Speech Coding Enhancement for Opus: Quality Requirements

Presenter: Jan Buethe (AWS)
jbuethe@amazon.com

IETF 118

draft-buethe-opus-speech-coding-enhancement

Opus (SILK) Speech Coding Enhancement

Algorithm Development

- Develop SOTA low-complexity speech coding enhancement methods
- First without side info, later with side info via extension mechanism
- Full optimization and integration into libopus

Standardization

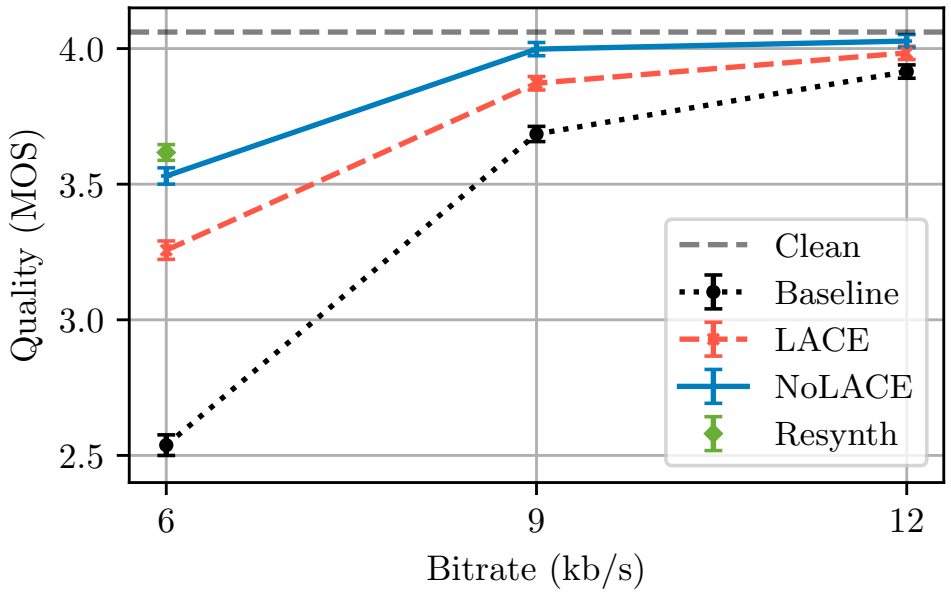
- Desirable to keep method open for improvement
- To achieve this: standardize requirements instead of methods regarding
  - Quality
  - Integration
  - Interoperability

Algorithm Development: Progress report

- New enhancement method NoLACE
- Higher quality than LACE
- Complexity: 600 MFLOPS
- Size: 1.7 M Parameters

Paper

Demo



Standardization: Evaluating Quality

- Goal: Ensure that enhancement method does not degrade Opus SILK
- Gold standard: subjective listening test,
  but
  - very costly
  - only practical for limited amount of data and operating points
- Alternative: objective metric,
  but
  - no single objective metric is perfect
  - metrics often don’t age well (overfitting on today’s conditions -> bad performance on tomorrow’s conditions)

Metrics under Consideration

1. PESQ: perceptual evaluation of speech quality (ITU-T P.862.2), a MOS predictor.
2. WARP-Q: distortion metric designed for neural speech codecs.
3. MOC: a modified version of opus_compare a simple psycho-acoustic distortion metric.
4. NOMAD: brand-new distortion metric based on neural embeddings (wav2vec 2.0). Originally designed as non-matching-reference method but used as full-reference method here.

Comparison to Listening Test Results (MOS)

- Metrics not directly comparable to MOS scale but we can look at the ordering of conditions from lowest to highest predicted quality.

