Opus Testing

Raymond Chen, Timothy Terriberry,
Jan Skoglund, Gregory Maxwell,
Hoang Thi Minh Nguyet

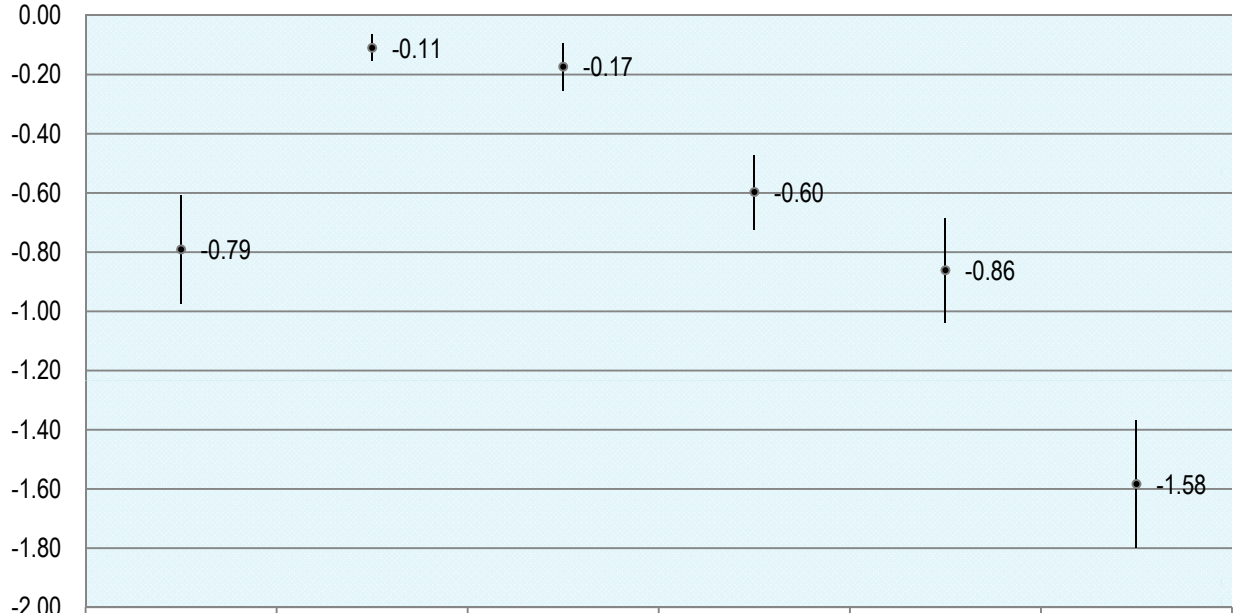
Broadcom Listening Test

- **In December 2010, Broadcom conducted an ITU-R BS.1116-style subjective listening test comparing different configurations of the CELT-only mode of the IETF Opus codec along with MP3 and AAC_LC.**
- **17 listeners participated.**
- **10 diverse full-band audio tracks with 44.1 kHz sampling used:**
  - 2 pure speech
  - 2 vocal
  - 2 solo instruments
  - 1 rock-and-roll
  - 1 pop
  - 1 classical orchestra
  - 1 jazz
- **The goal was to compare different configurations of the latest CELT at that time (version 0.9.1, as of November 2010), and compare them with the two reference codecs MP3 and AAC_LC at 96 kbps.**

Codec Conditions Tested

- **6 codec conditions were tested, all with constant bit-rate (CBR):**
  - Reference 1: MPEG-1, layer 3 (MP3) codec at 96 kbps
  - Reference 2: AAC Low Complexity profile (AAC_LC) codec at 96 kbps
  - 96 kbps CELT 0.9.1 with pitch prefilter/postfilter on, 23 ms frame size
  - 96 kbps CELT 0.9.1 with pitch prefilter/postfilter on, 5.8 ms frame size
  - 96 kbps CELT 0.9.1 with pitch prefilter/postfilter off, 5.8 ms frame size
  - 64 kbps CELT 0.9.1 with pitch prefilter/postfilter on, 5.8 ms frame size
- **It was later realized that CELT 0.9.1 was only optimized for 48 kHz and not for 44.1 kHz sampling.**
- **After the test, CELT was optimized for 44.1 kHz in addition to 48 kHz.**
- **After re-processing the audio test files with 44.1 kHz-optimized CELT, there was noticeable audio quality improvement.**
- **Later, CELT went through further audio quality enhancements.**
- **As a result, the CELT test scores (on a 5-point scale) presented on the next slide can be considered lower bounds of the scores of the current version (0.11) of CELT.**

CELT Listening Test Result (Audio Quality Degradation Relative to Uncoded Original)



	MP3_96kbps	AAC_LC_96kbps	CELT_96kbps_on_23ms	CELT_96kbps_on_5.8ms	CELT_96kbps_off_5.8ms	CELT_64kbps_on_5.8ms
95% CI-L	-0.61	-0.06	-0.09	-0.47	-0.68	-1.37
95% CI-U	-0.97	-0.16	-0.25	-0.72	-1.04	-1.80
• Mean	-0.79	-0.11	-0.17	-0.60	-0.86	-1.58

- "on" or "off" in the codec conditions above indicates whether the pitch prefilter and postfilter were on or off.
- 95% CI-L and 95% CI-U indicate the lower and upper bounds of the 95% confidence interval.

Conclusion

- 96 kbps CELT 0.9.1 with a frame size of 23 ms was rated significantly better than 96 kbps MP3.
- 96 kbps CELT 0.9.1 with a frame size of 23 ms was rated roughly equivalent to 96 kbps AAC_LC; with 44.1 kHz optimization and further enhancements, CELT 0.11 is expected to be no worse than AAC_LC.
- 96 kbps CELT 0.9.1 with a frame size of 5.8 ms was rated slightly better than 96 kbps MP3, even though its codec delay is much lower than that of MP3.
- The pitch prefilter/postfilter method provided statistically significant audio quality improvement for the 96 kbps CELT 0.9.1 with a frame size of 5.8 ms.

Hydrogen Audio

- Test set up and run by Igor Dyakonov on the online forum http://www.hydrogenaudio.org
  - Members are all audio enthusiasts, few are developers
  - Dedicated to repeatable, blind testing
    - TOS #8: “All members that put forth a statement concerning subjective sound quality, must -- to the best of their ability -- provide objective support for their claims. Acceptable means of support are double blind listening tests (ABX or ABC/HR) demonstrating that the member can discern a difference perceptually, together with a test sample to allow others to reproduce their findings. Graphs, non-blind listening tests, waveform difference comparisons, and so on, are not acceptable means of providing support.”
- http://listening-tests.hydrogenaudio.org/igorc/Public Multiformat Listening Test @ 64kbps.htm

Goals & Conditions

- Compare Opus to high-latency codecs
- Give them every advantage
  - All test codecs used VBR, unconstrained when available
- Low enough rate for listeners to give valid results without golden ears
  - 67-68 kbps stereo averaged over a large collection of CDs
  - Rates on individual files varied greatly (except for Opus)
- Broad coverage
  - For music, sample variety more important than more listeners
  - 30 samples
    - Mostly music, plus German, French speech, English singing
    - These are common sample sets used by HA for some time

Setup (1)

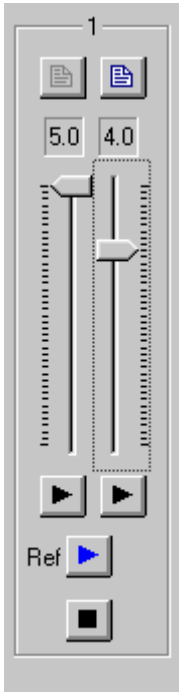
- Four codecs
  - CELT 0.11.2 [22.5ms latency, 67.5 kbps, VBR, complexity 10]
  - Vorbis (AoTuV 6.02 Beta) [-q 0.1]
  - HE-AAC (Apple QuickTime 7.6.9) [64 kbps constrained VBR, high quality]
    - Unconstrained VBR is buggy; constrained performs better
  - HE-AAC (Nero 1.5.4.0) [-q 0.245]
- Low anchor
  - AAC-LC (iTunes 7.6.9) [48 kbps, CBR]

Setup (2)

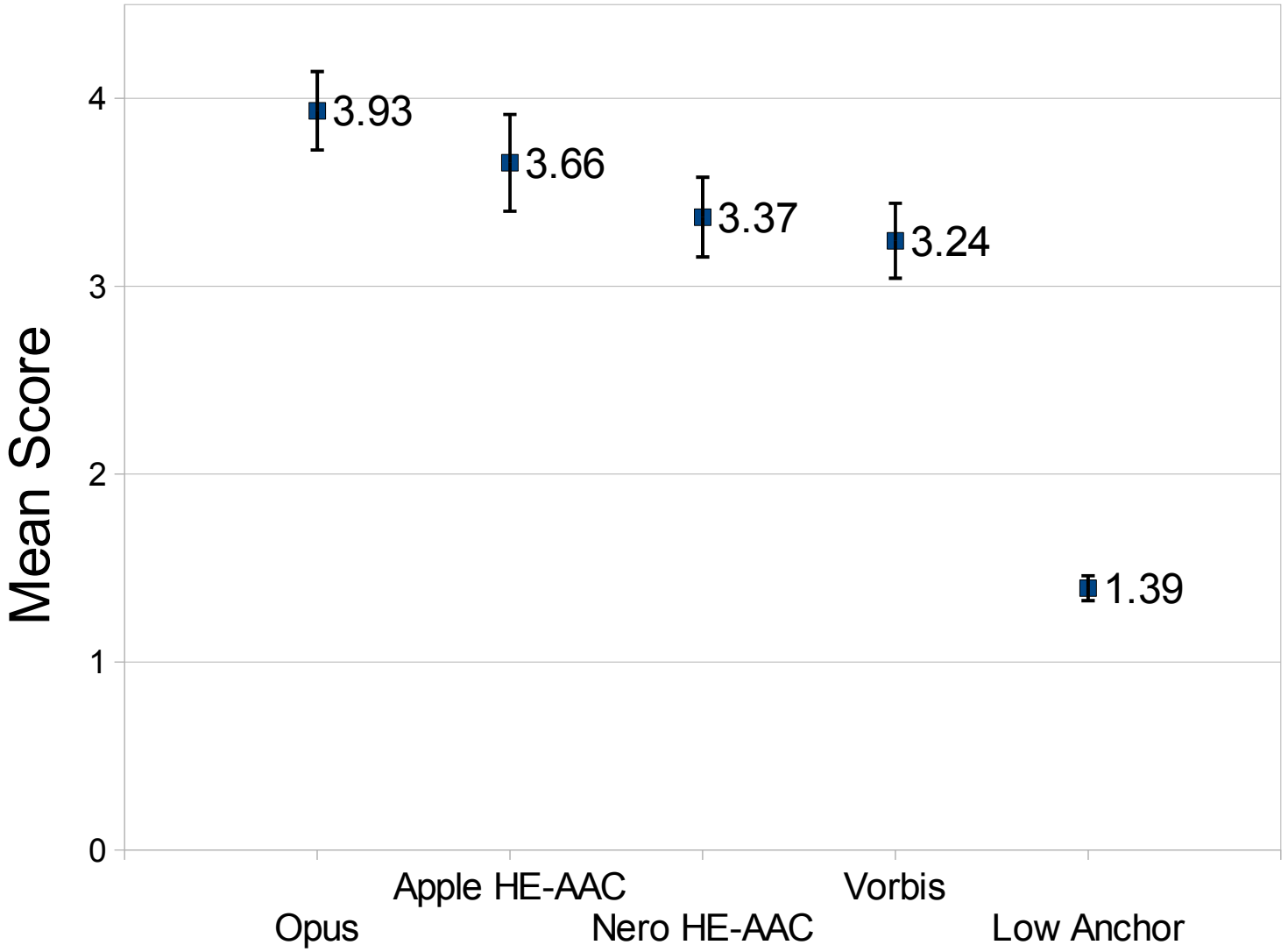
- Resampling
  - Opus input resampled from 44.1 kHz to 48 kHz
  - Non-Opus codecs run at 44.1 kHz, output resampled to 48 kHz
  - Used sox “very high” quality, plus noise-shaping dither
  - All samples checked by Igor to ensure no aliasing or other issues introduced
- Renormalizing: Used ABC-HR’s (unweighted RMS level)
  - Apple reduces volume, easily ABX’able without renormalizing
- Time alignment: ABC-HR’s correction applied
  - The HE-AAC encoders introduce a few ms shift

