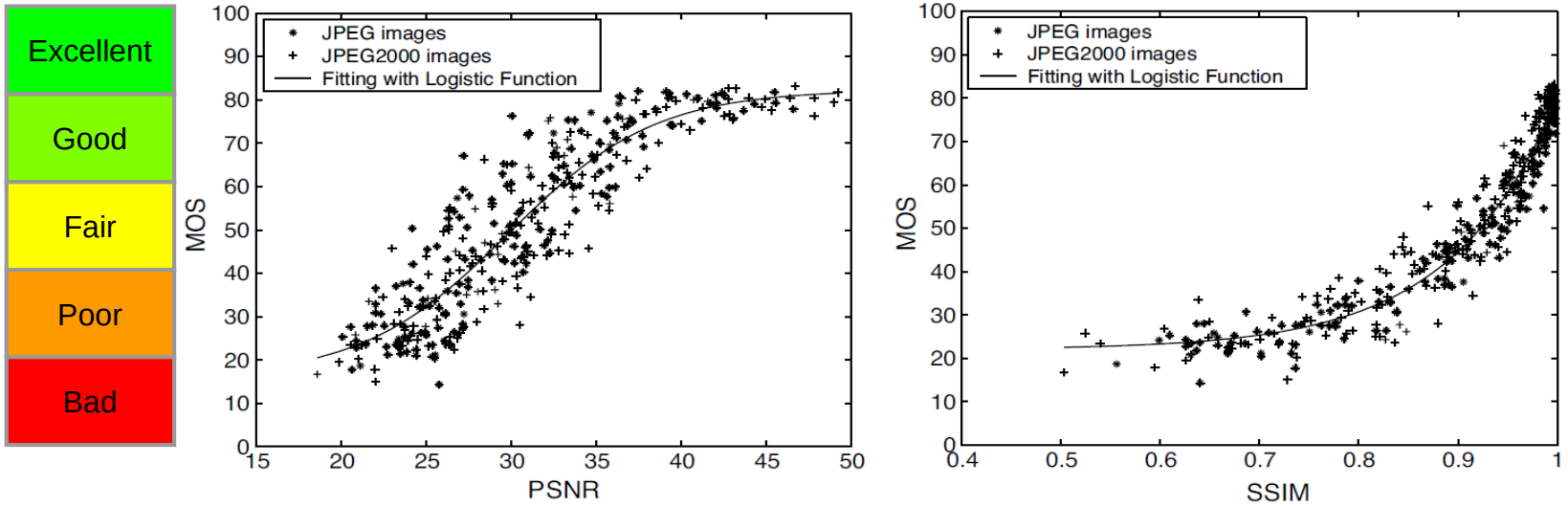
(Referenz)

(Testsequenz)



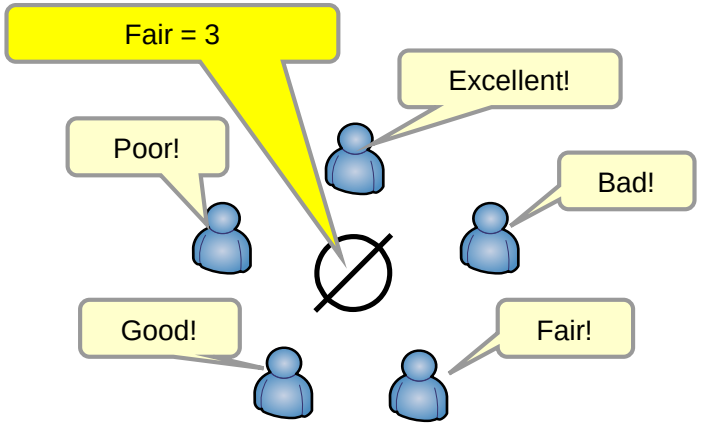
Von subjektiven zu objektiven Metriken

- Funktion bildet berechnete Werte auf wahrgenommene Qualitätsskala ab



(aus “Image quality assessment: From error visibility to structural similarity”, Z. Wang et. al.)

