**3GPP TSG-SA WG4 Meeting #132S4-251160**

**Japan, Fukuoka, 19 – 23 May 2025 Revision of S4-250850**

**Source: Dolby Laboratories Inc., Novamint**

**Title: [FS\_ULBC] On test methodologies for ULBC performance evaluation**

**Agenda item: 7.9**

**Document for: DISCUSSION and AGREEMENT**

# Background

3GPP SA4 has a long history of standardizing speech codecs that have been successfully deployed in 3GPP networks. Besides the most recently standardized IVAS codec, these are the AMR, AMR-WB and EVS codecs. SA4 adopted generally recognized ITU-T test methodologies for the quality evaluation of these codecs for the purpose of codec selection and characterization.

Similarly, one objective of the new FS\_ULBC study item is to identify suitable test methodologies for ultra-low bit rate speech codecs. The fact that the target bit rate is envisioned to be below 3 kbps along with recent technology advances in the area of deep-learning based speech coding means that test methodologies that were used for AMR, AMR-WB or EVS codec evaluations need to be revisited.

This contribution discusses the challenges of ultra-low bit rate codec quality evaluations in the light of potential properties of such codecs and proposes applicable test methodologies.

# Discussion

## Typical quality impairments of ultra-low bit rate speech coding

Speech codecs operating at ultra-low bit rates may impact aspects of speech communication in at least the following categories

* Loss of listening-only audio quality
* Audio bandwidth loss
* Impaired intelligibility
* Impaired speaker identifiability
* Prosodic impairments
* Hallucination, i.e. word and phone confusions

A related quality impairment category is sensitivity to non-speech input. Non-speech input in this context may mean any non-clean speech input such as background noise, music, but also noisy speech, interfering talker speech, reverberant speech.

Additionally, speech codecs operating at ultra-low bit rates may imply speech enhancement algorithms, such as noise suppression, gain normalization etc.

## Challenges of quality assessment of ultra-low bit rate speech codecs

Testing of ULBC introduces new challenges in comparison to signal processing-based codecs that may necessitate additional or alternative evaluation methods compared to previous 3GPP SA4 speech coding development activities.

Conversation-opinion tests are intended to reproduce, in a laboratory situation, the actual service conditions experienced by telephone customers. [ITU-T P.800 Section 6]. For practical reasons, and due to the need of strict test control, listening-opinion tests are often employed as an alternative in speech coding development.

According to ITU-T Rec. P.800, the recommended test method for listening-only tests is the "Absolute Category Rating" (ACR) method described in Annex B of ITU-T P.800 (see ITU-T P.800 Clause 6.2). An alternative to the Absolute Category Rating method is the Degradation Category Rating (DCR) method which is described in detail in Annex D of ITU-T P.800. As this is a comparative method it is suitable when the impairment (especially digital impairments) is small.

Based on the principles of ITU-T Rec. P.800, 3GPP established a practice of listening-only evaluations in the context of the standardization of its voice codecs. AMR, AMR-WB and EVS codecs were all evaluated using P.800 ACR and (modified) DCR test methodologies, the latter also applied for tests with larger impairments such as noisy speech over poor channel conditions. ACR was generally used for clean speech tests while P.800 DCR was used for SWB clean speech, mixed-bandwidth, speech + background noise and mixed/music quality evaluations. Other aspects relevant to speech communication, such as impaired intelligibility, impaired speaker identifiability and prosodic impairments were not the focus of testing in AMR, AMR-WB and EVS codec standardization.

For ULBC, these other aspects may need to be addressed more directly, through dedicated tests. Hallucination, for example, is a category that plays only a role in ML-based coding systems but not for signal-processing based codecs, which AMR, AMR-WB and EVS are.

While P.800 recommends ACR as the listening-only test method, it may not be the optimal test to quantify all potential impairments in speech communication such as hallucination, impaired intelligibility and prosodic impairments. These impairments are possible in ML-based coding systems and could be covered by alternative tests such as automatic speech recognition methods, modified rhyme tests and DCR tests (see below).

In contrast, DCR methodology generally focuses on differences to a reference. If such differences are small and pertain to prosodic differences, these may not directly impact the conversational capability of a communication system but can be relevant for other aspects such as identity recognition.

Besides the usage of ACR and DCR, further subjective or objective test methodologies are available that are specifically designed for the evaluation the above-listed impairment categories of ultra-low-rate speech codecs. A non-exhaustive list of such methods is:

* Diagnostic Rhyme Tests (DRT)
* Modified Rhyme Tests (MRT)
* MOS testing for speaker similarity
* Speaker verification/identification tests
* Prosodic naturalness MOS tests
* Intonation recognition tests
* Transcription tests involving testing for word and semantic equivalence
* Phoneme recognition tests
* Automatic speech recognition tests

Speech enhancing tool, which may be part of ultra-low bitrate-codecs, are typically evaluated by multi-dimensional rating scales provided by P.835. Here, the quality of speech and the noise suppression capability are evaluated separately.

