3GPP TSG-SA WG4 Meeting #131-bis-eS4-250565

Online, 11 – 17 April 2025

**Source: Fraunhofer IIS, Vivo, Qualcomm Incorporated, China Mobile**

**Title: Analysis of existing technologies for ULBC**

**Spec: 3GPP TR 26.940**

**Agenda item: 7.9**

**Document for: Agreement**

**1. Introduction**

At the recent SA meeting, the “Study on Ultra Low Bitrate Speech Codec” has been approved. According to the WID description [1], the study includes the following two objectives:

* Provide some evidence that the design criteria can be met, for example existing reference codecs
* Identify relevant reference codecs for comparison and evaluation purposes.

Furthermore, the WID references TR 22.887, where a total transmission data rate of [1-3] kbit/s is assumed. The present document aims to give an overview of existing technologies which may serve as reference codec or provide guidance regarding achievable design constraints and performance requirements. These codec parameters may also serve to study other objectives such as mouth-to-ear-delay.

**2. Reason for Change**

The present document provides a starting point for collecting information on existing technologies.

**4. Proposal**

It is proposed to agree the following changes to clause 7 of 3GPP TR 26.940

\* \* \* Start of Change \* \* \* \*

# 7 Existing technologies and feasibility evidence

## 7.1 Existing codec technologies

### 7.1.1 Overview

The present clause collects information on existing codec technologies and categorizes those using the following definitions:

**- 3GPP IMS codecs**: Even though the operating bitrates are out of scope for the envisioned services, these codecs can be considered as reference condition regarding the performance requirements.

**- Conventional Ultra Low Bitrate Codecs**: These codecs are based on conventional signal and speech processing algorithms (like the 3GPP IMS codecs) and can operate in the envisioned bitrate range.

**- AI-based postprocessor**: The bitstream or decoder of a conventional ultra low bitrate codec is decoded or enhanced by a AI based postprocessor

**- AI-based encoder and decoder**: Encoder and decoder implemented as Deep Neural Network (DNN) and can (almost) operate in the envisioned bitrate range. Due to fundamental design differences regarding the algorithmic delay, two more sub-categories are defined:

**- Causal systems**: Codecs which can operate in real-time, block based streaming applications

**- Non-causal systems:** Codecs which can only operate in non real-time applications due to large processing look-ahead or framing

The following codec properties are considered:

**- Audio bandwidth**: Support of NB, WB, SWB, FB and/or 12 kHz

**- Algorithmic delay**: Total codec delay, excluding framing

**- Frame duration**: Block size or framing delay, may be bit rate depended

**- Bitrates**: List of supported bitrates lower than 3 kbps or the lowest supported constant bitrate. Variable bitrate modes are not considered.

**- Specification / Access / Software**:

- Specification determines the level of development for a complete solution

- A: Fully specified and verified coding algorithm including all relevant system components such as channel resilience capability or discontinuous transmission

- B: Specification of codec algorithm only

- Access determines the availability of a reference implementation:

- A: Repository with open access

- B: No repository with open access known

- Software determines the deployment level of the reference software its level of optimization

- A: Software fully functional and optimized C-code

- B: Software fully functional and reference C-code

- C: Experimental software framework in Python only

- X: Software status unknown

NOTE: In the case of a DNN-based codec, both the implementation of the computational graph (i.e. the layer definitions and their order of execution) and the corresponding pretrained weights are needed for inference. For this reason, consider two sets of weights for the same computational graph as defining two different codecs.

Editor’s note: Documentation regarding complexity aspects (computational complexity, memory, ROM) is envisioned, however it requires a clear definition on the measurement first.

In the following a collection of speech codecs and the related parameter is presented in Table 7.1.1-1.

