voipfuture

VoIP Monitoring: The Codec Challenge

EVS forces VoLTE operators to act

2020

Introduction

Voice-over-IP (VoIP) services may encode audio signals in different ways. A recent study by Voipfuture found that around a dozen codecs are in active use for live VoLTE, VoWifi, international wholesale, and domestic VoIP services. The currently dominant codecs are G.711 for fixed voice services and AMR/AMR-WB for mobile services. However, the codec landscape is changing and, in both domains, new modern codecs are becoming more common.

Many new codecs offer multiple bitrates, which allow to adapt voice transmission to the available channel conditions and bandwidths. Such codec mode changes can occur on a packet-by-packet basis without any indication in the SIP signaling. VoIP monitoring needs to be aware of such mode changes, mainly because the user experience strongly depends on the bitrate.

This whitepaper provides an overview of codecs for VoIP communication services and their characteristics. Based on this, the paper discusses the impact of new wideband multi-mode codecs on service monitoring and the resulting requirements on VoIP monitoring systems.

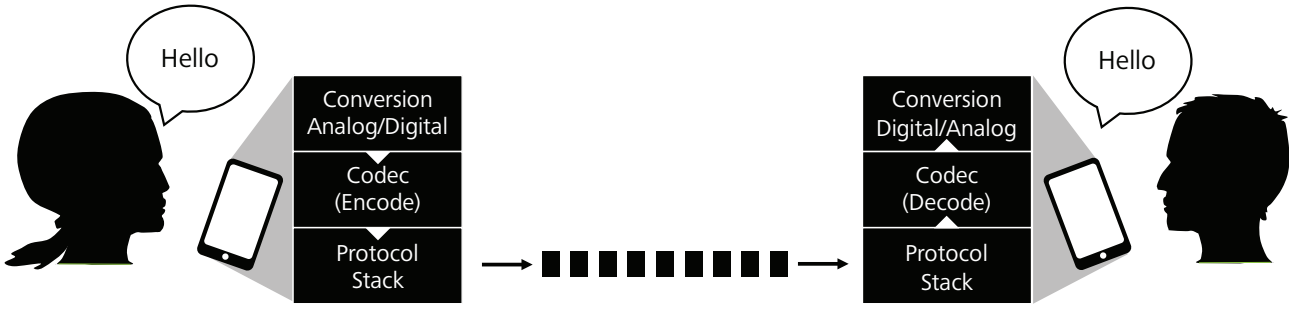
Audio Codecs Primer

A codec is software or hardware that implements an algorithm for encoding and decoding a signal. The coder function encodes signals for transmission or storage, while the decoder function recreates the signal for playback or editing.

In principle, this definition applies to any form of signal, but the word codec is most frequently used in the context of audio and video applications.

More specifically, an audio codec compresses and decompresses digital audio data according to a given coding format.

The purpose of an audio coder is to convert an audio signal into a representation with fewer number of bits while retaining quality. This effectively reduces the bandwidth required for transmission of audio data.



Codecs are often optimized to encode human speech as opposed to general audio signals. Human speech and human hearing have well-known characteristics.

For example, the human ear is capable of hearing frequencies in the range of 20 Hz to 20,000 Hz while human voice mainly – not exclusively – uses the frequency range from 300 Hz to 3,400 Hz. In telephony, the latter frequency range is referred to as narrowband.

By exploiting the properties of human voice and hearing, narrowband codecs can be used to efficiently transmit speech signals. This was long state-of-the-art for typical phone conversations but obviously has some limitations. For example, inbound call centers often play music to customers waiting to be served. Anyone who has had to listen to distorted music for longer periods of time knows that this can be very annoying.

Wideband codecs greatly extend the frequency range from 150 Hz to 7000 Hz for more natural voice transmission. Beyond this, super-wideband (or ultra-wideband) codecs cover the frequency range up to 16,000 Hz enabling the transmission of high definition audio including speech and music.

Finally, fullband codecs cover the entire frequency spectrum of human hearing and beyond. In simple terms, human hearing cannot distinguish between an original signal and its fullband encoded version; this is referred to as transparency. The voice quality delivered by codecs covering wideband and above will further be called high definition (HD) voice and respective codecs will further be summarized as HD codecs.