Methodology

- ABC/HR tests: http://ff123.net/abchr/abchr.html
  - A result where the reference was ranked or the low anchor was not ranked, is INVALID
  - A listener gets one chance to re-test an INVALID sample (must regrade all codecs)
- For these results, only included listeners who rated all 30 samples (7 listeners)
  - 7*30*5 = 1050 individual rankings
- Test results encrypted to prevent casual tampering before submission
- Test ongoing (current deadline April 10th)
  - Raw results will be posted so you can run your own analysis



Overall Results

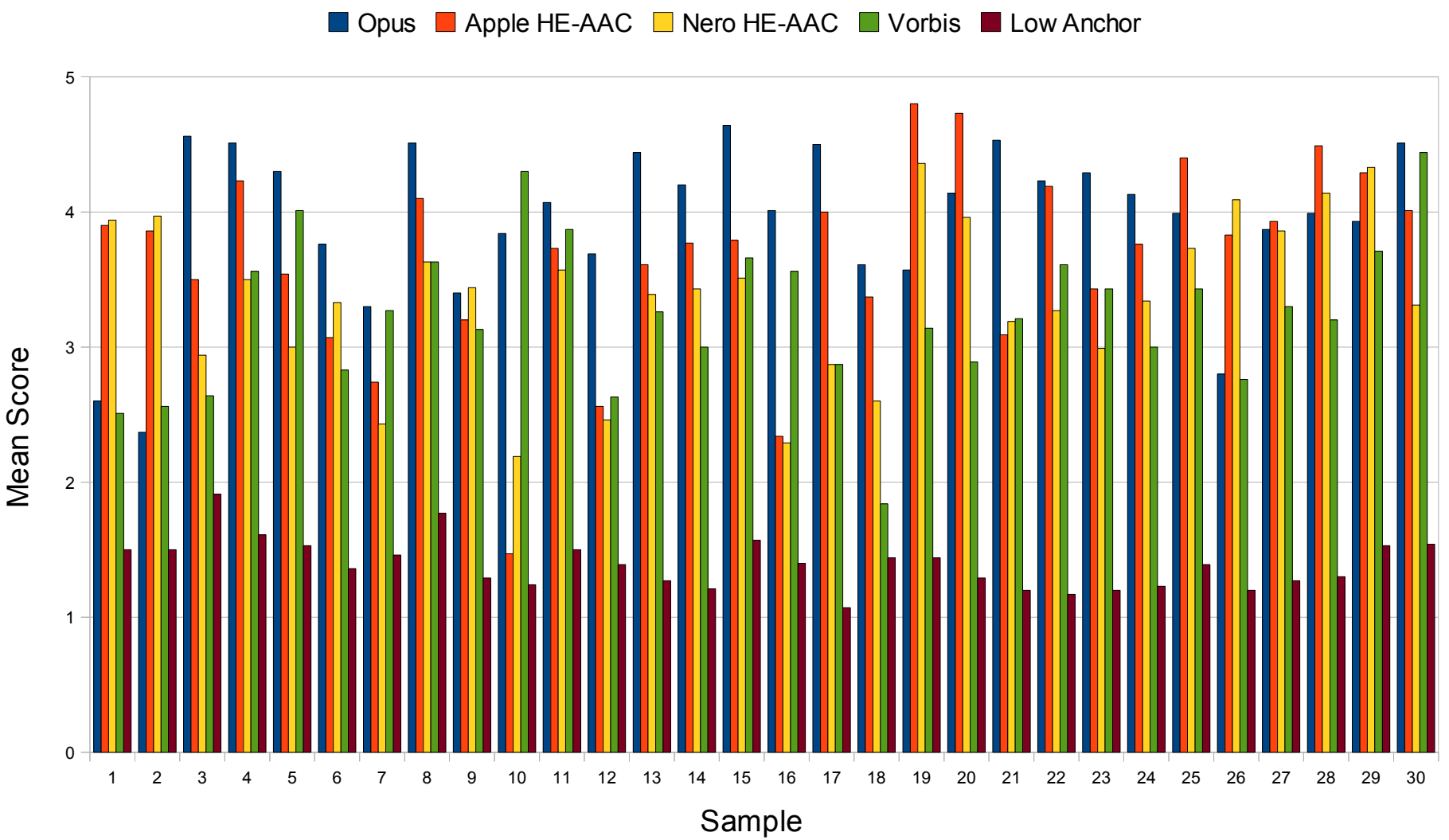


Pairwise Comparisons with Significance Level

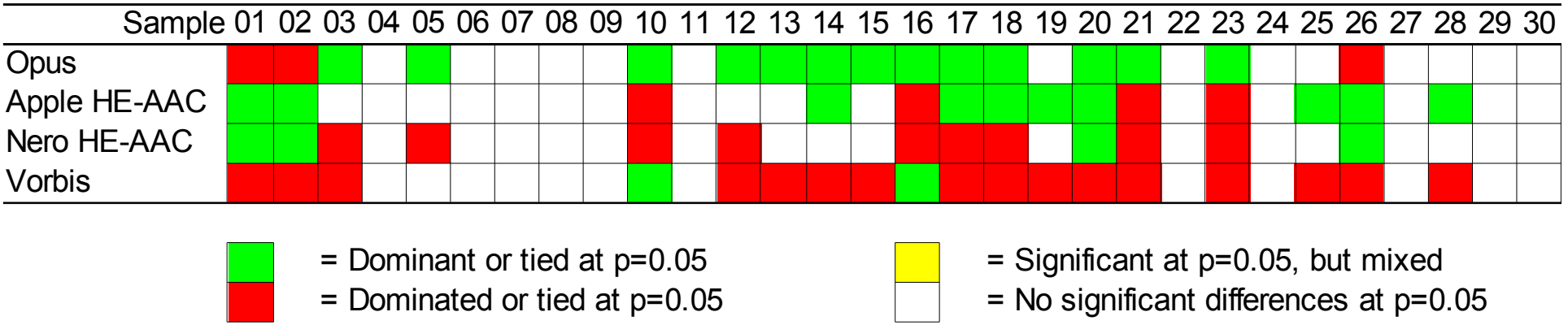
	Opus	Apple HE-AAC	Nero HE-AAC	Vorbis	Low Anchor
Opus		0.000317	0.000000	0.000000	0.000000
Apple HE-AAC	0.000317		0.000000	0.000006	0.000000
Nero HE-AAC	0.000000	0.000000		0.126480	0.000000
Vorbis	0.000000	0.000006	0.126480		0.000000
Low Anchor	0.000000	0.000000	0.000000	0.000000	

- Permutation tests (avoids assuming normality)
  - Two-sided, N=1,000,000 samples
  - Correction for multiple tests applied

Per-Sample Results



Per-Sample Significance (Condorcet winners)



- Do all pairwise comparisons (excluding low anchor)
  - All codecs beat or tied the low anchor
- Highlight codecs which were
  - Significantly better (worse) than at least one other
  - And not significantly worse (better) than any of the others

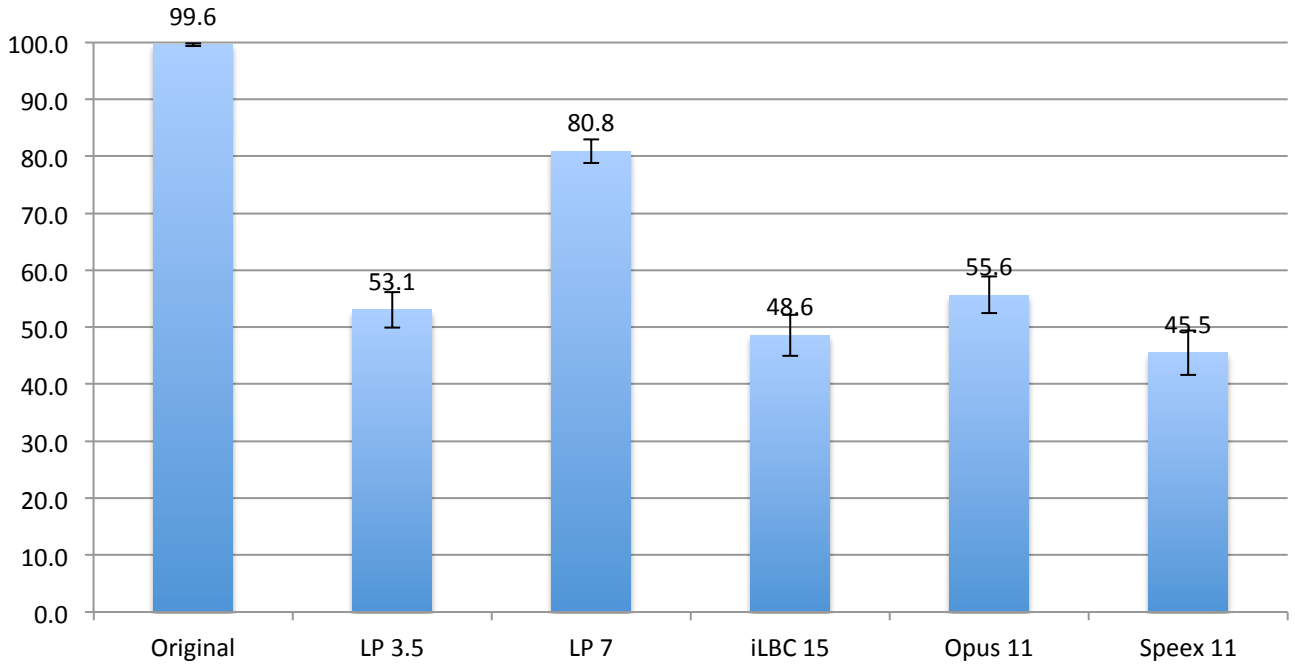
Introduction

- Three MUSHRA-type tests performed in March 2011 at Google
- Both trained and untrained listeners
- Tests presented on Windows PC with headphones

