Fraunhofer IIS

TECHNICAL PAPER

ENHANCED VOICE SERVICES (EVS) CODEC

Until now, telephone services have generally failed to offer a high-quality audio experience due to limitations such as very low audio bandwidth and poor performance on non-speech signals. However, recent developments in speech and audio coding now promise a significant quality boost in conversational services, providing the full audio bandwidth for a more natural experience, improved speech intelligibility and listening comfort.

The recently standardized Enhanced Voice Service (EVS) codec is the first 3GPP communication codec providing super-wideband (SWB) audio bandwidth for improved speech quality already at 9.6 kbps. At the same time, the codec's performance on other signals, like music or mixed content, is comparable to modern audio codecs. The key technology of the codec is a flexible switching scheme between specialized coding modes for speech and music signals. The codec was jointly developed by the following companies, representing operators, terminal, infrastructure and chipset vendors, as well as leading speech and audio coding experts: Ericsson, Fraunhofer IIS, Huawei Technologies Co. Ltd, NOKIA Corporation, NTT, NTT DOCOMO INC., ORANGE, Panasonic Corporation, Qualcomm Incorporated, Samsung Electronics Co. Ltd, VoiceAge and ZTE Corporation.

The objective of this paper is to provide a brief overview of the landscape of communication systems with special focus on the EVS codec. The main design constraints and features are highlighted, while some brief technology insights are also incorporated. Finally, listening test results conducted during the selection and characterization phase of the standardization process are presented and discussed.

Fraunhofer Institute for
Integrated Circuits IIS

Management of the institute
Prof. Dr.-Ing. Albert Heuberger
(executive)
Dr.-Ing. Bernhard Grill
Am Wolfsmantel 33
91058 Erlangen
www.iis.fraunhofer.de

Contact
Matthias Rose
Phone +49 9131 776-6175
matthias.rose@iis.fraunhofer.de

Contact USA
Fraunhofer USA, Inc.
Digital Media Technologies*
Phone +1 408 573 9900
codecs@dmt.fraunhofer.org

Contact China
Toni Fiedler
china@iis.fraunhofer.de

Contact Japan
Fahim Nawabi
Phone: +81 90-4077-7609
fahim.nawabi@iis.fraunhofer.de

Contact Korea
Youngju Ju
Phone: +82 2 948 1291
youngju.ju@iis-extern.fraunhofer.de

* Fraunhofer USA Digital Media Technologies, a division of Fraunhofer USA, Inc., promotes and supports the products of Fraunhofer IIS in the U. S.

COMMUNICATION SYSTEMS

When comparing the audio quality of a phone call with the sound of a movie on TV, the muffled audio of the standard telephone conversation becomes evident to everyone. This is mainly due to the limitation of the audio bandwidth in existing telephone systems. Figure 1 illustrates the different audio bandwidth capabilities present in typical communication/broadcast systems and the human auditory system.

