

IETF 123, 24 July 2025

draft-lechler-mlcodec-test-battery-00

A Comprehensive Listening Test Battery
for ML Audio Codec Evaluation

Presenter: Laura Lechler (Cisco Systems)

Motivation of draft-lechler-mlcodec-test-battery-00

- no specific evaluation procedure described for DRED
- evaluation of generative models faces new challenges compared to traditional codecs (both objective metric and human listening tests impacted)
- recent concerns [1] and recommendations [2] regarding listening tests addressed as follows:

Concern	Ref	Proposed Improvement
ACR reference-free (range-equalizing biases!)	[1]	A. MUSHRA 1S (single test stimulus) for clean input data B. DCR for (mild to moderate) real-world noise and reverberation
Both clean and noisy speech should be tested	[2]	Testing real-world speech in noisy and reverberant conditions
Intelligibility not typically assessed (e.g., in P.808)	[1]	A. add intelligibility test (DRT), B. qualification filters to recruit native speakers (large platform-specific differences)

[1] Marvin Sach, Yihui Fu, Kohei Saijo, Wangyou Zhang, Samuele Cornell, Robin Scheibler, Chenda Li, Anurag Kumar, Wei Wang, Yanmin Qian, Shinji Watanabe, Tim Fingscheidt. "P.808 Multilingual Speech Enhancement Testing: Approach and Results of URGENT 2025 Challenge". arXiv preprint, Jul 2025.

[2] Thomas Muller, Stéphane Ragot, Laetitia Gros, Pierrick Philippe, Pascal Scalart. "Speech quality evaluation of neural audio codecs." INTERSPEECH, Sep 2024, Kos Island, Greece.

Listening Test Battery

Test Data	Test Type	Assessment of
Clean	MUSHRA-1S ¹	Quality with high-quality inputs
Real-world	DMOS	Quality in real-world reverberation & noise
Clean	DRT	Phoneme-level intelligibility

Data to be open-sourced as part of the forthcoming Low-Resource Audio Codec Challenge.

¹MUSHRA 1S: MUSHRA test with a single test stimulus, anchor, and reference

Conditions tested to illustrate test capabilities

All inputs:

- single-channel, 16kHz, 16bit (i.e., for Opus silk is enabled, CELT is not used)
- level normalized for speech to be at approximately –23 to –26 RMS dB

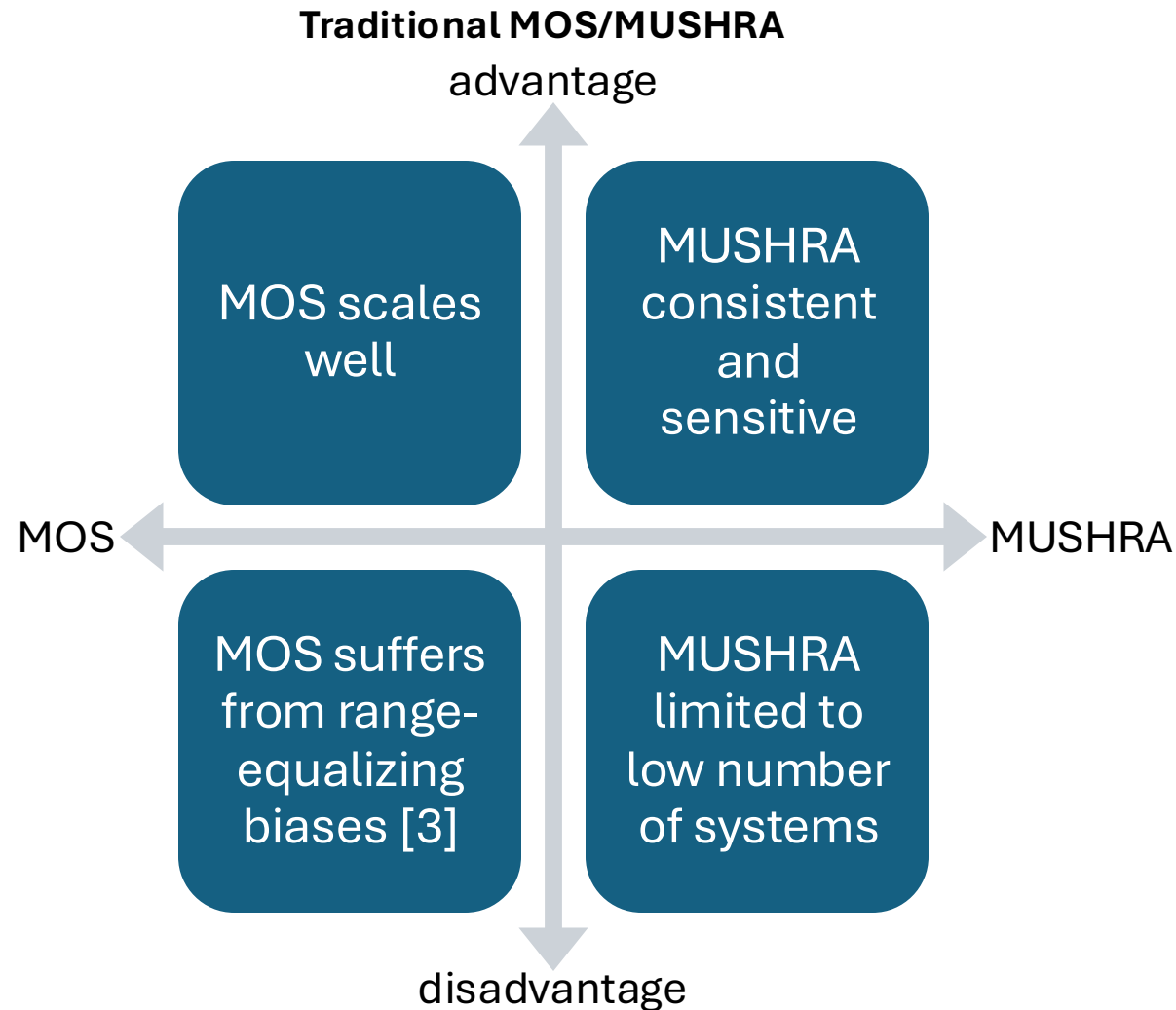
Codecs under test:

- Opus WB basic complexity at 6 kbps (as anchor/lower point of reference)
 - Opus WB basic complexity at 9 kbps
 - Opus WB LACE at 9 kbps
 - Opus WB NoLACE at 9 kbps
 - DRED Standalone¹ at 0.5 kbps
 - DRED Standalone at 1 kbps
 - DRED Standalone at 2 kbps
-
- Input files (clean speech or real-world reverberant speech in noise) used as reference
 - Output files were not level-adjusted.
 - Participants for listening tests were recruited on Prolific (details in draft).

¹The Python implementation of DRED Standalone was kindly contributed by Ahmed Mustafa.

Speech Quality Assessment

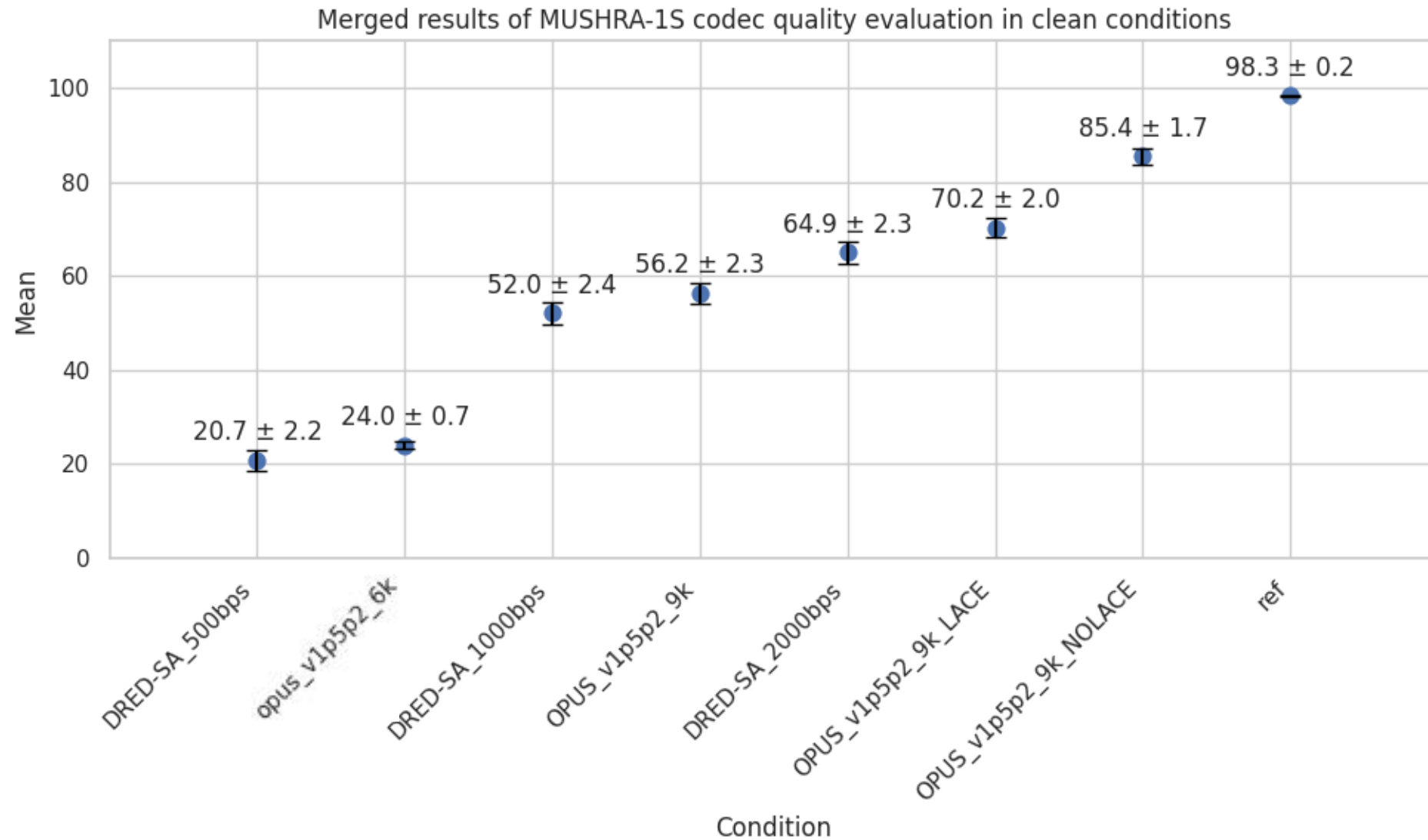
MUSHRA 1S: Evaluating one stimulus at a time



