ICASSP 2022 DEEP NOISE SUPPRESSION CHALLENGE

Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, Robert Aichner

Paper ID: 9406

May 7-13, 2022, Virtual

May 22-27, 2022, In-Person

2022 IEEE International Conference on
Acoustics, Speech, and Signal Processing
icassp 2022
Singapore
human-centric signal processing
May 7 - 13, 2022 - Virtual (all paper presentations)
May 22 - 27, 2022 - In-Person
Main In-Person Venue: Marina Bay Sands Expo & Convention Center, Singapore
Satellite In-Person Location: Shenzhen, China
IEEE

Motivation

- 4th DNS challenge
- Current DNS still far from achieving superior speech quality (DSIG >=0)
- Previous challenge results showed DSIG <0 with noticeable Word accuracy (WAcc) degradation resulting from over-suppression of noise/speech distortions

What is New?

- Full-band – 48kHz recordings
- Baseline model for Personalized DNS track
- Blind testset containing mobile device scenarios
- WAcc is new objective metric
- Final score defined as average of WAcc and P.835 SIG, BAK, and OVRL
- Opensource DNSMOS P.835 and WAcc APIs

ICASSP 2022 Challenge Tracks

Track 1: Real-Time non-personalized DNS for full band speech

❑ The noise suppressor must take less than the stride time Ts (in ms) to process a frame of size T (in ms) on an Intel Core i5 quad-core machine clocked at 2.4 GHz or equivalent processors. E.g., Ts = T/2 for 50% overlap between frames. The total algorithmic latency allowed including the frame size T, stride time Ts, and any lookahead must be <= 40ms. If a real-time system has a frame length of 20ms with a stride of 10ms, it results in an algorithmic latency of 30ms, and thus the latency requirements are satisfied. If a frame size of 32ms with a stride of 16ms is used, resulting in an algorithmic latency of 48ms, then the latency requirements are not met as the total algorithmic latency exceeds 40ms. If the frame size (T) plus stride (Ts) represented as T1 = T+Ts is less than 40ms, then up to (40 - T1) ms of future information can be used.

Track 2: Real-Time Personalized DNS for full band speech

❑ Satisfy Track 1 requirements.

❑ 2.5 minutes of clean speech for enrollment of each unique target speaker in the test set is provided for adopting DNS/speaker embedding extractor for personalized denoising. This track has a separate dev test set and blind test set.

Training Datasets

	Clean Speech (read speech, singing speech, emotional speech, and non-English speech)	Noise	Room Impulse Responses (RIR)
Source	Librivox, VocalSet, CREMA-D (Emotional data), Non-English clips from OpenSLR18, THCHS-30, OpenSLR33, AISHELL, OpenSLR39, OpenSLR61, OpenSLR71, OpenSLR73, OpenSLR74 and OpenSLR75, Spoken Wikipedia Corpora, German Corpus for Kinect, M-AILABS	Audioset, freesound and DEMAND database	3076 real and 115000 synthetic RIRs, OpenSLR26 and OpenSLR28
Size	760 hours	181 hours	
Synthesizer default config	SNR range = -5 to 25 dB Target levels = -35 to -15 dB FS		

Blind Testset

- Common blind test set for both tracks helps elucidate the benefits of personalized denoising.
- Enrollment: 2.5 minutes of clean speech
- Only English language
- Contains 859 real test clips, each 10s duration.
- Collected on various desktop (30%) and mobile (70%) platforms using mTurk.
- Several iterations of data validation based on unit tests and human listening.
- Each testclips have a unique speaker and background noise type.
- Transcribed the blind test set using a third-party data annotation service. To ensure high accuracy, expert listening was conducted to correct the speech transcription.

