STQ Workshop

New Horizons for practical metrics on speech intelligibility

Dr. Wolfgang Balzer

Joachim Pomy

17/11/2022

The QoE perspective – get the ball rolling

- We are not going to present solutions in the field of SpQ algorithms and models – there are people who are very good at that.
- Our perspective is that of a designer of metrics, testing strategies, and testing tools, with a focus on end to end, i.e. QoE perspective. The task is to translate strategic goals and requirements into methodologies which can produce the input needed for fact-based decisions.
- These methodologies need subsystems – and we are offering functional goals and assessment criteria for them.

Today…

- There are well-working tools for subjective assessment of speech quality (e.g. ITU-T P.863)
- These may however be called only the „low-hanging fruits“.
  - Speech lab-calibrated assessment of audio quality under ideal conditions
  - Deliberately avoiding emotional elements or even “meaning”

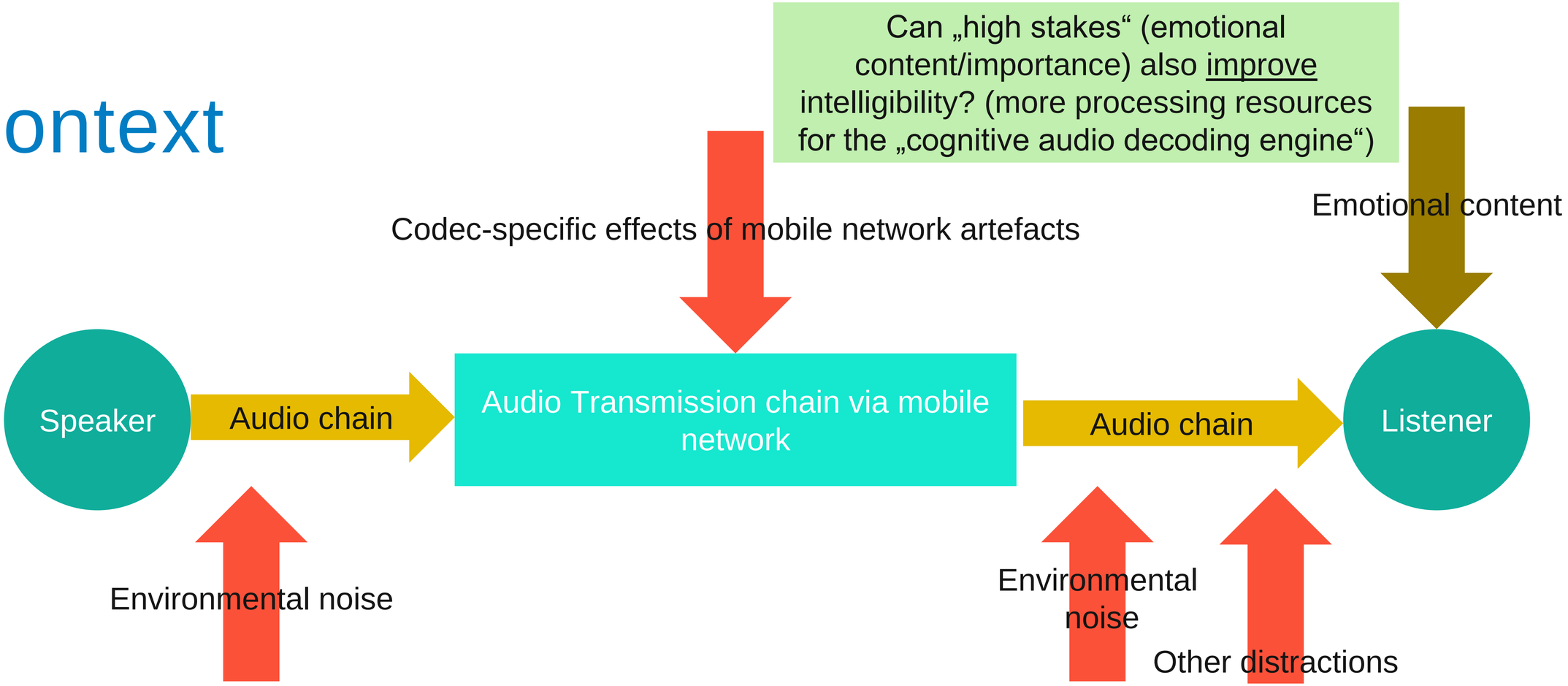
First set of questions

- Soft drop: technically the connection still exists, but audio is so poor that call parties agree to end the call
  - Today: assessment of „soft drops“ via MOS values of subsequent samples
  - Does the MOS scale really reflect the “annoyingness” of audio artefacts?
- How do isolated „unpleasant“ artefacts affect the listening experience?
  - Frequency and severity (beyond a fixed-time sampling interval); are there build-up effects?
  - „Stressful“ listening experience: nonlinear effects across longer periods of time
- Even with the modest level of “ non-emotional” content: how about non-native listeners? (courtesy Christian Schmidmer, Opticom)

Time for higher-hanging fruits

- True speech intelligibility
- Realistic conditions
  - Environmental noise and other distractions
  - Emotional involvement
- Yes, there has been and still is research, but is it using the full spectrum of available elements?
- How about the effect of “clustered” impairments (bursty artifacts: density vs. intensity) in conversational situations?