MUSHRA-1S



Quality Evaluation – Clean



DCR: Evaluating degradation in real-world conditions

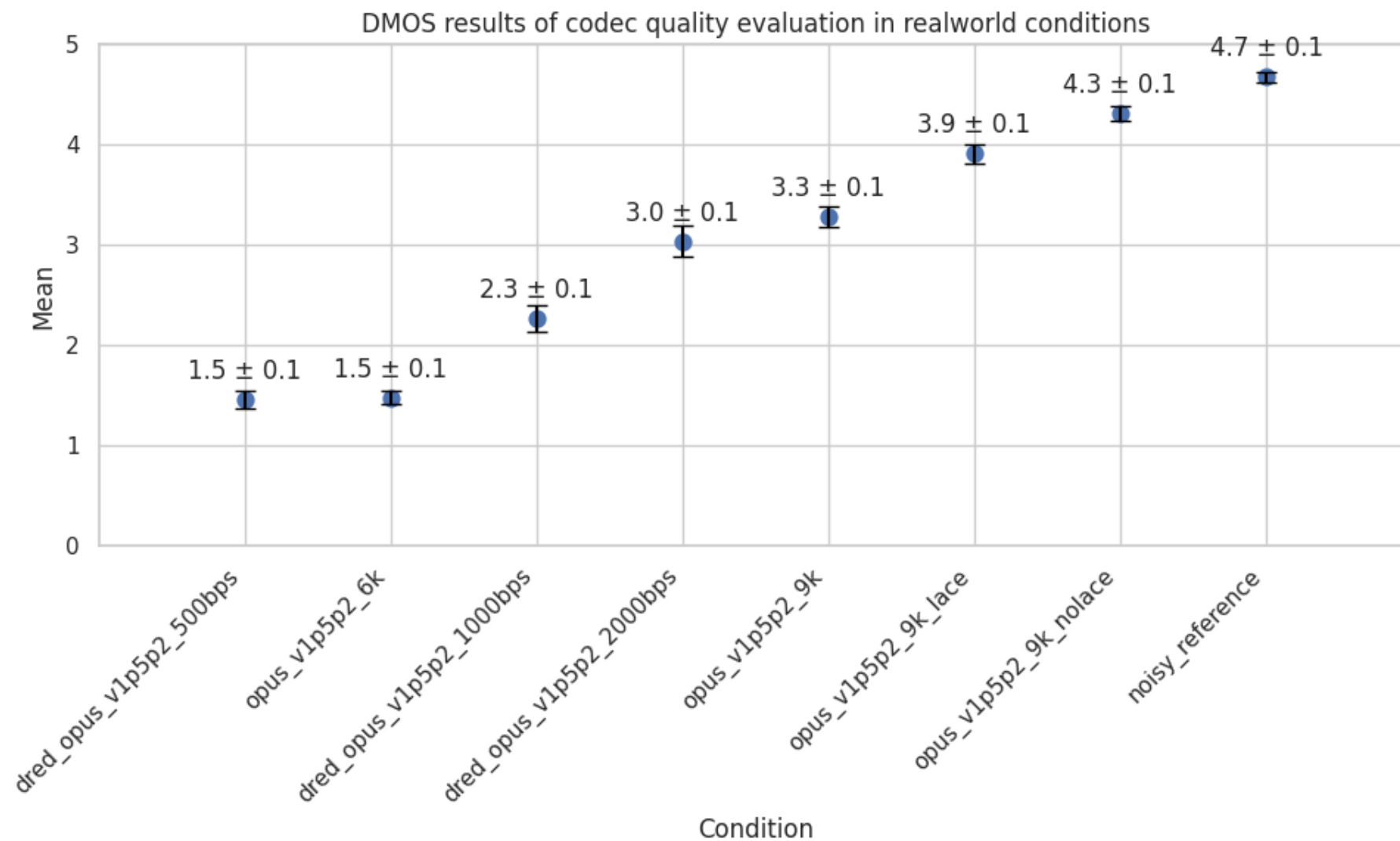
- Degradation Category Ratings (from which DMOS can be calculated)
- Evaluates degradation compared to a reference (as opposed to ACR, which is reference-free)
- Can be used also when no clean reference is available (e.g., noisy/reverberant real-world data)
- We used a controlled distributed test design, exposing each listener to one stimulus of each codec

DMOS scale according to ITU P.808 [4]:





















- 5 – Degradation is inaudible.
- 4 – Degradation is audible but not annoying.
- 3 – Degradation is slightly annoying.
- 2 – Degradation is annoying.
- 1 – Degradation is very annoying.

[4] ITU-T. "Subjective evaluation of speech quality with a crowdsourcing approach." ITU-T Recommendation P.808, 2021

Quality Evaluation - Mild reverberation and noise (real-world)



Listening Examples DCR

filename	real-world focus	DRED-SA 0.5kbps	DRED-SA 1kbps	DRED-SA 2kbps	Opus 1.5.2 9kbps
realworld_data_031.wav 4.75 	noisy child speech (fan/AC)	1.71 	1.71 	1.00 	3.8 
realworld_data_045.wav 4.25 	noisy male speech (traffic noise)	1.22 	1.94 	1.20 	3.15 
realworld_data_063.wav 4.29 	reverberant male speech	1.14 	1.40 	2.17 	3.50 
realworld_data_054.wav 4.80 	reverberant female speech	1.25 	2.65 	4.29 	2.25 

Speech Intelligibility Assessment

Diagnostic Rhyme Test (DRT)

What word do you hear?