QoE Measurement



Subjective User Ratings

Measuring of or mapping to key influence factors

Objective Measurements

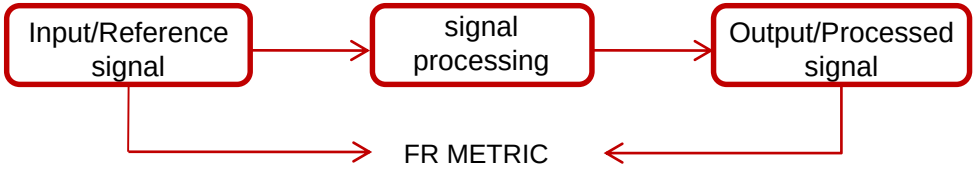
Identification of Key influence factors

Mapping objective quality to subjective QoE

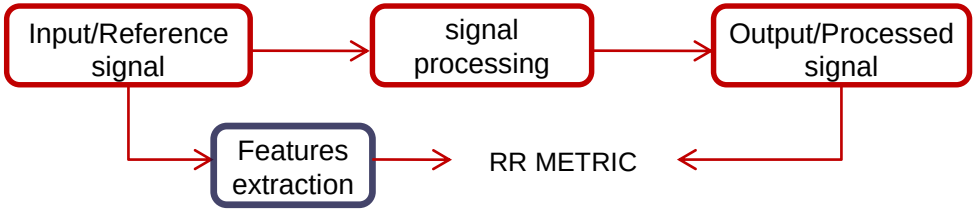
QoE Model

Objektive QoE Metriken

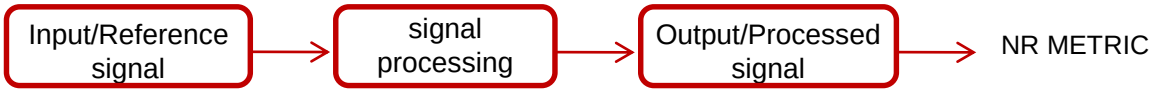
- Subjektive Tests zeitaufwändig und kompliziert
- Objektive Algorithmen und Metriken sollen einen subjektiven MOS mit hohem Korrelationsgrad abschätzen
  - Full reference metrics



