VOICE QUALITY EVALUATION OF VARIOUS CODECS

Anssi Rämö

Nokia Research Center, Tampere, Finland

ABSTRACT

In this paper a large amount of absolute mean opinion scores obtained within a single listening test are presented. Naïve listeners' preference on different speech signal properties such as mono/stereo and bandwidth was studied. Various codecs were ranked by their subjective voice quality. The listened speech sequences were recorded and selected to represent several realistic stereo audio capture and background noise configurations, where there are either one or several speakers. The number of conditions was selected to be as large as possible to be listenable in a single two hour session. Due to the test size, the results are divided into smaller graphs where interesting comparisons between different conditions can be easily evaluated.

Index Terms— speech coding, voice quality evaluation, listening test, MOS, stereo speech

1. INTRODUCTION

For over a century telephony has relied on narrowband (NB) voice quality, which is barely good enough to transport the most important elements of human speech. Wideband (WB) has been coming to services and devices for a few years, but the final breakthrough has not yet happened. In addition to even wider audio bandwidth such as superwideband (SWB) (see Table 1), improved user experience can be obtained through the introduction of new functionalities such as stereo and binaural coding. Multi-channel audio capture and reproduction has the ability to convey significant amount of information about the caller environment. For example, the audio scene can comprise of the position information of the teleconference participants. The positioning can be either naturally captured or artificially rendered at the teleconference bridge. The spatial audio capabilities can be used to convey the real audio environment of the caller to the listener. For example, grand parents can hear their grandchildren playing in their home environment. This will create much more intimate feeling for the duration of the call than with current telephony.

Abbr.	Meaning	Pass-band	Quality expectation
NB	Narrowband	300 - 3 400Hz	Telephone
WB	Wideband	50 - 7 000 Hz	AM-radio
SWB	Superwideband	50 - 14 000 HZ	FM-radio
FB	Fullband	20 - 20 000Hz	CD- quality

Table 1. Abbreviations used for different signal bandwidths

Overall, people are becoming more aware about quality of experience. For example, high definition television (HDTV) is gaining ground as well as high quality digital photography. Thus, it would be quite natural to follow the trend also in telephony service. The target should not be less than one equivalent to face-to-face conversational quality. This paper shows that current voice codecs capabilities are not good enough for this ultimate target.

This paper is organized as follows. Section 2 explains the used test material and listening test methodology. Section 3 shows various different results obtained from the listening test. Finally some concusions are drawn in section 4.

2. STEREO SPEECH

Stereo speech is a new concept for telecommunications. Initially one might think that capturing stereo is very difficult at least in mobile environment using only a traditional handset. However, stereo capture can be understood a bit more widely than "normal" stereo-image, where the orchestra players can be located from left to right in the stereo image. This is of course the current norm for HiFi-stereo. But for example a stereo image where speech is captured with one microphone near mouth and the environment with low level speech is captured with another microphone pointing outwards. When listened with stereo headset this near/far stereo sounds very pleasant especially when compared to listening monoaurally. Also attaching microphones over a stereo headset provides an almost binaural recording. This kind of thinking extends the speech capture possibilities enormously. Table 2 shows a set of configurations used in this paper to capture "stereo"- speech.

Set	Microphone setup	Arrangement	Background noise
1	ORTF [1]	4 people around a conference table	Quiet studio
2	Mid-Side stereo [1]	Female and male in a quiet room	Quiet with some reverb
3	Handset (near/far)	Male in a car	Car noise and music from CD
4	Near binaural headset	Female walking on sidewalk	Cars passing by, birds singing

Table 2. Sample sets used for listening test

The test material contained female and male voice samples with clean voice and voice in background noise. These samples were recorded in stereo with 48 kHz sampling rate using phonetically balanced sentences.