PLAY

moss

boss

- ITU-recommended method [5]
- phoneme-level intelligibility assessed
- no context effects
- no learning effect
- successful test adaptation for crowdsourcing [6]
- test vectors in several languages and software to set up test open-sourced [7]

Table 2 – ITU-T P.807 test-items

Distinctive Feature	Sample-1				Sample-2				Sample-3				Sample-4			
	VWL	CONS	Present	Absent	VWL	CONS	Present	Absent	VWL	CONS	Present	Absent	VWL	CONS	Present	Absent
Voicing	ɑ	b/p (i)	BOND	POND	o	v/f (i)	VOLE	FOAL	æ	g/k (i)	GAFF	CALF	ɛ	d/t (i)	DENSE	TENSE
	u	z/s (i)	ZOO	SUE	ɔ	d/t (i)	DAUNT	TAUNT	l	dʒ/tʃ (i)	GIN	CHIN	i	b/p (i)	BEAN	PEEN
	ɛ	v/f (f)	REV	REF	l	dʒ/tʃ (f)	RIDGE	RICH	o	g/k (f)	BROGUE	BROKE	ɑ	dʒ/tʃ (f)	HODGE	HOTCH
	i	v/f (f)	SHEAVE	SHEAF	æ	g/k (f)	BAG	BACK	ɔ	z/s (f)	LAWS	LOSS	u	b/p (f)	LUBE	LOOP
Nasality	ɛ	n/d (i)	NECK	DECK	ɑ	n/d (i)	KNOCK	DOCK	ɔ	m/b (i)	MOSS	BOSS	æ	m/b (i)	MAD	BAD
	i	m/b (i)	MEAT	BEAT	u	m/b (i)	MOOT	BOOT	u	n/d (i)	NOTE	DOTE	l	n/d (i)	NIP	DIP
	ɑ	m/b (f)	BOMB	BOB	ɛ	m/b (f)	GEM	JEB	æ	n/d (f)	FAN	FAD	ɔ	n/d (f)	BRAWN	BROAD
	u	n/d (f)	NOON	NUDE	i	n/d (f)	SCREEN	SCREED	l	m/b (f)	RIM	RIB	u	n/d (f)	MOAN	MODE
Sustention	æ	ð/d (i)	THAN	DAN	ɛ	f/p (i)	FENCE	PENCE	ɑ	v/b (i)	VOX	BOX	ɔ	θ/t (i)	THONG	TONG
	l	θ/t (i)	THICK	TICK	i	ʃ/tʃ (i)	SHEET	CHEAT	u	ʃ/tʃ (i)	SHOES	CHOOSE	u	ð/d (i)	THOUGH	DOUGH
	ɔ	f/p (f)	GOFF	GAWP	ɑ	v/b (f)	SLAV	SLOB	i	ð/d (f)	SEETHE	SEED	æ	f/p (f)	CALF	CAP
	u	ð/d (f)	LOATHE	LOAD	u	f/p (f)	GOOF	GOOP	ɛ	ʃ/tʃ (f)	FLESH	FLETCH	l	v/b (f)	LIVE	LIB
Sibilation	ɔ	dʒ/g (i)	JAWS	GAUZE	æ	s/θ (i)	SANK	THANK	ɛ	tʃ/k (i)	CHAIR	CARE	ɑ	tʃ/k (i)	CHOP	COP
	o	dʒ/g (i)	JOE	GO	l	dʒ/g (i)	JILT	GUILT	i	z/θ (i)	ZEE	THEE	u	tʃ/k (i)	CHEW	COO
	æ	tʃ/k (f)	PATCH	PACK	ɔ	s/θ (f)	ROSS	WROTH	ɑ	tʃ/k (f)	NOTCH	KNOCK	ɛ	dʒ/g (f)	EDGE	EGG
	l	s/θ (f)	MISS	MYTH	u	s/θ (f)	GROSS	GROWTH	u	z/θ (f)	SUES	SOOTHE	i	z/θ (f)	BREEZE	BREATHE
Graveness	i	p/t (i)	PEAK	TEAK	l	f/θ (i)	FIN	THIN	u	f/θ (i)	FORE	THOR	u	p/t (i)	POOL	TOOL
	ɛ	m/n (i)	MET	NET	æ	b/d (i)	BANK	DANK	ɔ	b/d (i)	BONG	DONG	ɔ	w/r (i)	WAD	ROD
	u	m/n (f)	LOOM	LOON	u	b/d (f)	STROBE	STRODE	l	m/n (f)	SHIM	SHIN	i	f/θ (f)	REEF	WREATH
	ɑ	p/t (f)	HOP	HOT	ɔ	f/θ (f)	TROUGH	TROTH	æ	p/t (f)	RAP	RAT	ɛ	b/d (f)	WEB	WED
Compactness	l	h/f (i)	HIT	FIT	u	ʃ/s (i)	SHOW	SO	u	j/r (i)	YOU	RUE	ɛ	k/p (i)	KEG	PEG
	æ	g/b (i)	GAT	BAT	ɔ	k/t (i)	CAUGHT	TAUGHT	ɔ	g/d (i)	GOT	DOT	i	j/w (i)	YIELD	WIELD
	u	k/t (f)	OAK	OAT	l	g/b (f)	BIG	BIB	ɛ	g/d (f)	BEG	BED	ɑ	g/b (f)	NOG	KNOB
	ɔ	g/d (f)	FLOG	FLAWED	æ	ʃ/s (f)	CLASH	CLASS	i	k/p (f)	SEEK	SEEP	u	k/p (f)	DUKE	DUPE