  - Reduced reference metrics



  - No reference metrics



Full Reference Metriken

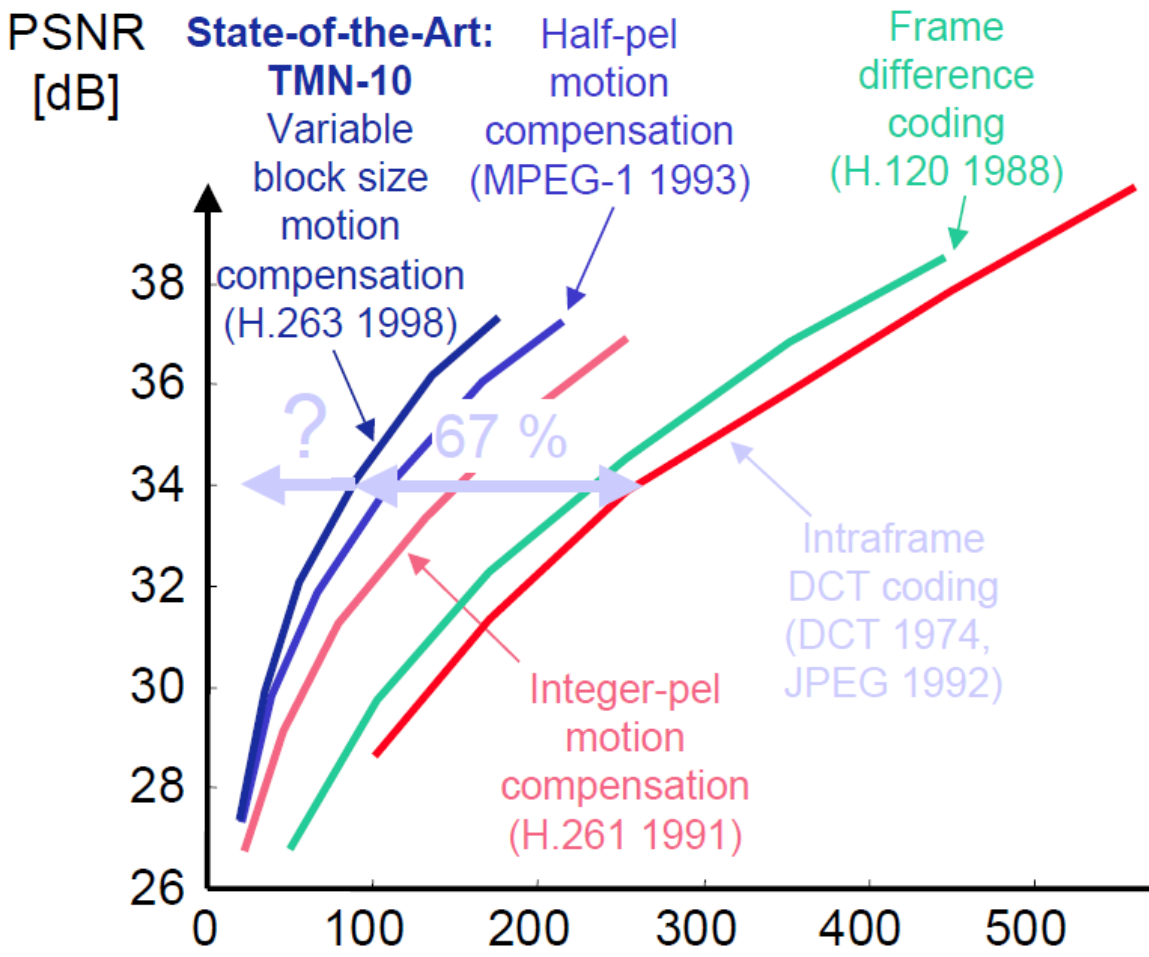
- Metrik zur Bestimmung der Wiedergabetreue im Vergleich zu einer Referenz
- Beispiele
  - Mean Square Error (MSE)
  - Peak Signal to Noise Ratio (PSNR)
  - Maximum Pixel Deviation (Linf)
  - Weighted PSNR
  - Masked PSNR
  - Structural SIMilarity (SSIM)
  - Multiscale Structural Similarity (MSSIM)
  - Visual Information Fidelity (VIF)
  - Video Quality Metric (VQM)

Full Reference Metriken

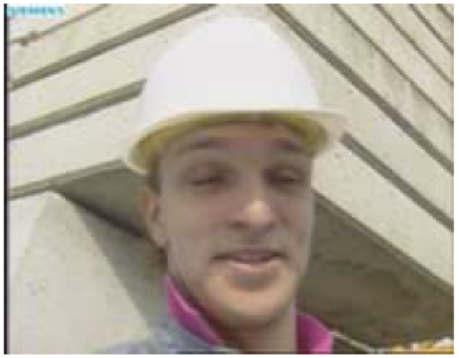
- Einfache und gebräuchliche Metriken für Bildkompression
  - mean square error (MSE) $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2$,

    mit den Ein- und Ausgabebilddaten X und Y.
  - peak signal to noise ratio (PSNR) $PSNR = 10log_{10}\left(\frac{MAX_{referenz}^2}{MSE}\right)$
    - in Dezibel (dB)
    - Weit verbreitete Verwendung durch die Einfachheit
    - Dadurch aber auch nur beschränkter Nutzen

PSNR



Foreman
10 Hz, QCIF
100 frames encoded



Structural Similarity (SSIM) Index

$$S(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

- Anwendung nur auf den Luminanz-Kanal
- Sliding Window Verfahren (z.B. auf 8x8 Blöcke)

σ_{xy} Kovarianz

μ_x Mittelwert

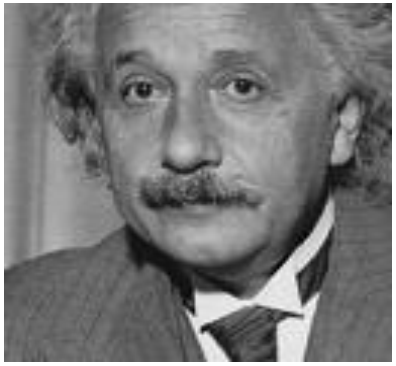
$C_1 = (k_1 L)^2; C_2 = (k_2 L)^2$

L Dynamikumfang (z.B. $2^{32} - 1$)