Test 1 – Narrowband Mono Speech

- 4 different male and 4 different female speakers
- Reference files sampled at 48 kHz in low background noise
- 2 anchors: lowpass-filtered at 3.5 kHz and 7.0 kHz
- 17 listeners, no post-screening
- 3 narrowband codecs, all using 20 ms frames
  - iLBC at 15.2 kbps, constant bit rate
  - Speex NB at 11 kbps, constant bit rate
  - Opus NB at 11 kbps, variable bit rate

Overall Results – Narrowband Speech

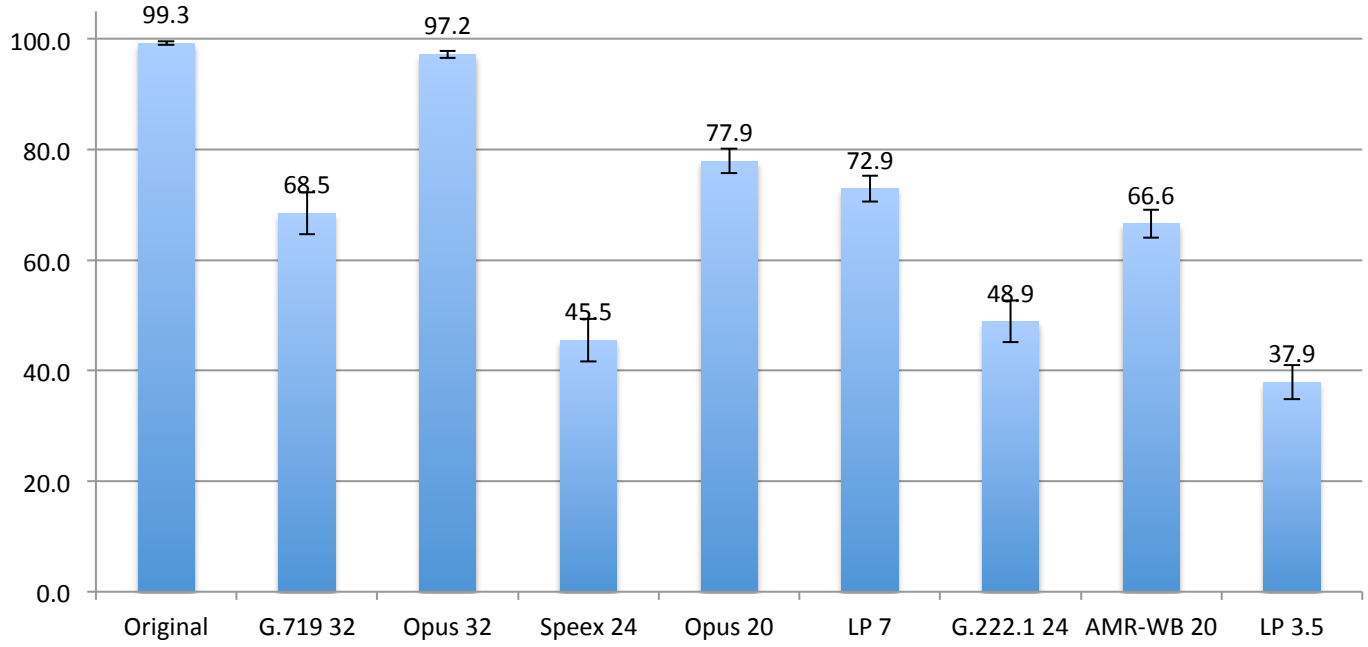


Opus at 11 kbps is better than iLBC at 15 kbps and Speex at 11 kbps

Test 2 – Wideband and Fullband Mono Speech

- 4 different male and 4 different female speakers
- Reference files sampled at 48 kHz in low background noise
- 2 anchors: lowpass-filtered at 3.5 kHz and 7.0 kHz
- 17 listeners, no post-screening
- 4 wideband codecs, all using 20 ms frames
  - G.722.1 at 24 kbps, constant bit rate
  - Speex WB at 23.8 kbps, constant bit rate
  - Opus WB at 19.85 kbps, variable bit rate
  - AMR-WB at 19.85 kbps, constant bit rate
- 2 fullband codecs, both using 20 ms frames
  - G.719 at 32 kbps, constant bit rate
  - Opus FB at 32 kbps, constant bit rate

Overall Results - Fullband and Wideband Speech



Opus at 32 kbps is almost transparent

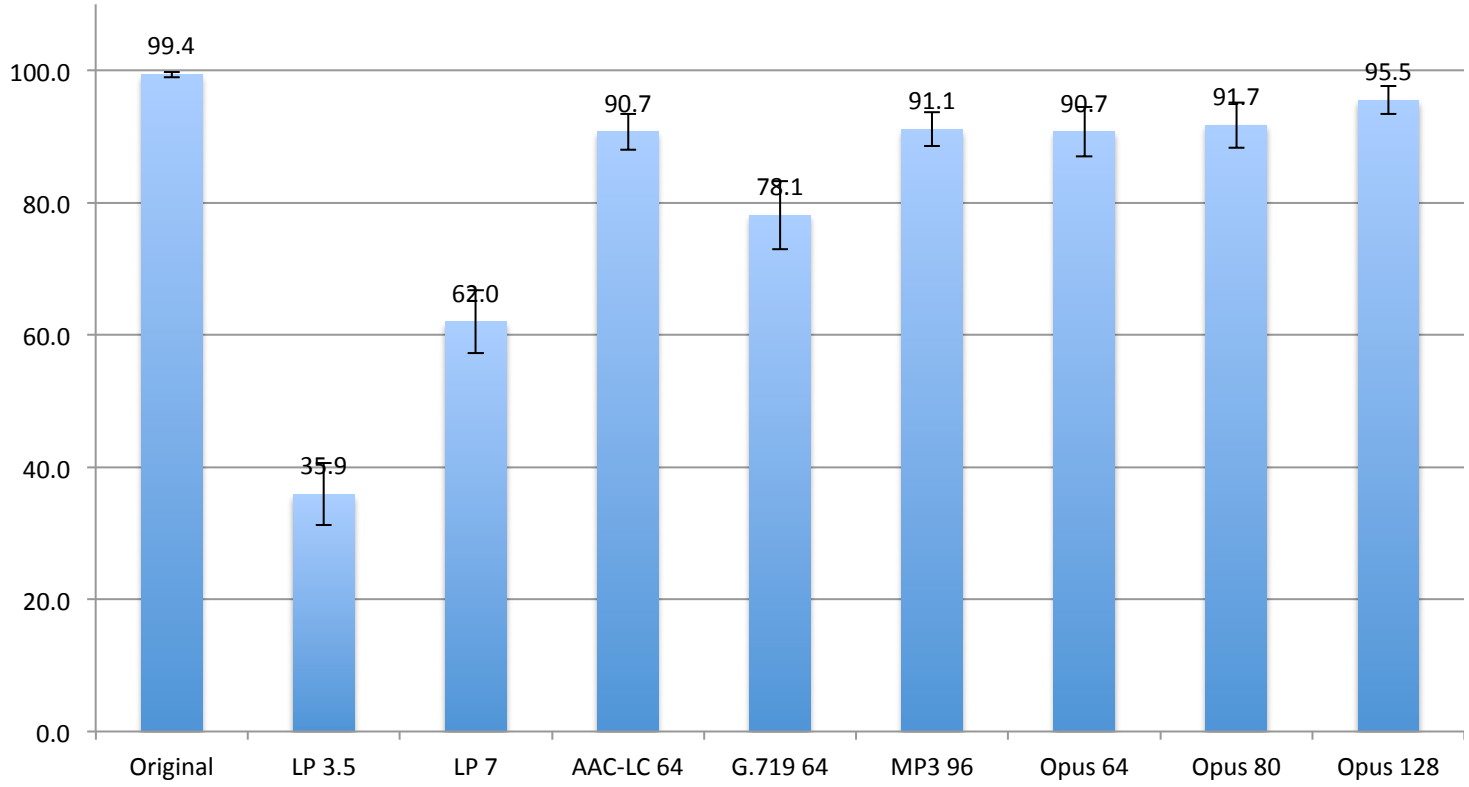
Opus at 20 kbps is better than LP filtered speech at 7 kHz

Test 3 –Fullband Stereo Music

- 10 stereo music files
  - Rock/R&B (Boz Scaggs)
  - Soft rock (Steely Dan)
  - Rock (Queen)
  - Jazz (Harry James Orchestra)
  - Classical (Purcell String Piece)
  - Electronica (Matmos)
  - Piano (Moonlight Sonata)
  - Vocals (Suzanne Vega)
  - Glockenspiel
  - Castanets

- Reference files sampled at 48 kHz and 44.1 kHz
- 2 anchors: lowpass-filtered at 3.5 kHz and 7.0 kHz
- 9 listeners, no post-screening
- 6 codecs
  - AAC-LC (Nero) at 64 kbps, 21 ms frame size, constant bit rate (bit reservoir)
  - G.719 at 64 (2 x 32) kbps, 20 ms frame size, constant bit rate
  - MP3 (Lame) at 96 kbps, <100 ms, constant bit rate
  - Opus at 64 kbps, 20 ms frame size, constrained variable rate
  - Opus at 80 kbps, 10 ms frame size, constrained variable rate
  - Opus at 128 kbps, 5 ms frame size, constrained variable rate