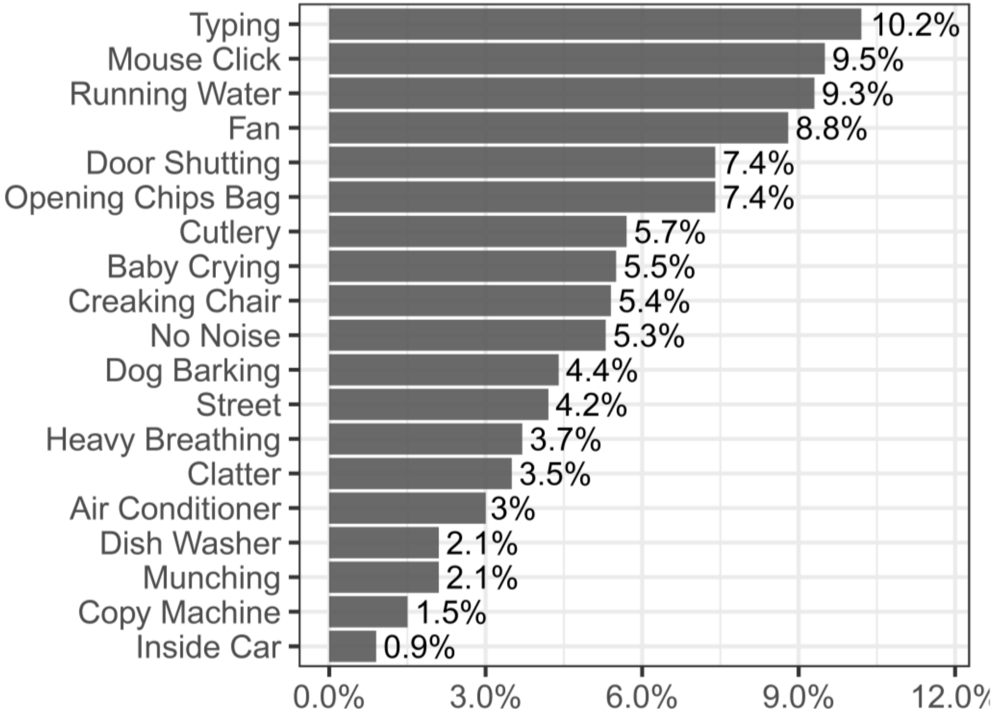
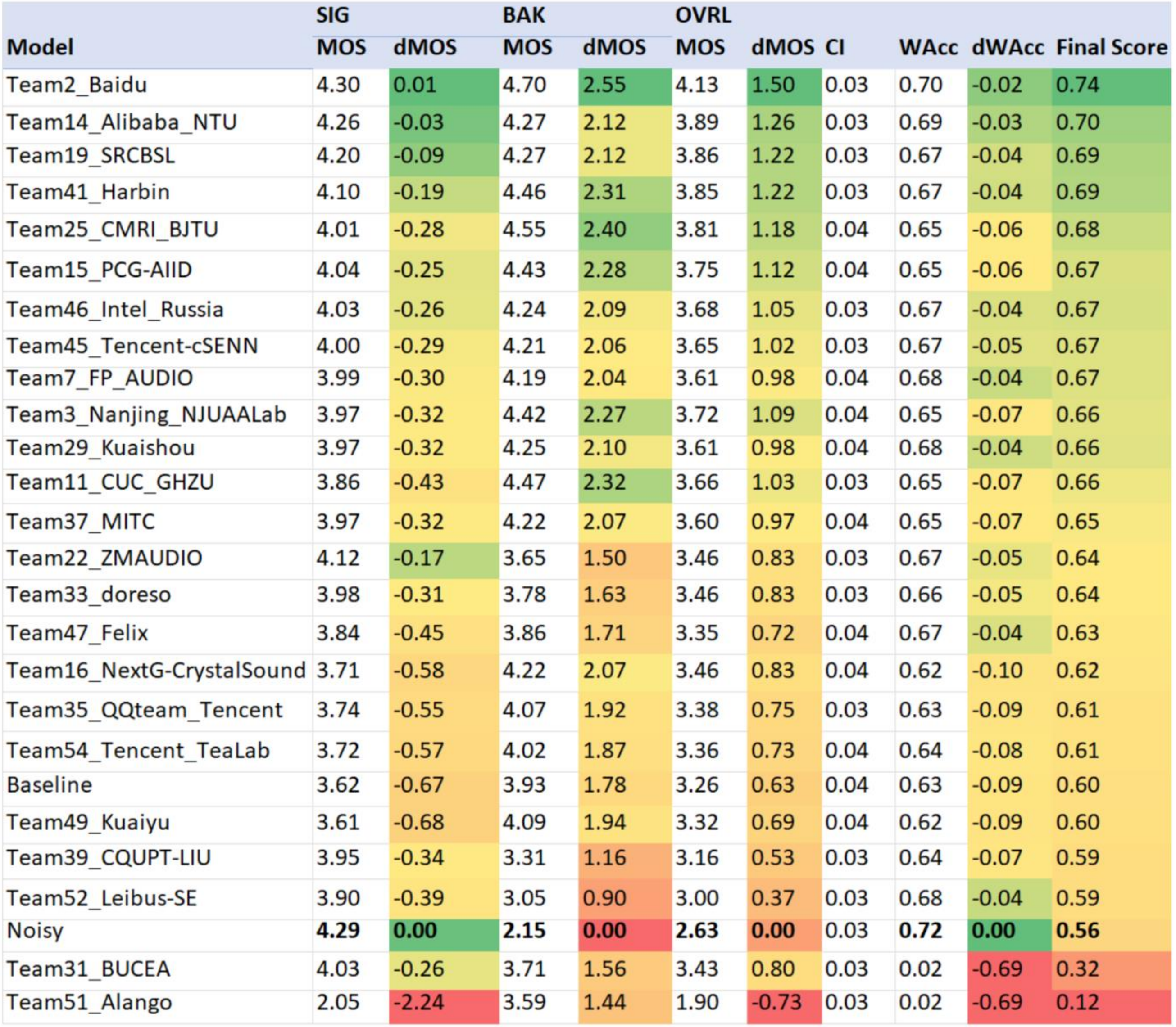


Fig. Distribution of noise types in our blind test set.

