IETF Audio Codec: Quality Testing

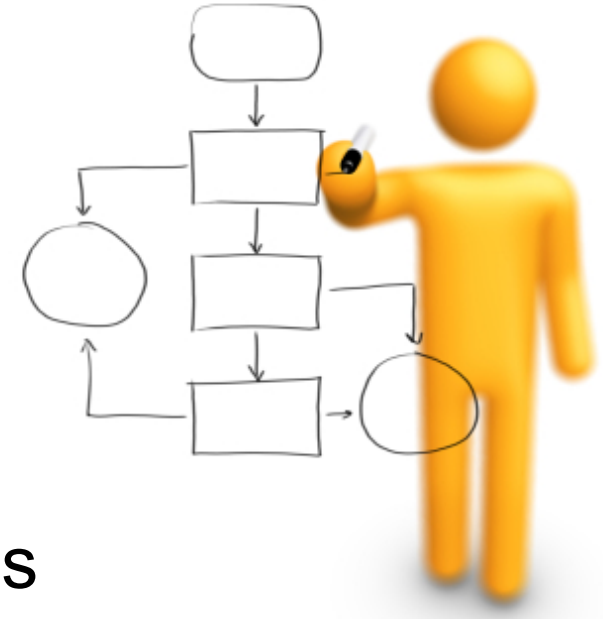
Christian Hoene, Universität Tübingen
Michael Knappe, Juniper Networks

Contents

- Introduction / Purpose of Presentation
- Background to Quality Testing and Codec Standardization
  - “Design of an IP Phone” – signal path walkthrough
  - Subjective Tests
  - Objective Tests
- ‘Realistic’ testing and potential pitfalls
- Recommendation for a streamlined workflow of required characterization testing
- Test house volunteers / recommended signal chain
- Potential future liaisons (outside WG scope)

Introduction

- In conjunction with codec development activities, the codec WG will also specify **a workflow for codec characterization**



- Worthwhile to broadly review current subjective and objective evaluation techniques
- Narrow the evaluation scope to tractable WG activities
  - **Goal 1: Agree upon characterization workflow**
  - **Goal 2: Sign up testing volunteers**

Why Codec Characterization?

Addresses the questions

- Does it fulfill the requirements?
- Is it free of major bugs?
- How does the codec perform in a real setting?
  - Needed for network planing
  - codec adaptation
  - selection amoung standardized codecs
  - advertising
  - …
- Ensure the high quality of the IIAC standard
- Do quality testing!

Where does quality testing have an impact?

During…

1. The requirement definition stage
   - Definition of scope and design goals
2. Codec development
   - Inventing and iterating codec algorithms
3. Codec selection
   - Comparing different codec contributions
4. Codec standardization
   - Describing the codec in absolute and/or relative terms
5. Qualification
   - Similar to 4, understanding the performance of the codec
6. Implementation Testing
   - Testing codec implementations for ‘correctness’
7. Conformance Testing
   - Checking codec implementations for interoperability

In our WG context...

1. ~~The requirement definition stage~~
   - *not required for current codec scope*
2. Codec development
   - Understanding the impact of different design decisions
3. ~~Codec selection~~
   - *emphasis on WG collaboration and consensus*
4. ~~Codec standardisation~~
   - *emphasis on WG collaboration and consensus*
5. Qualification
   - *important guidepost for codec 'advertising'*
6. Implementation Testing
   - ensure software quality
7. Conformance Testing
   - Ensure interoperability

Who needs the quality test results?

