QUALITY EVALUATION OF THE G.EV-VBR SPEECH CODEC

Anssi Rämö [*], *Henri Toukomaa* [*], *S. Craig Greer* [*], *Lasse Laaksonen* [*], *Jacek Stachurski* [#], *A. Erdem Ertan* [#], *Jonas Svedberg*[†], *Jon Gibbs*[‡], *Tommy Vaillancourt* [^]

[*] Nokia, [#] Texas Instruments, [†] Ericsson, [‡] Motorola, [^] VoiceAge

ABSTRACT

ITU-T has selected the candidate submitted by Ericsson, Nokia, Motorola, VoiceAge, and Texas Instruments as the baseline for the G.EV-VBR coding standard. G.EV-VBR is an embedded scalable speech codec that uses state-of-the-art technology to provide the most efficient encoded speech available for various real-time applications. EV-VBR encodes both narrowband (NB) and wideband (WB) speech signals starting at 8kbps. Near perfect wideband representation is achieved at 32kbps for all signal types. The bit stream is divided into five robust layers, providing sufficient granularity, in particular for VoIP applications. In addition, an extension to the codec will provide super-wideband and stereo capability by adding layers to the codec. Extensive listening tests were conducted during the ITU-T selection phase to support selection of the best-performing candidate. The selected EV-VBR candidate passed 69 of 70 required and 25 of 28 objective terms of reference [1].

Index Terms— Speech Coding, Listening Testing, Standardization, Embedded Coding, Scalable Coding

1 INTRODUCTION

The primary feature of the EV-VBR speech codec is embedded scalability. This means that additional encoded bits that are transmitted in addition to the core layer are appended to the bit-stream. The additional information is used to increase the quality and/or the robustness of the codec. The bit-rate of the EV-VBR core codec is 8kbps (L1). As the frame size is 20ms, the core encodes a frame into 160 bits. The baseline standard provides a total of five layers, known as L1-L5. The second and third layers are 80 bits in size, and in conjunction with the core layer provide encoded bit rates of 12kbps (L2) and 16kbps (L3) respectively. The fourth and fifth layers are 160 bits in size, providing encoded bit rates of 24kbps (L4) and 32kbps (L5). The first two layers (L1&L2) are based on embedded ACELP. L3 (16kbps) is a layer that contains additional redundant information for increased frame error robustness. Layer L3 also adds some MDCT coefficients for improved quality. The upper two layers (L4&L5) of the baseline codec contain only MDCT information, which provide excellent music and background noise performance at the higher L4 and L5 bit rates.

In this paper the results of both CuT2 and CuT4 results are averaged into "EV-VBR" in order to simplify the figures, since CuT2 and CuT4 use the same codec with slightly different tuning.

This paper presents the listening test results performed as part of the ITU-T selection testing. The paper is organized as follows. *Section 2* describes the listening test arrangements. *Section 3* presents the listening test results obtained with clean speech signals for both narrowband and wideband signals. The performance in case of various frame erasure patterns is also presented and discussed. *Section 4* presents results obtained with music signals. *Section 5* presents listening test results obtained in various background noise conditions. *Section 6* discusses EV-VBR performance as compared with G.729.1 [3]. *Section 7* provides conclusions.

2 TEST ARRANGEMENTS

Listening tests for the ITU-T selection phase were conducted in seven different laboratories during January-February 2007. There were a total of 9 different listening tests and each listening test was performed in two different listening laboratories in two different languages. Each experiment employed 32 naïve listeners for a total of 576 listeners. All together, they cast a total of 101,376 votes on 264 different conditions. In each laboratory each condition received 192 votes, and a mean opinion score (MOS) was calculated. The listening test conditions were conducted according to the test plan developed by Study Group 12 of the ITU-T and the conditions were distributed among tests as shown in *Table 1*.

Table 1 Condition division to separate listening tests

Exp.	Test description	Type
1	Input Level performance for NB speech	ACR
2a	Clean WB speech performance on L1, L2	ACR
2b	Clean WB speech performance on L3, L4	ACR
2c	Clean WB speech performance on L5	ACR
3	Music performance L3, L4 and L5	ACR
4	Car noise performance for NB speech	DCR
5	Street noise performance for NB speech	DCR
6	Interfering Talker performance for WB speech	DCR
7	Office noise performance for WB speech	DCR

All listening was conducted mono-aurally. Input signal filtering was according to P.341 for all conditions. The distribution of the listening tests among laboratories was as shown in *Table 2*. Clean conditions were tested using ACR listening method, where the processed sample is played to the listeners only once and listener makes his/her decision subjectively. The noisy conditions were tested using the DCR listening method, where the noisy direct signal is played first to the listener followed by the processed noisy sample.