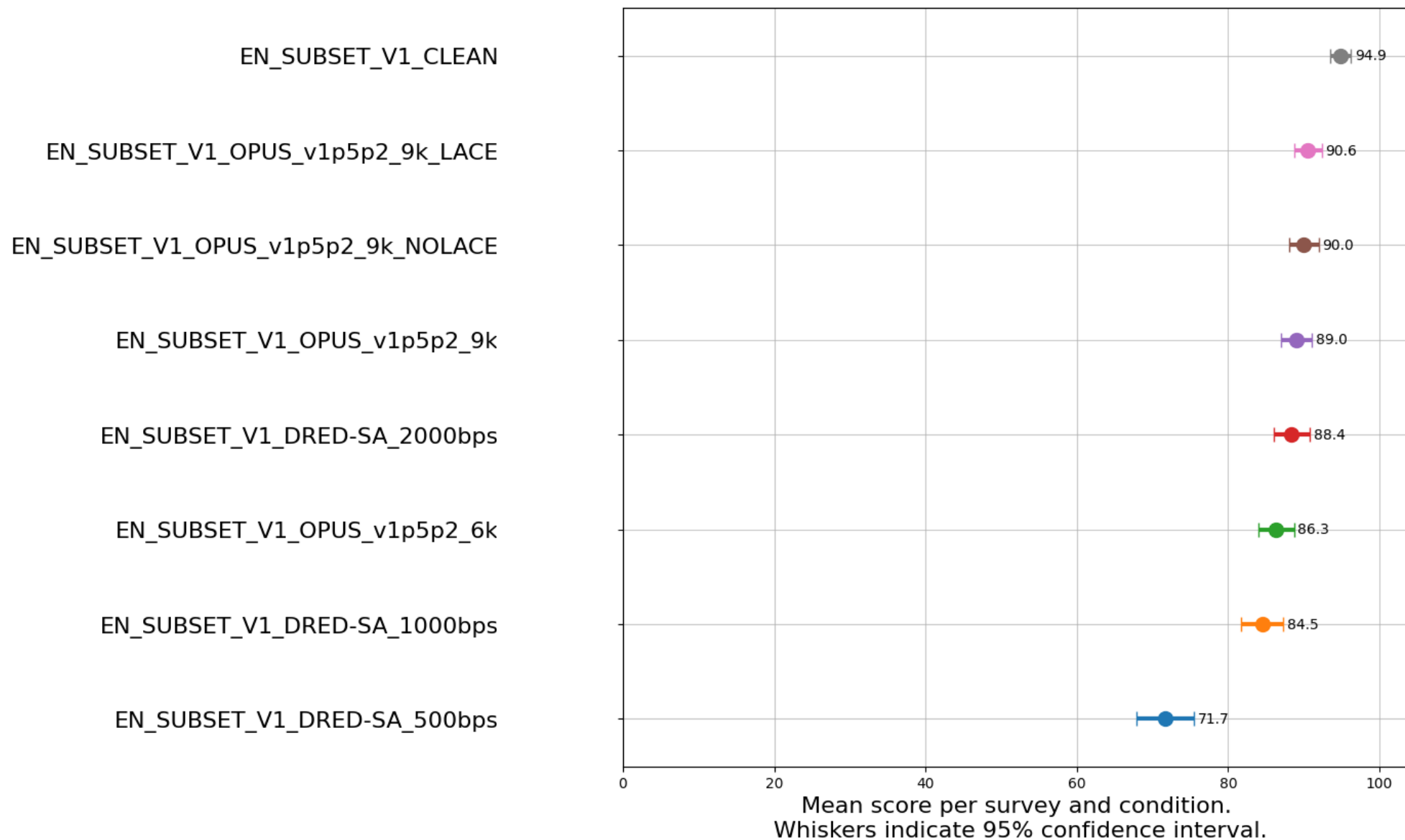
[5] ITU-T "Subjective test methodology for assessing speech intelligibility" ITU-T Recommendation P.807, 2016.

[6] Lechler, L. K. Wojcicki "Crowdsourced Multilingual Speech Intelligibility Testing." ICASSP 2024.

[7] Cisco Systems "Multilingual Speech Testing - Speech Intelligibility DRT" <https://github.com/cisco/multilingual-speech-testing/tree/main/speech-intelligibility-DRT>