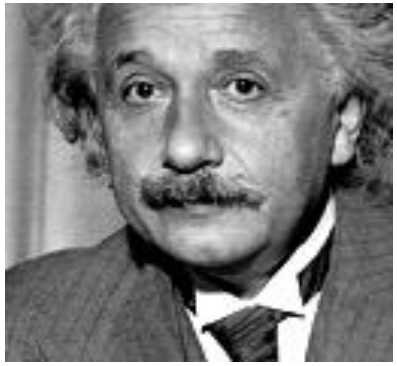
$k_1 = 0{,}01; k_2 = 0{,}03$

MSE vs. SSIM

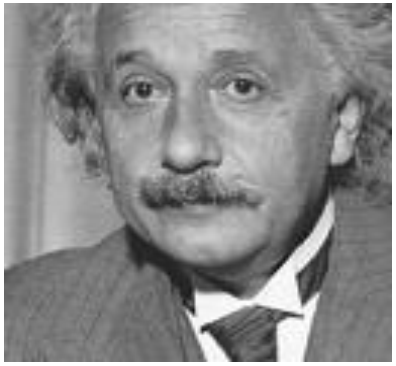
Referenz



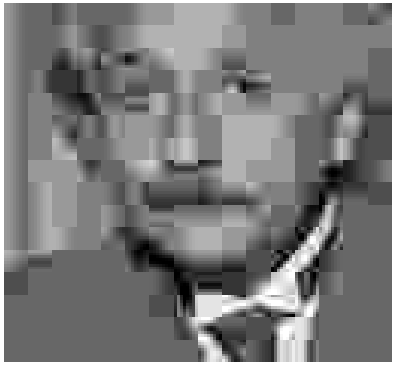
MSE=0, MSSIM=1



MSE=309, MSSIM=0.928

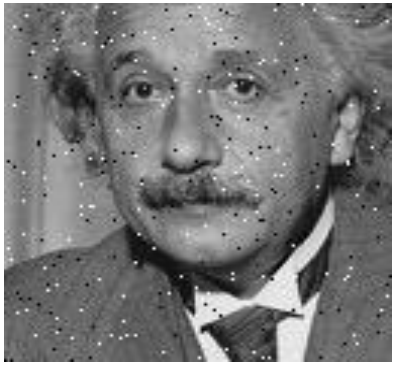


MSE=309, MSSIM=0.987



MSE=309, MSSIM=0.580

MSE=308, MSSIM=0.641



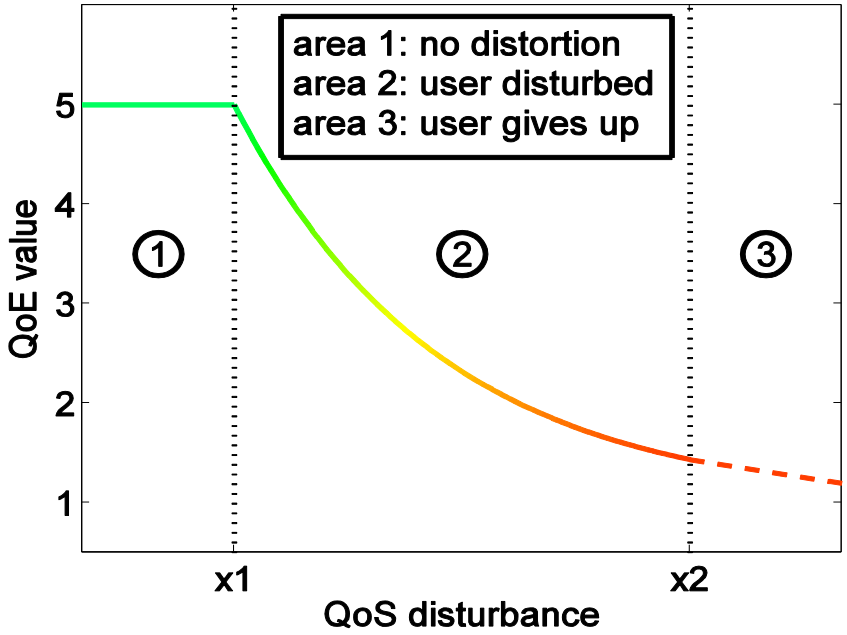
MSE=309, MSSIM=0.730

Sprachübertragungsmetriken in Standards

- PESQ (Perceptual Evaluation of Speech Quality)
  - ITU-T Standard P.862 (2001) zur objektiven Bewertung der Sprachqualität beim Testen von Übertragungssystemen
  - Einspeisen von Sprachproben an einem Ende der Verbindung und Vergleich mit Original am anderen Ende
  - Benutzt psychoakustisches Modell zur mathematischen Nachbildung des subjektiven Hörempfindens
- POLQA (Perceptual Objective Listening Quality Analysis)
  - ITU-T Standard P.863 (2011) zur instrumentellen Bewertung der Sprachqualität beim Testen von Übertragungssystemen
  - Erlaubt größere Frequenzbereiche (14kHz) als das ältere PESQ