Overall Results Fullband Stereo Music



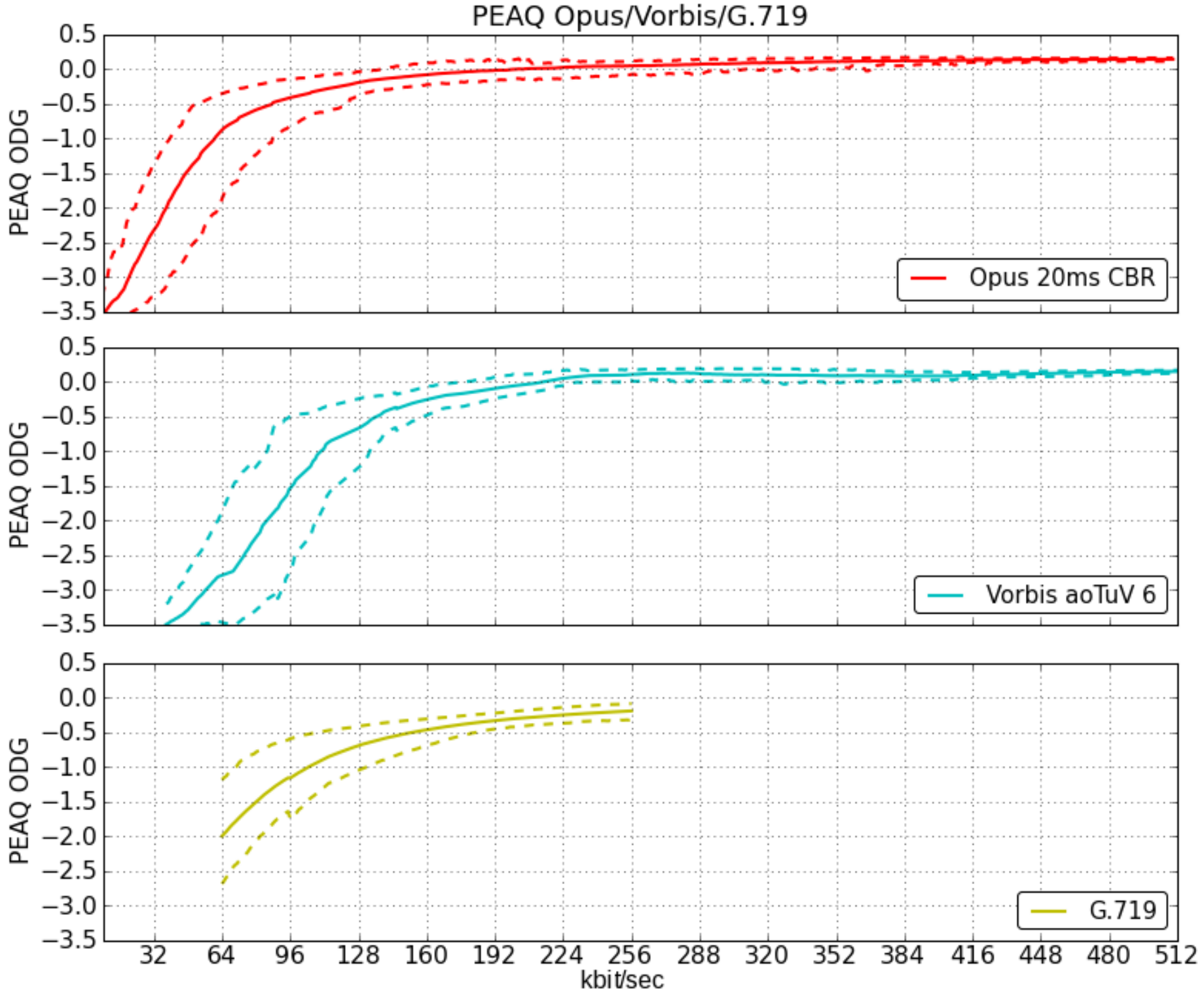
Opus (at 64 kpbs/20ms, 80 kbps/10 ms, and 128 kbps/5 ms) is

- equal in quality to MP3 at 96 kbps
- equal in quality to AAC-LC at 64 kbps
- better than G.719 at 64 kbps

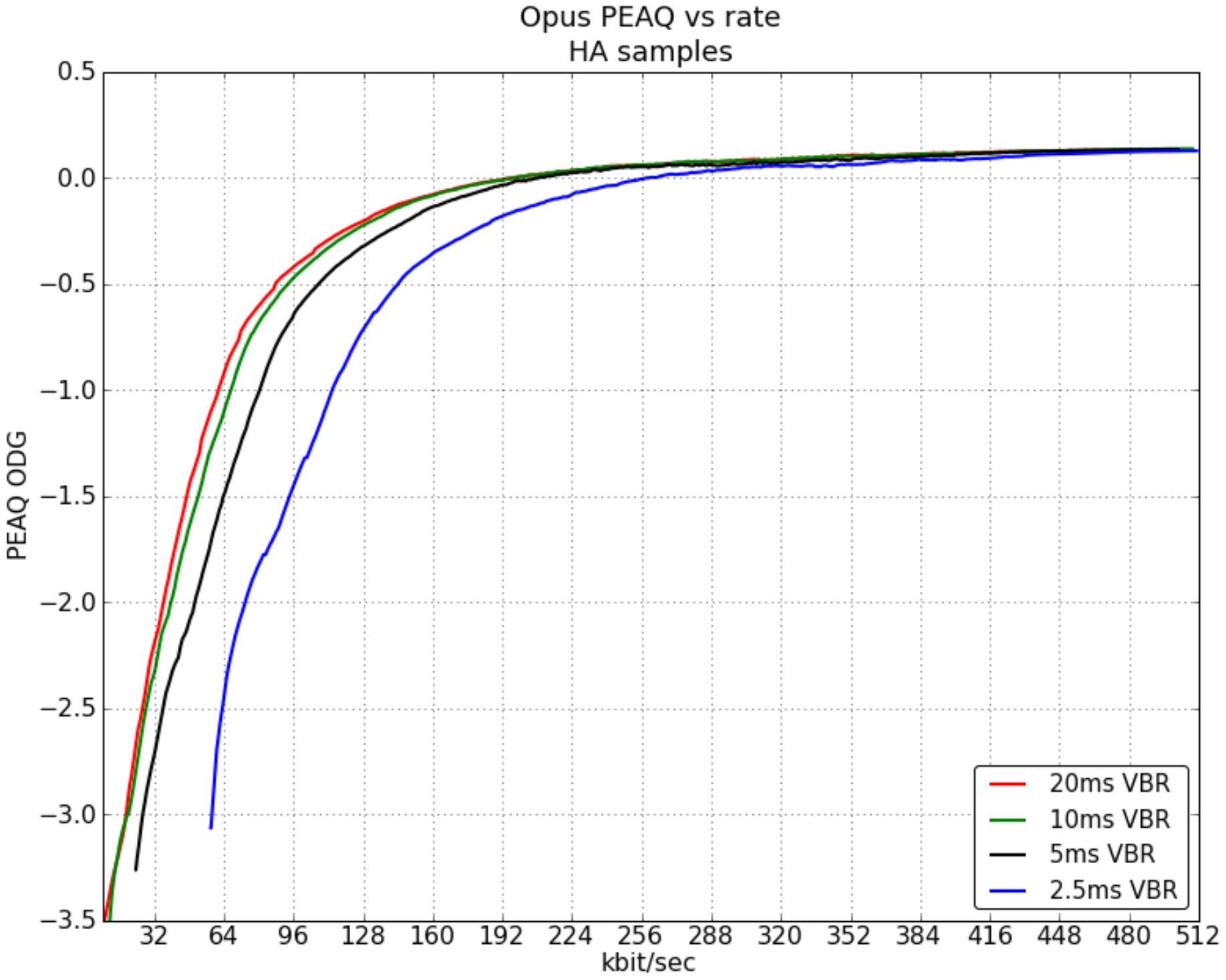
Opus objective validation

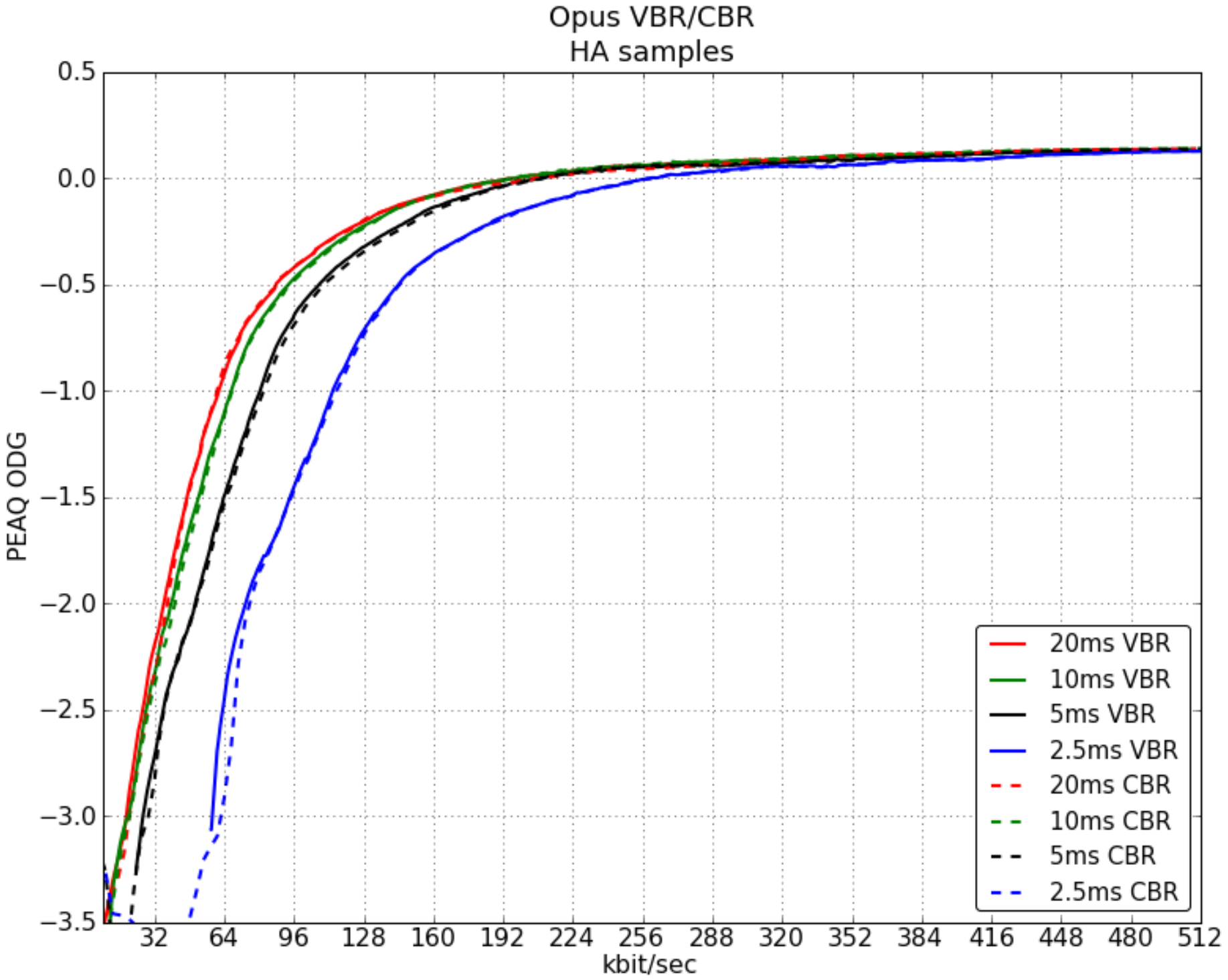
- Subjective testing is the gold standard
  - But… N modes * M rates * Z samples
  - Subjective testing isn't a option for all that
- Machine testing
  - Scales well
  - Sometimes more sensitive than subjective
- Dumb metrics aren't very useful
- Better tools apply psychoacoustics
- We've continually used objective testing during development.