Table 7.1.1-1 List of existing codec technologies

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Codec | Source | Audioband-width | Alg. Delay  [ms] | Frame duration  [ms] | Bitrates  [kbps] | Specification / Access / Software |
| 3GPP IMS codecs | | | | | | |
| AMR | 3GPP TS [26.071](https://www.3gpp.org/DynaReport/26071.htm) | NB | 5 | 20 | 4.75 | A / A / A |
| AMR-WB | 3GPP TS [26.171](https://www.3gpp.org/DynaReport/26171.htm) | WB | 5.9375 | 20 | 6.6 | A / A / A |
| EVS | 3GPP TS [26.445](https://www.3gpp.org/DynaReport/26445.htm) | NB | 12 | 20 | 7.2 | A / A / A |
| WB | 7.2 |
| SWB | 9.6 |
| Conventional Ultra Low Bitrate Codecs | | | | | | |
| MELP / MELPe  (Note1) | <https://www.compandent.com/about-melpe/> and melp\_faq | NB | 36.25 | 90 | 0.6 | A / B / A |
| 27.25 | 67.5 | 1.2 |
| 20.125 | 22.5 | 2.4 |
| AMBE-LR | https://www.dvsinc.com/software/technology.shtml#ambelr | NB |  |  | 1.6 – 1.8 | A / B / A |
| MPEG--HVXC | https://www.iso.org/obp/ui/en/#iso:std:iso-iec:14496:-3:ed-5:v1:en:sec:1.3 | NB |  |  | 2 – 4 | B / B / B  (Note2) |
| TWELP MR  (Note 1) | <https://dspini.com/vocoders/lowrate/twelp-lowrate/twelp300-3600-mr> | NB | 20 | 40 | 3.2 | A / B / A  (Note5) |
| 20 | 40 | 2.4 |
| 40 | 60 | 1.6 |
| 40 | 60 | 1.2 |
| 80 | 100 | 0.7 |
| 80 | 100 | 0.6 |
| 100 | 120 | 0.48 |
| 100 | 120 | 0.3 |
| Codec2 | https://github.com/drowe67/codec2 | NB | 40  (Note 3) | [TBD] | 0.45 | A / A / A |
| 40 | 0.7 |
| 40 | 1.2 |
| 40 | 1.3 |
| 40 | 1.4 |
| 40 | 1.6 |
| 20 | 2.4 |
| AI based decoders | | | | | | |
| WaveNet Codec2 | [Paper](https://arxiv.org/pdf/1712.01120) | WB | See Codec2 | 20 | 2.4 | B / B / X |
| CQNV  Codec2 | [Paper](https://arxiv.org/abs/2307.13295) | WB | 40 | 40-60 | 1.0  1.1 | B / B / X |
| AI based encoder and decoder (causal) | | | | | | |
| LPCnet | [Paper](https://jmvalin.ca/papers/lpcnet_codec.pdf), [Code](https://github.com/xiph/LPCNet), [Demo](https://jmvalin.ca/demo/lpcnet_codec/) | WB | 25 | 40 | 1.6 | B / A / C |
| LyraV21.3.2  (Note1) | [Paper (SoundStream)](https://arxiv.org/pdf/2107.03312), [Code](https://github.com/google/lyra), [Demo (SoundStream)](https://research.google/blog/soundstream-an-end-to-end-neural-audio-codec/) | WB | [TBD] | 20 | 3.2, 6, 9.2 | B / A / A |
| EnCodec | [Paper](https://arxiv.org/pdf/2210.13438), [Code](https://github.com/facebookresearch/encodec), [Demo](https://audiocraft.metademolab.com/encodec.html) | 24kHz | 0 | 13.3 | 1.5, 3, … | B / A / C |
| FB | 1000 | 13.3 | 6, 12, 24 |
| Mimi-Codec | [Paper(moshi)](https://kyutai.org/Moshi.pdf), [Code](https://huggingface.co/kyutai/mimi), [Demo (moshi)](https://moshi.chat/) | 24kHz | 0 | 80 | 0.55, 1.1 | B / A / C |
| TS3 | [Paper](https://arxiv.org/pdf/2411.18803), [Code](https://github.com/ga642381/speech-trident), Demo | WB | 0 | 20 | 0.64, 0.8 | B / B / X |
| TAAE | [Paper](https://arxiv.org/pdf/2411.19842v1), [Code](https://github.com/Stability-AI/stable-codec), [Demo](https://stability-ai.github.io/stable-codec-demo/) | WB | 0 | 20, 40 | 0.4, 0.7 | B / B / X |
| AI based encoder and decoder (non-causal) | | | | | | |
| DAC | [Paper](https://arxiv.org/pdf/2306.06546), [Code](https://github.com/descriptinc/descript-audio-codec/tree/main), [Demo](https://descript.notion.site/Descript-Audio-Codec-11389fce0ce2419891d6591a68f814d5) | WB | 800 | 20 | 0.5, 1.0, 1.5, … | B / A / C |
| 24kHz | 500 | 13.3 | 0.75 1.5, 3, … |
| DAC-IBM | Paper, Code, Demo | 24kHz | 500 | 13.3 | 0.75, 1.5, 3 | B / A / C |
| SNAC | [Paper](https://arxiv.org/pdf/2410.14411), [Code](https://github.com/hubertsiuzdak/snac) | 24 kHz | 1000 | 80 | 0.98 | B / A / C |
| SpeechTokenizer | [Paper](https://arxiv.org/pdf/2308.16692), [Code](https://github.com/ZhangXInFD/SpeechTokenizer), [Demo](https://0nutation.github.io/SpeechTokenizer.github.io/) | WB | full-signal | 20 | 0.5, 1.0 | B / A / C |
| SemantiCodec | [Paper](https://arxiv.org/pdf/2405.00233), [Code](https://github.com/haoheliu/SemantiCodec-inference), [Demo](https://haoheliu.github.io/SemantiCodec/) | WB | full-signal | 10 | 1.25, …, 1.4 | B / A / C |
| 20 | 0.63, …, 0.70 |
| 40 | 0.31, …, 0.35 |
| FunCodec  (Note4) | [Paper](https://arxiv.org/pdf/2309.07405), [Code](https://github.com/modelscope/FunCodec), [Demo](https://funcodec.github.io/index.html) | WB | [TBD] | 20 | 0.5, 1.0, … | B / A / C |
| 40 | 0.25, 0.5, … |
| WavTokenizer (Note4) | [Paper](https://arxiv.org/pdf/2408.16532), [Code](https://github.com/jishengpeng/WavTokenizer), [Demo](https://wavtokenizer.github.io/) | 24kHz | [TBD] | 40 | 0.25, 0.5, … | B / A / C |
| 25 | 0.9 |
| BigCodec  (Note4) | [Paper](https://arxiv.org/html/2409.05377v1), [Code](https://github.com/Aria-K-Alethia/BigCodec), [Demo](https://aria-k-alethia.github.io/bigcodec-demo/) | WB | [TBD] | 12.5 | 1.04 | B / A / C |
| FocalCodec | [Paper](https://arxiv.org/pdf/2502.04465), [Code](https://github.com/lucadellalib/focalcodec), [Demo](https://lucadellalib.github.io/focalcodec-web/) | WB | [TBD] | 20 | 0.65 | B / A / C |
| 40 | 0.33 |
| 80 | 0.16 |