2.1. Extended Range MOS Test Method

A modified version of the traditional ACR (Absolute Category Rating) MOS [2] method was used for the listening test. The ACR MOS scale was extended to be 9 categories wide. Only the extreme categories (9 "excellent" and 1 "very bad") were defined with verbal description. We have noticed that the 9-scale ACR saturates less easily than the standard 5-scale ACR MOS. In practise this new scale is somewhat between MUSHRA and 5-scale ACR. The assessment is not free sliding, but nine different values still provide listener more ways to discriminate the samples. In practise 9-scale ACR test is also much faster to conduct with naïve listeners than MUSHRA.

2.2. Test Description

The listening test was conducted in Nokia Research Center listening test facilities [3]. The main research question was: How do naïve listeners prefer NB vs WB vs SWB and mono vs stereo speech signals without any preparatory information. 64 naïve listeners took part in the listening test. Each listener evaluated 123 conditions with 4 different stereo or mono voice samples from all scenarios shown in Table 2. Thus each listener scored 492 individually processed samples in random order. Since each sample took about 10 seconds to listen and evaluate, and there are mandatory comfort breaks every twenty minutes, the listening took about two hours per listener. Each condition obtained 256 votes. In order to have some initial scale to the listeners, the test started with 16 introductory (practice) samples, which represented the full scale of the conditions. These preparatory test results were omitted from the final results.

The samples were summed (from stereo to mono), and downsampled with high quality filters to be used as lower quality references or as input signals for various codecs. Since the test contains both mono and stereo samples intermixed, stereo headphones (Sennheiser HD-580) were used for the listening test. Diotic listening was conducted for mono conditions.

3. RESULT ANALYSIS

There are so many MOS scores that normal bar graphs or numerical tables are almost impossible to read so a new method of representing listening test results is introduced. Sub-sets of codec and/or reference conditions are collected to a X-Y line graph, where bullets point to individual MOS results and interpolated line connects the bullets, when relevant scalable codec or codec family result is shown. 95% confidence intervals are presented with dotted lines so the reader may consider how relevant the quality difference actually is. On the left side of the table MOS scale is shown. On the bottom either bitrate or the bandwidth is shown. All results are represented in linear scale, however minimum and maximum vary with each figure, in order to show only the most interesting area of the MOS scores.

3.1. Direct Mono and Stereo Results

Mono and stereo references in Figure 1 show that there is a drastic increase in the user experience when going from narrowband to wideband audio representation, and again a significant improvement when going from wideband to superwideband. Widening the audio bandwidth further from superwideband to fullband has less impact and the perceived quality does not improve significantly. The fullband stereo quality reaches MOS score of 7.5. One interesting note is that SWB mono (MOS 7.25) is preferred over WB stereo (MOS 6.79). This means that for telephony applications it is more important to increase signal bandwidth before introducing stereo or binaural presentation. Note also that NB mono receives a MOS score of 4.79 in this 9-scale ACR test, which is quite typical value for NB direct also in 5-scale ACR MOS tests.

3.2. 3GPP Codec Voice Quality

Fig. 2 shows how AMR and AMR-WB scale in quality with increasing bitrate. Both AMR and AMR-WB are based on the same ACELP paradigm, thus the increased bandwith is the main reason for the improvement in voice quality. As can be seen AMR-WB at 8.85 kbit/s provides better quality than direct NB. AMR-WB at 12.65 kbit/s appears to be the sweet spot bitrate wise. It is a knee-point after which

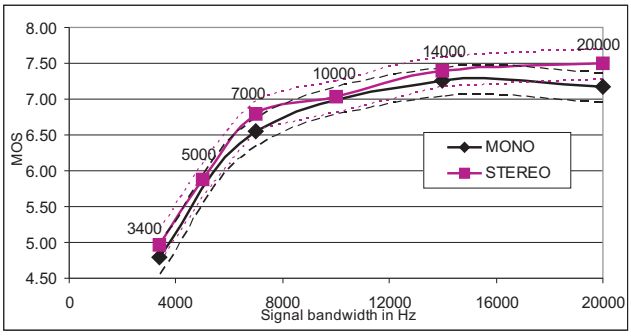

Fig. 1. Scaling of speech quality with increasing signal bandwidth

the quality increase with bitrate becomes less significant. Previous generation codecs such as GSM-HR (5.6 kbit/s Half Rate) and GSM-FR (13.0 kbit/s Full Rate) are also included to the same figure. Results really show that the newer generation speech codecs provide significantly improved speech quality at similar bitrate or reduced bitrate for the same quality.