Codecs can be characterized in the following ways:

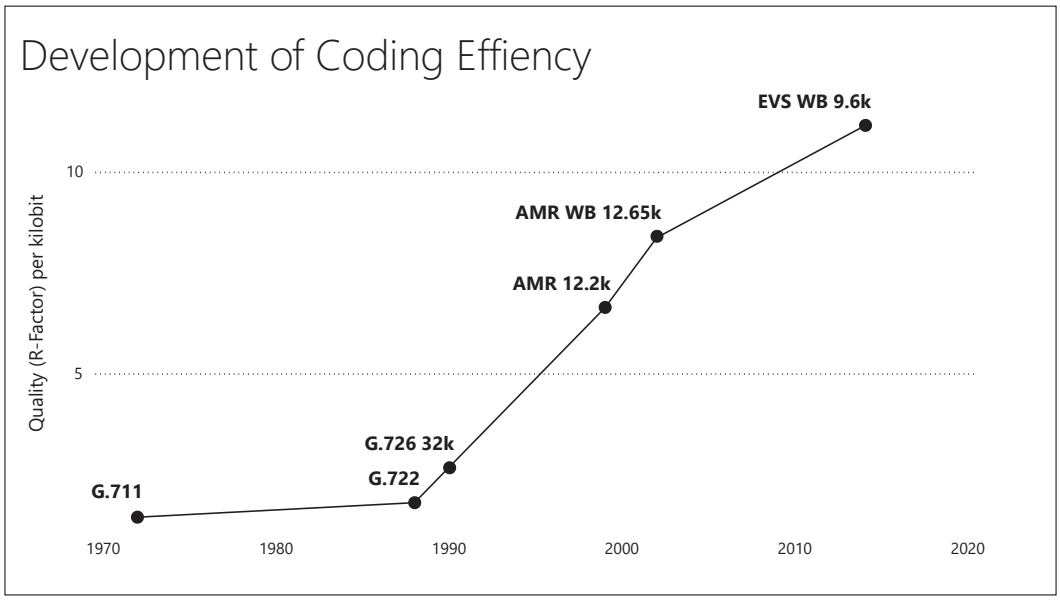
- SAMPLE RATE/BANDWIDTH: The majority of codecs used today sample the audio signal at a rate of 8,000 Hz. The reason is that the sampling frequency must be at least twice the highest component of the voice frequency (3,400 Hz) for effective reconstruction of the signal. Wideband codecs, such as G.722, have a sample rate of 16,000 Hz.
- BITS PER SAMPLE: The value range of a sample, typically 8 or 16 bit allowing for 256 respectively 65536 sample values.
- BITRATE: The nominal bitrate of a codec is determined by the size of the compressed audio signal. The bitrate is constant for most codecs although there are exceptions from this rule. Codecs typically process collections of audio samples known as frames. For example, a G.729 frame covers 10 milliseconds (ms) of speech and has a size of 10 bytes, which yields a bitrate of 8 kbit/s. The required network bandwidth for transmission is higher than a codec's bitrate, since headers for application, transport, and network protocols need to be added.
- LATENCY: Codecs also differ in the amount of algorithmic delay or latency they introduce. This paper only considers so-called conversational codecs, i.e. codecs with a latency that is sufficiently low to facilitate bi-directional conversations.
- COMPLEXITY: Codecs have different requirements in terms of the processing power needed to encode audio. Codecs with a high complexity are less suitable for battery-powered devices, such as mobile phones.
- SPECIAL FEATURES: Modern codec often include built-in functionality, such as voice activity detection, silence suppression/discontinued transmission, comfort-noise generation and the ability to automatically adapt to current network conditions. Most modern codecs implement functionality that goes beyond plain encoding and decoding of audio. This includes error concealment mechanisms to mitigate the impact of packet loss, voice activity detection, silence suppression/discontinued transmission, comfort-noise generation and the ability to automatically adapt to current network conditions.
- USER EXPERIENCE: The codec quality in terms of the user experience under optimal conditions varies considerably, mainly depending on the codecs' bitrate and effectiveness. User experience is typically measured using the Mean Opinion Score (MOS) or R-factor.
- COST: Not all codecs are free of charge. Indeed, most modern codecs require a license.
- SUPPORT: Numerous codecs have been proposed and standardized over the past decades. Few are widely supported by telecommunications equipment, such as mobile handsets, IP phones, MGWs, and SBCs. Even fewer are in active use.