- Codec developers
  - Which algorithm/parameters to select?
- Equipment manufacturerrs
  - Which codec shall we implement or include?
- Network planning
  - How much bandwidth do we need for good quality?
- VoIP applications/rate control
  - How to parameterize the codec to work ideally under the current transmission conditions?
- End users
  - Ingredient branding: „IETF Codec Inside“

Subjective Listening-Only Tests

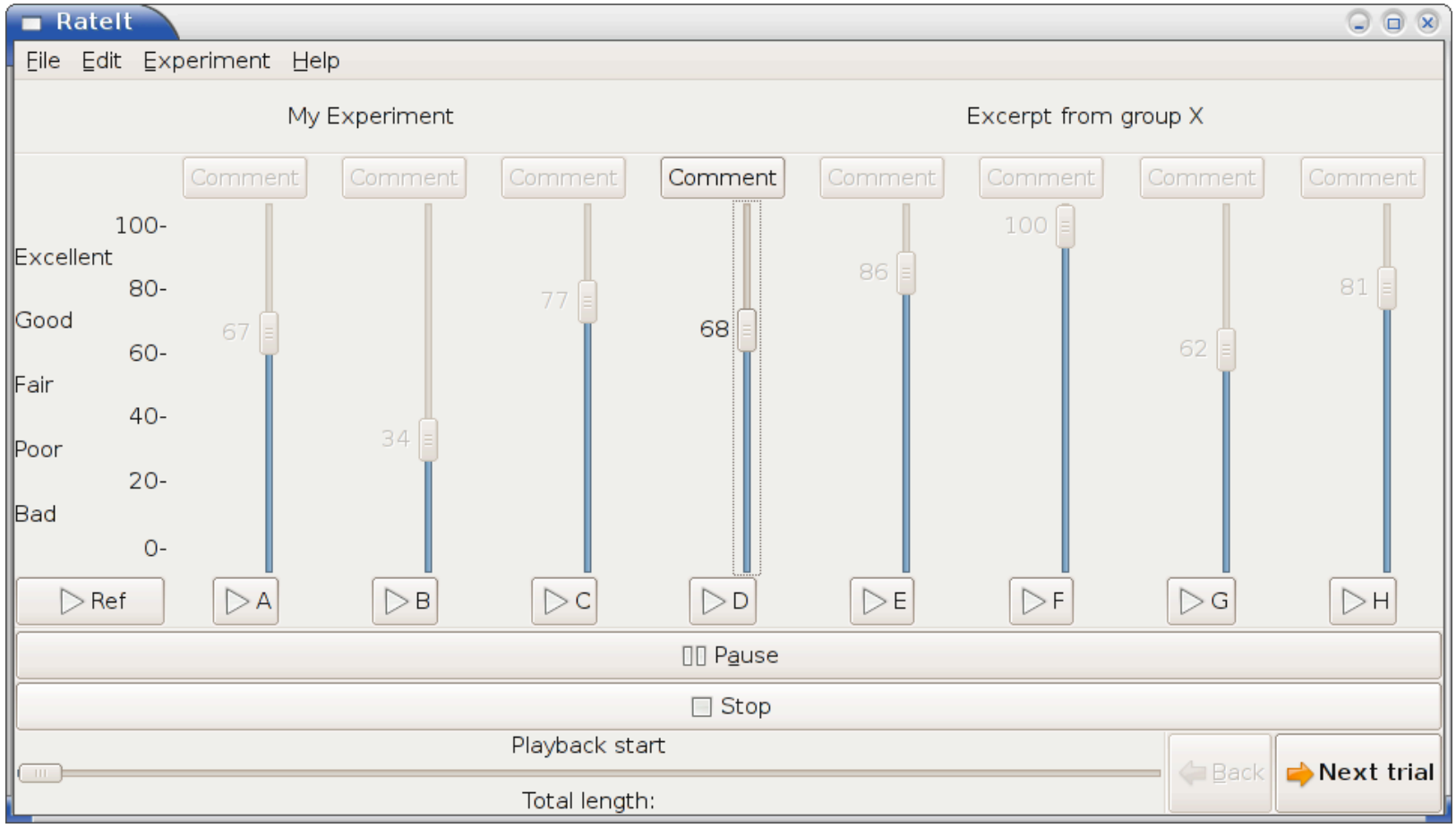
- Standardized at ITU-R and ITU-T SG12
- ITU-T P.800: Absolute Category Scale (ACS)
  - Having 5, 11, or more categories.
  - Classically used for speech (ACR-5, MOS) and video (ACR-11)
  - Fastest method
- ITU-R BS.1116-1:
  - Most precise (for high quality audio tests)
  - Used for development of G.719
- ITU-R BS.1534-1: Mushra Testing
  - For intermediate quality
  - Faster than BS.1116

