Towards Consistent Assessment of Audio Quality of Systems with Different Available Bandwidth

Yu Jiao  Sławek Zieliński  Francis Rumsey

Institute of Sound Recording, University of Surrey
Outline

- New challenges in quality assessment
- Attributes to assess
- Standard scale
- Summary
- Demo I: an example of attributes identification and assessment
- Demo II: assessing envelopment, an example of direct anchoring
Where should I go?
New Challenges in Speech and Audio Quality Assessment

Is there a need for the development of a new, more universal standard for audio quality assessment, regardless of application or bandwidth?
Towards a Universal Method

Two challenges:

- Which perceptual attributes to use?
- How to calibrate the scale so that the results obtained from assessment of audio quality in different applications can be sensibly compared?
Outline

- New challenges in quality assessment
- Attributes to assess
- Standard scale
- Summary
- Demo I: an example of attributes identification and assessment
- Demo II: assessing envelopment, an example of direct anchoring
Which Attributes to Use?

Most commonly used attributes:

- Speech Quality (ACR, DCR, CCR) – ITU-T Recommendations
- Basic Audio Quality (Continuous scale), Stereophonic Image Quality, Front Image Quality, Impression of Surround Quality – ITU-R Recommendations
- What about other attributes? (see next slide)
Which Attributes to Use?

Preferences

- Preference
- Basic Audio Quality
  - Timbral Quality
  - Spatial Quality
  - Front Image Accuracy
  - Spatial Accuracy
  - Localisation error
- Judgements
  - Timbral Accuracy
  - Loudness
  - Dynamics
  - Envelopment
  - Distortions
  - Noise presence
  - Width change
  - Distance change

This set is not exhaustive.
Which Attributes to Use?

- It is relatively easy to agree upon and standardise high level attributes
- It is more difficult to standardise low-level attributes
- Usefulness of low-level attributes is application specific

However, a pool of standardised attributes, together with associated anchors and scale system, may be of help.
Example of Spatial Attributes Pool – Rumsey(2002)

One of the most systematic attribute pools for spatial audio assessment
Outline

- New challenges in quality assessment
- Attributes to assess
- Standard scale
- Summary
- Demo I: an example of attributes identification and assessment
- Demo II: assessing envelopment, an example of direct anchoring
Do we need a standard audio quality scale?

- It may help to reduce bias in listening tests
- Essential for calibration of objective models
- It may help to compare results across different applications with different available bandwidth

Poulton (1989)
Zielinski, Rumsey and Bech (2008)
Range Equalising Bias

Hypothetical Distribution – Stimulus Set A

Response Distribution (Scores)

Hypothetical Distribution – Stimulus Set B

Perceptually Linear Scale

Assessment Scale

Perceptually Linear Scale

e.g. Wide Band Stimuli

Lannion, France, 10-12 September 2008

International Telecommunication Union

e.g. Narrow Band Stimuli
Range Equalising Bias = “Rubber Ruler Effect”

e.g. Narrow Band Stimuli

e.g. Wide Band Stimuli

e.g. Full Band Stimuli

- Means and 95% CIs
- Systematic upward shift
- Max diff ≈ 20
- Absolute scores?

Conclusion: Do not put confidence in the labels

- Means and 95% CIs
- Systematic upward shift
- Max diff = 13
- Absolute scores?
Range Equalising Bias – Cheer (2008)

Test based on ITU-T P.800

Questions:

- Is “Absolute Category Rating” (ACR) really absolute?
- Do we need a better calibrated scale?
Towards Consistent Assessment across Narrow-, Wide- and Full-Band Applications

Use a standard scale:

If possible, use physical units that are familiar to the listeners (Poulton, 1989)

• For example, width of frontal image can be assessed as an angle expressed in degrees

• Distance between listener and apparent source could be assessed in metres

• Use an open-ended ratio scale – biased (Narens, 1996)

• Use verbal anchors along the scale (ineffective – see the previous slide)