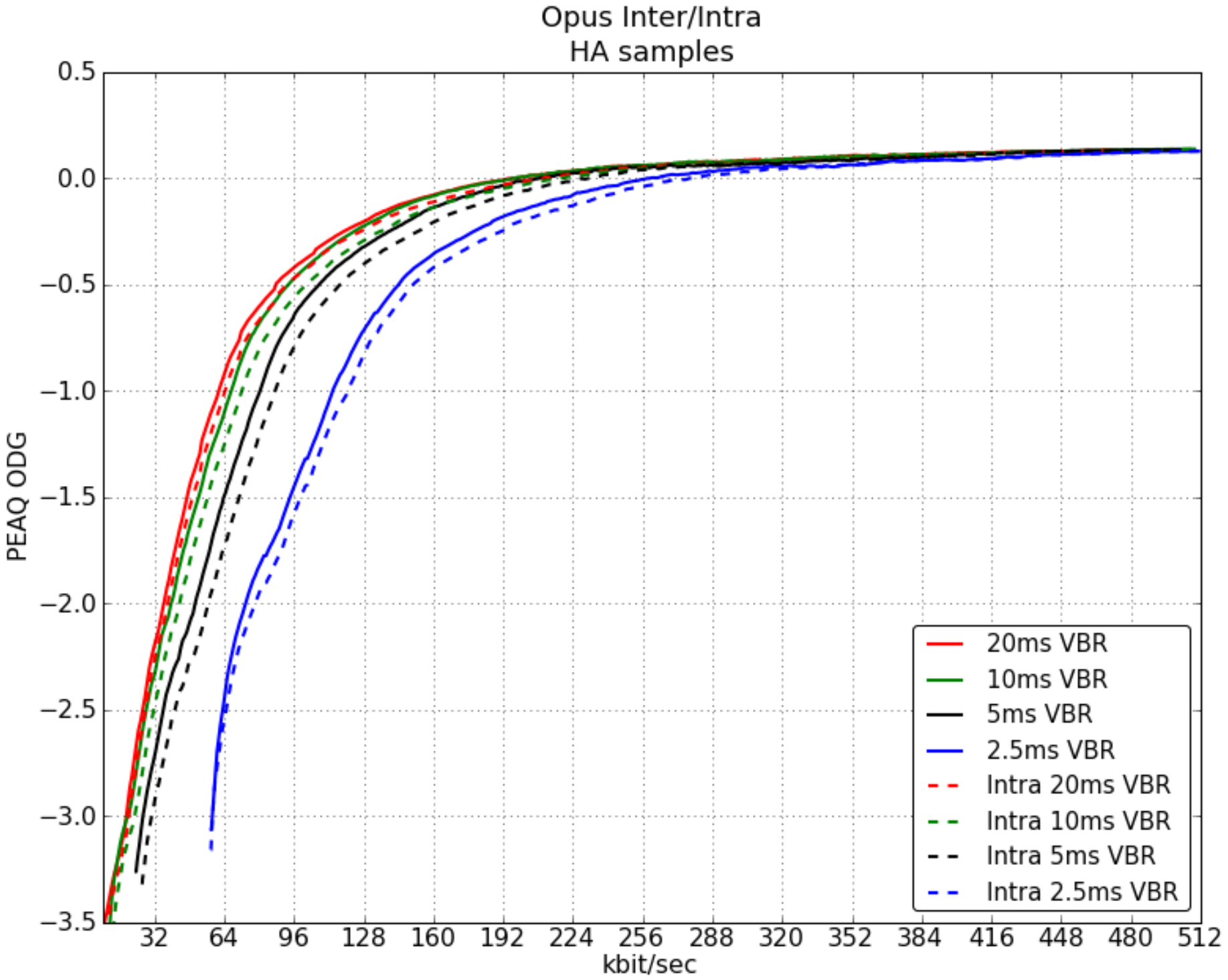
- These results use PEAQ basic
  - McGill University's AFsp PQevalaudio
- Other codecs provided to show the shape of their PEAQ results
  - Don't compare across multiple codecs
- The Hydrogen-audio samples were used
- 2ch, FB, 48kHz, 2.5/5/10/20ms, Intra, VBR/CBR
- For opus, these results reflect:
  - 4 frames sizes * 4 settings * 504 rates * 30 samples * 32 overlaps = 7,741,440 cases tested
  - **~2.5 years of audio**



Note: The PEAQ results are probably not very representative of the real differences between the codecs.


Opus PEAQ vs rate
HA samples
PEAQ ODG
kbit/sec
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−3.5
32
64
96
128
160
192
224
256
288
320
352
384
416
448
480
512
20ms VBR
10ms VBR
5ms VBR
2.5ms VBR


Opus VBR/CBR
HA samples
PEAQ ODG
kbit/sec
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−3.5
32
64
96
128
160
192
224
256
288
320
352
384
416
448
480
512
20ms VBR
10ms VBR
5ms VBR
2.5ms VBR
20ms CBR
10ms CBR
5ms CBR
2.5ms CBR


Opus Inter/Intra
HA samples
PEAQ ODG
kbit/sec
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−3.5
32
64
96
128
160
192
224
256
288
320
352
384
416
448
480
512
20ms VBR
10ms VBR
5ms VBR
2.5ms VBR
Intra 20ms VBR
Intra 10ms VBR
Intra 5ms VBR
Intra 2.5ms VBR

Results show expected performance:

- Smooth ~log-linear improvements with bitrate, converging on very high ODG

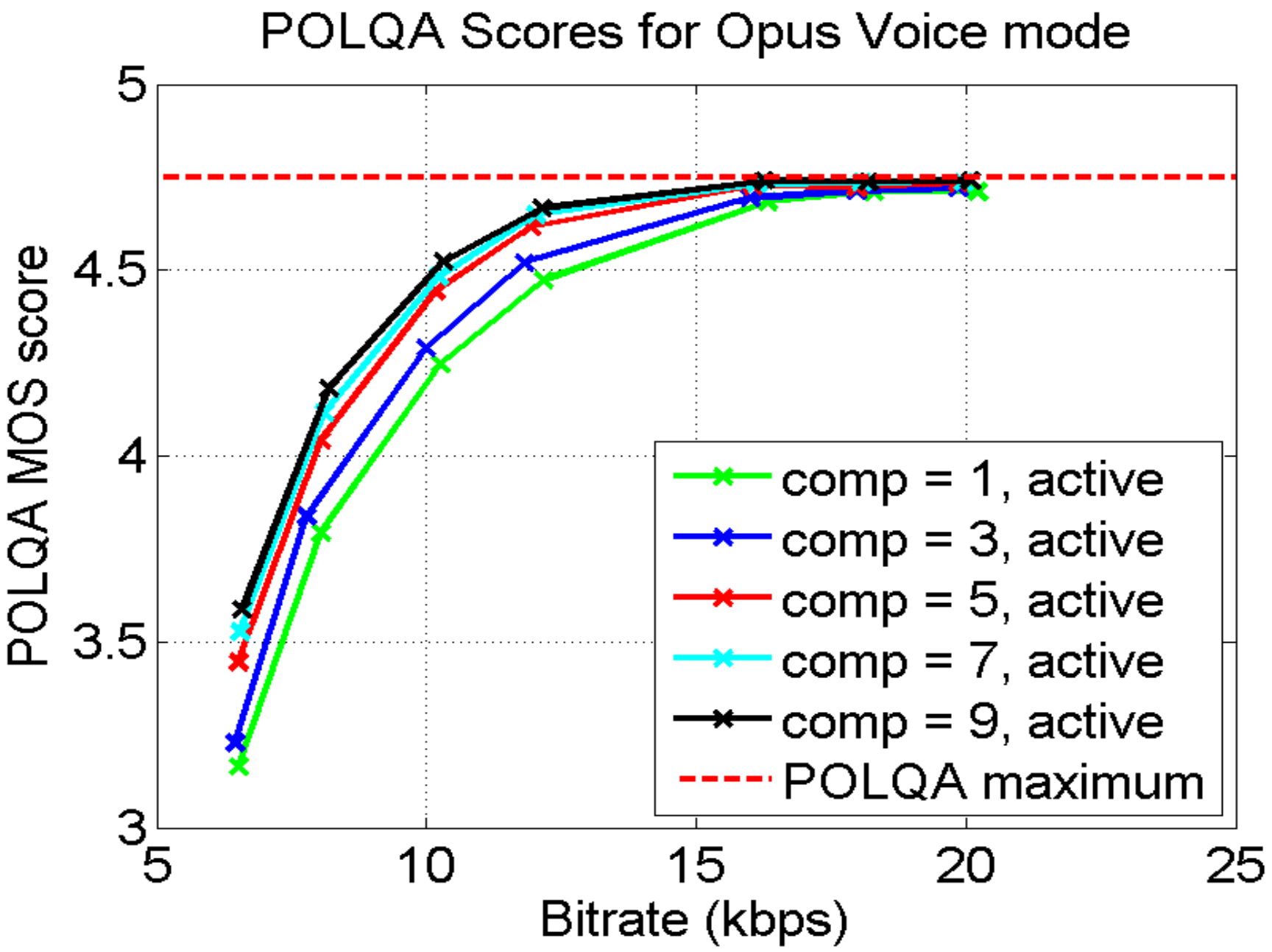
Additional objective testing opportunities:

- More samples
- Loss and error conditions
- Other metrics
- *(Some of these have been posted in parts on the list in the past)*