Note1: These codecs may include a noise suppression module

Note2: Decoder and reference encoder are available in the MPEG reference software and for MPEG members

Note3: Encoder with 20ms overlapping FFT and iFFT at decoder with same overlap-add size

Note4: Only non-causal version publicly available

Note5: Although TWELP does not have an open reference implementation, a complete quality assessment testbench is available at <https://dspini.com/files/downloads/speechsamples/C2_vs_TWELP.zip>.

### 7.1.2 Observations regarding codec parameter

The following observations are extracted from the Table 7.1.1-1

**- Audio bandwidth:** While conventional ultra-low bitrate codecs operate on NB only, modern AI based codecs offer an audio bandwidth of WB or higher

**- Algorithmic delay:** Considering the total codec delay (algorithmic delay plus framing),

- IMS codecs operate in range between 25ms (AMR) and 32ms (EVS)

- Conventional audio codecs operate in the delay range of 60ms – 80ms (Codec2) up to 126.25ms (MELPe).

- Causal AI based codecs can operate between 20ms (TS3) and up to 80 ms (Mimi/Fun-Codec)

- Non-Causal AI based codecs require several 100ms (min. 500ms for DAC/DAC-IBM) or the entire input signal

**- Frame duration:** For conventional ultra-low bitrate codecs, the frame duration is increased compared to the regular 20ms VoIP framing. A similar tendency can be observed for AI based codecs (MimiCodec), even though some designs maintain the frame duration parameter but increase other parameter values such as bitrate (LyraV2).

**- Bitrate:** All listed codecs (except the IMS codecs and LyraV2) offer at least one bitrate mode lower than 3kbps.

### 7.1.3 Complexity Considerations:

The procedure on measuring complexity of AI based codecs and how to compare those to conventional codecs is an independent objective of the study item. Therefore, the source just wants to share some observations regarding computational complexity, RAM and ROM demand:

**- Complexity**: While the complexity of the IMS codecs and conventional ultra low bitrate codecs can considered as low, almost all AI based codecs may not run on a mobile processor in real-time. One exception is LyraV2, which only requires 35% of an ARM A53 core according to internal measurements on RaspberryPi 3+.