Intelligibility Evaluation – Clean

Overall results of the Diagnostic Rhyme Test for English

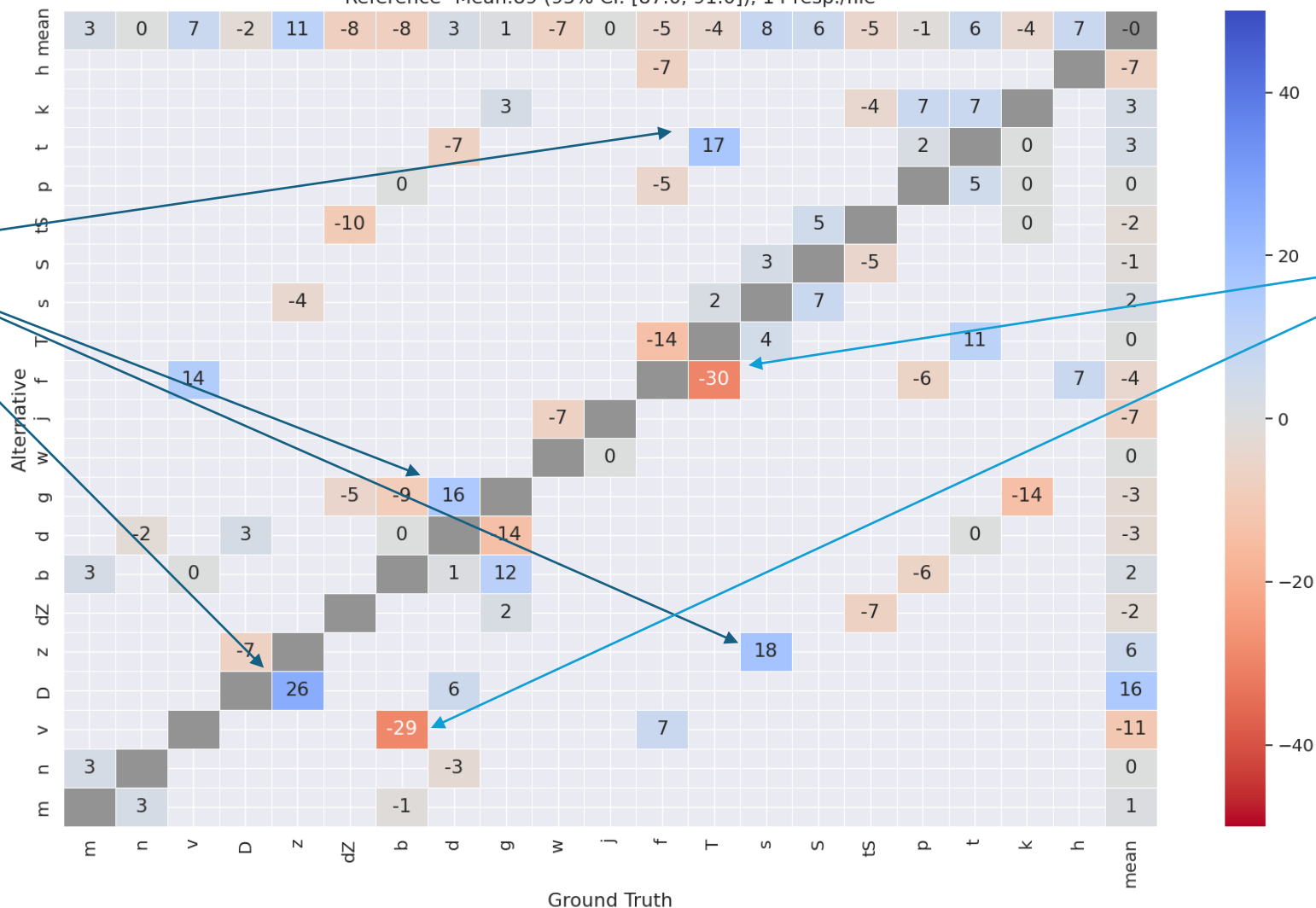






















Delta Phoneme Scores

Delta phoneme score matrix
 EN_SUBSET_V1_DRED-SA_2000bps_PROLIFIC vs EN_SUBSET_V1_OPUS_v1p5p2_9k_PROLIFIC [EN]
 Delta: -0.6; 95% CI: [-2.8, 1.6];
 Treatment--Mean: 88 (95% CI: [86.0, 90.8]), 15 resp./file
 Reference--Mean:89 (95% CI: [87.0, 91.0]), 14 resp./file

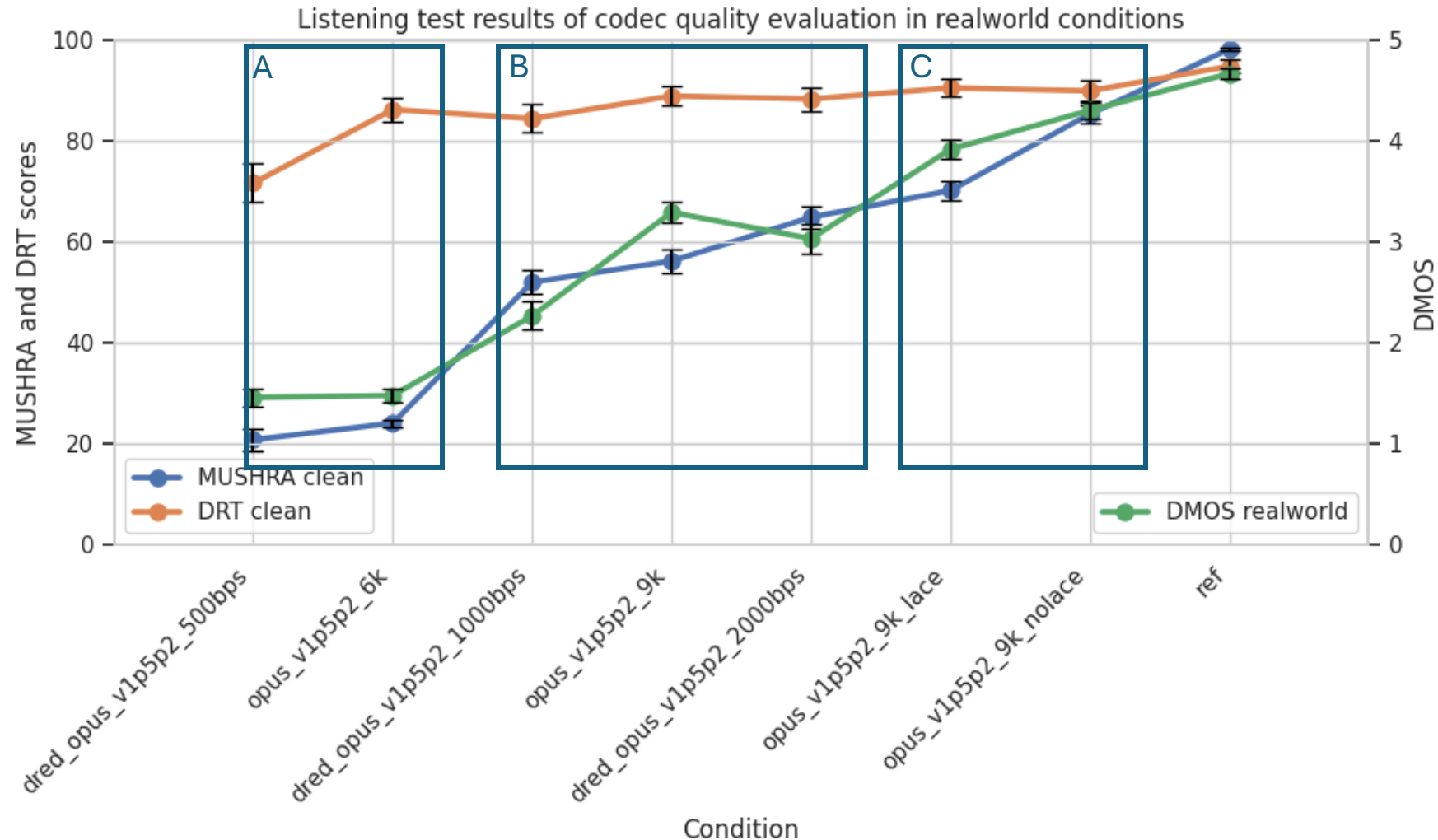
Some improvements
over Opus!