  - Gleiche Skalen und Bewertungsmechanismen wie PESQ, aber größere Genauigkeit im VoIP Bereich

Netzwerkeinflüsse auf QoE

- QoS-Faktoren beeinflussen QoE
  - QoE = f(QoS_1, QoS_2, …)
- Bei Tests (Messungen, Befragungen, ...)
  - Beschränkung auf immer nur einen QoS-Faktor



Fundamentale QoE-QoS Beziehungen

- Linear $QoE \pentagram QoS_i$
- Logarithmisch $QoE \pentagram \log(QoS_i)$
- Exponentiell $\log(QoE) \pentagram QoS_i$
- (Potenz $\log(QoE) \pentagram \log(QoS_i)$)

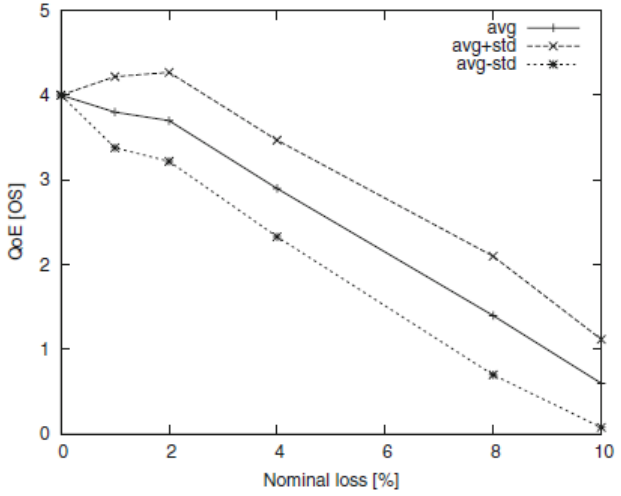
- Beispiele aus:
  Shaikh, J., Fiedler, M., & Collange, D. (2010). Quality of Experience from user and network perspectives. *annals of telecommunications-annales des télécommunications*, *65*(1-2), 47-57.

Lineare Beziehung

QoE		QoS axis
Lineare Skala	←	Lineare Skala
Additive Änderung	←	Additive Änderung

- QoE Steigung unabhängig von aktuellen QoE/QoS Werten
- i.d.R. Modelliert durch Lineare Regression

Beispiel: Wahrnehmung der Downloadzeit im Verhältnis zu Paketverlust (Shaikh et al., 2010)



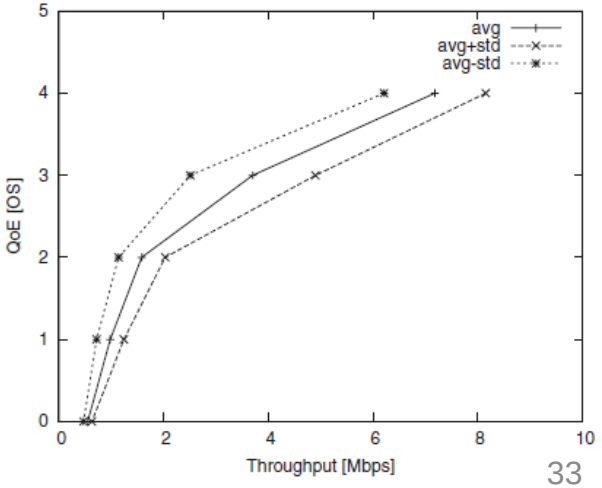
$$QoE \approx 4.3 - 31PLR \quad (\Re^2 > 0.99)$$