MUSHRA

- MUltiple Stimuli with Hidden Reference and Anchor
- ITU-R BS.1534-1
- Recommended for assessing ‘intermediate audio quality’
- Uses both a known reference and hidden reference, along with hidden anchors, including a 3.5 kHz bandlimited version of reference to pull the scale closer to an absolute measure
- Requires statistically fewer subjective participants to generate a significant score

MUSHRA screenshot

- RateIt tool - thanks Jean-Marc!



Conversational Tests

- Telephony is bidirectional.
  Listening-only tests do not cover interactivity.
- Conversational Tests are more realistic as compared to listening-only tests
  - Because they also consider delay, echos, …
- Thus, conversational tests might needed
  - Defined in ITU-T P.800 for speech only
  - Uses ACR-5 (MOS)

BUT

- No tests for distributed ensemble performances
- No tests for teleconferencing scenarios, yet

Instrumental Testing Methods

- Subjective tests are expensive and time consuming
- Objective (instrumental) tests try to predict human rating

- PESQ: Perceptual Evaluation of Speech Quality
  - ITU-T P.862
  - listening-only tests (MOS)
  - Correlation R=0.94 for known kinds of distortions
- POLQA: Perceptual Objective Listening Quality Analysis
  - Updating PESQ
  - From narrowband till superwideband
  - Also time stretching/shrinking
- PEAQ: Perceptual Evaluation of Audio Quality
  - ITU-R BS.1387-1
  - Listening-only tests (ACR)
  - Packet loss?
  - No time variations!

Instrumental Testing Methods (cont)

- Objective testing unreliable for unknown distortions
  - Without subjective testings and mapping to subjective ratings.
  - New codec introduces a kind of new distortion
- After successful verification, objective algorithms are assumed to give stable and reliable ratings
- Define mapping from objective to subjective ratings