Most codecs relevant to landline and mobile VoIP telephony are standardized by ITU-T and 3GPP. One notable exception is the Opus codec, which was introduced by the IETF. The table lists some codecs and their key characteristics.
The table (next page) is by no means complete and much more could be said about each individual codec. Please refer to the respective standards for more detailed information.

Codec	Bandwidth	Net Bitrate	Usage
AMR (Adaptive Multi-Rate)	NB	4.75 - 12.2 kbit/s	Widely used multi-mode codec in 2G/3G mobile networks
EVS (Enhanced Voice Services)	NB, WB, SWB, FB	5.9 - 128 kbit/s	Currently being deployed in 4G mobile networks; mandatory codec for 5G voice services
G.711	NB	64 kbit/s	Standard codec for fixed line voice services
G.711.1	NB,WB	64, 80, 96 kbit/s	Not widely supported, but used as HD codec by some fixed voice operators
G.722	WB	64 kbit/s	Widely supported HD codec for fixed line services
G.722.1	WB	24, 32 kbit/s	Mainly used by videoconferencing systems
G.722.2 (AMR-WB)	WB	6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85 kbit/s	Widely used multi-mode codec in 3G/4G mobile networks
G.723.1	NB	5.3, 6.3 kbit/s	Widely supported by fixed network devices, but not widely used
G.726	NB	16, 24, 32, 40 kbit/s	Widely supported by fixed network devices, but not widely used
G.729	NB	8 kbit/s	Widely supported by fixed network devices and frequently used
G.729.1	NB	8 kbit/s, 12–32 kbit/s in 2 kbit/s steps	Not widely supported
iLBC	NB	13.33, 15.20 kbit/s	Widely supported by fixed network devices, but not widely used
Opus	NB, WB, SWB, FB	6–510 kbit/s	Widely supported by fixed network devices, but not widely used Multi-mode audio codec defined by the IETF in RFC 6716; required by WebRTC implementations and slowly gaining support by fixed network devices

NB = narrowband, WB=wideband, SWB=super-wideband, FB=fullband

Modern HD codecs do not only provide a better user experience but are also extremely efficient. This can be seen if one divides a codec's maximum achievable R-Factor value by its nominal bitrate. This quality-per-bit ratio has steadily increased over the years.



Coding Efficiency: R-factor per Bit

Currently, the industry is in a transition phase and moving towards HD voice. Advantages typically cited by users are:

- Clearer overall sound quality
- Easier to recognize voices, distinguish confusing sounds and understand accented speakers
- Reduced listening effort, resulting in increased productivity and lessened listener fatigue

The biggest driver for this development is certainly the introduction of AMR-WB and EVS by VoLTE operators.

But also fixed network operators are – at least for on-net calls – slowly moving to G.722 and G.711.1. Finally, IPX wholesale services are enabling HD voice continuity in international calls.

(Estimating) Better User Experience

9

The rise of wideband codecs creates new challenges for monitoring voice quality and user experience. The Mean Opinion Score (MOS) is considered the key metric for voice quality, yet few are aware of its many different flavors.