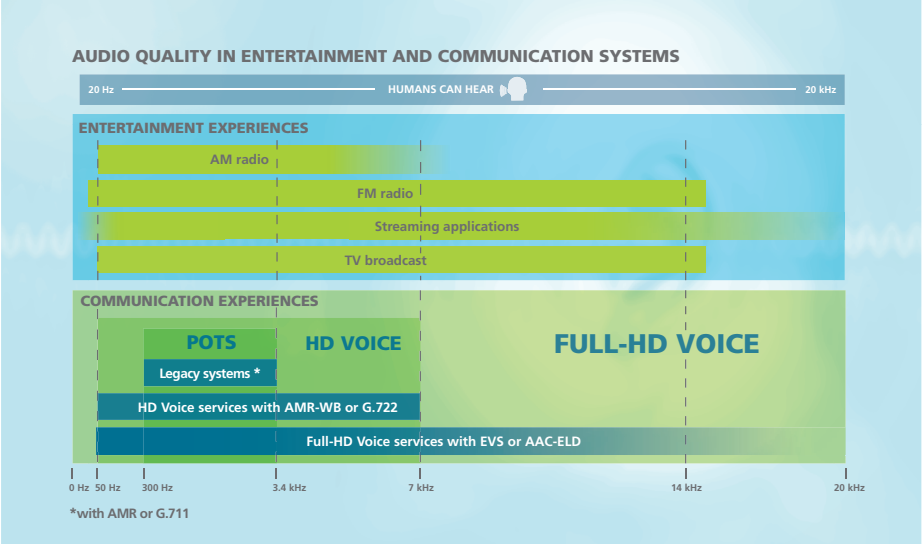

Figure 1: Audio bandwidth in broadcast and communication systems

Plain old telephone systems (POTS) provide narrow band (NB) audio signals, meaning frequencies of up to 3.4 kHz of audio bandwidth only. HD Voice services deliver wideband (WB) quality, where WB stands for an audio bandwidth of 7 kHz. Considering the capability of the human auditory system, higher frequencies up to 20 kHz – relevant to high-fidelity sound – are still missing. Therefore, HD Voice is further extended to Full-HD Voice, including the quality levels of super-wideband (SWB) and full-band (FB). SWB stands for an audio spectrum of 16 kHz, while FB contains all frequency components up to 20 kHz.

Landline telephone services today provide either NB or WB quality. The codecs used in these systems are G.711 [1] or G.722 [2] both operating at 64 kbit/s. In the mobile world, NB is the default quality level; however, an increasing number of WB services are now emerging. The codecs used for NB and WB mobile services are AMR-NB [3] and AMR-WB [4] both usually operating at bit rates around 12 kbit/s. Some mobile networks even allow higher rates for AMR-WB, i.e. 23.85 kbit/s, although the quality improvement compared to the default rates is rather limited. The codecs for mobile communications are highly optimized for speech signals and, as a consequence, their capability for coding other signals like music is not satisfying.

Specialized communication systems for telepresence or dedicated video conferencing systems can already provide Full-HD Voice quality today. The de-facto codec standard for such systems is AAC-(E)LD [5]. The codec operates on a wide range of bit rates, starting from 24 kbit/s up to 64 kbit/s, while being able to transmit speech and music signals.

AAC-(E)LD is also utilized for so called over the top (OTT) services. Typical OTT applications include Skype or Facetime, where the IP packet transmission is handled without the network management of an operator.

The 3GPP EVS codec [6,7] overcomes the two major problems existing in mobile and landline telephone systems, i.e. limited audio bandwidth and poor performance on non-speech signals. At the same time, the codec is able to operate at typical bit rates for mobile services. This paves the way for a new standard of user experience in communication quality, applicable to all kinds of networks, including landline, mobile and OTT.
In the following sections, this paper outlines the key technical aspects of the EVS codec design which result in this significant step forward in service quality. It should be noted, that many more aspects beside the audio coder may have a significant influence on the end-to-end service user experience. These factors include the audio-front-end processing, including echo cancellation, noise suppression, automatic gain control, wind noise filtering and de-reverberation, or the network behavior causing delay jitter and packet loss.

OVERVIEW OF THE EVS CODEC

Design objectives

The EVS codec, as standardized by 3GPP in September 2014 [6], provides a wide range of functionalities enabling unprecedented versatility and efficiency in communication. It has been primarily designed for Voice over LTE (VoLTE) and fulfills all objectives defined by 3GPP, namely:

1. Enhanced quality and coding efficiency for narrowband (EVS-NB) and wideband (EVS-WB) speech services;
2. Enhanced quality by the introduction of super-wideband (EVS-SWB) speech;
3. Enhanced quality for mixed content and music in conversational applications;
4. Robustness to packet loss and delay jitter;
5. Backward compatibility to the AMR-WB codec [20].

As pointed out previously, this paper focuses on aspects 2) and 3) of the design objectives. For the sake of completeness, the quality enhancements for the legacy NB and WB services in 1) are discussed later in this paper as well. Besides the improvements listed above, EVS comes along with a full set of system functions required for communication systems such as voice activity detection (VAD), discontinuous transmission (DTX), comfort noise generation (CNG) or jitter buffer management (JBM). The codec operates at a wide span of bit rates starting from 5.9 kbit/s up to 128 kbit/s, and as a consequence always provides the optimal rate for each network. All definitions of the design constraints, developed during the EVS standardization, are given in [6].