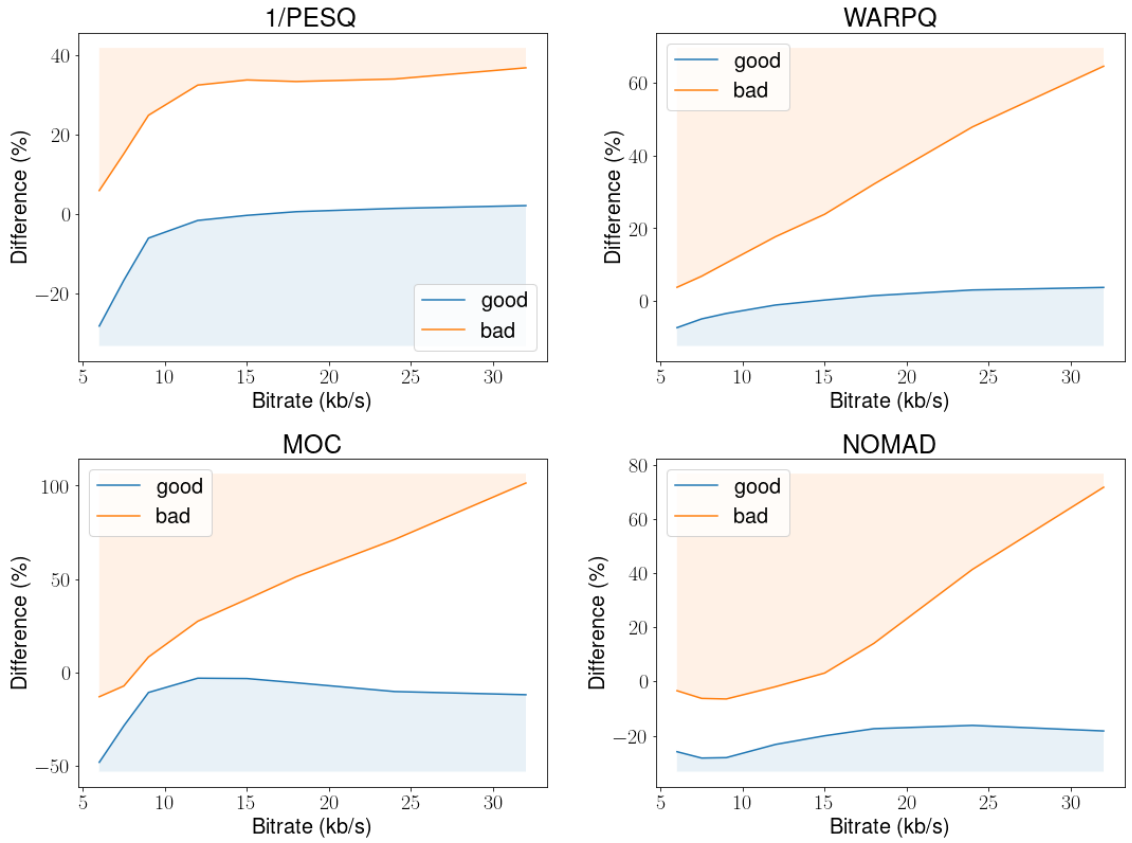
$$Error(method) := \sum_{i=1}^{num\_conditions} |Position_{metric}(condition_i) - Position_{MOS}(condition_i)|$$

Rank	Method	Error
0	MOS	0
1	NOMAD	2
2	MOC	4
2	WARP-Q	4
3	PESQ	8

Detecting Degradation

- Goal: distinguish good from bad enhancement models
- Idea: use almost untrained versions of LACE and NoLACE as examples of bad models
- LACE and NoLACE are initialized around the identity function
- Bad versions still reproduce the signal but clearly degrade quality for medium to high bitrates

Separating the Good from the Bad



$$\frac{Metric\big(Enhanced(x)\big) - Metric(x)}{Metric(x)}$$

Tight thresholds for which LACE and NoLACE would pass

Bitrate	1/PESQ	WARP-Q	MOC	NOMAD
6000	< -28.29 %	< -7.19 %	< -48.05 %	< -25.91 %
7500	< -16.72 %	< -4.80 %	< -28.49 %	< -28.25 %
9000	< -6.13 %	< -3.31 %	< -10.69 %	< -28.09 %
12000	< -1.68 %	< -1.00 %	< -3.05 %	< -23.22 %
15000	< -0.40 %	< 0.37 %	< -3.23 %	< -20.02 %
18000	< 0.52 %	< 1.57 %	< -5.49 %	< -17.41 %
24000	< 1.34 %	< 3.12 %	< -10.19 %	< -16.19 %
32000	< 2.05 %	< 3.82 %	< -11.89 %	< -18.25 %

Summary

- All four tested metrics were capable of separating good models from bad models.
- Test based on metric-dependent thresholds seems likely to catch issues with enhancement models
- Depending on the metric, thresholds would need to allow for slightly worse scores at higher bitrates
- NOMAD seems favorable to other metrics but is difficult to standardize (neural network with ~ 95 M parameters)
- WARP-Q and MOC easier to standardize, MOC slightly preferable
- PESQ already standardized (ITU-T P.862.2 in 2005) but performs worst

