

ON TRAINING TARGETS AND OBJECTIVE FUNCTIONS FOR DEEP-LEARNING-BASED AUDIO-VISUAL SPEECH ENHANCEMENT

SUPPLEMENTARY MATERIAL

Daniel Michelsanti¹, Zheng-Hua Tan¹, Sigurdur Sigurdsson², Jesper Jensen^{1,2}

¹ Aalborg University, Department of Electronic Systems, Denmark

² Oticon A/S, Denmark

{danmi,zt,jje}@es.aau.dk {ssig,jesj}@oticon.com

Table 1: Convolutional layers of the audio and video encoders.

Video Encoder				Audio Encoder		
Layer	Filters	Kernel	Stride	Filters	Kernel	Stride
1	128	5×5	1×1	64	5×5	2×2
2	128	5×5	1×1	64	4×4	2×1
3	256	3×3	1×1	128	4×4	2×2
4	256	3×3	1×1	128	2×2	2×1
5	512	3×3	1×1	128	2×2	2×1
6	512	3×3	1×1	128	2×2	2×1

Table 2: PESQ and ESTOI results in the audio-visual setting where the scores are averaged across all the SNRs.

PESQ	Seen Speakers						Unseen Speakers					
Noise Type	BBL	BUS	CAF	PED	STR	SSN	BBL	BUS	CAF	PED	STR	SSN
Unproc.	1.32	1.44	1.30	1.30	1.30	1.29	1.31	1.44	1.31	1.30	1.30	1.28
STSA-DM	1.67	1.97	1.77	1.72	1.79	1.58	1.53	1.84	1.64	1.59	1.67	1.50
LSA-DM	1.85	2.22	1.98	1.90	2.03	1.71	1.65	2.05	1.81	1.75	1.88	1.57
MSA-DM	1.65	1.94	1.78	1.72	1.79	1.58	1.52	1.82	1.64	1.60	1.68	1.50
LMSA-DM	1.72	1.95	1.83	1.78	1.86	1.60	1.53	1.77	1.63	1.60	1.67	1.45
PSSA-DM	1.61	1.91	1.71	1.66	1.76	1.60	1.49	1.78	1.58	1.55	1.65	1.51
STSA-IM	1.60	1.90	1.72	1.66	1.73	1.54	1.47	1.78	1.58	1.54	1.62	1.46
LSA-IM	1.71	2.00	1.76	1.72	1.83	1.57	1.56	1.89	1.66	1.62	1.73	1.46
MSA-IM	1.67	1.96	1.78	1.72	1.77	1.56	1.52	1.85	1.64	1.60	1.66	1.48
LMSA-IM	1.78	2.04	1.85	1.79	1.89	1.61	1.58	1.88	1.69	1.64	1.74	1.48
PSSA-IM	1.64	1.96	1.74	1.70	1.79	1.58	1.51	1.83	1.61	1.58	1.67	1.51
STSA-MA	1.84	2.09	1.94	1.90	1.98	1.64	1.64	1.93	1.76	1.72	1.81	1.51
PSSA-MA	1.85	2.19	1.93	1.90	2.02	1.68	1.71	2.07	1.82	1.79	1.92	1.59
ESTOI	Seen Speakers						Unseen Speakers					
Noise Type	BBL	BUS	CAF	PED	STR	SSN	BBL	BUS	CAF	PED	STR	SSN
Unproc.	0.33	0.46	0.38	0.33	0.37	0.33	0.33	0.45	0.37	0.33	0.37	0.32
STSA-DM	0.51	0.62	0.57	0.54	0.57	0.49	0.42	0.57	0.50	0.47	0.51	0.42
LSA-DM	0.52	0.63	0.58	0.55	0.58	0.49	0.42	0.58	0.51	0.48	0.52	0.42
MSA-DM	0.52	0.63	0.58	0.55	0.58	0.50	0.44	0.58	0.51	0.48	0.52	0.44
LMSA-DM	0.54	0.64	0.59	0.57	0.59	0.49	0.44	0.58	0.52	0.48	0.52	0.42
PSSA-DM	0.49	0.61	0.54	0.51	0.54	0.47	0.41	0.56	0.49	0.45	0.49	0.42
STSA-IM	0.51	0.62	0.56	0.53	0.56	0.48	0.42	0.57	0.50	0.47	0.51	0.43
LSA-IM	0.51	0.61	0.56	0.53	0.56	0.48	0.42	0.56	0.49	0.46	0.50	0.40
MSA-IM	0.53	0.63	0.58	0.55	0.58	0.50	0.44	0.58	0.51	0.48	0.52	0.43
LMSA-IM	0.54	0.64	0.59	0.56	0.59	0.50	0.43	0.58	0.51	0.48	0.52	0.41
PSSA-IM	0.49	0.61	0.55	0.51	0.55	0.48	0.42	0.57	0.50	0.46	0.50	0.43
STSA-MA	0.55	0.65	0.60	0.57	0.60	0.52	0.45	0.60	0.53	0.49	0.54	0.44
PSSA-MA	0.50	0.62	0.56	0.52	0.56	0.48	0.44	0.59	0.52	0.48	0.52	0.44

Table 3: PESQ and ESTOI results in the audio-visual setting where the scores are averaged across all the noise types.