Technical Overview

Coding paradigms

In general, the world of audio coding can be divided into two paradigms:

- Speech coding: Approach to model the vocal tract of human beings
- Perceptual coding: Approach exploiting the limitations of the perception of human auditory system

As described in [8], efficient speech coding schemes, such as AMR-NB and AMR-WB, have typically three major components: (1) a short-term linear prediction (LP) filter modeling the vocal tract; (2) a long-term prediction (LTP) filter, which models the periodicity in the excitation signal from the vocal chords; and (3) an innovation codebook, for encoding the non-predictive part of the speech signal.

Perceptual coding schemes, such as AAC [9], are based on three primary steps: (1) a time/frequency conversion; (2) irrelevance reduction composed by a subsequent quantization stage, in which the quantization error is controlled using information from a psychoacoustic model; and (3) redundancy reduction, featuring an encoding stage in which the quantized spectral coefficients and corresponding side information are entropy-encoded using code tables. This results in a source-controlled codec adapting to the input signal statistics as well as to the characteristics of human perception.

In general, the speech coding approach offers best performance on pure, clean speech signals at low bit rates, while the perceptual coding approach delivers better performance on generic content, e.g. music, and provides up to perceptual transparent quality.

The first codec combining these two major coding approaches was the Unified Speech and Audio Codec (USAC) [8]. USAC exceeds an algorithmic delay of more than 100ms, which is not acceptable for bi-directional communication applications. However, motivated by USAC's excellent coding performance, the unified coding approach has been adopted and further optimized to complement the challenging demands of the EVS codec.

Switched Speech/Audio Coding at Low Delay

The EVS codec is the first mobile communications codec to deploy content-driven, on-the-fly switching between speech and audio compression at low algorithmic delay of 32 ms, leading to significantly improved coding of generic content such as music signals.

The speech codec is an improved variant of Algebraic Code-Excited Linear Prediction (ACELP), extended with specialized LP-based modes for different speech classes. For audio coding, frequency domain (MDCT) coding is used. Special attention was paid to increase the efficiency of MDCT based coding at low delay/low bitrates and on obtaining seamless and reliable switching between the speech and the audio cores. Figure 2 depicts a high-level block diagram of the EVS encoder and decoder.

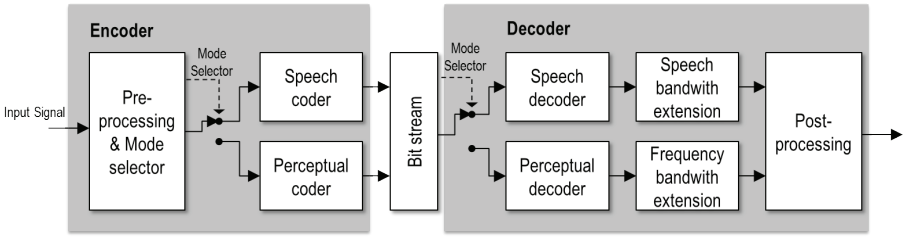

Figure 2: EVS codec structure

Super-wideband Coding and Beyond

EVS is able to provide SWB and even FB quality level and therefore overcomes the muffled sound known from today's telephony. Technically, the codec achieves this benefit by utilizing bandwidth extensions. Depending on whether the speech or audio mode is active, either a time-domain bandwidth extension (TBE) technology or an integrated frequency domain solution is deployed. The latter provides several sub-modes, e.g. harmonic model coding, which can cope with typical music signals. EVS is the first codec providing differently optimized bandwidth extensions that are utilized and switched in a source-controlled manner. Due to the dedicated content optimization, a very natural and clean sound quality can be offered even at very low bit rates.