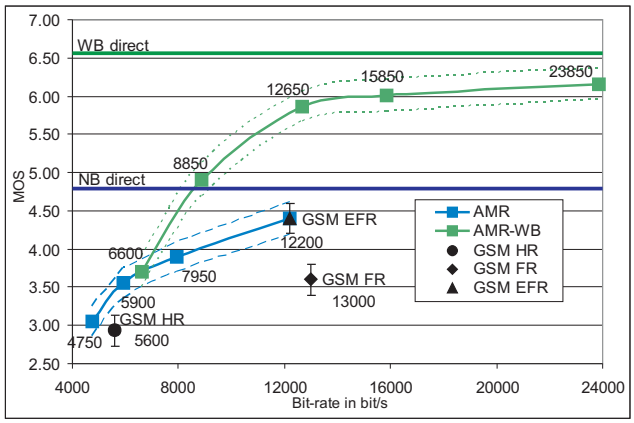

Fig. 2. Older 3GPP codecs compared to AMR-WB

3.3. Other Narrowband Codecs Compared to AMR

Very low bitrate codecs such DoD MELP, MPEG4 HXVC and LPC10 were also included into the test as historically interesting low quality references. As can be seen in Fig. 3 their voice quality is not useable for modern telephony.

iLBC is a speech codec developed to be patent free and to work well in VoIP environment with frame erasures [4]. iLBC is more recent than AMR or EFR, but its performance still lags behind AMR at similar bit-rates [5]. iLBC supports both 20 ms (13.3 kbit/s) and 30 ms (15.2 kbit/s) frame sizes. We tested 30 ms variant and its voice quality is statistically same as AMR at 7.95 kbit/s. Speex supports NB, WB and SWB bandwidths with many bitrates. Speex's NB voice quality is significantly below that of AMR at similar bitrates.

3GPP2 has standardized several speech codecs such as SMV, EVRC and VMR-WB for conversational servives. However, currently only EVRC-line of the codecs is in use. Both narrowband codecs EVRC-A and EVRC-B (Fig. 3) and wideband codec EVRC-C (Fig. 5) obtained results comparable to AMR. Narrowband EVRC codecs (A and B) are about the same quality as AMR at 5.15 kbit/s or 5.9 kbit/s. EVRC's actual bitrate is hard to measure, due to its

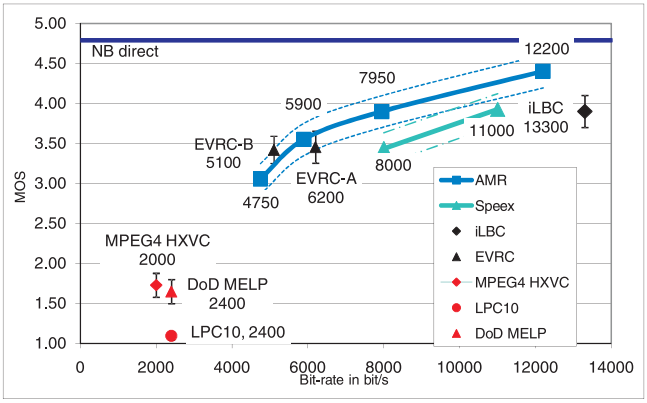

Fig. 3. NB low bitrate codecs compared to AMR

source adaptive nature. However, with active speech and normally used operation points narrowband EVRC codecs use 5- 6 kbit/s. All EVRC codecs have strong noise suppression integrated, this causes somewhat synthetic voice quality when the noise is reduced.