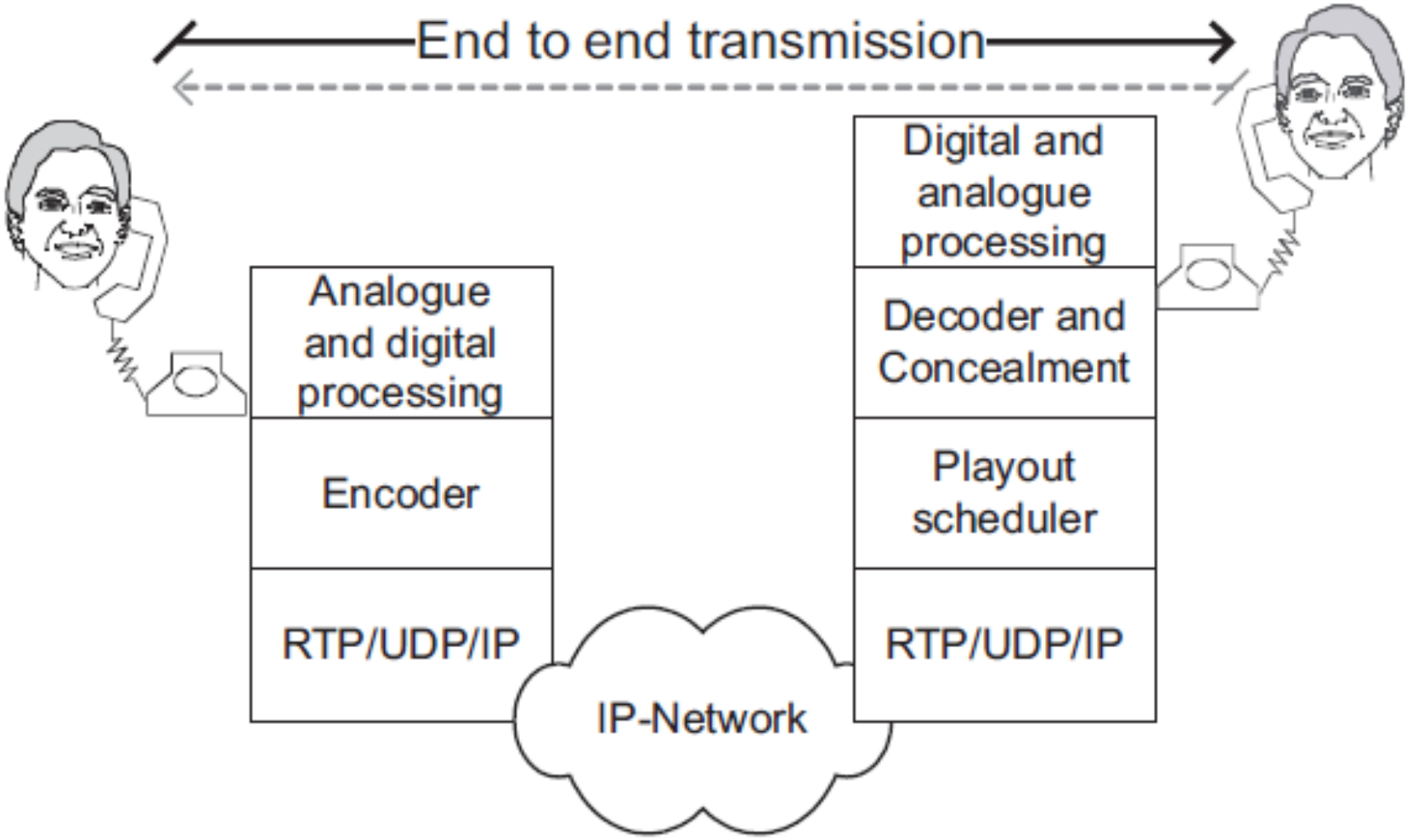
Measuring Quality of Experience

- "The overall acceptability of an application or service, as perceived subjectively by the end-user." [ITU-T P.10/G. 100]