Table 2 Listening Laboratories

ID	Listening Lab	Language	Experiments
Lab_A	**A**rcon	NA English	2c, 6
Lab_B	**B**IT	Chinese	1, 2a, 7
Lab_D	**D**ynastat	NA English	2a, 4
Lab_F	**F**rance Telecom	French	2c, 3
Lab_J	NTT-AT	**J**apanese	2b, 4, 5
Lab_N	**N**okia	Finnish	2b, 5, 7
Lab_V	**V**oiceAge	Canadian French	1, 3, 6

The terms of reference (ToR) [2] document contains selection criteria for determining the winner from the tested codecs. For the speech tests the reference codec was either G.729 (NB conditions) or G.722.2 (WB conditions). Similar bit-rates were chosen from the references when possible. Only in the music test was G.722 used as the reference. G.729.1, a more recent ITU-T speech coding standard that also has an embedded layer structure, was included in the listening tests for informational purposes and its results are therefore also included in the analysis of the test results in [1] as well as in this paper. The result figures are ordered so that direct reference is always at the top, ToR requirements are next, EV-VBR codec results follow and finally G.729.1 results are also shown. Full result including MNRU and missing conditions can be found in [1].

3 RESULTS IN CLEAN SPEECH AND WITH FER

Clean speech performance has traditionally been very important in speech coding research. For noisy conditions, the EV-VBR terms of reference [2] specifically state that background noise should not be considered as noise but part of the signal. The performance of the selected EV-VBR candidate in clean speech is excellent when compared to other state-of-the art narrowband and wideband speech codecs despite the fact that EV-VBR has to support both signal types (NB&WB), provides embedded scalability, and is required to encode background noise signals.

Robustness against frame erasures was also given a high priority in the requirements. The reasoning behind this was that in future packet based networks frame erasure is most likely to be the biggest problem. Random frame erasure patterns were used for all frame erasure conditions.

3.1 Narrowband

Narrow band clean speech signals were tested in one experiment. The experiment also included a 3% frame erasure pattern.

Clean speech results in *Figure 1* show that the EV-VBR codec is almost transparent, even at 8kbps. The 12kbps condition was a special case, where the input signal was wideband and the output was rendered in narrowband by the codec. The results show transparent quality for EV-VBR even in this difficult condition. Error robustness of EV-VBR is also state-of-the-art when compared to G.729 in similar error conditions. It was noted by ITU-T that Lab_B results showed some inconsistencies, however full result are shown here for completeness. [1]

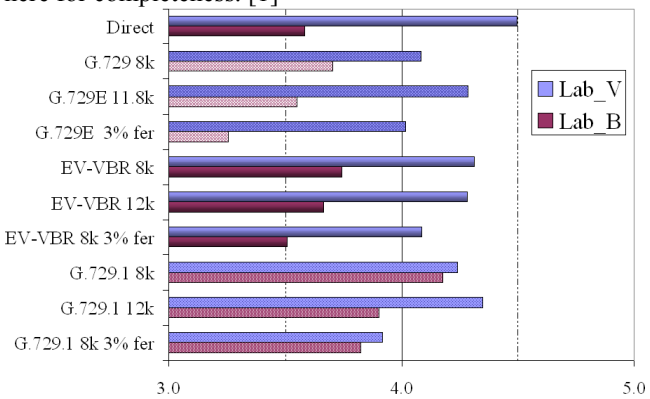

Figure 1 NB performance in clean speech (Exp 1)

3.2 Wideband

Codec performance with clean wideband speech was tested in three different tests (2a, 2b and 2c) each of which concentrated on different layers/bit rates. The first test in *Figure 2* clearly shows that EV-VBR outperforms G.722.2 at 8.85kbps, at similar bit rates and especially shows the superiority of its error robustness. It is noteworthy that the performance of EV-VBR when subjected to a 3% frame erasure rate is equivalent to G.722.2 at 8.85kbps in a clean channel. Fast switching means that the bit rate always remains constant for 10 consecutive 20ms frames and then the rate changes randomly to another bit rate from the set for the next 10 frames.