ITU-T P.835 framework for Subjective Evaluation

- P.835 provides three scores for each audio clips for overall speech quality (OVL), standalone quality scores of speech (SIG) and noise (BAK).
- Standalone ratings aim to narrow down areas of improvement to achieve better overall speech quality. It enables prioritizing speech quality (SIG) over suppression of background noise (BAK).
- Each test clip was rated by 5 qualified raters, which gave the maximum 95% CI of 0.05 DMOS per model
- Participants ranked based on Final Score given it satisfy real-time requirements. Participants are required to submit the number of operations per second of their model. This could be used as a tie-breaker.

Track 1 Non-PDNS Subjective Evaluation



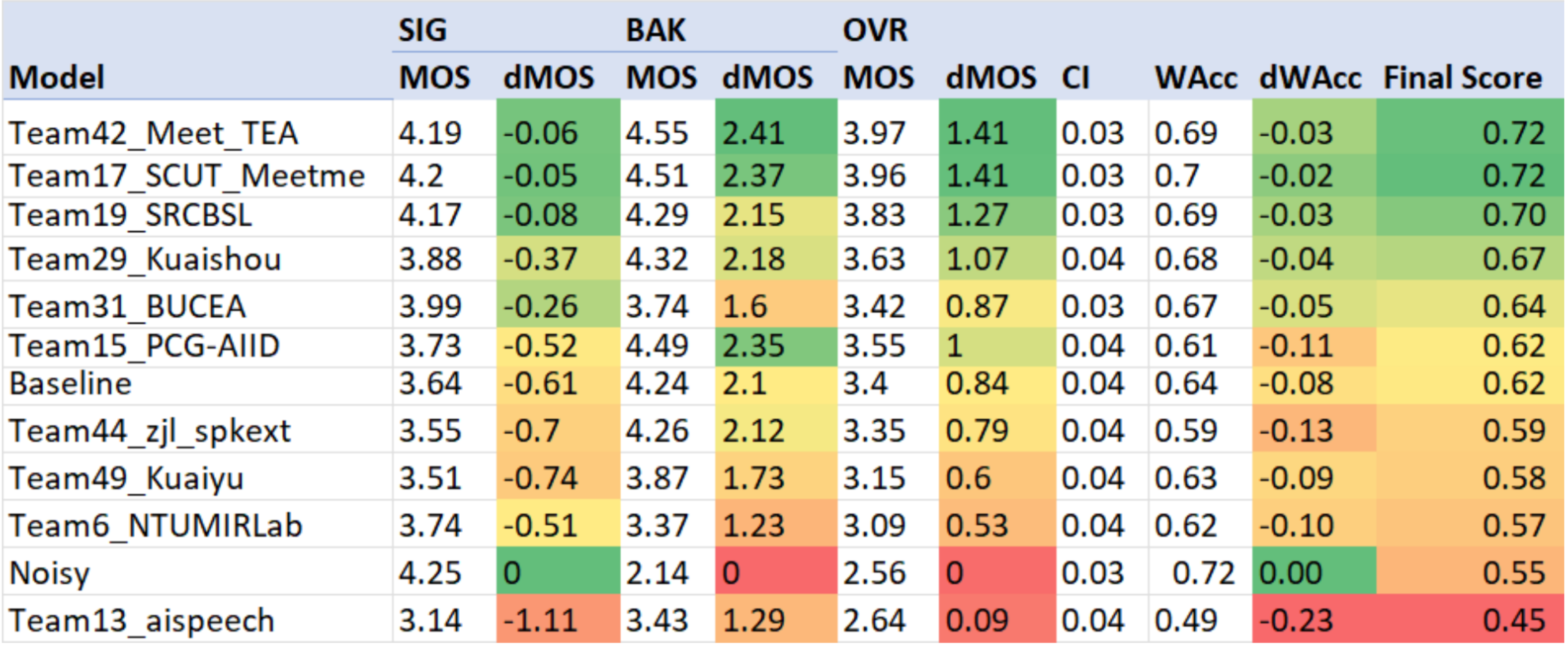
	SIG		BAK		OVRL					
Model	MOS	dMOS	MOS	dMOS	MOS	dMOS	CI	WAcc	dWAcc	Final Score
Team2_Baidu	4.30	0.01	4.70	2.55	4.13	1.50	0.03	0.70	-0.02	0.74
Team14_Alibaba_NTU	4.26	-0.03	4.27	2.12	3.89	1.26	0.03	0.69	-0.03	0.70
Team19_SRCBSL	4.20	-0.09	4.27	2.12	3.86	1.22	0.03	0.67	-0.04	0.69
Team41_Harbin	4.10	-0.19	4.46	2.31	3.85	1.22	0.03	0.67	-0.04	0.69
Team25_CMRI_BJTU	4.01	-0.28	4.55	2.40	3.81	1.18	0.04	0.65	-0.06	0.68
Team15_PCG-AIID	4.04	-0.25	4.43	2.28	3.75	1.12	0.04	0.65	-0.06	0.67
Team46_Intel_Russia	4.03	-0.26	4.24	2.09	3.68	1.05	0.03	0.67	-0.04	0.67
Team45_Tencent-cSENN	4.00	-0.29	4.21	2.06	3.65	1.02	0.03	0.67	-0.05	0.67
Team7_FP_AUDIO	3.99	-0.30	4.19	2.04	3.61	0.98	0.04	0.68	-0.04	0.67
Team3_Nanjing_NJUAALab	3.97	-0.32	4.42	2.27	3.72	1.09	0.04	0.65	-0.07	0.66
Team29_Kuaishou	3.97	-0.32	4.25	2.10	3.61	0.98	0.04	0.68	-0.04	0.66
Team11_CUC_GHZU	3.86	-0.43	4.47	2.32	3.66	1.03	0.03	0.65	-0.07	0.66
Team37_MITC	3.97	-0.32	4.22	2.07	3.60	0.97	0.04	0.65	-0.07	0.65
Team22_ZMAUDIO	4.12	-0.17	3.65	1.50	3.46	0.83	0.03	0.67	-0.05	0.64
Team33_doreso	3.98	-0.31	3.78	1.63	3.46	0.83	0.03	0.66	-0.05	0.64
Team47_Felix	3.84	-0.45	3.86	1.71	3.35	0.72	0.04	0.67	-0.04	0.63
Team16_NextG-CrystalSound	3.71	-0.58	4.22	2.07	3.46	0.83	0.04	0.62	-0.10	0.62
Team35_QQteam_Tencent	3.74	-0.55	4.07	1.92	3.38	0.75	0.03	0.63	-0.09	0.61
Team54_Tencent_TeaLab	3.72	-0.57	4.02	1.87	3.36	0.73	0.04	0.64	-0.08	0.61
Baseline	3.62	-0.67	3.93	1.78	3.26	0.63	0.04	0.63	-0.09	0.60
Team49_Kuaiyu	3.61	-0.68	4.09	1.94	3.32	0.69	0.04	0.62	-0.09	0.60
Team39_CQUPT-LIU	3.95	-0.34	3.31	1.16	3.16	0.53	0.03	0.64	-0.07	0.59
Team52_Leibus-SE	3.90	-0.39	3.05	0.90	3.00	0.37	0.03	0.68	-0.04	0.59
Noisy	**4.29**	**0.00**	**2.15**	**0.00**	**2.63**	**0.00**	0.03	**0.72**	**0.00**	**0.56**
Team31_BUCEA	4.03	-0.26	3.71	1.56	3.43	0.80	0.03	0.02	-0.69	0.32
Team51_Alango	2.05	-2.24	3.59	1.44	1.90	-0.73	0.03	0.02	-0.69	0.12