Next Steps

Algorithm Development

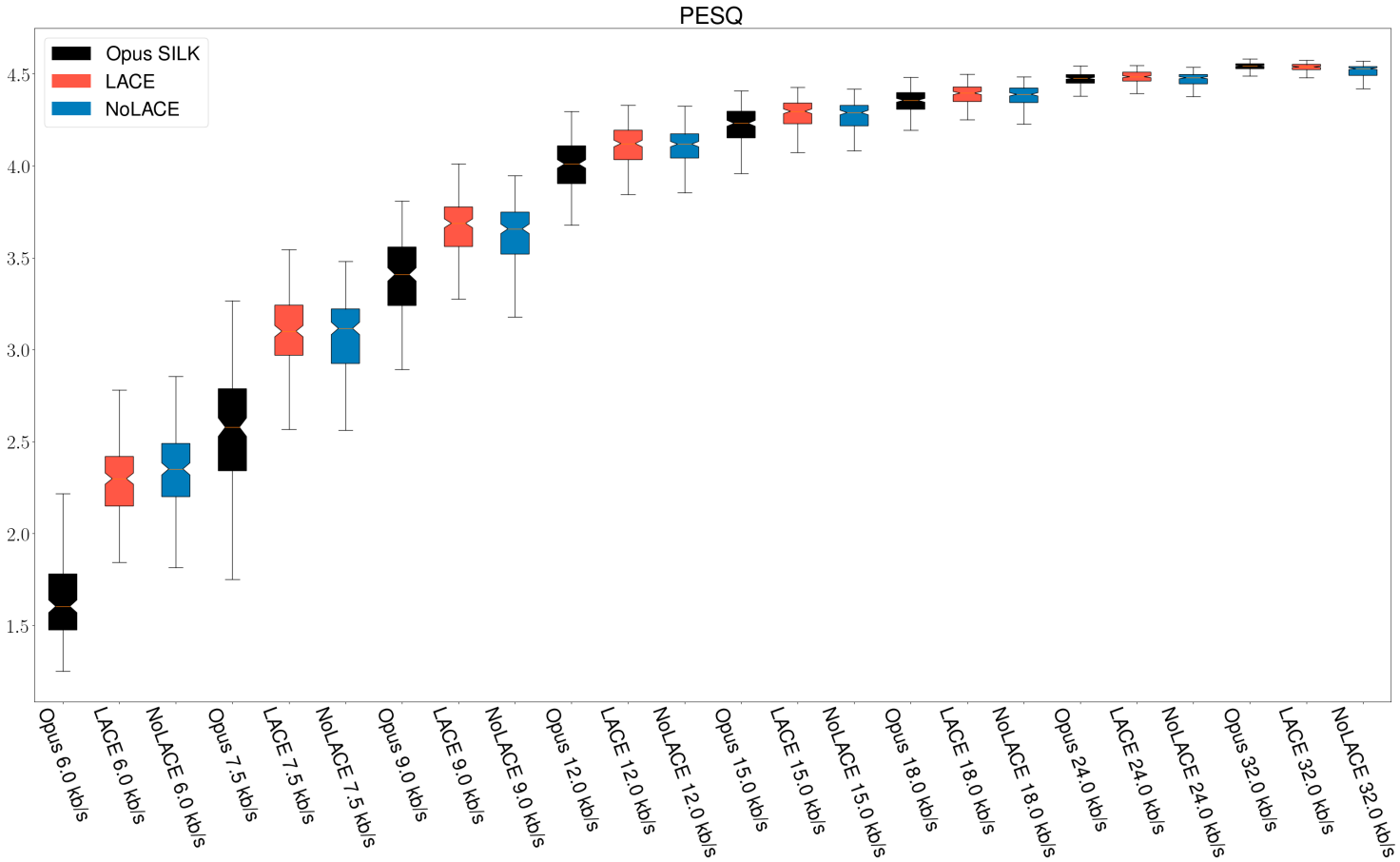
- Integration of LACE and NoLACE into Opus (opus-ng branch)
- Size and complexity optimization
- Investigate noisy speech performance (in progress)
- Add Bandwidth Extension (in progress)
- Add side info

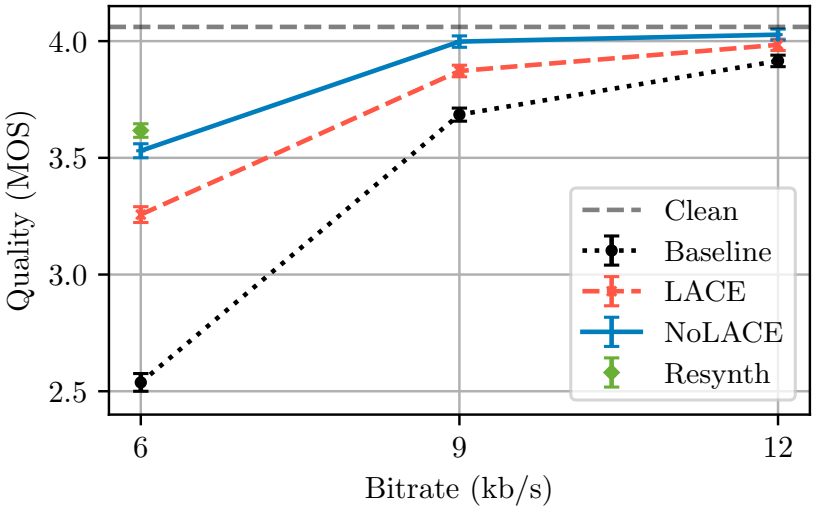
Standardization

- Assemble testvectors from open datasets
- Spell out requirements for clean-speech / noisy-speech performance
- Questions:
  - How strict / lenient should we be?
  - Single metric or multiple metrics?

Thank you!