# Proposal

We propose to document the consideration of test methodologies for ultra-low bit rate codecs provided above in a pCR to 3GPP TR 26.940 as shown below.

# pCR to 26.940

\* \* \* First Change \* \* \* \*

# 9 Test methodologies

Editor’s Note:

5b. Identify appropriate test methodologies, regarding speech quality, intelligibility, conversational quality, in particular taking into account:

- Clean speech and noisy speech

- Tandeming with existing IMS voice codecs

- Clean channel and GEO channel conditions

### General

### 9.1.1 Typical quality impairments of ultra-low bit rate speech coding

Speech codecs operating at ultra-low bit rates may impact aspects of speech communication in at least the following categories

* Loss of listening-only audio quality
* Audio bandwidth loss
* Impaired intelligibility
* Impaired speaker identifiability
* Prosodic impairments
* Hallucination, i.e. word and phone confusions

A related quality impairment category is sensitivity to non-speech input. Non-speech input in this context may mean any non-clean speech input such as background noise, music, but also noisy speech, interfering talker speech, reverberant speech.

Additionally, speech codecs operating at ultra-low bit rates may imply speech enhancement algorithms, such as noise suppression, gain normalization etc.

Editor’s note: Investigate impairment relevance in context of use case, e.g. emergency call.

### 9.1.2 Challenges of quality assessment of ultra-low bit rate speech codecs

Testing of ULBC introduces new challenges in comparison to signal processing-based codecs that may necessitate additional or alternative evaluation methods compared to previous 3GPP SA4 speech coding development activities.

Conversation-opinion tests are intended to reproduce, in a laboratory situation, the actual service conditions experienced by telephone customers. [ITU-T P.800 Section 6]. For practical reasons, and due to the need of strict test control, listening-opinion tests are often employed as an alternative in speech coding development.

According to ITU-T Rec. P.800, the recommended test method for listening-only tests is the "Absolute Category Rating" (ACR) method described in Annex B of ITU-T P.800 (see ITU-T P.800 Clause 6.2). An alternative to the Absolute Category Rating method is the Degradation Category Rating (DCR) method which is described in detail in Annex D of ITU-T P.800. As this is a comparative method it is suitable when the impairment (especially digital impairments) is small.

Based on the principles of ITU-T Rec. P.800, 3GPP established a practice of listening-only evaluations in the context of the standardization of its voice codecs. AMR, AMR-WB and EVS codecs were all evaluated using P.800 ACR and (modified) DCR test methodologies, the latter also applied for tests with larger impairments such as noisy speech over poor channel conditions. ACR was generally used for clean speech tests while P.800 DCR was used for SWB clean speech, mixed-bandwidth, speech + background noise and mixed/music quality evaluations. Other aspects relevant to speech communication, such as impaired intelligibility, impaired speaker identifiability and prosodic impairments were not the focus of testing in AMR, AMR-WB and EVS codec standardization.

For ULBC, these other aspects may need to be addressed more directly, through dedicated tests. Hallucination, for example, is a category that plays only a role in ML-based coding systems but not for signal-processing based codecs, which AMR, AMR-WB and EVS are.

While P.800 recommends ACR as the listening-only test method, it may not be the optimal test to quantify all potential impairments in speech communication such as hallucination, impaired intelligibility and prosodic impairments. These impairments are possible in ML-based coding systems and could be covered by alternative tests such as automatic speech recognition methods, modified rhyme tests and DCR tests (see below).

In contrast, DCR methodology generally focuses on differences to a reference. If such differences are small and pertain to prosodic differences, these may not directly impact the conversational capability of a communication system but can be relevant for other aspects such as identity recognition.

Besides the usage of ACR and DCR, further subjective or objective test methodologies are available that are specifically designed for the evaluation the above-listed impairment categories of ultra-low-rate speech codecs. A non-exhaustive list of such methods is:

* Diagnostic Rhyme Tests (DRT)
* Modified Rhyme Tests (MRT)
* MOS testing for speaker similarity
* Speaker verification/identification tests
* Prosodic naturalness MOS tests
* Intonation recognition tests
* Transcription tests involving testing for word and semantic equivalence
* Phoneme recognition tests
* Automatic speech recognition tests

Speech enhancing tool, which may be part of ultra-low bitrate-codecs, are typically evaluated by multi-dimensional rating scales provided by P.835. Here, the quality of speech and the noise suppression capability are evaluated separately.

Editor’s note: Other options for listening methods are invited for documentation.

\* \* \* End of Changes \* \* \* \*