Some degradations



filename/ target word/ alternative	clean	Opus 1.5.2 9kbps	DRED-SA 2kbps	DRED-SA 1kbps	DRED-SA 0.5kbps
sues_A18Q27VK32D8M7_e76bf705fab c4ec38c78d4452491b5ec.wav sues soothe	75 	0 	73 	73 	73 
lib_AVLSFS5SYL5WG_f602fd78431a43 f6be02a222b7f1e539.wav lib live	100 	100 	-60 	-29 	-14 
net_A1NQOR4RQJIDMW_bc55648935 76480e9735215e4f27fb41.wav net met	100 	100 	47 	-60 	60 
got_A3942QGFYCSDRM_eea5384f918 540939b71d5d5cbb5baca.wav got dot	86 	87 	33 	-14 	-100 

Summary Results



DRED 0.5 kbps & Opus 6kbps:

- comparable quality
- Opus 6kbps better intelligibility

DRED 1kbps, Opus 9kbps, DRED 2kbps:

- DRED 2kbps best quality in clean
- Opus 9kbps highest resilience to real-world noise and reverberation
- Opus 9kbps slightly better intelligibility

Opus 9kbps LACE & NoLACE:

- NoLACE superior quality in clean and real-world conditions
- both LACE and NoLACE improve over Opus 9kbps in quality tests
- comparable intelligibility

Next steps?

Test Data	Test Type	Assessment of
Clean	MUSHRA-1S ¹	Quality with high-quality inputs
Real-world	DMOS	Quality in real-world reverberation & noise
Clean	DRT	Phoneme-level intelligibility
Noisy	DRT	Phoneme-level intelligibility in SSN
Multi-talker conversation	DMOS	Preservation of simultaneous talkers
Lossy	TBD	PLC and blending