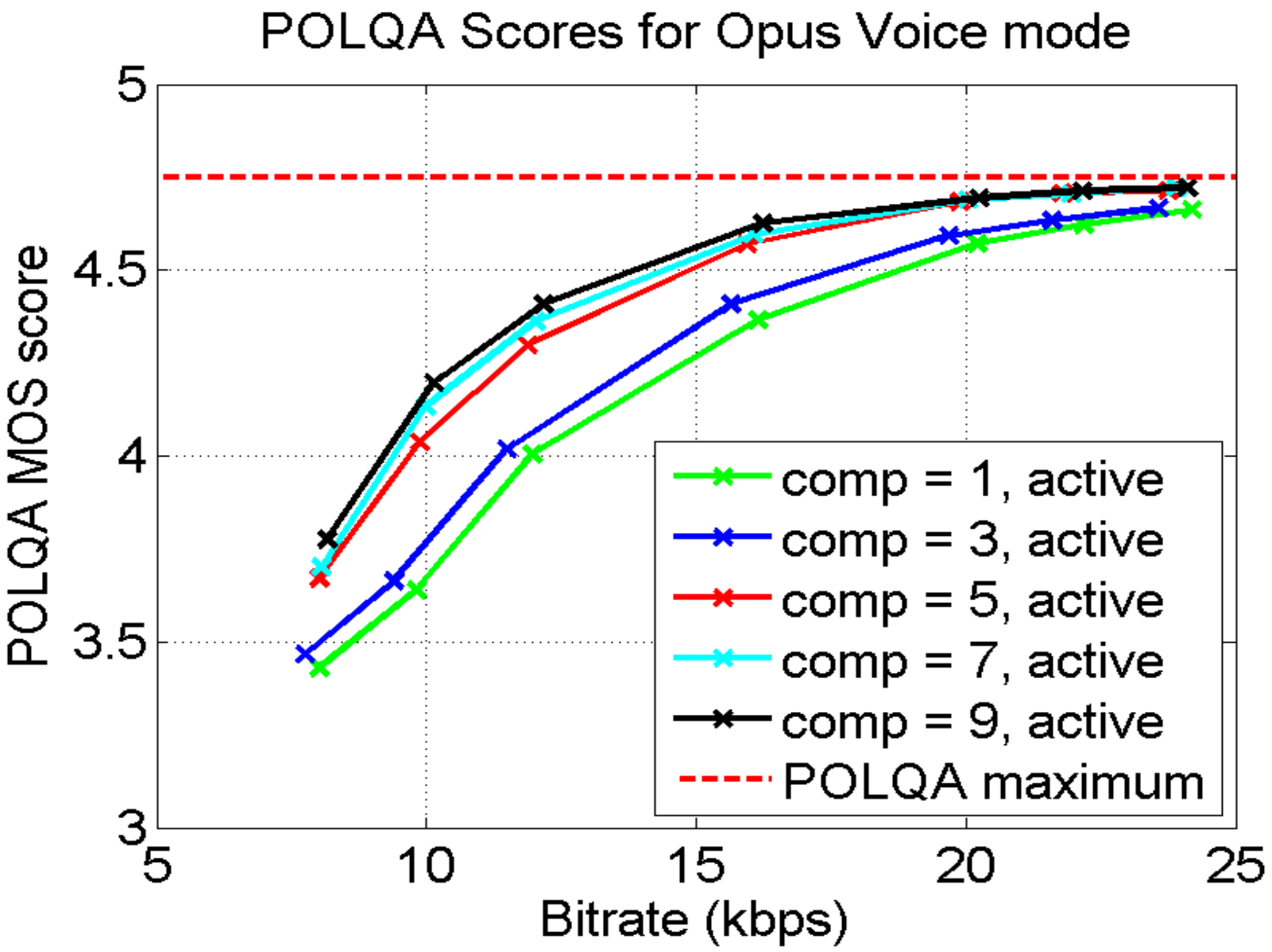
Objective Voice Mode Testing

- Used POLQA (successor to PESQ)
- Speech samples only
- Tested:
  - Bitrates
  - Frame sizes
  - Complexity
  - DTX
  - In-band FEC

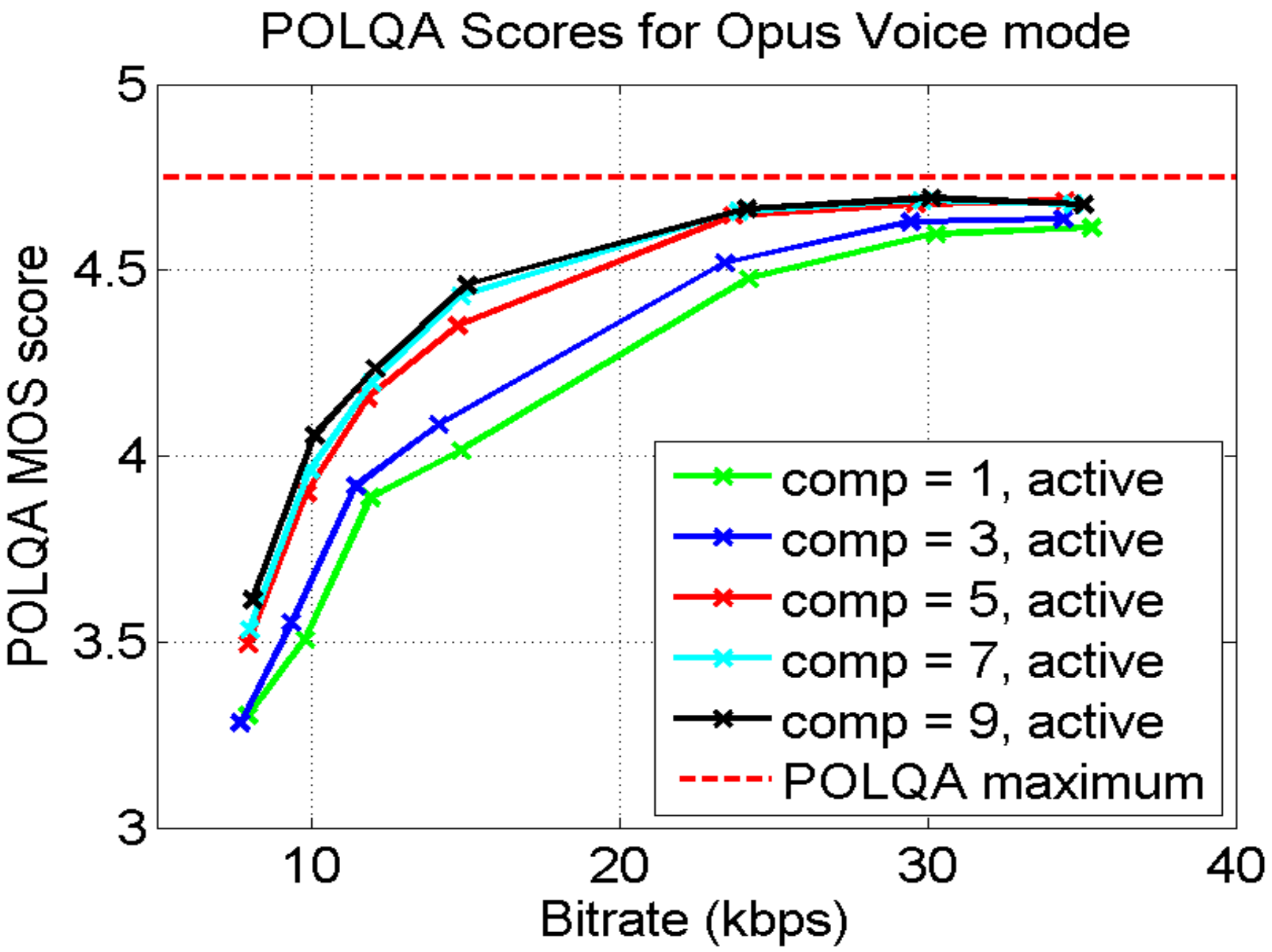
Narrowband


POLQA Scores for Opus Voice mode
POLQA MOS score
Bitrate (kbps)
5
4.5
4
3.5
3
5
10
15
20
25
comp = 1, active
comp = 3, active
comp = 5, active
comp = 7, active
comp = 9, active
POLQA maximum

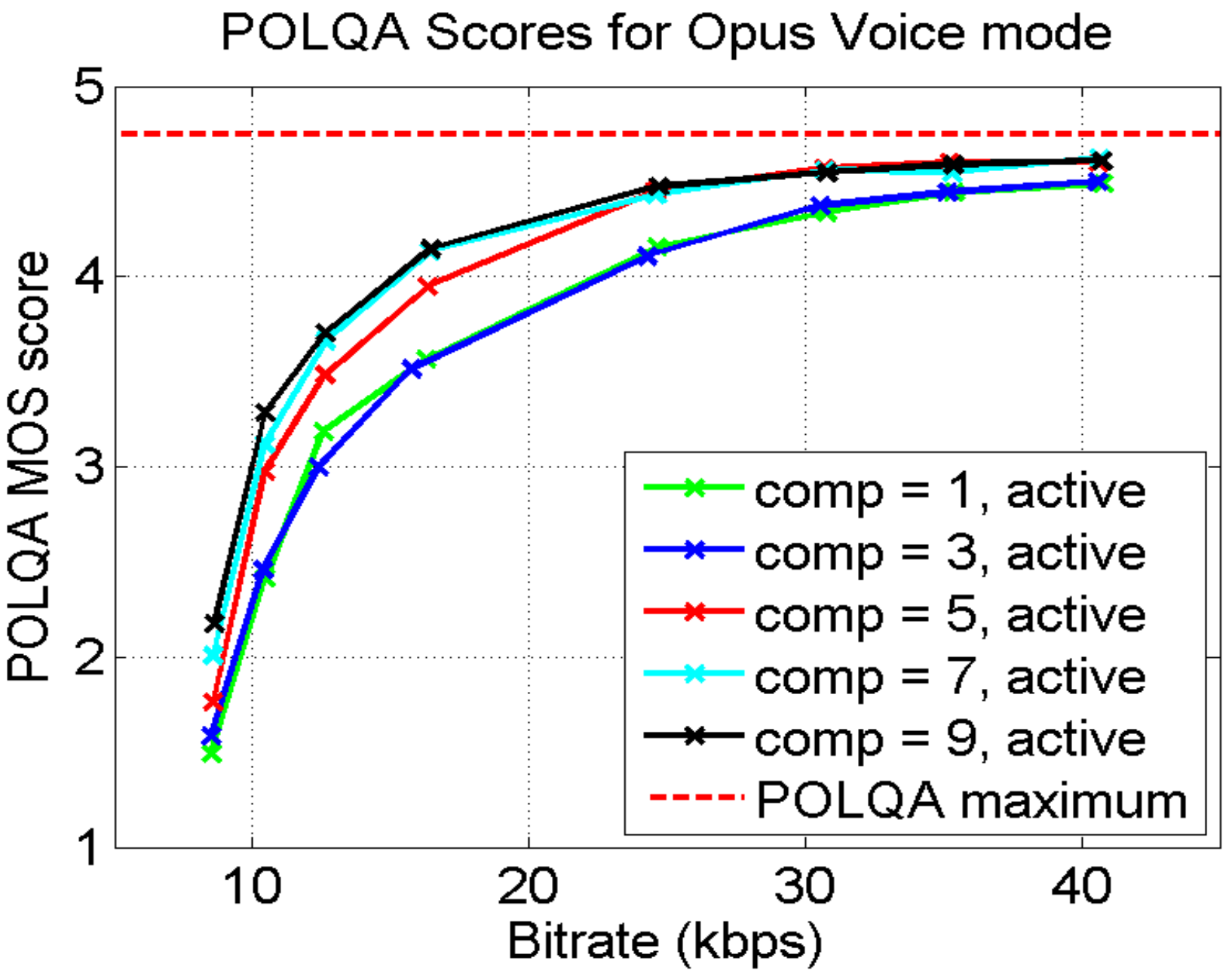
Mediumband


POLQA Scores for Opus Voice mode
POLQA MOS score
Bitrate (kbps)
5
4.5
4
3.5
3
5
10
15
20
25
comp = 1, active
comp = 3, active
comp = 5, active
comp = 7, active
comp = 9, active
POLQA maximum