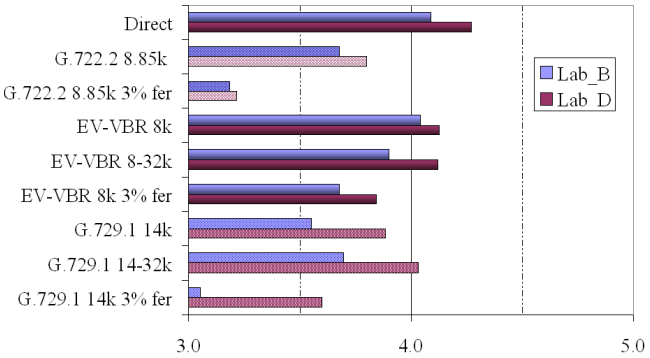

Figure 2 WB L1, L1 @ 3%FER, fast switching (Exp2a)

The second WB test, experiment 2b (*Figure 3*), contains layers L3 and L4, where L3 concentrates on codec robustness to frame erasures. The robustness is evident, as EV-VBR at 6% FER maintains equivalence to G.722.2 at 15.85kbps without errors. Layer three (L3) in EV-VBR was specifically designed to conceal frame losses. In addition, a special transient mode was developed for the most critical speech onsets, where errors are the most difficult to conceal.

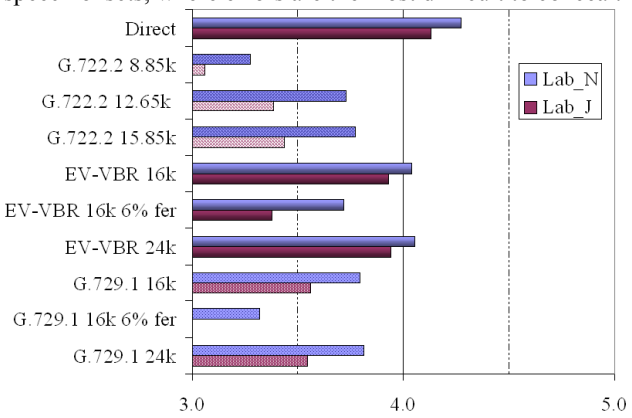

Figure 3 WB speech in L3/L4 and L3 6% FER (Exp 2b)

As shown in *Figure 4*, the third WB clean speech test, experiment 2c, demonstrates that EV-VBR performs noticeably closer to direct (uncoded) speech than the required reference of G.722.2 at 23.85kbps. A special layered error case where the core layer has no frame erasures and layers L2 to L5 have 2%, 4%, 6% and 10% FER respectively nicely illustrates that the FER on higher layers have very little effect on speech performance as long as the core layer is received without errors. In addition, the EV-VBR codec tolerates a 3% FER on all layers very well.

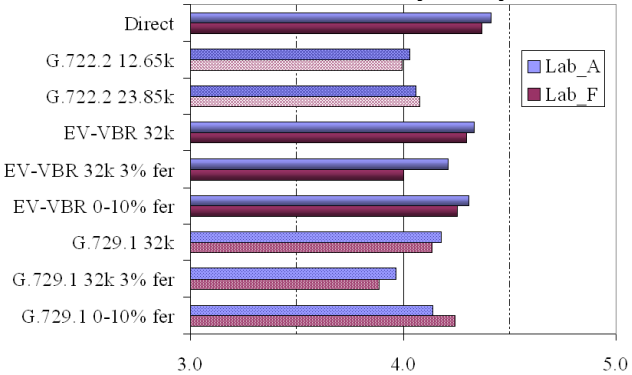

Figure 4 WB speech in L5, ACR test (Exp 2c)

4 RESULTS FOR MUSIC SIGNALS

Although EV-VBR is predominantly a speech codec, emphasis was also placed on music performance, since in the terms of reference some of the primary applications such as multimedia streaming may include a lot music signals[2]. Therefore, the target quality was set quite high and a specific high rate audio codec was used for the reference.