Track 1: 24 submissions

$$\text{Final score} = 0.5[\text{WAcc} + 0.25(\text{OVRL} - 1)]$$

Track 2 PDNS Subjective Evaluation

Track 2: 10 submissions

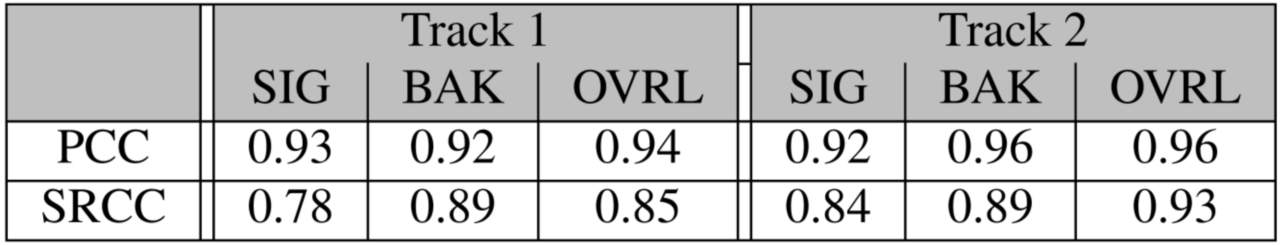


	SIG		BAK		OVR					
Model	MOS	dMOS	MOS	dMOS	MOS	dMOS	CI	WAcc	dWAcc	Final Score
Team42_Meet_TEA	4.19	-0.06	4.55	2.41	3.97	1.41	0.03	0.69	-0.03	0.72
Team17_SCUT_Meetme	4.2	-0.05	4.51	2.37	3.96	1.41	0.03	0.7	-0.02	0.72
Team19_SRCBSL	4.17	-0.08	4.29	2.15	3.83	1.27	0.03	0.69	-0.03	0.70
Team29_Kuaishou	3.88	-0.37	4.32	2.18	3.63	1.07	0.04	0.68	-0.04	0.67
Team31_BUCEA	3.99	-0.26	3.74	1.6	3.42	0.87	0.03	0.67	-0.05	0.64
Team15_PCG-AIID	3.73	-0.52	4.49	2.35	3.55	1	0.04	0.61	-0.11	0.62
Baseline	3.64	-0.61	4.24	2.1	3.4	0.84	0.04	0.64	-0.08	0.62
Team44_zjl_spkext	3.55	-0.7	4.26	2.12	3.35	0.79	0.04	0.59	-0.13	0.59
Team49_Kuaiyu	3.51	-0.74	3.87	1.73	3.15	0.6	0.04	0.63	-0.09	0.58
Team6_NTUMIRLab	3.74	-0.51	3.37	1.23	3.09	0.53	0.04	0.62	-0.10	0.57
Noisy	4.25	0	2.14	0	2.56	0	0.03	0.72	0.00	0.55
Team13_aispeech	3.14	-1.11	3.43	1.29	2.64	0.09	0.04	0.49	-0.23	0.45

Results: DNSMOS, Model size

- Performance of DNSMOS: The high correlation between subjective scores and DNSMOS P.835 in both tracks shows the efficacy of DNSMOS P.835 in ranking the DNS models.

Table 1. DNSMOS PCC and SRCC

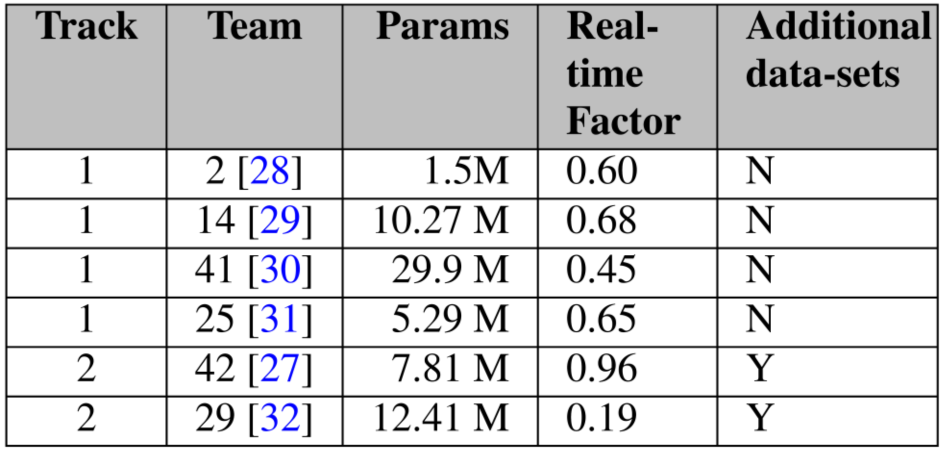


	Track 1			Track 2		
	SIG	BAK	OVRL	SIG	BAK	OVRL
PCC	0.93	0.92	0.94	0.92	0.96	0.96
SRCC	0.78	0.89	0.85	0.84	0.89	0.93

- Comparison of top teams

https://arxiv.org/abs/2202.13288

Table 2. Comparison of top performing models.



Track	Team	Params	Real-time Factor	Additional data-sets
1	2 [28]	1.5M	0.60	N
1	14 [29]	10.27 M	0.68	N
1	41 [30]	29.9 M	0.45	N
1	25 [31]	5.29 M	0.65	N
2	42 [27]	7.81 M	0.96	Y
2	29 [32]	12.41 M	0.19	Y

Results: ANOVA

- For the top performing teams, we ran an ANOVA test to determine statistical significance
- The 2nd, 3rd and 4th place are tied for Track 1. Likewise, the 1st and 2nd place for Track 2 are tied. Teams 17, 19, and 42 did not submit a paper so were disqualified per the challenge rules.

ANOVA results for Track-1

	Team2_Baidu	Team14_Alibaba_NTU	Team19_SRCBSL	Team41_Harbin	Team25_CMRI_BJTU
Team2_Baidu	1	0	0	0	0
Team14_Alibaba_NTU	0	1	0.21	0.10	0
Team19_SRCBSL	0	0.19	1	0.79	0.03
Team41_Harbin	0	0.15	0.89	1	0.05
Team25_CMRI_BJTU	0	0	0.04	0.07	1