Context



- A codec „translates“ effects of mobility/inadequate coverage into audio artefacts
  - Essentially packet loss/packet retransmission due to insufficient RF level, plus effects of handover/reconfiguration (peaks in latency)

focusinfocom
Member of NET CHECK

Value of audio quality

- The actual use of an audio interaction is to exchange information, with some purpose in mind
  - QoE: how stressful is the conversation?
    - Intelligibility: How hard is it to extract the correct information from incoming audio? Or worse – can errors go unnoticed?
      - How much secondary communication is needed to transfer the intended content?
        - „Can you repeat“; meta communication about poor quality

The building blocks and tools should be there

- Sensors present in today‘s smartphones
  - Environmental noise, light, acceleration (indicators of current environment)
- „Crowdworking“ to access actual user experience
- Pattern recognition (machine-learning) technologies
  - Classification as well as data cleansing/detection of unreliable samples
- Modern concepts to create deeper involvement of subjects, e.g. gamification

Critical assessment – TS 103 558 (2021-07)

- “Communication in noisy environments may be extremely stressful for the person located at the near-end side. Since the background noise is originated from the natural environment, it can usually not be reduced for the listener.” (Section 4, first paragraph)
  - Annex D presents also ANC headphones. In practice, phone calls over fully isolating ANC are unpleasant due to missing own voice (that’s why such headphones use sidetone functions when in a call)
- “In contrast to "classical" intelligibility tests, the auditory assessment of listening effort collects opinion scores instead of "measuring" the word error rate of multiple test subjects.” (Section 4, third paragraph)

  - Does this consider the effect of accelerated learning by professional subjects?
  - Cross-validation against crowdsourcing/crowdworking with real situations
    - Is the effect of a real environment equivalent/calibratable?
- Annex D (D.5) show considerable spread of correlation (0.5 – 1.0 MOS)
- Does TS 103 558 have a „takeway“? (conclusion/result in a nutshell)

Looking at labs

- Can a lab actually create the full extent of a target situation?
  - How do we create an emotionally stressful situation?
  - How do we create conditions beyond ambient noise (i.e. beyond audio channel only)?
- → Gamification/reward/loss situations; How can we determine when the goal is reached?
- How can e measure when a test user is „used up“ (listening experience >> „normal listener“)?

Speech quality vs speech intelligibility

- Speech quality: looking from the outside – how does it sound?
- Speech intelligibility: looking at the inside
  - Listening effort: how easy or difficult is it to get at the information content?
  - How high is the risk of information loss or distortion of information?
  - How difficult/tiring/emotionally draining is it to listen to audio?
- Evolved time scales
  - Is it possible to leave the strict „fixed sample length“ pattern? E.g. find „audio events“ which significantly impact perceived quality

Speech Intelligibility

- Subjective: ask people how hard or easy is it to understand the content?
  - Problem: Subjects get used (or „professional“) in listening; scales are changing
  - Better to have some objective indicators
    - Physiological indicators: Brainwave patterns, Eye movement, Blood pressure, Heart rate,...
    - Cognitive indicators, e.g. Answer delay (in a challenge/response situation), audio instruction to do something on the device; Quality of response (e.g. a „dictation“ scenario)
- Objective: correlate measurable quantities to actual listening effort even if it is „subconscious“
- Pattern recognition: Find correlation between audio patterns and subjective indicators for stress/high listening effort

Some work has been done…

- (by TU Berlin)
  - Effect of external stimuli (environmental noise) on perceived audio quality
  - Extend to evolved listening effort/speech intelligibility indicators
- There is various research work on speech intelligibility and listening effort (see e.g. McGarrigle-et-al-2014-Listening-effort-and-fatigue-discussion.pdf)
- There are also products claiming to deliver Listening Effort scores. However, it appears that nothing is in sight which is comparable to P.863/POLQA in terms of widespread acceptance and/or usability.

…but there is some way to go

- Objective/quantitative measurement of delay effects in conversational situations
- Look at new methodologies
  - (courtesy Christian Schmidmer, Opticom): Using minimal pairs to construct testing situations (Minimal pairs: words which differ in only one phonological element and have distinct meanings)
  - Further bio- and cognitive indicators for listening effort/audio degradation-induced stress
- And some even more ambitious goals
  - Is 100% audio quality really the limit? E.g. some kind of „forward shaping“ to improve intelligibility in noisy environments – characterization of respective methods by objective measurements
- And of course translate academia results into actual fieldworthy products

Thank you for your attention!

Contact:

Dr. Wolfgang Balzer
wfk@focus-infocom.de

Dipl. Ing. Joachim Pomy
consultant@joachimpomy.de

focusinfocom
Member of NET CHECK