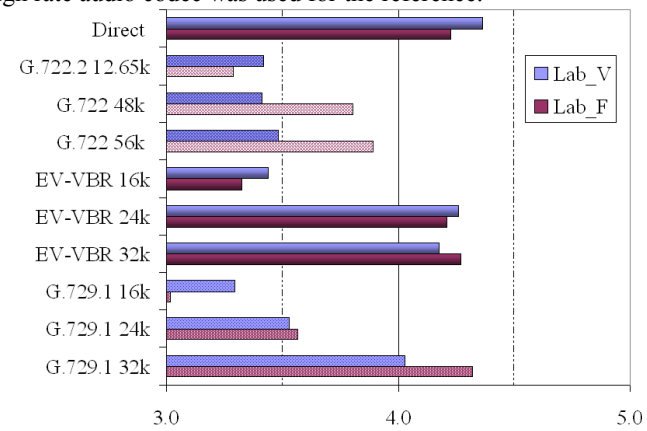

Figure 5 WB music performance at L3 - L5 (Exp 3)

As can be seen from the results in *Figure 5*, the performance of EV-VBR at both layers L4 and L5, compared to the original, is almost transparent. At 16kbps EV-VBR is statistically equivalent to the requirement of G.722.2 at 12.65kbps. In contrast, G.722 shows somewhat poorer scores due to the excessive noise it exhibits at high frequencies.

5 RESULTS IN BACKGROUND NOISE

As discussed earlier in *Introduction*, the terms of reference require the codec to encode, to the extent possible, any signals associated with the speaker's environment, i.e. background noise or ambient signals. Therefore the EV-VBR speech codec does not use noise suppression for the purpose of reducing noise in speech.

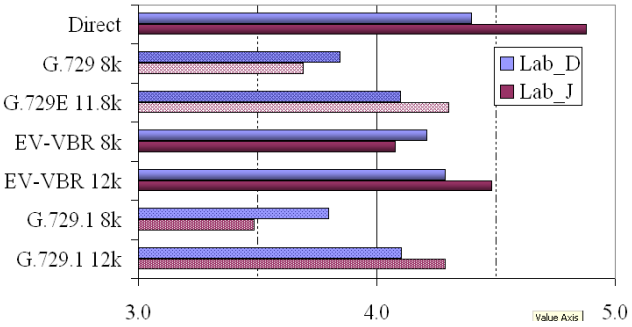

Figure 6 Car noise at -15dB, NB, DCR test (Exp 4)

5.1 Narrowband

Two experiments (4 and 5) shown in *Figure 6* and *Figure 7* were conducted using -15dB car noise and -20 dB street noise respectively. The overall performance of the selected EV-VBR codec remains excellent in these background noise cases. Across all test laboratories and at both narrowband supported bit rates of 8kbps and 12kbps, EV-VBR significantly outperforms the required performance of G.729

at 8kbps and 11.8kbps. In fact, EV-VBR at 8kbps approaches the quality of G.729E at 11.8kbps.

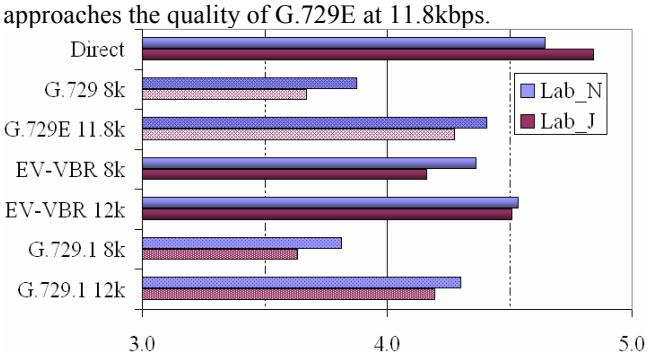

Figure 7 Street noise -20dB, NB, DCR test (Exp5)

5.2 Wideband

Wideband performance under background noise was tested in two experiments using DCR methodology.

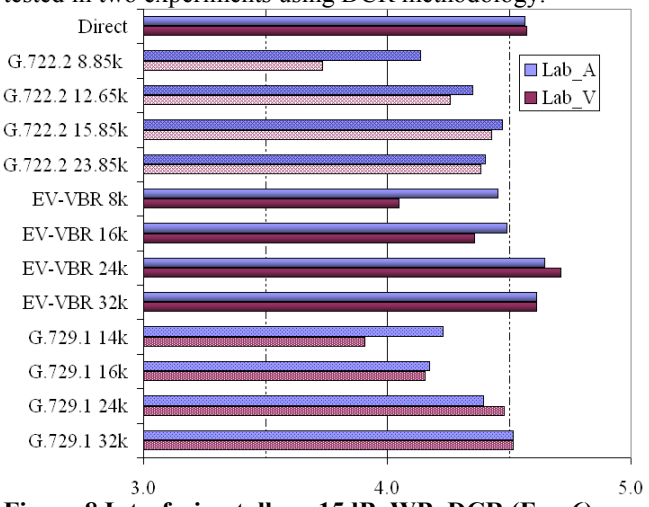

Figure 8 Interfering talker -15dB, WB, DCR (Exp 6)