MOS Comparison: PESQ

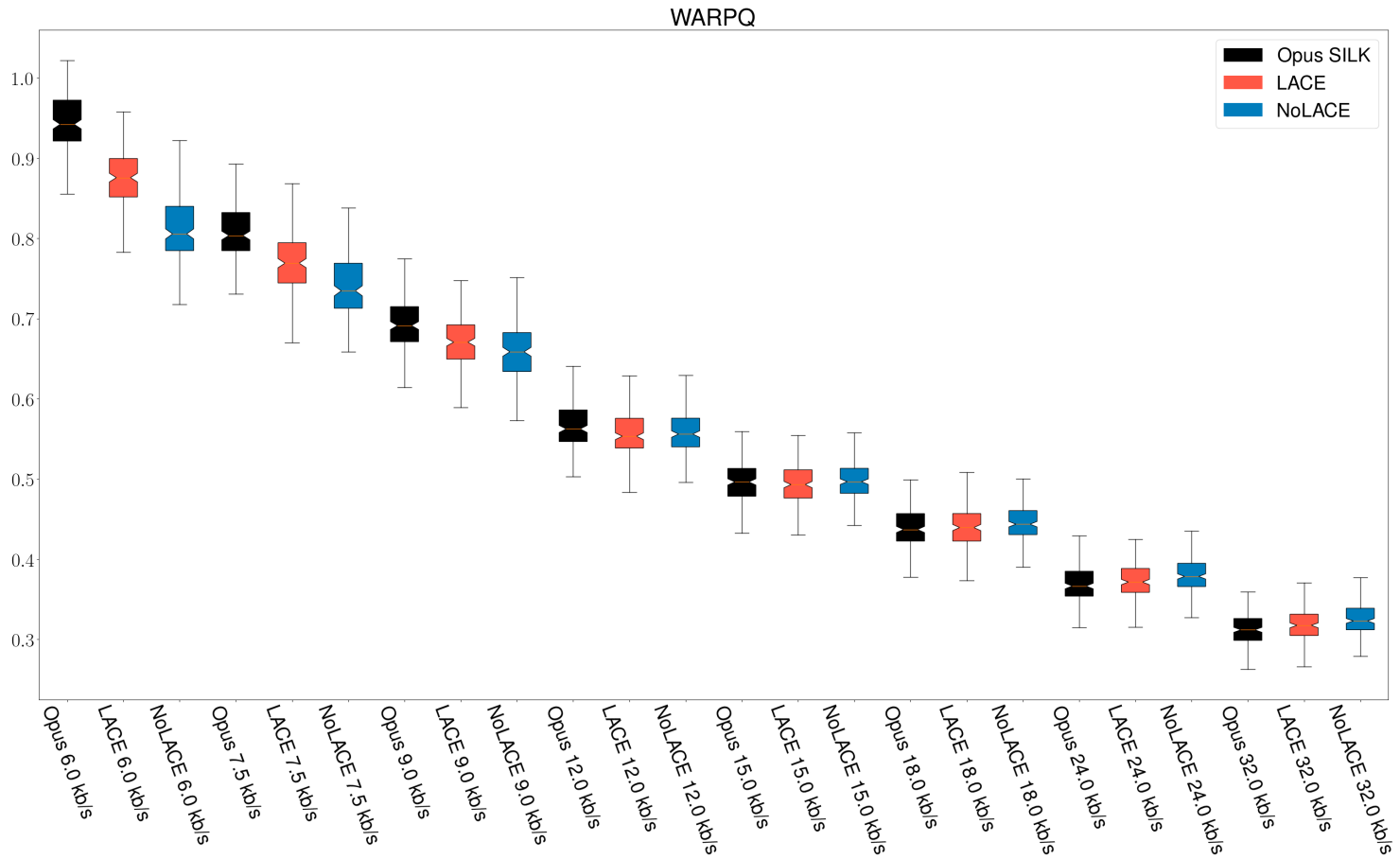


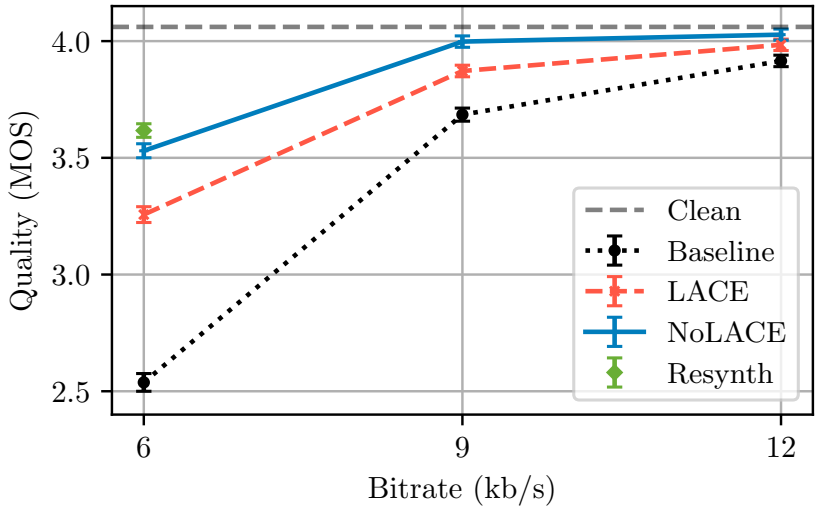


Mean (std. deviation)