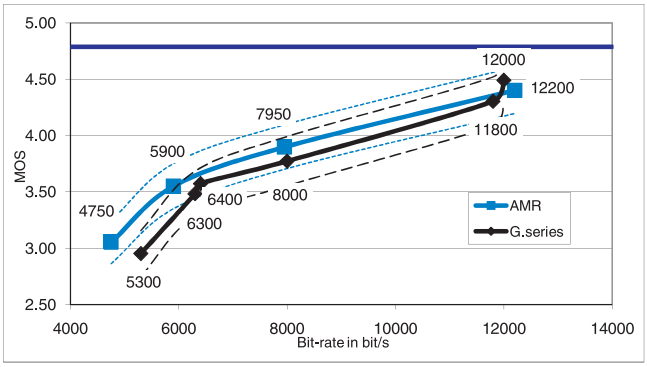

Fig. 4. ITU-T G.723.1, G.729, G.729.1 and AMR compared

The comparison of AMR and ITU-T G.-series NB codecs (Fig. 4) shows that AMR codec modes are slightly better than the ITU-T codecs G.723.1, G.729 and G.729.1 at similar bit-rates. One explanation for difference is that G.729 has 10 ms, G.723.1 30 ms and AMR 20 ms frame size. The ITU-T G.-series codec results consist of G.723.1 (5.3 kbit/s, 6.3 kbit/s), G.729 (8 kbit/s), G.729 annex D (6.4 kbit/s), G.729 annex E (11.8 kbit/s) and finally the embedded G.729.1 (12 kbit/s).

3.4. Wideband Codecs Compared

When AMR-WB is compared against embedded G.718 and G.729.1 codecs and G.722.1 audio codec we can see that G.729.1 is about one layer (8 kbit/s) behind G.718's quality like in [6]. The main difference between G.729.1 and G.718 is that G.729.1 core layer is compatible with older narrowband codec G.729. G.718 on the other hand supports wideband signals already at the base bitrate of 8 kbit/s [7]. G.729.1 supports wideband signals starting from 14 kbit/s. AMR-WB compares very well against these newer embedded codecs. Embeddedness and robustness causes some performance penalty at the 12- 16 kbit/s bitrates for G.718.

MLT (Modulated Lapped Transform) based G.722.1 is surprisingly good also with these relatively noisy and realistic speech signals. It is very close to G.718 at 16 kbit/s. However with very clean speech G.722.1 has some audible transform coding artifacts. Also

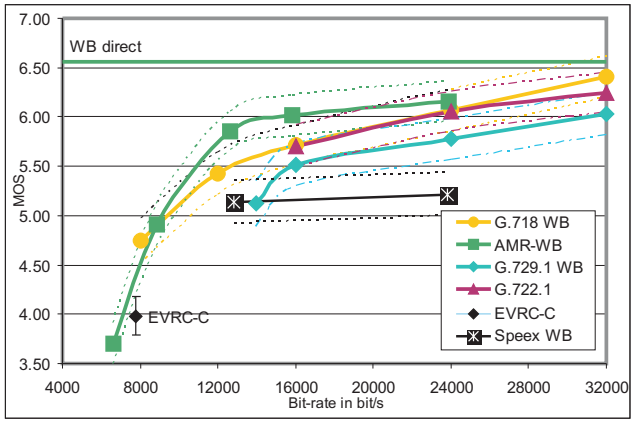

Fig. 5. G.718, G.729.1, G.722.1, Speex, EVRC-C and AMR-WB compared

the lowest bitrate of the G.722.1 is a bit on the high side for being useful as a wireless speech codec.

With EVRC-C (WB) the situation is a bit worse. Its bitrate is not simple to measure, but in general it uses with active speech most of the time with its highest mode (8.55 kbit/s) and the average bitrate was around 7.5 kbit/s in our testing. At that bitrate it achieves quality that is somewhat better than AMR-WB at 6.6 kbit/s, but it is significantly worse than AMR-WB at 8.85 kbit/s. This means that EVRC-C's absolute wideband voice quality is too low, and thus it is not useable as a generic wideband speech codec. However, it compares well with narrowband codecs at similar bitrates. It is statistically equivalent to AMR at 7.95 kbit/s.

