**3GPP TSG-SA WG4 Meeting #132S4-250xxx**

**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 typically impact speech quality in at least the following categories

* Loss of general 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.

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

AMR, AMR-WB and EVS codecs were evaluated using P.800 ACR and (modified) DCR test methodologies. ACR was generally used for clean speech tests while modified DCR was used for speech + background noise quality evaluations. Other categories than general audio quality were not tested since it could be assumed that impairment categories such as impaired intelligibility, impaired speaker identifiability and prosodic impairments played a minor role. Hallucination 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.

Given the quality impairment categories of ultra-low bit rate speech coding other than loss of general audio quality and audio bandwidth loss, ACR may be problematic as it does not provide opinion scores associated with the deviation from the uncoded (direct) reference. This may mean that degradation impairments in the other categories remain undetected in such tests.

In contrast, DCR methodology covers degradation impairments. A problematic property of DCR methodology is however that any deviation from the uncoded (direct) reference is rated as a degradation. This may be undesirable if the coding system incorporates a speech enhancer, which may enhance non-clean speech input. CCR test methodology may address this problem.

Besides the usage of ACR, DCR and CCR, 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

Notably, these methods may require substantial effort and SA4 has no own experience with them. It may thus be difficult to prepare corresponding tests in the short timeframe of the expected ULBC standardization timeframe.

It is also expected that subjective test methods that rate the degradation of the coded speech signal in comparison with an uncoded (direct) reference would give at least indicative answers to impairments in the relevant quality degradation categories. One expected downside is that degradation category ratings are an overall quality metric, and answers could not be expected to which of the degradation dimensions a degradation score could specifically be attributed.

# 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 typically impact speech quality in at least the following categories

* Loss of general 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.

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

AMR, AMR-WB and EVS codecs were evaluated using P.800 ACR and (modified) DCR test methodologies. ACR was generally used for clean speech tests while modified DCR was used for speech + background noise quality evaluations. Other categories than general audio quality were not tested since it could be assumed that impairment categories such as impaired intelligibility, impaired speaker identifiability and prosodic impairments played a minor role. Hallucination 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.

Given the quality impairment categories of ultra-low bit rate speech coding other than loss of general audio quality and audio bandwidth loss, ACR may be problematic as it does not provide opinion scores associated with the deviation from the uncoded (direct) reference. This may mean that degradation impairments in the other categories remain undetected in such tests.

In contrast, DCR methodology covers degradation impairments. A problematic property of DCR methodology is however that any deviation from the uncoded (direct) reference is rated as a degradation. This may be undesirable if the coding system incorporates a speech enhancer, which may enhance non-clean speech input. CCR test methodology may address this problem.

Besides the usage of ACR, DCR and CCR, 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

Notably, these methods may require substantial effort and SA4 has no own experience with them. It may thus be difficult to prepare corresponding tests in the short timeframe of the expected ULBC standardization timeframe.

It is also expected that subjective test methods that rate the degradation of the coded speech signal in comparison with an uncoded (direct) reference would give at least indicative answers to impairments in the relevant quality degradation categories. One expected downside is that degradation category ratings are an overall quality metric, and answers could not be expected to which of the degradation dimensions a degradation score could specifically be attributed.

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