bitrate (bps)	Opus	LACE	NoLACE
6000	1.630 (0.21)	2.273 (0.23)	2.338 (0.22)
7500	2.546 (0.34)	3.078 (0.24)	3.059 (0.24)
9000	3.379 (0.26)	3.661 (0.17)	3.600 (0.25)
12000	3.999 (0.14)	4.115 (0.11)	4.068 (0.21)
15000	4.215 (0.11)	4.283 (0.08)	4.233 (0.20)
18000	4.352 (0.06)	4.388 (0.06)	4.330 (0.22)
24000	4.469 (0.04)	4.481 (0.04)	4.411 (0.23)
32000	4.536 (0.03)	4.532 (0.03)	4.446 (0.25)

MOS Comparison: WARPQ

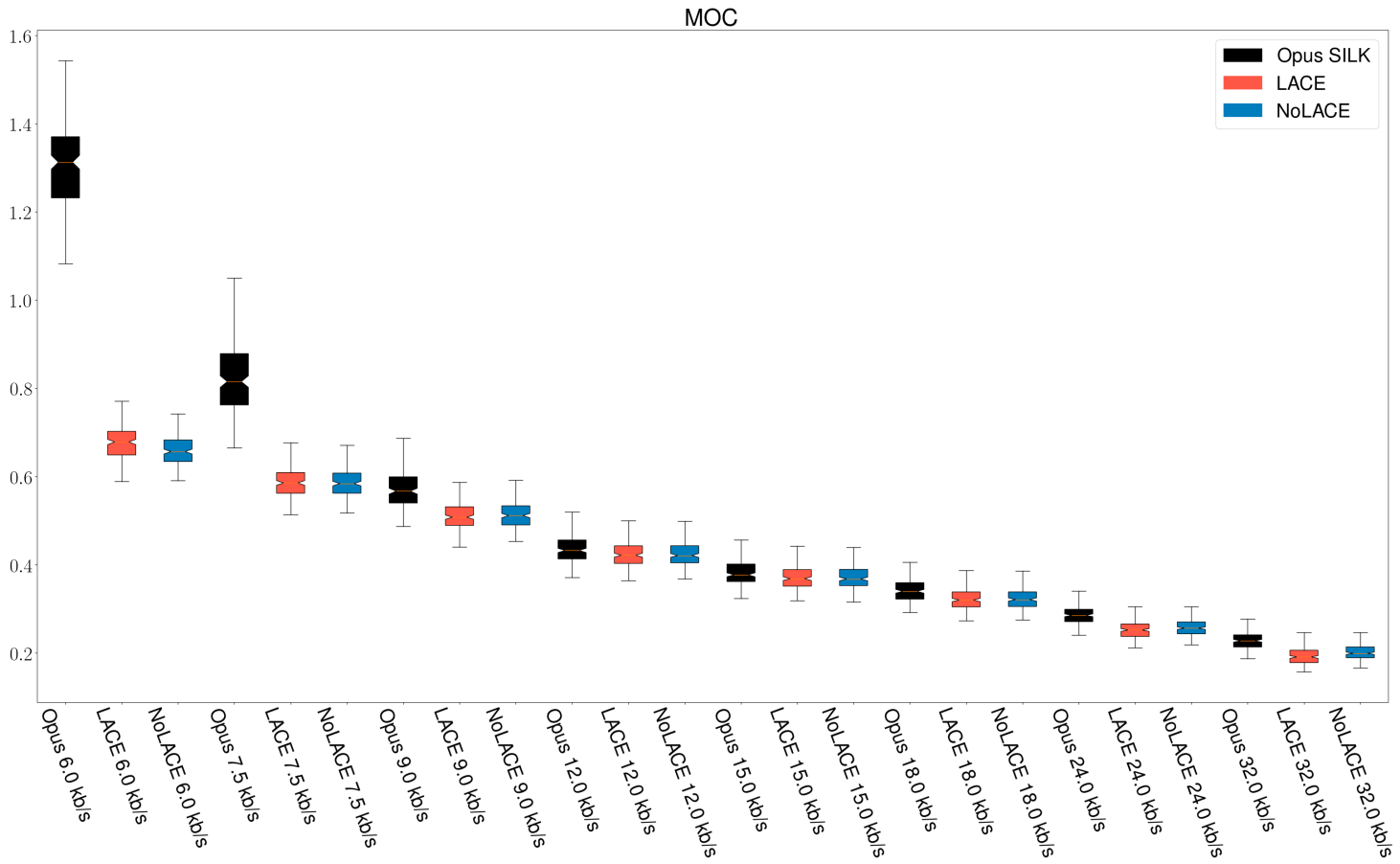


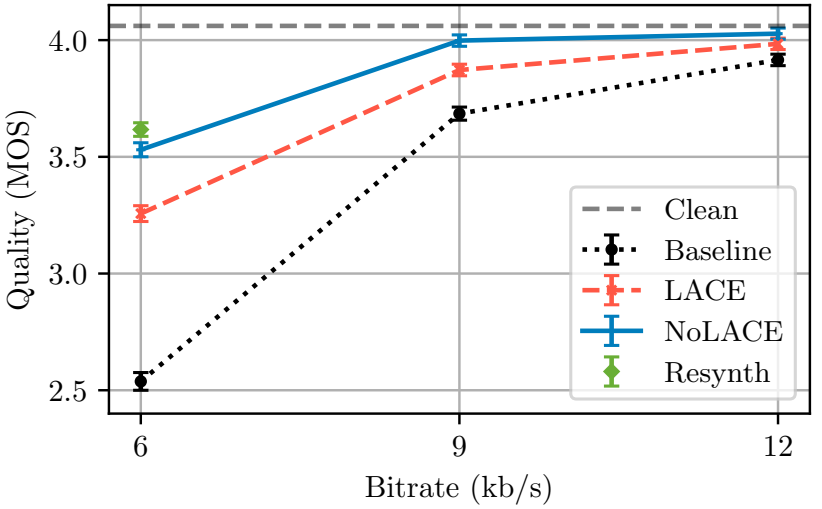


Mean (std. deviation)

bitrate (bps)	Opus	LACE	NoLACE
6000	0.944 (0.04)	0.876 (0.04)	0.811 (0.04)
7500	0.809 (0.04)	0.770 (0.04)	0.744 (0.04)
9000	0.695 (0.03)	0.671 (0.03)	0.663 (0.04)
12000	0.567 (0.03)	0.558 (0.03)	0.561 (0.03)
15000	0.497 (0.03)	0.493 (0.03)	0.499 (0.03)
18000	0.440 (0.03)	0.440 (0.03)	0.447 (0.02)
24000	0.368 (0.02)	0.372 (0.02)	0.380 (0.02)
32000	0.313 (0.02)	0.318 (0.02)	0.325 (0.02)