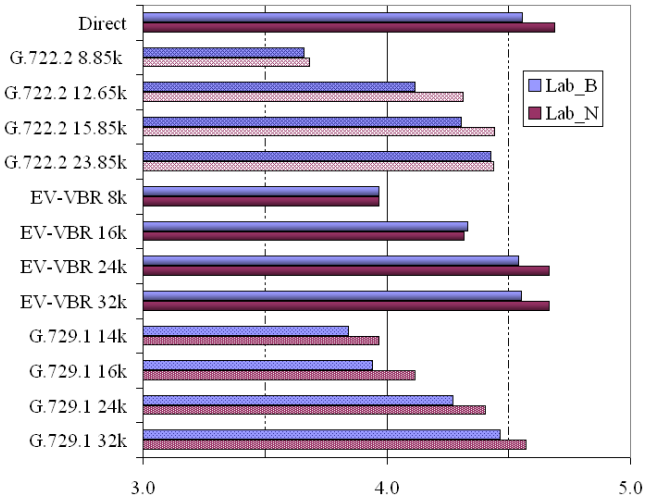

Figure 9 Office noise in -20dB, WB, DCR (Exp7)

The first WB background noise test used interfering talker noise at -15 dB level (*Figure 8*) and the second office noise at -20dB (*Figure 9*). Under background noise, the EV-VBR codec at 8kbps shows better performance than G.722.2 at 8.85kbps. At 24kbps (L4) the quality already saturates to that of the direct input.

6 COMPARISON WITH G.729.1

In addition to reviewing the EV-VBR performance against the reference conditions set for the codec by the ITU-T, it is interesting to compare the selected codec with a recent ITU-T standard, G.729.1. This codec was included in the selection testing after a withdrawal by one of the original candidates both to eliminate the need to develop a new test plan and to provide an impartial analysis of each embedded codec's strengths. G.729.1 is particularly interesting in comparison to EV-VBR, because it utilizes a similar overall approach of embedded scalability with quite a similar bit rate set for the individual layers as EV-VBR. One major difference between the codecs is that EV-VBR is a wideband codec in all of its rates, while G.729.1 is a narrowband-only codec below 14kbps. It was therefore necessary to compare EV-VBR at 8kbps and 12kbps with G.729.1 at 14kbps in the wideband conditions.

It is clear from the results that the EV-VBR performs significantly better than G.729.1 for wideband signals at every bit rate and in all the tests and also outperforms G.729.1 in narrowband conditions with background noise. In clean narrowband speech G.729.1 and EV-VBR are equivalent. Also in all frame erasure cases EV-VBR performs significantly better.

7 CONCLUSIONS

EV-VBR is a work-in-progress codec that shows state-of-the-art quality in all areas of tested conditions. EV-VBR was specifically designed to be robust against frame loss and that is indeed demonstrated by the results. The embedded scalability of the codec is also shown through the results, since in every test, quality increases with the additional layers, despite saturation effects being evident in some tests (1, 3, 6 and 7).

8 REFERENCES

[1] ITU-T Q7/SG12, AH-07-03 Rev.1, "Global Analyses for the Selection Phase for the Embedded-VBR Speech Codec," Geneva, 19 - 23 March 2007

[2] ITU-T Q9/SG16, TD-157/WP2, Annex A of Q9/16 Report, "Terms of Reference for the Embedded VBR (EV) Audio Coding Algorithm", Geneva, April 2006

[3] Ragot, S et al., "ITU-T G.729.1 an 8-32 Kbit/S Scalable Coder Interoperable with G.729 for Wideband Telephony and Voice Over IP," *ICASSP 2007*, Page(s):IV-529 - IV-532

[4] ITU-T SG16, COM16–C199R1, "Extended high-level description of the Q9 EV-VBR baseline codec," Source VoiceAge, Nokia, Geneva, 26 June – 6 July 2007