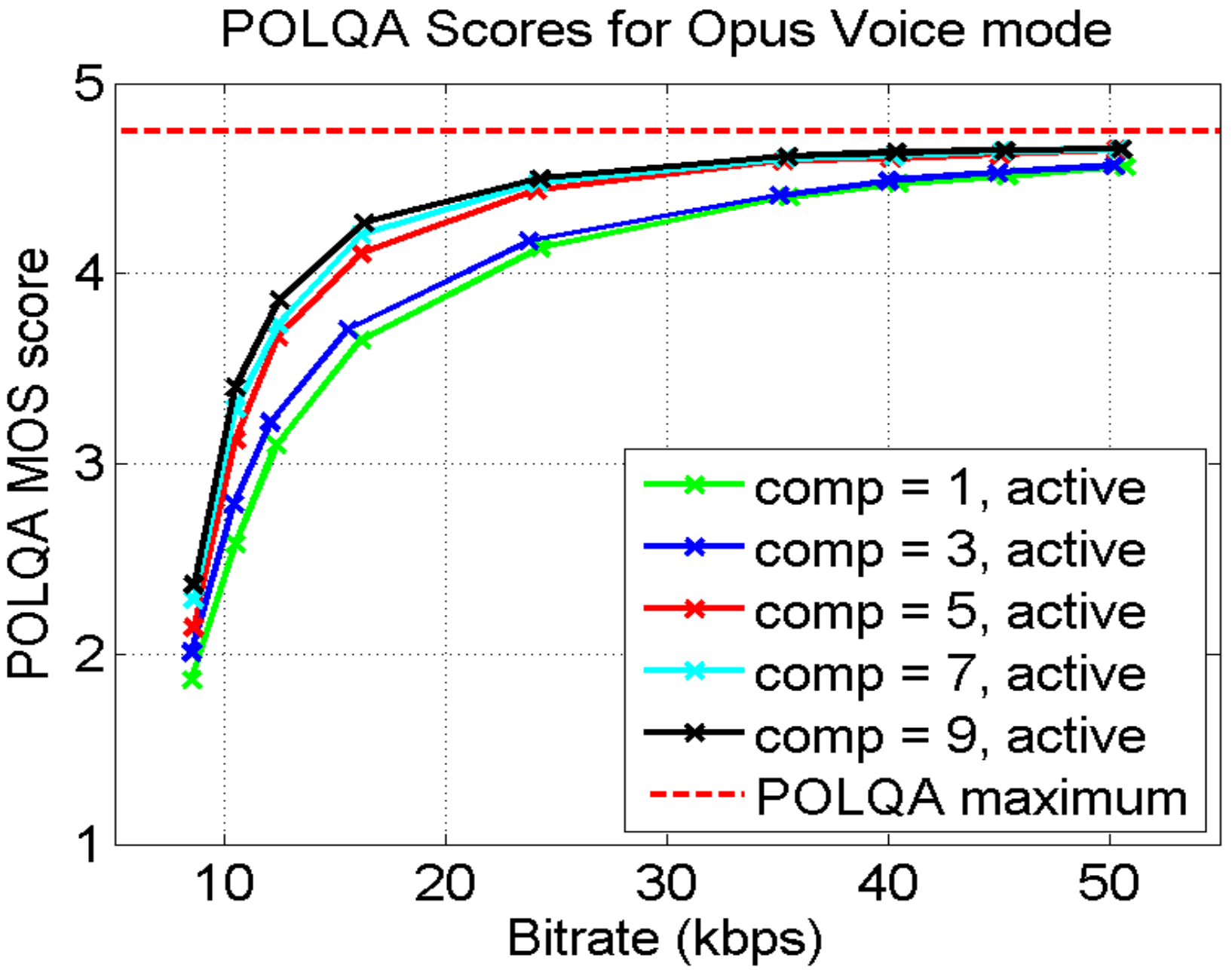
Wideband


POLQA Scores for Opus Voice mode
POLQA MOS score
5
4.5
4
3.5
3
10
20
30
40
Bitrate (kbps)
comp = 1, active
comp = 3, active
comp = 5, active
comp = 7, active
comp = 9, active
POLQA maximum