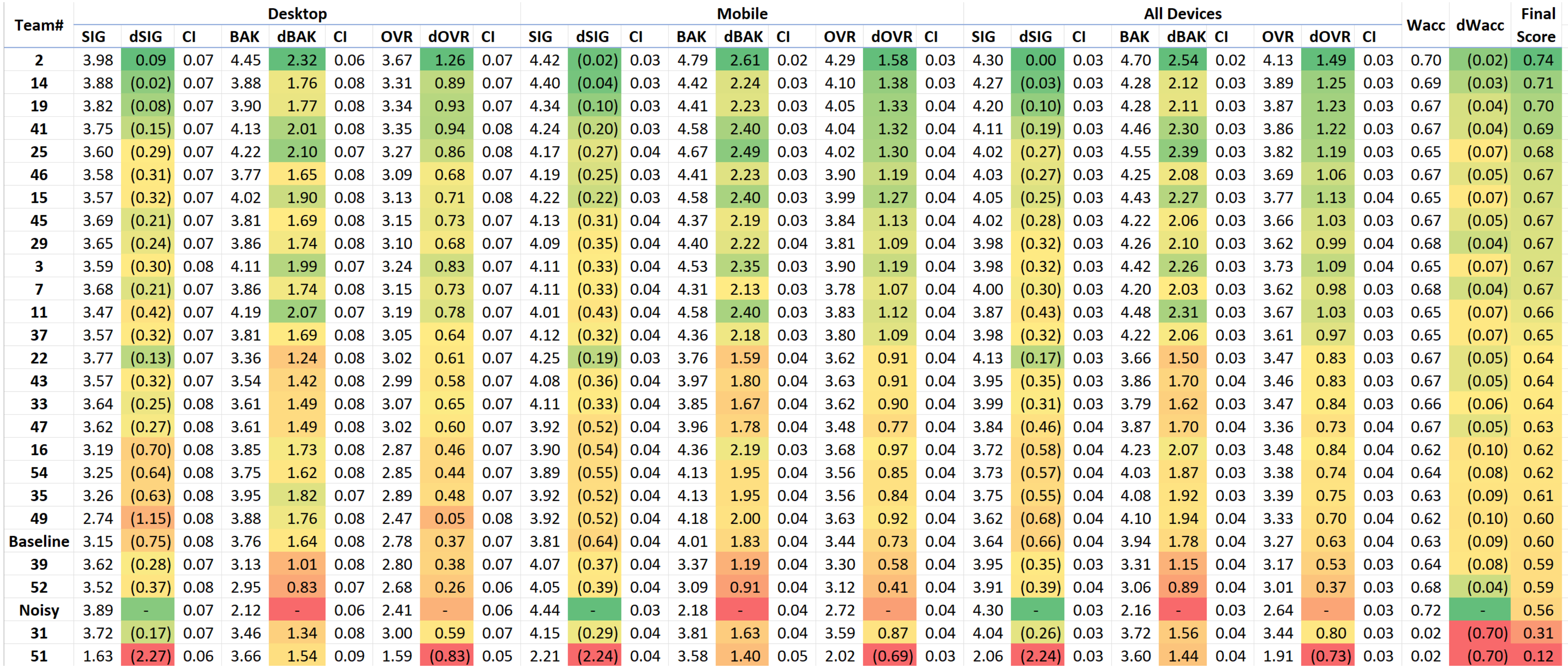
ANOVA results for Track-2

	Team42_Meet_TEA	Team17_SCUT_Meetme	Team19_SRCBSL	Team29_Kuaishou	Team15_PCG-AIID	Team31_BUCEA_Yu
Team42_Meet_TEA	1.00	0.70	0.00	0.00	0.00	0.00
Team17_SCUT_Meetme	0.70	1.00	0.00	0.00	0.00	0.00
Team19_SRCBSL	0.00	0.00	1.00	0.00	0.00	0.00
Team29_Kuaishou	0.00	0.00	0.00	1.00	0.00	0.00
Team15_PCG-AIID	0.00	0.00	0.00	0.00	1.00	0.00
Team31_BUCEA_Yu	0.00	0.00	0.00	0.00	0.00	1.00

https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-icassp-2022/results/

Results: Mobile vs Desktop Track 1

- MOS scores for clips recorded on mobile devices is higher than those from desktop devices suggesting that mobile had better acoustic devices or environments than the desktop scenarios.