MOS Comparison: MOC

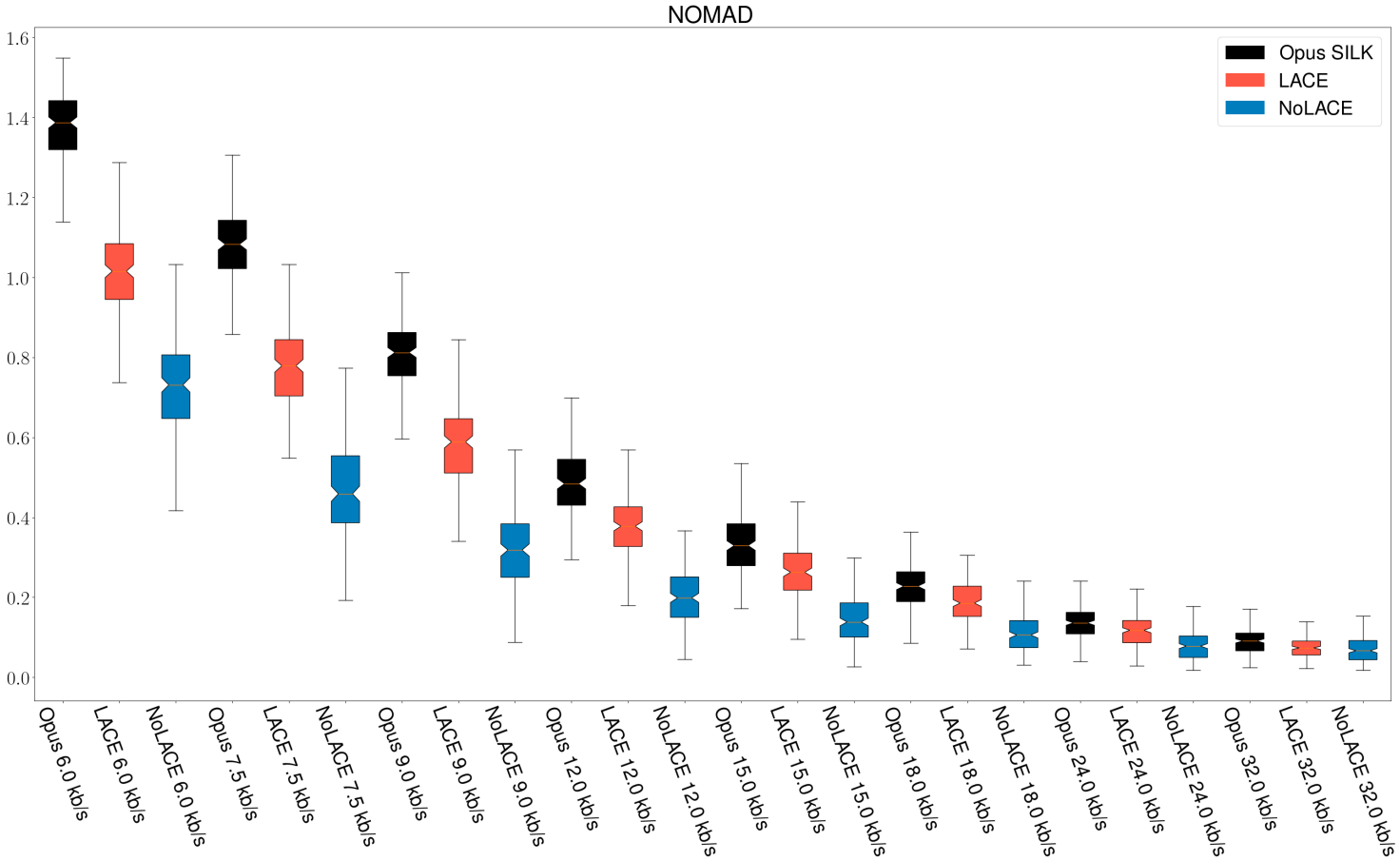


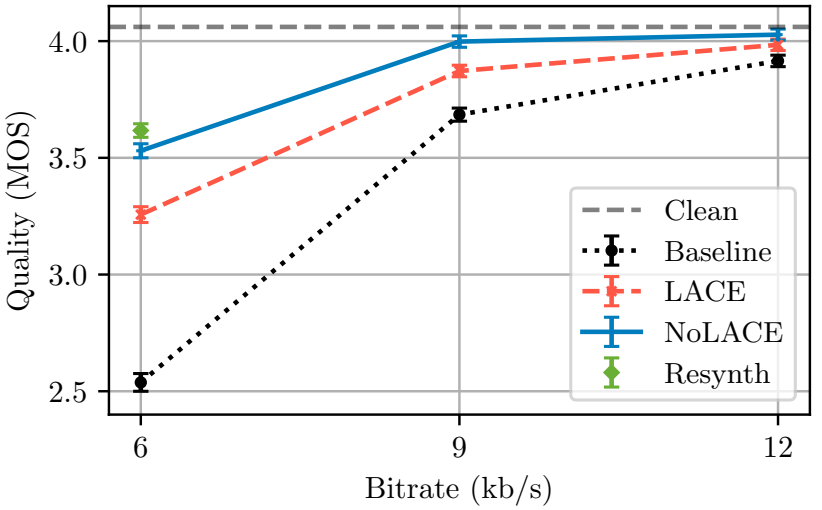


Mean (std. deviation)

bitrate (bps)	Opus	LACE	NoLACE
6000	1.305 (0.10)	0.678 (0.04)	0.658 (0.03)
7500	0.823 (0.09)	0.588 (0.04)	0.588 (0.03)
9000	0.577 (0.05)	0.513 (0.03)	0.515 (0.03)
12000	0.439 (0.04)	0.426 (0.03)	0.426 (0.03)
15000	0.385 (0.03)	0.372 (0.03)	0.372 (0.03)
18000	0.342 (0.03)	0.323 (0.02)	0.323 (0.02)
24000	0.287 (0.02)	0.255 (0.02)	0.257 (0.02)
32000	0.229 (0.02)	0.195 (0.02)	0.202 (0.02)

MOS Comparison: NOMAD





Mean (std. deviation)

bitrate (bps)	Opus	LACE	NoLACE
6000	1.370 (0.10)	1.015 (0.12)	0.723 (0.13)
7500	1.075 (0.10)	0.771 (0.11)	0.471 (0.11)
9000	0.808 (0.09)	0.580 (0.10)	0.319 (0.09)
12000	0.486 (0.08)	0.373 (0.08)	0.203 (0.08)
15000	0.328 (0.07)	0.262 (0.07)	0.147 (0.06)
18000	0.228 (0.06)	0.189 (0.05)	0.113 (0.05)
24000	0.137 (0.04)	0.114 (0.04)	0.081 (0.04)