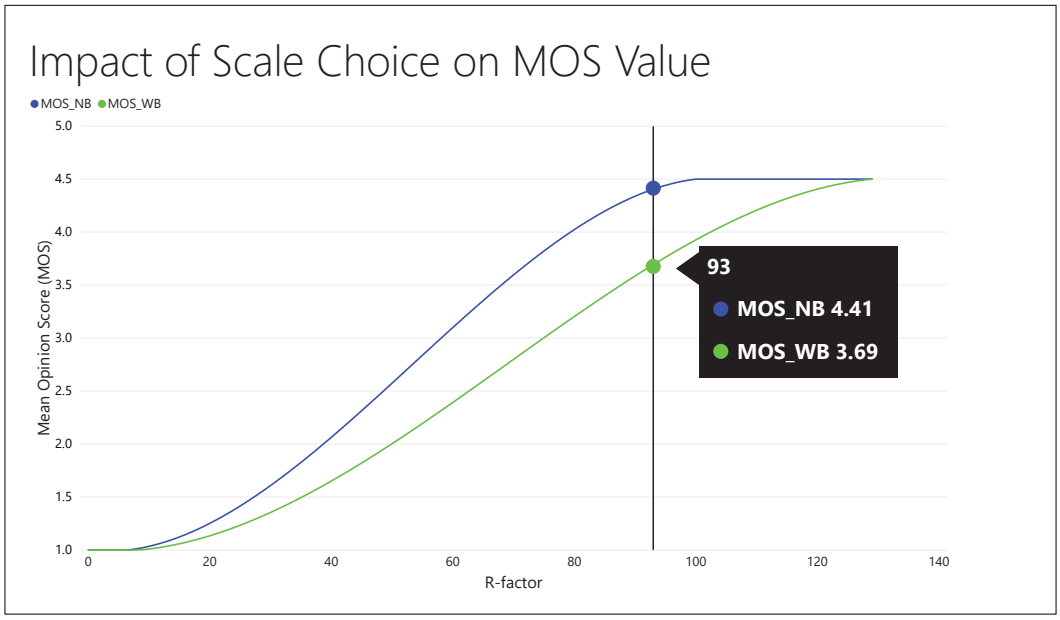
Historically, MOS is a subjective measurement. A set of listeners in a "quiet room" score the quality of a call on a scale of 1 ("bad") to 5 ("excellent"). The result is an average, i.e. the mean opinion of all listeners. All monitoring tools that provide a MOS essentially attempt to estimate the outcome of such an empirical study.

ITU-T Recommendation P.800.1 specifies terminology to distinguish different types of MOS. In effect, P.800.1 defines nine different types, but only two are relevant in the context of VoIP monitoring:

- MOS_{LQE} (listening quality estimate), i.e. the MOS provided by passive (no reference single-ended) VoIP monitoring systems and
- MOS_{LQO} (objective listening quality), i.e. the MOS calculated by full reference end-to-end test systems, e.g. based on ITU-T P.862 (PESQ) or P.863 (POLQA).

The 2006 revision of P.800.1 introduced notation to specify the reference audio bandwidth, i.e. the MOS scale. Based on this the letters N, W, S and F for narrowband, wideband, super-wideband and fullband should be appended to denote the quality reference. For example, a MOS listening quality estimate with a wideband reference should be denoted as MOS_{LQEW}. Unfortunately, this precise notation is rarely used in practice.

Stating the quality reference is however necessary because narrowband codecs no longer define the user expectations. With the introduction of wideband, super-wideband and even fullband codecs user expectations have started to shift. What was "excellent" before, is now considered mediocre.



As a result, new MOS scales had to be introduced, which take the varying user contexts and expectations into account. The traditional MOS scale is now often referred to as narrowband MOS (MOS_{NB}). It is complemented by the wideband (MOS_{WB}) and super-wideband (MOS_{SWB}) scales. Because of this, the same codec now has different maximum MOS values, depending on which scale is used. For example, G.711 has a MOS_{LQE} of 4.41 on the narrowband scale, but only 3.69 on the wideband scale.

An anecdote illustrates why new MOS scales are needed. Some years ago, a large German bank deployed a new VoIP system for the entire company. The selected phones supported standard narrowband G.711 as well as the G.722 wideband codec. It was therefore decided to use G.722 where possible to benefit from higher call clarity, i.e.

- G.722 was used for internal calls between the new IP phones;
- G.711 was used for external calls because this was the only codec supported for the interconnection with the provider.

What initially seemed like a good idea turned out to be problematic as employees started complaining at the bank's IT help desk. The users were perceiving 'bad quality' when talking with customers, even though the quality of those calls was perfect landline toll quality. Apparently, the employees accepted G.722 wideband audio as the new normal and adjusted their view on G.711. Without objective quality becoming worse, the subjective quality perception dropped – after a few months it was decided to use only G.711 to ensure a consistent narrowband user experience. It can be expected that the introduction of HD codecs for VoLTE calls will cause the same habituation effect as people quickly get used to new levels of quality.