**- RAM**: RAM numbers may highly depend on the implementation esp. for AI based solutions. However, it should be noted that AI based codecs in available implementation show a significant higher RAM demand compared to conventional codecs, e.g. LyraV2 show ca. 54 Mbyte while EVS was characterized with 294 kByte.

**- ROM** It is expected that ROM number for AI based solutions are significantly higher compared to legacy codecs, e.g. EVS was characterized with ca. 2MB while for TAAE 950M parameters are reported. In general, the number of parameters for AI based solution correlates to the ROM demand. Assuming 8bit integer values for the parameters, this corresponds to ca. 900 Mbyte. Other solution like SNAC operate with 19M parameters which correspond to 18 Mbyte. Implementations may also use 32bit floating point parameters which leads to an ROM increase by a factor four.

### 7.1.4 Performance Evaluation

In order to get a first impression on the performance of the codecs listed in Table 1, an ITU-T P.808 ACR listening test has been conducted in English clean speech (4 talker x 6 samples). The direct signal was 32kHz sampled with SWB, normalized to -26 dBoV, to accommodate the SWB conditions. The following conditions were included:

- Conventional codecs:

- Codec2 at 0.7, 1.2, 2.4 kbps

- AMR at 4.75 kbps

- AMR-WB at 6.65 kbps

- EVS-WB at 7.2 kbps

- EVS-SWB at 9.6 kbps

- AI based codecs:

- LPCNET at 1.6 kbps

- LyraV2 at 3.2 kbps

- Mimi 0.55, 1.1 kbps

- SemantiCodec 0.34, 0.68, 1.35 kbps

- DAC\_ibm 1.5 kbps

- SNAC 0.98 kbps

- Excluded codecs

- EnCodec was not considered due the poor P.800 scoring in [2]

- DAC was excluded in favour of the better optimized DAC-ibm for this bitrate range

- FunCodec will be added in a potential follow up test.

Figure 7.1.4-1 shows the Mean Opinion Scores and 95% confidence intervals of 24 subjects.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 7.1.4-1 P.808 Mean Opinnion Scores and 95% confidence intervals

The following observations can be extracted from the results in Figure 7.1.4-1:

- Codec2 at any rate performs significantly worse than AMR at 4.75 kbps

- SemantiCodec, LyraV2, LPCnet and Mimi-Codec at 0.55 kbps score comparable to AMR-WB at 6.65 kbps

- Three conditions show a promising performance on par or slightly better than EVS at 9.6 kbps and therefore, might reach a level quality considerable for voice services. At least one is considered as causal codec

- Mimi-Codec at 1.1 kbps

and two more as non-causal codecs

- DAC-ibm at 1.5 kbps

- SNAC at 0.98 kbps

- Comparing conventional ultra-low bitrate codecs to AI based solutions, a significant quality gain of 2 MOS or more can be observed

### 7.1.5 Conclusion on existing technology

The following conclusions on existing technologies can be drawn based on this sub-clause 7.1:

- While conventional ultra-low bitrate codecs do not provide a sufficient quality level, the listening test in Figure 7.1.4-1 has shown that AI-based solutions can deliver voice quality comparable to IMS codecs.

- [At least one causal solution operates within the expected bitrate range (Mimi-Codec at 1.1 kbps) and functions with a frame duration of 80 ms and provides speech quality comparable to EVS 9.6 kbps.]

- As can also be observed with other codecs, increasing the frame duration to larger blocks than the standard 20 ms VoIP size seems to be beneficial for ultra-low bitrate coding.

- [Most existing AI codecs have high complexity. whether the codecs fit into complexity constraints of current chipset is FFS.]

- none of the presented solutions can be considered as a complete candidate for ULBC.

- Besides an optimal match on the basic parameters such as frame length or bitrate, the ULBC candidate needs to be implementable on mobile devices. Furthermore, the candidate needs to support required system aspects such as channel resilience capability or discontinuous transmission, which is currently not supported by any AI-based solution.

\* \* \* End of Change \* \* \* \*

# References

[1] Tdoc [SP-250378](https://www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_107_Incheon_2025-03/Docs/SP-250378.zip), “Study on Ultra Low Bitrate Speech Codec”

[2] T. Muller et al., “Speech quality evaluation of neural audio codecs”, Interspeech 2024, https://www.isca-archive.org/interspeech\_2024/muller24c\_interspeech.pdfers, this corresponds to ca. 900 Mbyte. Other solution like SNAC operate with 19M parameters which correspond to 18 Mbyte. Implementations may also use 32bit floating point parameters which leads to an ROM increase by a factor four.