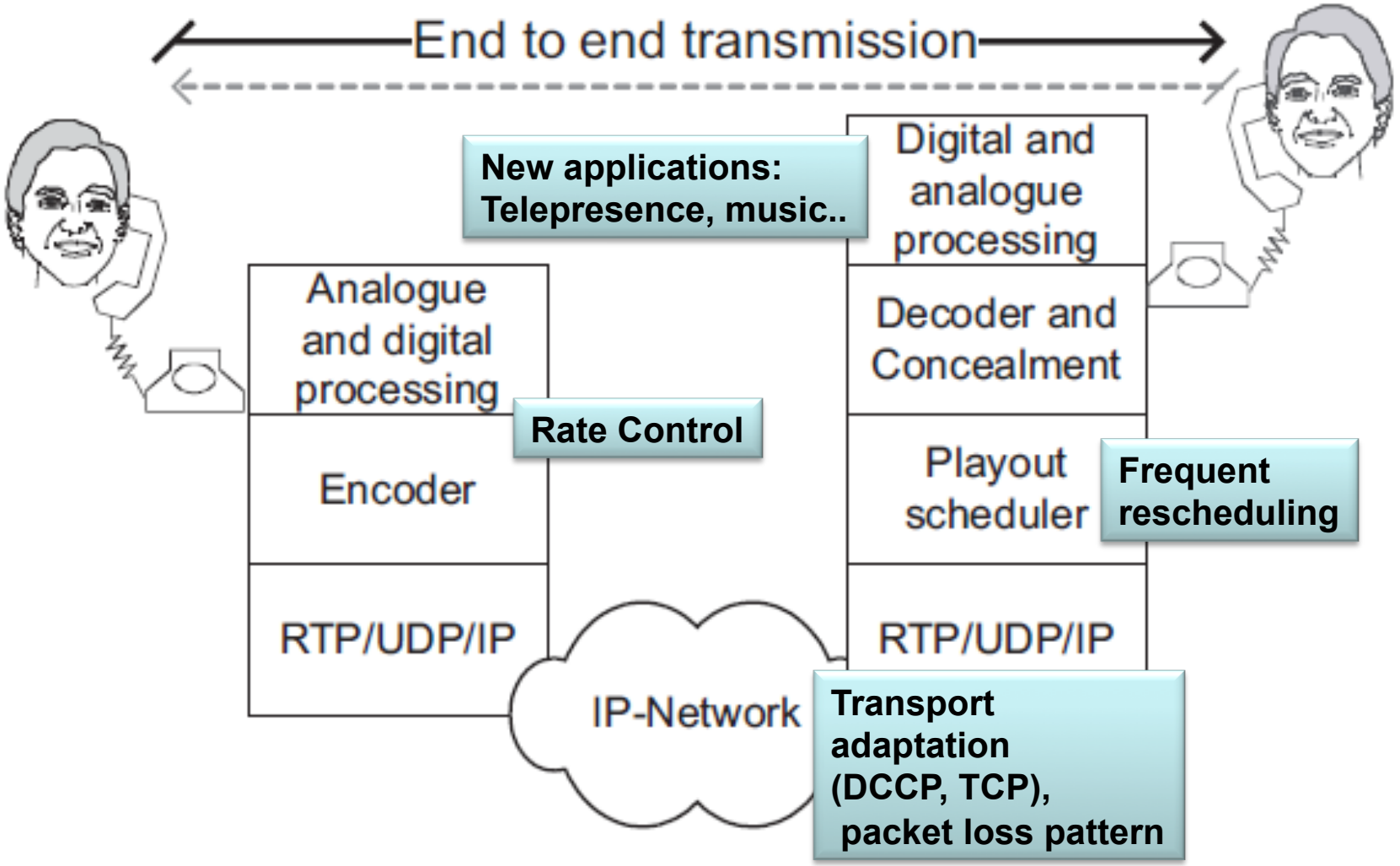
Thus:

1. The acceptability and subjective quality impression of **end-users** have to be measured.
2. The IIAC codec has to be tested as part of entire **telecommunication systems**. It is not sufficient to just the codec's performance in a stand-alone setup.
3. The circumstances of particular **communication scenarios** have to be considered and controlled because they might have impact of the human rating behavior.

Telecommunication System: IP Phone



Modern IP Phone



Codec Testing in Realistic Environments

- Applications
  - New applications will have different requirements
  - Use cases require different qualities
- Acoustic processing
  - Presence of echo cancelation, automatic gain control
- Playout Buffering
  - Fixed or adaptive? Stretching, shrinking?
- IP Transmission
  - Impact packet transmission
  - Loss patterns
  - Delay distribution
  - Interaction between rate control and network/other flows?

Available/Unavailable Test Methods

- Applications
  - Speech conversation (different degrees of interactivity)
  - Audio listening test
  - But: No test methods for music playing or telepresence…
- Acoustic processing
  - Typical ignored
  - Reference room / headset / headphones standardized
- Playout Buffering
  - P.OLQA can measure playout time adjustments
  - But: No agreed standard playout buffering algorithm
- IP Transmission
  - ITU-T G.1050/TIA-921 simulates loss and delays
    - modifies packet traces (PCAP) to consider delay and loss
  - But: No interaction with rate-control
  - But: No simulation of DCCP

What to do?