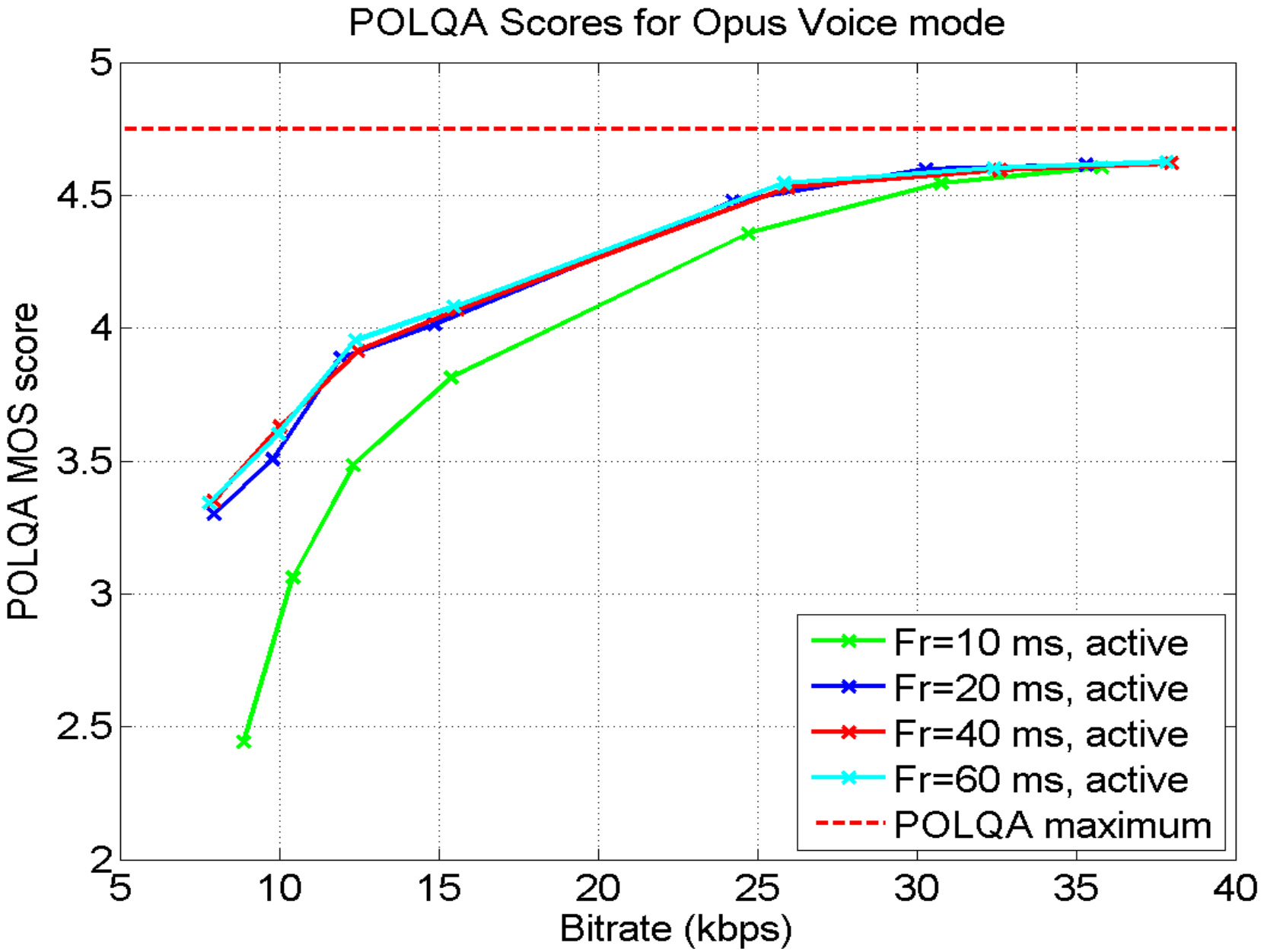
Super Wideband



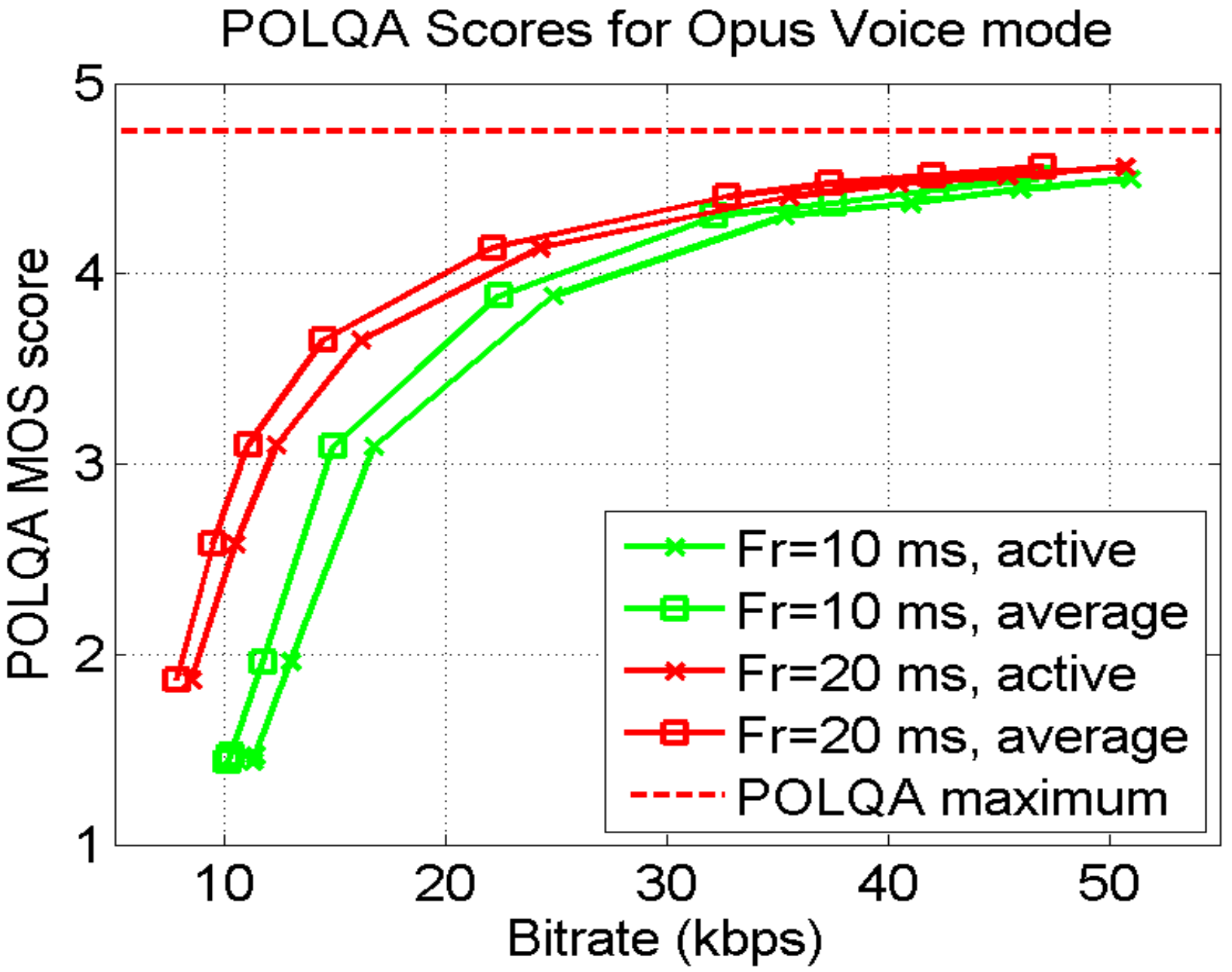
Fullband


POLQA Scores for Opus Voice mode
POLQA MOS score
5
4
3
2
1
10
20
30
40
50
Bitrate (kbps)
comp = 1, active
comp = 3, active
comp = 5, active
comp = 7, active
comp = 9, active
POLQA maximum

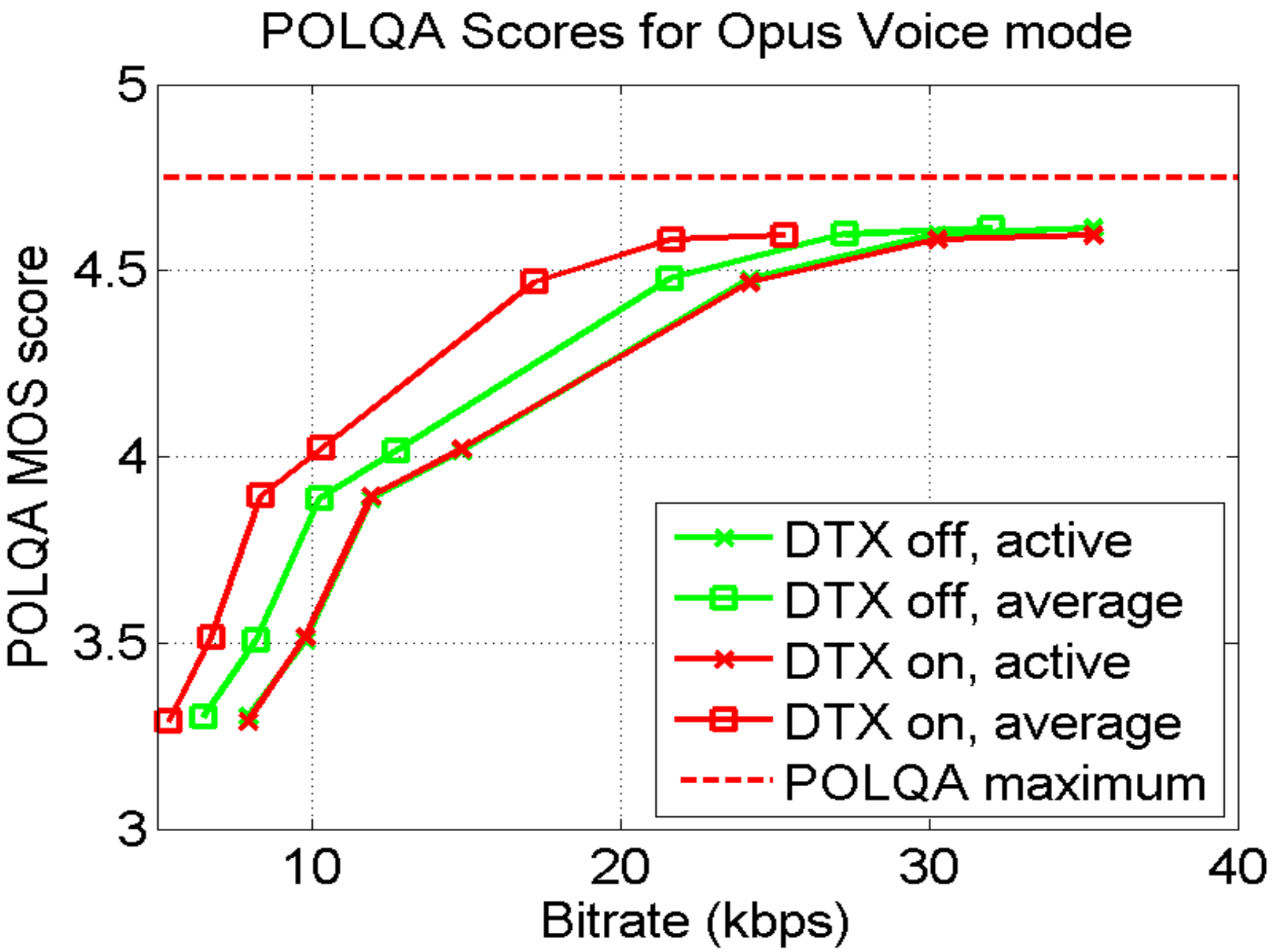
Wideband: 10/20/40/60 ms



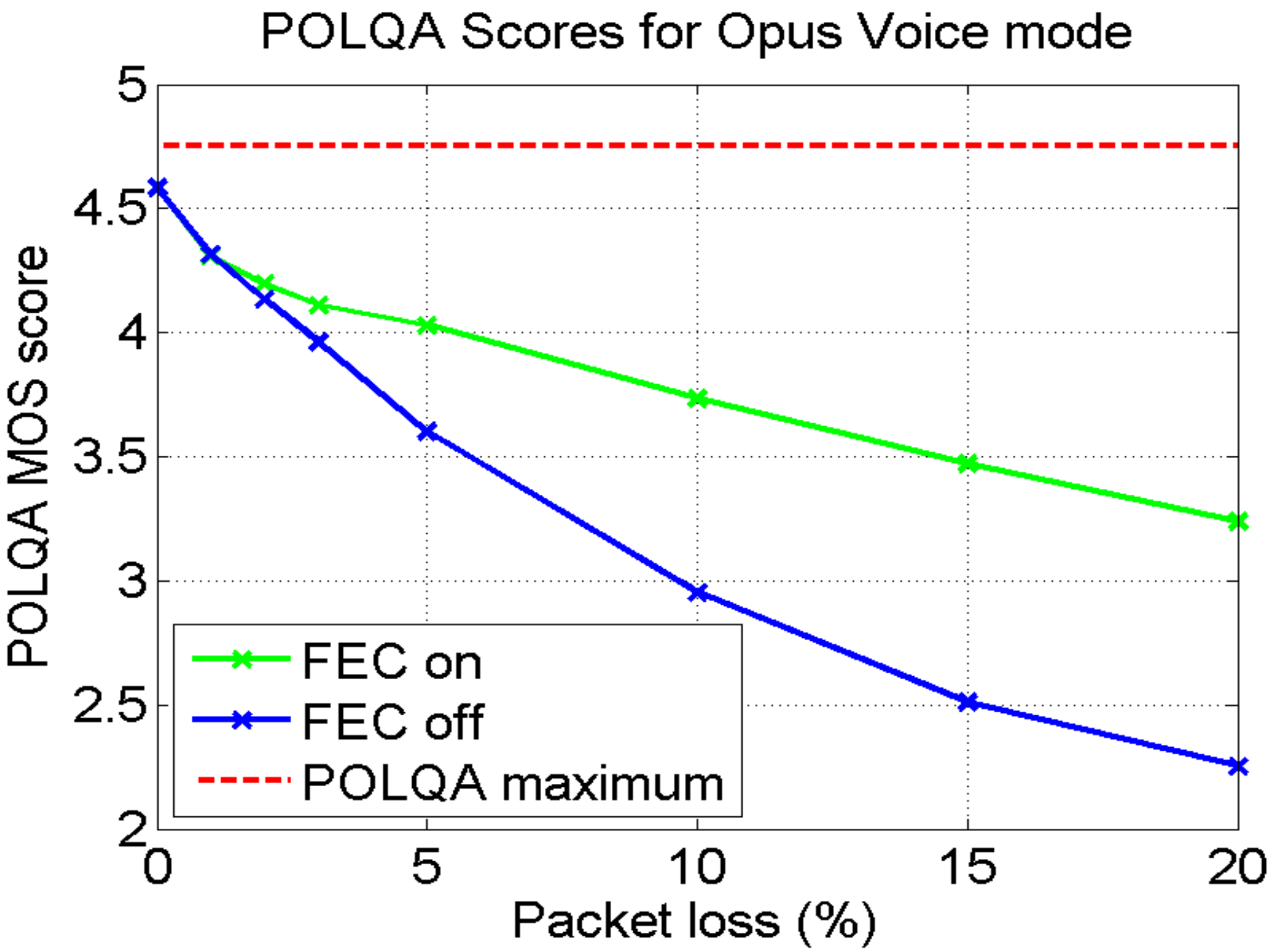
Fullband 10/20 ms



Wideband DTX on/off



Wideband FEC on/off



Wideband Floating/Fixed Point