• Use auditory anchors (effective but difficult to implement)
### Three Types of Auditory Anchors:

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Direct Anchors</strong></td>
<td>Listeners are instructed how to use the scale relative to two or more auditory anchors.</td>
</tr>
<tr>
<td></td>
<td>• Help to define a “frame of reference”</td>
</tr>
<tr>
<td></td>
<td>• Examples of partial use: ITU-R BS.1116 and ITU-R BS.1534 MUSHRA</td>
</tr>
<tr>
<td><strong>2. Indirect Anchors</strong></td>
<td>Anchors are included in the set of stimuli under assessment. Listeners are not instructed how to assess them. They are unaware of their purpose.</td>
</tr>
<tr>
<td></td>
<td>• Effective bias diagnostic tool</td>
</tr>
<tr>
<td></td>
<td>• May help to define a “frame of reference” if used properly</td>
</tr>
<tr>
<td></td>
<td>• Examples of use: ITU-T P.800 (MNRU reference quality impairments). Also 3.5kHz anchor in ITU-R BS.1534 MUSHRA</td>
</tr>
<tr>
<td><strong>3. Background Anchors</strong></td>
<td>Anchors are only presented during familiarisation phase prior to a listening test but they are not included in the proper listening test. Listeners are not instructed as to how these anchors relate to the scale.</td>
</tr>
<tr>
<td></td>
<td>• They do not calibrate the scale but “calibrate” listeners</td>
</tr>
<tr>
<td></td>
<td>• Used very rarely</td>
</tr>
</tbody>
</table>


Example of a Scale with Direct Anchors (A & B)
– Conetta (2008)

Only two anchors in this example but more than 2 can also be used.

Lannion, France, 10-12 September 2008
Challenges of Direct Anchoring

- How to choose them in terms of levels of quality?
- How to make them similar to stimuli under assessment in terms of perceptual properties?

Hypothetical Standard Quality Scale

Anchor 1: Full bandwidth full surround sound quality?
Anchor 2: CD quality?
Anchor 3: FM radio quality?
Anchor 4: Narrow-bandwidth telephone quality?
Example of Diagnostics with Indirect Anchors

- Indirect anchors – useful diagnostic tool to check for bias
- Scores “float” along the scale
- Do not put confidence in labels
Outline

- New challenges in quality assessment
- Attributes to assess
- Standard scale
- Summary
- Demo I: an example of attributes identification and assessment
- Demo II: assessing envelopment, an example of direct anchoring
Summary

• A need for the development of a new, more universal standard for audio quality assessment, regardless of application or bandwidth.
• More attributes are needed to reveal the nature of quality degradation.
• Comparing audio quality across different applications, e.g. with different audio bandwidth, is problematic due to potentially inconsistent use of a scale (ill-defined frame of reference).
• Standard scale needed.
• Direct anchoring technique could be used but difficult to identify suitable auditory anchors.
Demo I: An example of attributes identification & assessment - Jiao et al. (2007)

- New Developed Codec
  - Attribute Identification and Selection
    - Basic Audio Quality
      - Timbral Distortion
      - Spatial Distortion
        - Dynamic Spatial Distortion (DSD)
          - Level of DSD
          - Dynamicity of DSD

- Scales & test method design
\[ BAQ = -0.668 \times TD - 0.350 \times LDSD - 0.179 \times DDSD + 86.45 \]
Demo II: Assessing envelopment, an example of direct anchoring - George et al. (2008)

How enveloping are these recordings?

R1 R2 R3 R4 R5

Interface allowing assessment of envelopment arising from surround recordings.
References


Conetta, R. (2007) Scaling and predicting spatial attributes of reproduced sound using an artificial listener, MPhil/PhD Upgrade Report, University of Surrey, Institute of Sound Recording.


Acknowledgment

Some of this work was carried out with the financial support of the Engineering and Physical Sciences Research Council, UK (Grant EP/C527100/1).