PERFORMANCE EVALUATION

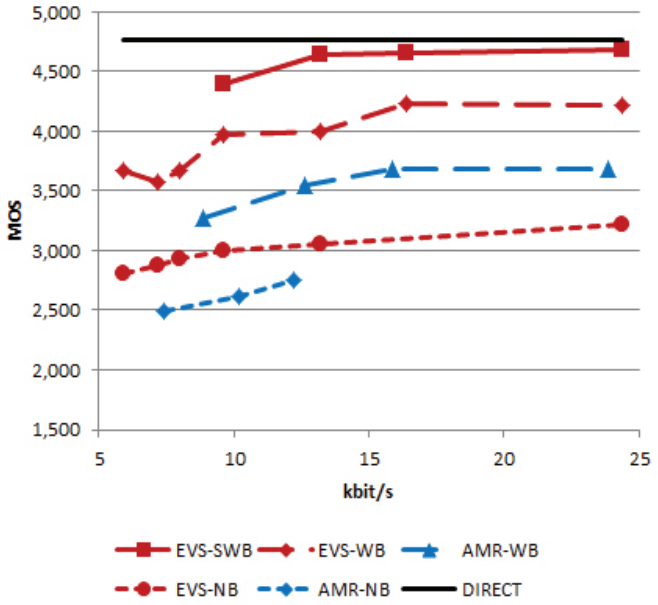

Figure 3: 3GPP EVS characterization test results for multi-bandwidth clean speech

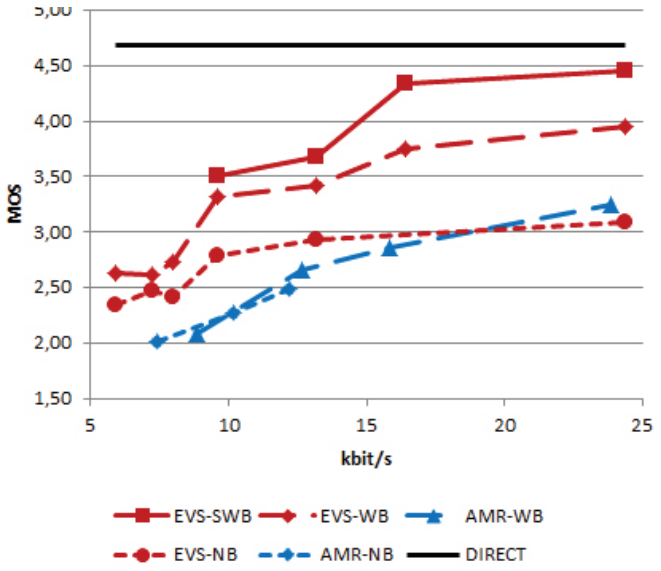

Figure 4: 3GPP EVS characterization test results for multi-bandwidth mixed content and music

Extensive testing has been performed within 3GPP to verify the performance of the EVS codec over a wide range of operating points and content types [11], including multi-bandwidth tests conducted with the P.800 DCR method [12]. Figures 3 and 4 provide a high-level impression of the quality (in DMOS score) for clean speech (English) and for mixed content and music. The results include the quality levels narrow band, wideband and super wideband at bit rates regarded as typical for mobile cellular services. The results are discussed in the following:

- For EVS (operating in SWB mode), the clean speech quality level is already very high for 9.6 kbit/s, outperforms AMR-WB 23.85 kbit/s significantly, and increases further with bitrate. Starting from 13.2 kbit/s, the EVS-SWB clean speech quality already approaches that of the "Direct Source" (original) quality.

- EVS (operating in SWB mode) outperforms AMR-WB even more significantly at mixed & music content. Their grades differ in average more than 1.2 Mean Opinion Score (MOS) at comparable bitrates. At 24.4 kbit/s, EVS mixed content & music quality approaches that of the "Direct Source" (original) quality.

- Even when operated in WB mode (e.g. in devices or services only supporting WB mode), EVS 9.6 kbit/s outperforms AMR-WB at 23.85 kbit/s. Furthermore it offers much higher quality for clean speech and music when using an equivalent bit rate (24.4 kbit/s).

- At first glance it may be surprising that AMR-WB cannot outperform AMR-NB for mixed and music at same bit rate despite exploiting the double audio bandwidth. Thanks to EVS, this weakness has been overcome.

- In the case of NB input signals, the EVS codec performs significantly better than AMR-NB, for both clean speech and mixed and music content. This mode may be useful in instances of interconnection to other NB networks such as landline.

It is well-known that test results and their interpretation vary with language and material chosen. However, the EVS codec has been examined with regard to 10 languages, 6 different background noises and various music materials in the 3GPP Selection Phase, showing excellent performance and improvement over earlier standards on a broad basis. These results, combined with further extensive performance characterization of the EVS codec, have been published in the 3GPP Technical Report (TR) 26.952 [11].