Logarithmische Beziehung

QoE		QoS
Lineare skala	←	Logarithmische skala
Additive Änderung	←	Multiplikative Änderung

- QoE Steigung proportional zum Kehrwert der QoS
- Weber-Fechner-Gesetz (1834)
  - QoS-Änderungen bei kleinen QoS-Werten haben größeren Einfluss auf QoE als bei hohen Werten
  - Schall, Temperatur, Helligkeit, ...

Beispiel: Wahrnehmung der Downloadzeit im Verhältnis zur Bandbreite (Shaikh et al., 2010)



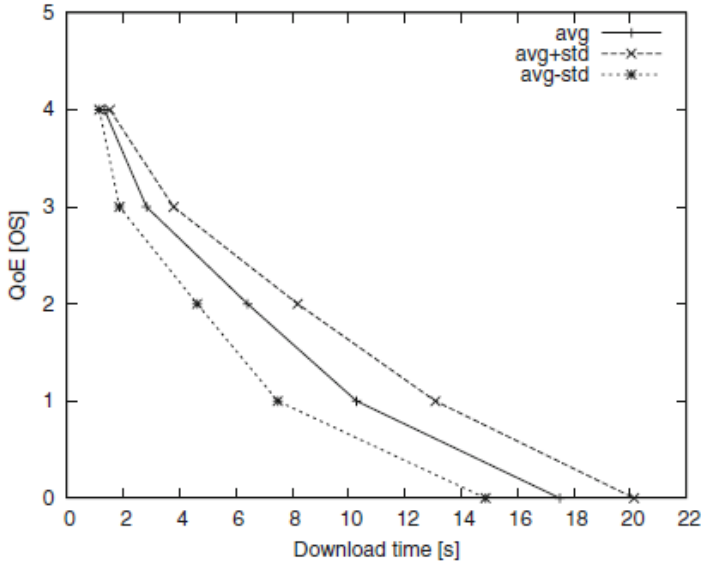
$$QoE \approx 1.2 + 3.3 \lg(R/\mathrm{Mbps}) \quad (\Re^2 > 0.99)$$

Exponentielle Beziehung

QoE		QoS
Logarithmische Skala	←	Linear skala
Multiplikative Änderung	←	Additive Änderung

- QoE Steigung proportional zur eigentlichen QoE
- IQX Hypothese

Beispiel: Wahrnehmung der Downloadzeit (Shaikh et al., 2010)



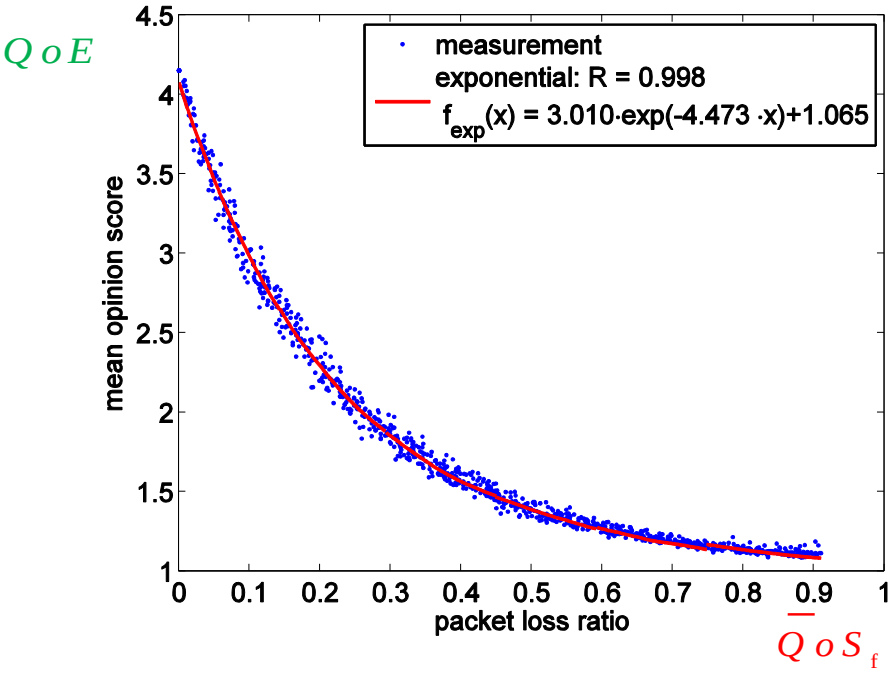
$$QoE \approx 4.8 \exp(-0.15 RT/\mathrm{s}) \quad (\Re^2 > 0.99)$$