Appendix

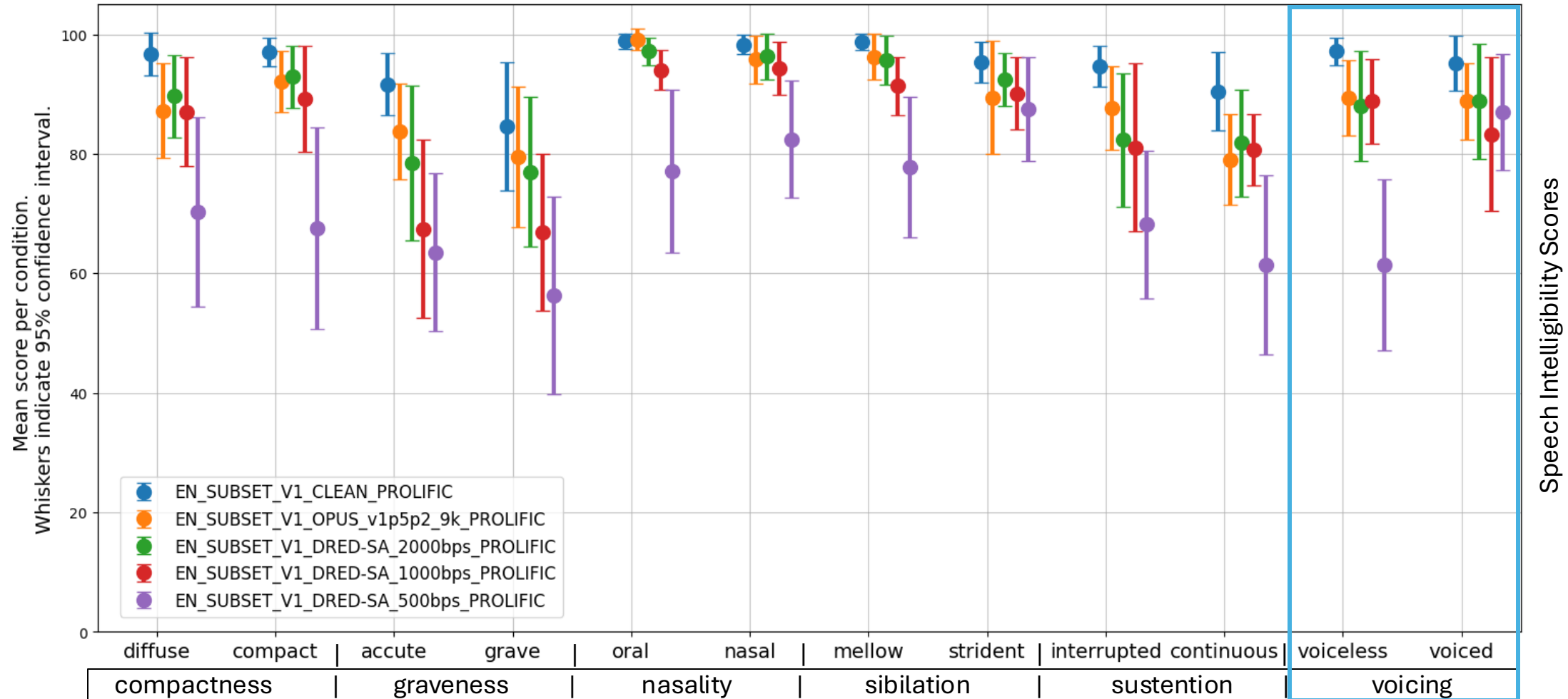
Results by Phonetic Feature State

Example for DRED 0.5 kbps:

In **voicing** category:

voiced consonants (e.g., /b/) generally **correct**

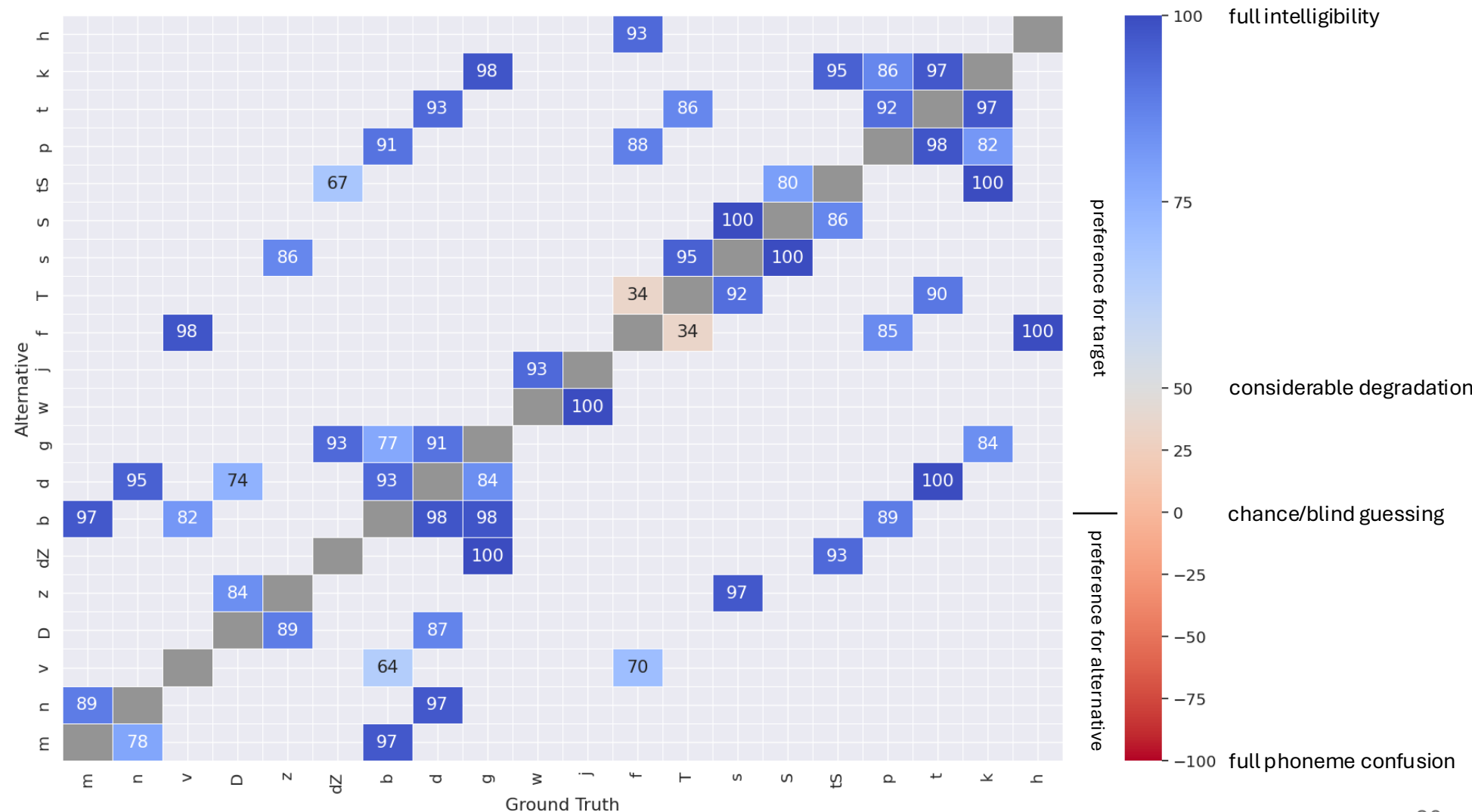
voiceless consonants (e.g., /p/) often **confused**



Results by Phoneme Pair

DRT scores (%_{correct} adjusted for guessing)

Phoneme score matrix for EN_SUBSET_V1_DRED-SA_2000bps_PROLIFIC [EN] (scores adjusted for guessing).
Overall mean: 88.4; 95% confidence interval: [86.0, 90.8]; average number of responses per file: 15.



Delta Phoneme Scores

DRT delta scores ($\text{DRT_score}_{\text{DRED2}} - \text{DRT_score}_{\text{CleanBaseline}}$)