APPLICATIONS

Since Long Term Evolution (LTE), the fourth generation of mobile network standards, has been introduced, cellular phone networks are starting to switch to IP-based transmission. LTE is based on the older, established GSM and UMTS standards, offering an all-IP architecture and low latencies. It requires the deployment of all-IP voice services or Voice-over-LTE (VoLTE) and in turn opens up the prospect of moving all voice services onto IP networks, eventually phasing out the legacy-switched services based on GSM, UMTS and CDMA networks.

With the help of Full-HD Voice technologies, service providers can shake off the limitations of these legacy services, including very limited audio bandwidth and the use of speech-centric codecs. Since VoLTE is providing Quality Of Service (QoS) in a managed network, EVS heralds the opportunity to outperform OTT services such as Skype or Viber, not only in audio quality but also in terms of robustness and service availability. Hence, mobile operators will be able to regain lost ground with respect to voice minutes.

EVS, due to its outstanding error robustness [10], is also well-suited for usage in best effort networks such as VoWifi (Voice over Wifi), and may also be available for 3G/circuit switched systems in the future.

CONCLUSION

Various new features accompanied by unmatched speech and audio quality make the EVS codec, the latest 3GPP codec for enhanced voice services, the most efficient and versatile codec for high quality communication in any type of network, in particular cellular LTE and Voice over WiFi networks. EVS opens up a completely new experience for users everywhere, delivering an audio quality that is close to transparency even for mobile communication services. The imminent introduction of the EVS codec will therefore bring lasting benefits for both mobile operators and their customers.

REFERENCES

[1] ITU-T Rec. G.711, "Pulse code modulation of (PCM) of voice frequencies"

[2] ITU-T Rec. G.722, "7 kHz audio-coding within 64 kbit/s"

[3] K. Järvinen. "Standardisation of the Adaptive Multirate Codec," Proc. EUSIPCO, Sept. 2000.

[4] B. Bessette, et al., "The adaptive multi-rate wideband speech codec (AMR-WB)," IEEE Trans. on Speech and Audio Processing, vol. 10, no. 8, pp. 620-636, Nov. 2002.

[5] M. Schnell, et al., "MPEG-4 Enhanced Low Delay AAC - a new standard for high quality communication", 125th AES Convention, Oct. 2008.

[6] S. Bruhn, et al., "Standardization of the new EVS Codec", Proc. ICASSP, Apr. 2015.

[7] Martin Dietz et al., "Overview of the EVS Codec Architecture," Proc. ICASSP, April 2015.

[8] M. Neuendorf, et al.: "Unified Speech and Audio Coding Scheme for high quality at low bitrates", Proc. ICASSP, Apr. 2009

[9] M. Bosi, et al., "ISO/IEC MPEG-2 Advanced Audio Coding", paper 4382, 101st AES Convention, Nov. 1996.

[10] V. Atti, et al., „Improved error resilience for VOLTE and VOIP with 3GPP EVS channel aware coding", Proc. ICASSP, Apr. 2015.

[11] 3GPP TR 26.952, "Universal Mobile Telecomm-unications System (UMTS); LTE; Codec for Enhanced Voice Services (EVS); Performance characterization," http://www.3gpp.org/DynaReport/26952.htm

[12] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," Aug. 1996.

ABOUT FRAUNHOFER IIS

The Audio and Media Technologies division of Fraunhofer IIS has been an authority in its field for more than 25 years, starting with the creation of mp3 and co-development of AAC formats. Today, there are more than 10 billion licensed products worldwide with Fraunhofer’s media technologies, and over one billion new products added every year. Besides the global successes mp3 and AAC, the Fraunhofer technologies that improve consumers’ audio experiences include Cingo® (spatial VR audio), Symphoria® (automotive 3D audio), xHE-AAC (adaptive streaming and digital radio), the 3GPP EVS VoLTE codec (crystal clear telephone calls), and the interactive and immersive MPEG-H TV Audio System.

With the test plan for the Digital Cinema Initiative and the recognized software suite easyDCP, Fraunhofer IIS significantly pushed the digitization of cinema. The most recent technological achievement for moving pictures is Realception®, a tool for light-field data processing.

Fraunhofer IIS, based in Erlangen, Germany, is one of 69 divisions of Fraunhofer-Gesellschaft, Europe’s largest application-oriented research organization.

For more information, contact amm-info@iis.fraunhofer.de, or visit www.iis.fraunhofer.de/amm.