Typen von QoE und QoS Parametern

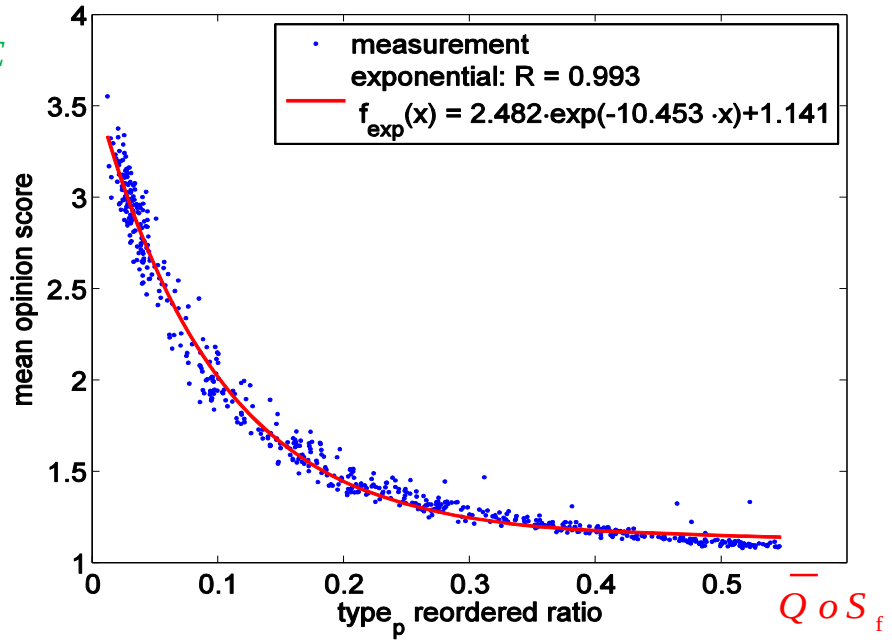
- Success rating QoE
  - Größer ist besser
  - MOS (1..5)
- Failure rating $\overline{QoE}$
  - Größer ist schlechter
  - Abbruchrate
  - Churnrate

- Resource measure QoS_r
  - Größer ist besser
  - Durchsatz
- Success measure QoS_s
  - Verfügbarkeit (e.g. 99.99 %)
  - Erfolgsquote
- Failure measure $\overline{QoS}_f$
  - Größer ist schlechter
  - Paketverlustrate
  - Jitter
  - Paket-Reordering

QoE Beispiele: UDP Audio (Skype)