The applicable scale depends on the (assumed) expectations of a user group. VoLTE users will quickly become used to wideband codecs, i.e. their user experience should be judged on the MOS_{WB} scale. In contrast, the reference quality of fixed line users is still standard G.711 quality, therefore MOS_{NB} should be applied. Of course, this is a simplification as user groups are not mutually exclusive and expectations will generally rise as users are becoming more exposed to HD voice quality. In the current transition period, the different MOS scales present an irritating source of confusion to the industry, specifically if values based on different reference qualities are compared.

Technically, passive monitoring systems calculate the MOS_{LQE} using the E-Model defined in G.107. The E-Model specifies how to determine the so-called R-Factor, which is then transformed into a MOS value. Originally, the R-Factor scale went from 0 to 100, with an R-Factor of 93 for G.711. For wideband codecs this scale was extended to 129. Two different transformation rules provide a mapping from the R-Factor to either a MOS_{NB} or MOS_{WB}. In June 2019 Recommendation ITU-T G.107.2 introduced the latest update to the E-Model by extending the R-Factor scale to 148 and providing a transformation rule for fullband MOS.

Monitoring Multi-Mode Codecs

12

Looking more closely at the Enhanced Voice Service (EVS) codec illustrates the challenge of monitoring user experience when using multi-mode audio codecs. EVS supports all four bandwidth modes NB, WB, SWB and FB and each bandwidth offers a set of bitrate modes.

kbit/sec	5.9	7.2	8	9.6	13.2	16.4	24.4	32	48	64	96	128
NB	x	x	x	x	x	x	x	-	-	-	-	-
WB	x	x	x	x	x	x	x	x	x	x	x	x
SWB	-	-	-	x	x	x	x	x	x	x	x	x
FB	-	-	-	-	x	x	x	x	x	x	x	x

The modes are controlled in different ways and can be switched every 20 ms.

CONTROL OF MODES VIA SDP

The codec to be used for a call is negotiated at call setup time via the Session Description Protocol (SDP). For EVS this entails that the call parties also agree on a bandwidth and bitrate (or range thereof). This is done via the codec-specific parameters 'bw' and 'br' as shown in the following examples:

bw=nb-swb;

br=16.4;

br=13.2-24.4;

Ranges may initially be specified by the caller. The callee may then accept or not accept the EVS codec as a whole. If he accepts, he shall only answer with ranges within the constraints given by the caller, as in this example:

Caller SDP: bw=nb-wb;

Callee SDP: bw=wb;

Here, the call parties agree to restrict bandwidth to wideband mode.

CONTROL VIA CMR BYTE

The Codec Mode Request (CMR) byte within the EVS codec payload header can be used to request a certain bandwidth/bitrate mode from the remote call party. The requested modes shall be within the range agreed upon via SDP. Once a CMR has been received, the receiving party should use this mode, and may use lower modes within the SDP constraints.

CONTROL BY ENCODER

Finally, the EVS encoder performs bandwidth detection on the input signal to apply a bandwidth decision logic. This means that the encoder may decide to encode the audio in a lower bandwidth than determined by the input sampling rate. For example, assume the input sampling frequency is 32 kHz. If the bandwidth detection logic determines that there is no "energetically meaningful" spectral content above 8 kHz, then the codec is operated in the WB mode.