Team#	Desktop									Mobile									All Devices									Wacc	dWacc	Final
	SIG	dSIG	CI	BAK	dBAK	CI	OVR	dOVR	CI	SIG	dSIG	CI	BAK	dBAK	CI	OVR	dOVR	CI	SIG	dSIG	CI	BAK	dBAK	CI	OVR	dOVR	CI			Score
2	3.98	0.09	0.07	4.45	2.32	0.06	3.67	1.26	0.07	4.42	(0.02)	0.03	4.79	2.61	0.02	4.29	1.58	0.03	4.30	0.00	0.03	4.70	2.54	0.02	4.13	1.49	0.03	0.70	(0.02)	0.74
14	3.88	(0.02)	0.07	3.88	1.76	0.08	3.31	0.89	0.07	4.40	(0.04)	0.03	4.42	2.24	0.03	4.10	1.38	0.03	4.27	(0.03)	0.03	4.28	2.12	0.03	3.89	1.25	0.03	0.69	(0.03)	0.71
19	3.82	(0.08)	0.07	3.90	1.77	0.08	3.34	0.93	0.07	4.34	(0.10)	0.03	4.41	2.23	0.03	4.05	1.33	0.04	4.20	(0.10)	0.03	4.28	2.11	0.03	3.87	1.23	0.03	0.67	(0.04)	0.70
41	3.75	(0.15)	0.07	4.13	2.01	0.08	3.35	0.94	0.08	4.24	(0.20)	0.03	4.58	2.40	0.03	4.04	1.32	0.04	4.11	(0.19)	0.03	4.46	2.30	0.03	3.86	1.22	0.03	0.67	(0.04)	0.69
25	3.60	(0.29)	0.07	4.22	2.10	0.07	3.27	0.86	0.08	4.17	(0.27)	0.04	4.67	2.49	0.03	4.02	1.30	0.04	4.02	(0.27)	0.03	4.55	2.39	0.03	3.82	1.19	0.03	0.65	(0.07)	0.68
46	3.58	(0.31)	0.07	3.77	1.65	0.08	3.09	0.68	0.07	4.19	(0.25)	0.03	4.41	2.23	0.03	3.90	1.19	0.04	4.03	(0.27)	0.03	4.25	2.08	0.03	3.69	1.06	0.03	0.67	(0.05)	0.67
15	3.57	(0.32)	0.07	4.02	1.90	0.08	3.13	0.71	0.08	4.22	(0.22)	0.03	4.58	2.40	0.03	3.99	1.27	0.04	4.05	(0.25)	0.03	4.43	2.27	0.03	3.77	1.13	0.04	0.65	(0.07)	0.67
45	3.69	(0.21)	0.07	3.81	1.69	0.08	3.15	0.73	0.07	4.13	(0.31)	0.04	4.37	2.19	0.03	3.84	1.13	0.04	4.02	(0.28)	0.03	4.22	2.06	0.03	3.66	1.03	0.03	0.67	(0.05)	0.67
29	3.65	(0.24)	0.07	3.86	1.74	0.08	3.10	0.68	0.07	4.09	(0.35)	0.04	4.40	2.22	0.04	3.81	1.09	0.04	3.98	(0.32)	0.03	4.26	2.10	0.03	3.62	0.99	0.04	0.68	(0.04)	0.67
3	3.59	(0.30)	0.08	4.11	1.99	0.07	3.24	0.83	0.07	4.11	(0.33)	0.04	4.53	2.35	0.03	3.90	1.19	0.04	3.98	(0.32)	0.03	4.42	2.26	0.03	3.73	1.09	0.04	0.65	(0.07)	0.67
7	3.68	(0.21)	0.07	3.86	1.74	0.08	3.15	0.73	0.07	4.11	(0.33)	0.04	4.31	2.13	0.03	3.78	1.07	0.04	4.00	(0.30)	0.03	4.20	2.03	0.03	3.62	0.98	0.03	0.68	(0.04)	0.67
11	3.47	(0.42)	0.07	4.19	2.07	0.07	3.19	0.78	0.07	4.01	(0.43)	0.04	4.58	2.40	0.03	3.83	1.12	0.04	3.87	(0.43)	0.03	4.48	2.31	0.03	3.67	1.03	0.03	0.65	(0.07)	0.66
37	3.57	(0.32)	0.07	3.81	1.69	0.08	3.05	0.64	0.07	4.12	(0.32)	0.04	4.36	2.18	0.03	3.80	1.09	0.04	3.98	(0.32)	0.03	4.22	2.06	0.03	3.61	0.97	0.03	0.65	(0.07)	0.65