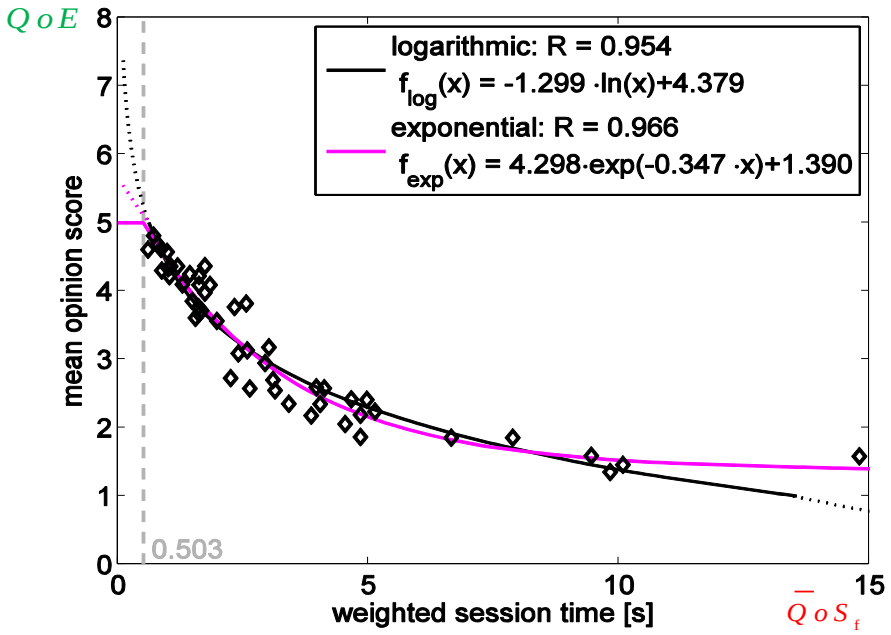


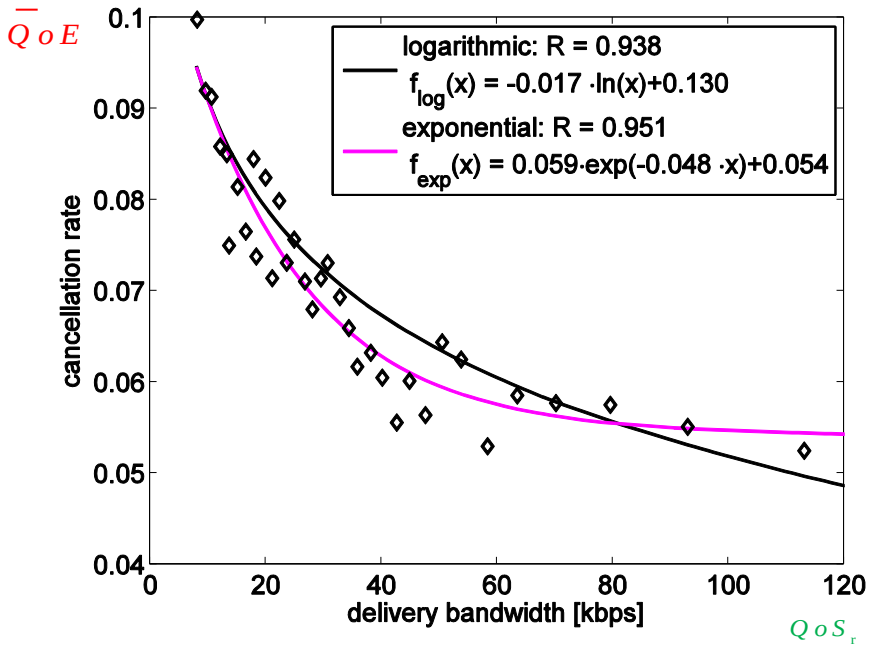
MOS = *f*(packet loss ratio)



MOS = *f*(reordered ratio)

QoE Beispiele: Web





MOS = *f*(session time)

Cancel-rate = *f*(delivery bandwidth)

Streaming Metriken

- UDP
  - Implizit adaptiv
  - Einige FR/RR Videoqualitätsmetriken funktionieren weiterhin
  - Dediziert z.B. Media Delivery Index (Delay Factor, Media Loss Rate) [RFC 4445]
- Progressive TCP
  - Nur Stalling (Dauer L, Anzahl N, „Initial Stall“)
  - Anzahl der Stall-Phasen „schlimmer“ als Dauer
  - Typischerweise No-Reference (Continuity Index, Pause Index, …)
    - Rebuffering Artifact Value $v = 1 - \left(\frac{L}{dur_{video}}\right)^{0,0737}$ [ITU-T P1201.1]
    - $f(L, N) = 3{,}5e^{-(0{,}15L+0{,}19)N} + 1{,}5$ [Hoßfeld, 2013]
- Adaptive TCP
  - Zusätzliche Parameter-Dimension
  - QoE-Metriken? (FR, RR, NR?)

Beispiel: Video Streaming QoE Hysterese