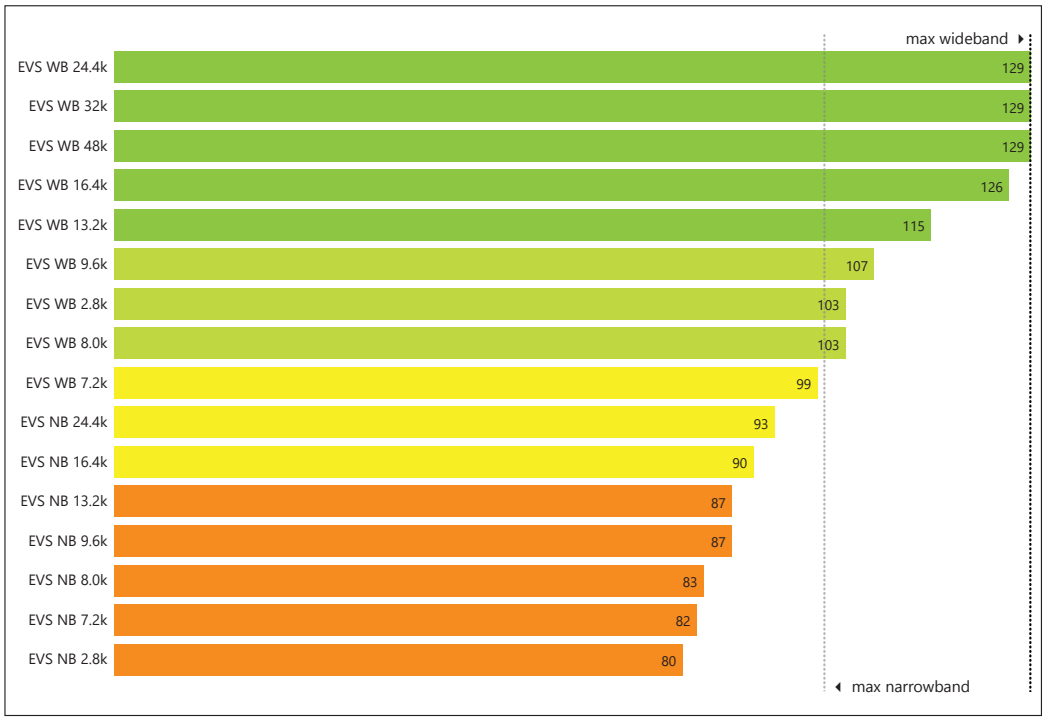
Interestingly, this codec decision does not seem to be constrained by the result of the SDP negotiation. Voipfuture analysed live VoLTE traffic and found many examples of calls where SDP negotiations settled on EVS WB as bandwidth mode, but the encoder delivered EVS NB encoded frames.

IMPACT ON MONITORING

All of this would be of little importance to VoIP monitoring, if the used bandwidth and bitrate modes had no impact on the user experience.

The opposite is true. For example, the R-Factor of EVS WB bitrate modes ranges from 99 (7.2k) to 129 (24.4k), which translates to a MOS_{WB} range of 3.91 to 4.5. According to ITU-T G.107 this leaves user satisfaction between "some dissatisfied" to "very satisfied".

AMR-WB presents an even more extreme example, where satisfaction (based on wideband reference quality) between the different bitrates ranges from "nearly all users dissatisfied" to "very satisfied".



This must be seen in light of the fact that very different factors contribute to codec, codec bandwidth and bitrate selection, which are not all under full control of a service provider. Quality can change on a packet by packet basis – put to the extreme, user satisfaction can change every 20ms.

For modern multi-mode HD voice codecs it is therefore not sufficient to determine the codec based on the result of the SDP negotiation.

Instead monitoring tools must inspect all RTP packets to reliably measure the user experience. Yet, most VoIP monitoring tools focus on the signaling and are unable to cope with the large amounts of RTP packets.

Real RTP monitoring systems, such as Voipfuture's Qrystal, analyze the RTP flows and continuously determine the used codecs and their parameters.

Qrystal uses intelligent timeslicing technology to summarize the characteristics, such as the codec and its modes, of every five second time slice. This ensures efficient data storage while providing a high temporal resolution and accurate estimation of the user experience.

CONCLUSION

For decades narrowband codecs, such as G.711, have been considered the gold standard of telephony. The current rise of modern HD codecs is about to change this, which has a number of implications. For example, HD codecs change the user expectations, because users quickly get used to new quality standards. This in turns requires to speed up the transition to wideband codecs.

Technically, HD audio – and especially multi-mode codecs – create new challenges for service monitoring. Quality can change at any time; not only because of network impairments, but also because of codec mode changes. Different codecs and modes may deliver widely different quality and thus VoIP monitoring systems must be designed to analyze every RTP packet.

Most VoIP monitoring tools on the market are not capable of detecting the codec based on analysis of the RTP payload. Qrystal is a notable exception and provides service providers worldwide with accurate information on the user satisfaction.

voipfuture

Voipfuture is a premium voice quality analytics vendor providing tools for assessing, aggregating, analyzing, and visualizing voice quality information. Voipfuture products offer a precise view on media and control plane to communication service providers, VoLTE carriers, wholesalers and enterprises.

DOC-ID 144, v1.1

+49 40 688 90 01 0
info@voipfuture.com
www.voipfuture.com

Wendenstr. 4
20097 Hamburg
Germany