- We need to perform subjective testing
- BUT exhaustive formal subjective tests are not possible from either a cost or „time-to-market“ perspective...
- We need an iterative and continous test methodology based on shared testing responsibilities and broad user feedback

Recommendation

Continuous testing workflow:

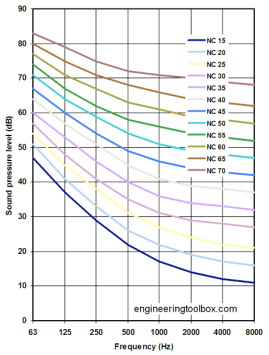
- Phase: Development
  - Iterative design decisions based on expert opinion and informal subjective listening tests (MUSHRA)
- Phase: Characterization *(using reference implementation at 3-5 volunteer ‚test houses')*
  - Use one method for all listening-only tests, e.g. MUSHRA
  - Latency measure [ms] to cover conversations impact re G.114
  - Conduct professional tests on a few codec operational points (e.g. complexity estimation, tone passthrough)
  - ***Important to note that in testing we are not mandating specific performance for acceptance, but as a benchmarking tool to guide consensus, or re-iteration as the WG deems necessary***
  - ***Also encourage ‚alpha' implementation for in-situ network testing***
- Phase: Implementation and Conformance
  - Use objective tools (PESQ, PEAQ, P.OLQA) for bug finding and conformance tests (after mapping to MUSHRA values)

Codec ‘Test Houses’

- Asking for 3-5 volunteer companies to become codec ‘test houses’
- Agree to provide recommended testing signal chain and audio environment
  - Expected 5 to 10K budget
- Agree on audio test material (speech, music)
- Agree to sign up subjective test volunteers and perform codec tests at designated testing periods and provide results to the codec WG in a timely manner
- Work in a committee fashion to generate a collaborative test report that identifies test discrepancies and an overall composite result

Recommended test chain

- Quiet listening environment at NC25 (approx 35 dBA) – e.g. ISOBOOTH
- Standardized sample preparation
  - 8, 16, 24, 32 etc to 48 kHz / 16 bit
  - SecretRabbitCode
- MUSHRA assessment tool
  - RateIt
  - MUSHRAM (Matlab based)
- High quality D/A
  - e.g. Benchmark DAC, Metric Halo ULN-2, Apogee MiniDAC
- High quality headphone amp and playback level calibration
  - Decent headphone amp frequently included with good D/A
  - Playback levels measured via Etymotic in-ear mic
- High quality headphone (e.g. AKG 240DF, Senn HD600)



Metric Halo ULN-2

Sennheiser HD600

Potential future IETF liaison activities

- Cooperation with TIA to developed a realistic, real-time IP packet/loss simulation/emulator
  - Especially, the interactivity (between IP simulator and rate control) is still a missing feature
  - Might be easy added in the next version of TIA-921 aka ITU-G. 1050
- Define reference playout buffer
  - Used for tests with IP traces, simulation
  - Bound in respect of lower quality
- ITU-T Study Group 16 has started to defined playout scheduler for their codecs

Potential future IETF liaison activities (cont)

- Ask (members of) study group 12 for help to evaluate perceptual quality
  - Supporting time varying quality
  - Supporting playout rescheduling
  - Supporting speech and audio
- Use in-situ tests as early as possible.
  - To find bugs
  - To get quality feedback
  - To test codec under realistic conditions
- However, cannot be applied for formal qualitfication or conformance testings
- Ask Study Group 12 for help on formal in-situ testing…

Summary

- Comprehensive testing of codec is challenging
  - Potential new requirements
  - Need for realistic operational settings
- Streamlined codec development and qualification workflow
  - Test the „running systems“ with real users and experts
  - Qualify via multi-site MUSHRA, latency, and complexity estimates at 3-5 volunteer companies, using a reference implementation. Also, in-situ implementation testing desired.
  - *results used to assist consensus or reiteration, not as a process gating mechanism*
- Future cooperation with TIA and ITU-T
  - To develop formal testing and listening procedures
  - Long term relationship and knowledge sharing (e.g. network impairments)