32000	0.091 (0.03)	0.075 (0.03)	0.070 (0.03)

Not-Worse-On-Average Test (Clean Speech)

PESQ

bitrate	LACE	NoLACE
6000	pass	pass
7500	pass	pass
9000	pass	pass
12000	pass	pass
15000	pass	pass
18000	fail	fail
24000	fail	fail
32000	fail	fail

WARP-Q

bitrate	LACE	NoLACE
6000	pass	pass
7500	pass	pass
9000	pass	pass
12000	pass	pass
15000	pass	fail
18000	pass	fail
24000	fail	fail
32000	fail	fail

MOC

bitrate	LACE	NoLACE
6000	pass	pass
7500	pass	pass
9000	pass	pass
12000	pass	pass
15000	pass	pass
18000	pass	pass
24000	pass	pass
32000	pass	pass

NOMAD

bitrate	LACE	NoLACE
6000	pass	pass
7500	pass	pass
9000	pass	pass
12000	pass	pass
15000	pass	pass
18000	pass	pass
24000	pass	pass
32000	pass	pass

Not-Worse-On-Average Test (Noisy Speech)

PESQ

bitrate	LACE	NoLACE
6000	pass	pass
7500	pass	pass
9000	pass	fail
12000	pass	fail
15000	pass	fail
18000	pass	fail
24000	pass	fail
32000	fail	fail