PESQ	Seen Speakers										Unseen Speakers									
SNR (dB)	-20	-15	-10	-5	0	5	10	15	20		-20	-15	-10	-5	0	5	10	15	20	
Unproc.	1.10	1.09	1.08	1.08	1.11	1.20	1.39	1.71	2.14		1.10	1.10	1.09	1.08	1.11	1.20	1.39	1.70	2.13	
STSA-DM	1.21	1.27	1.35	1.48	1.65	1.86	2.08	2.31	2.53		1.11	1.13	1.19	1.30	1.48	1.73	1.99	2.24	2.47	
LSA-DM	1.17	1.24	1.37	1.57	1.84	2.14	2.45	2.74	3.01		1.12	1.15	1.23	1.37	1.59	1.91	2.25	2.57	2.86	
MSA-DM	1.21	1.27	1.36	1.49	1.67	1.87	2.07	2.28	2.47		1.11	1.14	1.20	1.32	1.51	1.75	1.99	2.21	2.41	
LMSA-DM	1.20	1.27	1.39	1.56	1.78	2.01	2.18	2.31	2.40		1.12	1.15	1.22	1.34	1.53	1.77	1.98	2.14	2.25	
PSSA-DM	1.19	1.24	1.32	1.44	1.61	1.82	2.04	2.25	2.46		1.11	1.13	1.18	1.28	1.45	1.70	1.94	2.17	2.38	
STSA-IM	1.19	1.24	1.33	1.45	1.61	1.77	1.95	2.19	2.49		1.10	1.13	1.18	1.28	1.44	1.65	1.87	2.11	2.43	
LSA-IM	1.14	1.17	1.25	1.39	1.60	1.89	2.19	2.49	2.77		1.11	1.13	1.17	1.28	1.46	1.72	2.02	2.34	2.64	
MSA-IM	1.20	1.26	1.34	1.47	1.64	1.85	2.07	2.30	2.56		1.11	1.13	1.19	1.29	1.47	1.71	1.98	2.24	2.51	
LMSA-IM	1.16	1.21	1.32	1.48	1.72	1.99	2.26	2.53	2.77		1.11	1.13	1.19	1.30	1.49	1.76	2.06	2.35	2.61	
PSSA-IM	1.23	1.29	1.37	1.50	1.68	1.87	2.05	2.22	2.41		1.13	1.16	1.22	1.33	1.51	1.74	1.96	2.15	2.36	
STSA-MA	1.24	1.31	1.42	1.57	1.78	2.02	2.29	2.58	2.88		1.12	1.15	1.21	1.32	1.52	1.81	2.15	2.48	2.80	
PSSA-MA	1.21	1.28	1.38	1.54	1.78	2.08	2.40	2.71	2.98		1.14	1.18	1.25	1.38	1.61	1.95	2.31	2.63	2.90	