Speex has serious voice quality problems with WB bandwidths.

3.5. Superwideband Mono Codecs Compared

Currently there exists only one standardized superwideband mono codec (G.722.1C) that is optimized for real time usage. AMR-WB+ was also included to the test, althought it has too high delay for realtime telephony [8]. Soon to be standardized ITU-T G.718 / G.729.1 superwideband extension was included into listening test with G.729.1 core [9]. Speex also provides SWB support. Results in Fig. 6 indicate that quite high quality (better than WB direct) can be achieved with 24 kbit/s with AMR-WB+. Results also indicate that switching from wideband to superwideband seems to be feasible at bitrates around 16- 24 kbit/s with these AMR family codecs. Direct SWB mono quality is not achieved with AMR-WB+ even at its maximum tested mono bitrate of 32 kbit/s, which indicates that perfect SWB mono speech would require even higher bitrate. A new standardization study is already on going in 3GPP where one of the quality targets is to have SWB with good quality and low delay at around 20 kbit/s.

3.6. Stereo Quality Scalability

Currently there is no standardized stereo optimized low delay voice codec. AMR-WB+ is not a conversational codec, but it provides an excellent quality target for all high bandwidth voice codecs, since it is optimized for both speech and music signals and supports parametric stereo. Figure 7 indicates bitrate point at which one should switch from SWB mono speech to SWB/FB stereo. With AMR-WB+ the transition from mono to stereo seems to happen when the bitrate is above 32 kbit/s. However, both mono and stereo qualities

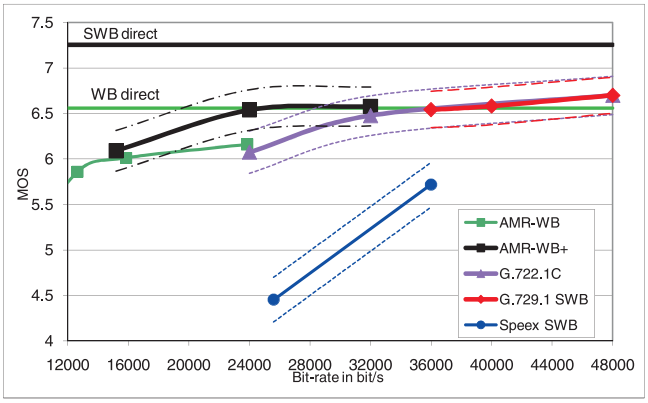

Fig. 6. G.729.1 SWB annex, G.722.1C, Speex and AMR-WB+ superwideband codecs compared. AMR-WB is also shown for comparison

are still quite far away from the direct quality. Currently on going 3GPP standardization study also includes a possibility to require a low delay stereo or spatial coding capabilities for the future EVS (Enhanced Voice Services) codec.