22	3.77	(0.13)	0.07	3.36	1.24	0.08	3.02	0.61	0.07	4.25	(0.19)	0.03	3.76	1.59	0.04	3.62	0.91	0.04	4.13	(0.17)	0.03	3.66	1.50	0.03	3.47	0.83	0.03	0.67	(0.05)	0.64
43	3.57	(0.32)	0.07	3.54	1.42	0.08	2.99	0.58	0.07	4.08	(0.36)	0.04	3.97	1.80	0.04	3.63	0.91	0.04	3.95	(0.35)	0.03	3.86	1.70	0.04	3.46	0.83	0.03	0.67	(0.05)	0.64
33	3.64	(0.25)	0.08	3.61	1.49	0.08	3.07	0.65	0.07	4.11	(0.33)	0.04	3.85	1.67	0.04	3.62	0.90	0.04	3.99	(0.31)	0.03	3.79	1.62	0.03	3.47	0.84	0.03	0.66	(0.06)	0.64
47	3.62	(0.27)	0.08	3.61	1.49	0.08	3.02	0.60	0.07	3.92	(0.52)	0.04	3.96	1.78	0.04	3.48	0.77	0.04	3.84	(0.46)	0.04	3.87	1.70	0.04	3.36	0.73	0.04	0.67	(0.05)	0.63
16	3.19	(0.70)	0.08	3.85	1.73	0.08	2.87	0.46	0.07	3.90	(0.54)	0.04	4.36	2.19	0.03	3.68	0.97	0.04	3.72	(0.58)	0.04	4.23	2.07	0.03	3.48	0.84	0.04	0.62	(0.10)	0.62
54	3.25	(0.64)	0.08	3.75	1.62	0.08	2.85	0.44	0.07	3.89	(0.55)	0.04	4.13	1.95	0.04	3.56	0.85	0.04	3.73	(0.57)	0.04	4.03	1.87	0.03	3.38	0.74	0.04	0.64	(0.08)	0.62
35	3.26	(0.63)	0.08	3.95	1.82	0.07	2.89	0.48	0.07	3.92	(0.52)	0.04	4.13	1.95	0.04	3.56	0.84	0.04	3.75	(0.55)	0.04	4.08	1.92	0.03	3.39	0.75	0.03	0.63	(0.09)	0.61
49	2.74	(1.15)	0.08	3.88	1.76	0.08	2.47	0.05	0.08	3.92	(0.52)	0.04	4.18	2.00	0.04	3.63	0.92	0.04	3.62	(0.68)	0.04	4.10	1.94	0.04	3.33	0.70	0.04	0.62	(0.10)	0.60
Baseline	3.15	(0.75)	0.08	3.76	1.64	0.08	2.78	0.37	0.07	3.81	(0.64)	0.04	4.01	1.83	0.04	3.44	0.73	0.04	3.64	(0.66)	0.04	3.94	1.78	0.04	3.27	0.63	0.04	0.63	(0.09)	0.60
39	3.62	(0.28)	0.07	3.13	1.01	0.08	2.80	0.38	0.07	4.07	(0.37)	0.04	3.37	1.19	0.04	3.30	0.58	0.04	3.95	(0.35)	0.03	3.31	1.15	0.04	3.17	0.53	0.03	0.64	(0.08)	0.59
52	3.52	(0.37)	0.08	2.95	0.83	0.07	2.68	0.26	0.06	4.05	(0.39)	0.04	3.09	0.91	0.04	3.12	0.41	0.04	3.91	(0.39)	0.04	3.06	0.89	0.04	3.01	0.37	0.03	0.68	(0.04)	0.59
Noisy	3.89	-	0.07	2.12	-	0.06	2.41	-	0.06	4.44	-	0.03	2.18	-	0.04	2.72	-	0.04	4.30	-	0.03	2.16	-	0.03	2.64	-	0.03	0.72	-	0.56
31	3.72	(0.17)	0.07	3.46	1.34	0.08	3.00	0.59	0.07	4.15	(0.29)	0.04	3.81	1.63	0.04	3.59	0.87	0.04	4.04	(0.26)	0.03	3.72	1.56	0.04	3.44	0.80	0.03	0.02	(0.70)	0.31
51	1.63	(2.27)	0.06	3.66	1.54	0.09	1.59	(0.83)	0.05	2.21	(2.24)	0.04	3.58	1.40	0.05	2.02	(0.69)	0.03	2.06	(2.24)	0.03	3.60	1.44	0.04	1.91	(0.73)	0.03	0.02	(0.70)	0.12