WARP-Q

bitrate	LACE	NoLACE
6000	pass	pass
7500	pass	pass
9000	pass	pass
12000	pass	pass
15000	pass	fail
18000	pass	fail
24000	fail	fail
32000	fail	fail

MOC

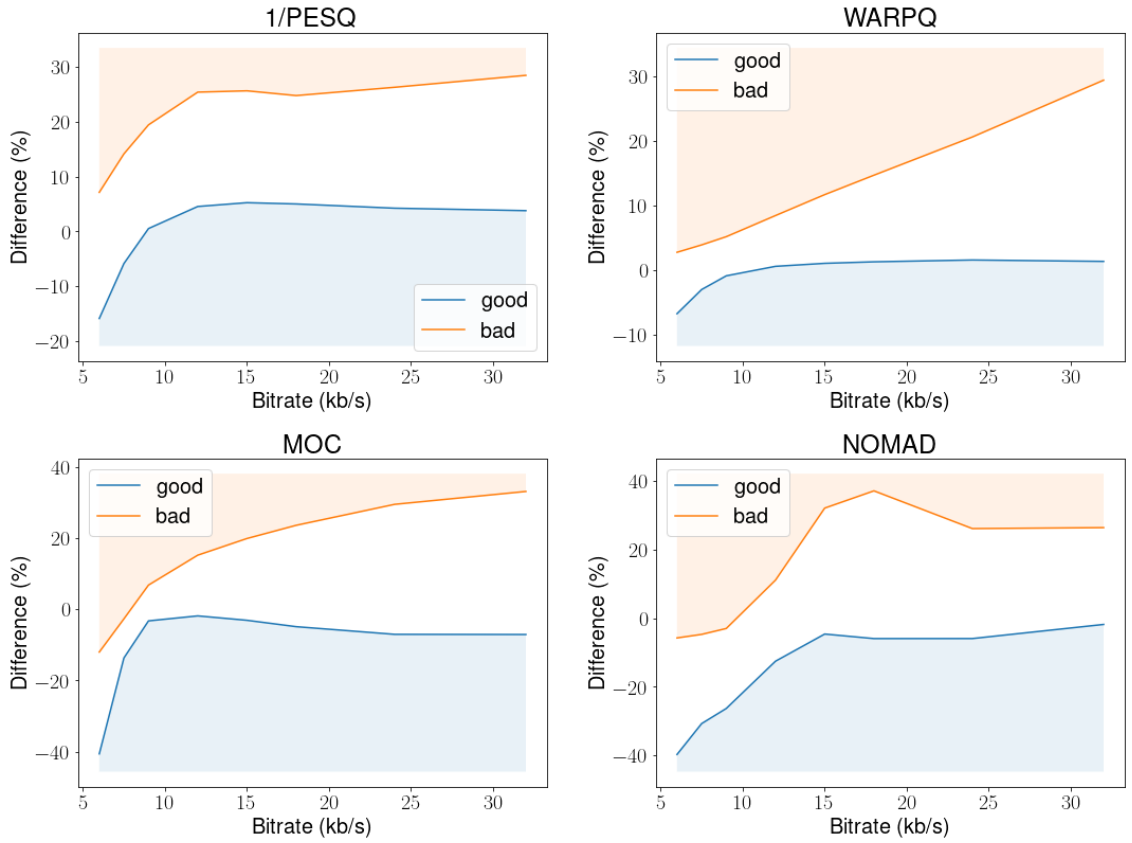
bitrate	LACE	NoLACE
6000	pass	pass
7500	pass	pass
9000	pass	pass
12000	pass	pass
15000	pass	pass
18000	pass	pass
24000	pass	pass
32000	pass	pass

NOMAD

bitrate	LACE	NoLACE
6000	pass	pass
7500	pass	pass
9000	pass	pass
12000	pass	pass
15000	pass	pass
18000	pass	pass
24000	pass	pass
32000	pass	pass

Separating the Good from the Bad (Noisy Speech)

Disclaimer: NOMAD not designed for evaluating noisy speech



$$\frac{Metric\big(Enhanced(x)\big) - Metric(x)}{Metric(x)}$$

Tight thresholds for which LACE and NoLACE would pass

Bitrate	1/PESQ	WARP-Q	MOC	NOMAD
6000	< -15.84 %	< -6.76 %	< -40.62 %	< -39.64 %
7500	< -5.82 %	< -3.01 %	< -13.64 %	< -30.66 %
9000	< 0.57 %	< -0.89 %	< -3.20 %	< -26.26 %
12000	< 4.59 %	< 0.59 %	< -1.79 %	< -12.50 %
15000	< 5.28 %	< 1.05 %	< -3.03 %	< -4.59 %
18000	< 5.05 %	< 1.29 %	< -4.83 %	< -5.91 %
24000	< 4.28 %	< 1.57 %	< -6.98 %	< -5.90 %
32000	< 3.83 %	< 1.35 %	< -7.03 %	< -1.79 %