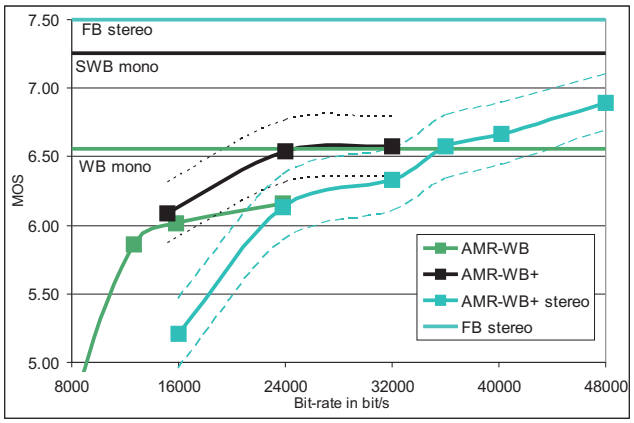

Fig. 7. AMR-WB and AMR-WB+ stereo quality scalability

4. CONCLUSIONS

As can be expected also naïve listeners prefer wider signal bandwidths over narrower and stereo over mono. Likewise, higher bitrate correlates with better voice quality. From the results it appears that 3GPP standardized AMR-line of codecs provides the most efficient coding solutions for narrowband (AMR), wideband (AMR-WB) and superwideband (AMR-WB+) qualities. ITU-T standardized codecs such as G.729, G.723.1, G.729.1, G.718, G.722.1C are, however, quite close to AMR-line in quality at respective bitrates. Codecs done without thorough standardization effort like Speex and iLBC offer significantly reduced efficiency, probably due to much lesser optimization, listening tests and IPR free design. EVRC-line of codecs suffers from too low bitrates and heavy noise reduction. The most interesting result is that with superwideband significant voice quality improvements can be brought to consumers at reasonable bitrates of around 20 kbit/s. Likewise stereo is promising at bitrates of 30 kbit/s or more with SWB and fullband signals. Figure 8 and Table 3 summarize, what kind of quality can be expected at each bitrate with different signal bandwidths. Also the optimal transition points from category to next are approximated based on the listening test results.

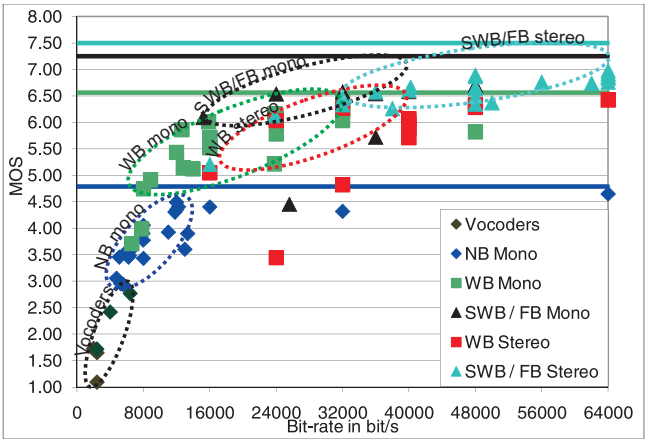

Fig. 8. All results grouped into six categories

Codec category	Efficient bitrate range	MOS range	First useable bitrate
Vocoders	1- 4 kbit/s	< 3.0	< 1 kbit/s
NB mono	5- 13 kbit/s	3.0- 4.5	4 kbit/s
WB mono	8- 32 kbit/s	4.5- 6.5	8 kbit/s
SWB /FB mono	16- 48 kbit/s	6.0- 7.0	16 kbit/s
WB stereo	16- 48 kbit/s	5.5- 6.5	never
SWB / FB stereo	24- 80 kbit/s	6.5- 7.5	32 kbit/s

Table 3. All tested codecs summarized to six categories

5. REFERENCES

[1] Wikipedia article, "Stereophonic sound," `http://en.wikipedia.org/wiki/Stereophonic_sound`.

[2] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," 1996.

[3] M. Kylliäinen et al., "Compact high performance listening spaces," in *Proc. of Euronoise*, Italy, 2003.

[4] S.V. Andersen et al., "ilbc - a linear predictive coder with robustness to packet losses," in *Proc. of the IEEE Speech Coding Workshop*, Tsukuba, Japan, 2002.

[5] A. Rämö and H. Toukomaa, "On comparing speech quality of various narrow- and wideband speech codecs," in *Proc. of ISSPA*, Sydney, Australia, 2005.

[6] A. Rämö et al., "Quality evaluation of the g.ev-vbr speech codec," in *Proc. of ICASSP*, Las Vegas, NV, USA., 2008.

[7] M. Jelinek et al., "G.718: A new embedded speech and audio coding standard with high resilience to error-prone transmission channels," *IEEE Communications Magazine*, vol. 46, no. 10, pp. 117 – 123, October 2009.

[8] R. Salami et al., "Extended amr-wb for high-quality audio on mobile devices," *IEEE Communications Magazine*, vol. 44, pp. 90–97, May 2006.

[9] M. Tammi et al., "Scalable superwideband extension for wideband coding," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.