Summary

- V4 challenge models provided feasibility of superior DNS performance
- Most successful DNS Challenge yet, both in terms of number of participants and quality of the models
- DSIG >= 0 seems must for winning, it almost eliminates WAcc degradations. Models are ranked using Final scores.
- Winning model shows new interesting test case for headset scenarios where neighboring speaker is in far-field
- For the top performing teams, we ran an ANOVA test to determine statistical significance (see https://aka.ms/dns-challenge). The 2nd, 3rd and 4th place are tied for Track 1. Likewise, the 1st and 2nd place for Track 2 are tied. Teams 17, 19, and 42 did not submit the ICASSP paper hence disqualified.
- Organizing team conducted the reviews of papers. Only top models were invited to submit paper.

What is Next for 5th DNS Challenge?

- Detecting faked/spoofed neighboring speakers and noise is essential to ensure representative testset
- Creating new spec for DNS testset- headset, personalized etc.
- To add diversity in testset – more speakers, more languages, accents, and devices, scenarios (emotional, paralinguistics), device & language mis-match in personalized DNS
- Create approach for model validation of challenge participants. Strong indications that some teams utilize non-causal models or different models for different scenarios.
- Create inference engine for computing the model complexity/inference time for all challenge models. Further, include a validation of the lookahead to ensure fair comparison.
- Adding CCR MOS in addition to ACR MOS to detect suppression of emotional/paralinguistic speech etc.