ESTOI	Seen Speakers										Unseen Speakers									
SNR (dB)	-20	-15	-10	-5	0	5	10	15	20		-20	-15	-10	-5	0	5	10	15	20	
Unproc.	0.04	0.08	0.15	0.24	0.35	0.47	0.58	0.67	0.74		0.04	0.08	0.14	0.23	0.34	0.46	0.57	0.66	0.73	
STSA-DM	0.30	0.35	0.41	0.49	0.57	0.64	0.70	0.74	0.77		0.19	0.23	0.29	0.39	0.49	0.59	0.67	0.72	0.76	
LSA-DM	0.30	0.35	0.41	0.49	0.58	0.65	0.71	0.76	0.79		0.20	0.24	0.30	0.39	0.49	0.60	0.68	0.73	0.76	
MSA-DM	0.31	0.36	0.42	0.49	0.57	0.64	0.70	0.74	0.77		0.20	0.24	0.31	0.40	0.51	0.61	0.68	0.73	0.76	
LMSA-DM	0.33	0.37	0.44	0.51	0.60	0.66	0.71	0.75	0.77		0.21	0.25	0.31	0.40	0.51	0.61	0.68	0.72	0.75	
PSSA-DM	0.23	0.29	0.36	0.46	0.56	0.64	0.70	0.74	0.77		0.14	0.19	0.27	0.37	0.49	0.60	0.68	0.72	0.76	
STSA-IM	0.28	0.33	0.40	0.48	0.56	0.64	0.69	0.74	0.78		0.19	0.23	0.29	0.39	0.50	0.60	0.67	0.72	0.76	
LSA-IM	0.29	0.33	0.38	0.46	0.55	0.63	0.70	0.75	0.79		0.19	0.22	0.28	0.36	0.46	0.57	0.66	0.73	0.77	
MSA-IM	0.32	0.36	0.42	0.50	0.58	0.65	0.70	0.75	0.78		0.21	0.25	0.31	0.40	0.50	0.60	0.68	0.73	0.77	
LMSA-IM	0.32	0.36	0.42	0.50	0.59	0.66	0.72	0.76	0.79		0.20	0.24	0.30	0.38	0.49	0.60	0.68	0.73	0.77	
PSSA-IM	0.24	0.29	0.37	0.46	0.56	0.64	0.70	0.75	0.78		0.16	0.21	0.28	0.38	0.50	0.61	0.68	0.73	0.77	
STSA-MA	0.35	0.39	0.45	0.52	0.60	0.67	0.72	0.77	0.80		0.22	0.26	0.32	0.41	0.51	0.62	0.70	0.75	0.78	
PSSA-MA	0.24	0.29	0.36	0.46	0.57	0.66	0.72	0.77	0.80		0.17	0.22	0.29	0.40	0.52	0.63	0.70	0.75	0.79	

Table 4: PESQ and ESTOI results in the audio-only setting where the scores are averaged across all the noise types.

PESQ	Seen Speakers										Unseen Speakers									
SNR (dB)	-20	-15	-10	-5	0	5	10	15	20		-20	-15	-10	-5	0	5	10	15	20	
Unproc.	1.10	1.09	1.08	1.08	1.11	1.20	1.39	1.71	2.14		1.10	1.10	1.09	1.08	1.11	1.20	1.39	1.70	2.13	
STSA-DM	1.11	1.14	1.20	1.32	1.51	1.74	1.97	2.21	2.44		1.11	1.13	1.19	1.30	1.49	1.72	1.95	2.18	2.40	
LSA-DM	1.10	1.14	1.22	1.37	1.61	1.94	2.28	2.60	2.87		1.10	1.14	1.21	1.34	1.58	1.89	2.22	2.52	2.77	
MSA-DM	1.10	1.14	1.20	1.33	1.51	1.74	1.96	2.17	2.37		1.11	1.13	1.19	1.31	1.49	1.72	1.94	2.15	2.34	
LMSA-DM	1.09	1.12	1.18	1.32	1.54	1.82	2.08	2.29	2.42		1.10	1.12	1.18	1.30	1.52	1.78	2.01	2.19	2.31	
PSSA-DM	1.12	1.16	1.23	1.35	1.55	1.77	2.01	2.24	2.45		1.11	1.14	1.21	1.32	1.50	1.74	1.98	2.20	2.41	
STSA-IM	1.10	1.14	1.20	1.32	1.50	1.71	1.95	2.23	2.57		1.10	1.13	1.18	1.29	1.47	1.69	1.92	2.20	2.54	
LSA-IM	1.09	1.11	1.15	1.25	1.43	1.70	2.02	2.36	2.70		1.09	1.11	1.15	1.25	1.42	1.69	2.01	2.34	2.67	
MSA-IM	1.11	1.14	1.21	1.34	1.56	1.83	2.10	2.36	2.63		1.11	1.14	1.20	1.32	1.53	1.80	2.08	2.34	2.61	
LMSA-IM	1.09	1.11	1.16	1.28	1.47	1.71	1.97	2.21	2.43		1.09	1.11	1.17	1.27	1.46	1.70	1.95	2.17	2.38	
PSSA-IM	1.13	1.17	1.25	1.39	1.60	1.83	2.05	2.27	2.51		1.13	1.16	1.23	1.36	1.56	1.80	2.02	2.23	2.47	
STSA-MA	1.09	1.12	1.17	1.27	1.47	1.76	2.11	2.47	2.82		1.10	1.12	1.17	1.26	1.45	1.75	2.10	2.45	2.80	
PSSA-MA	1.13	1.17	1.24	1.38	1.62	1.96	2.34	2.68	2.98		1.12	1.16	1.23	1.36	1.61	1.96	2.33	2.65	2.94	

ESTOI	Seen Speakers										Unseen Speakers									
SNR (dB)	-20	-15	-10	-5	0	5	10	15	20		-20	-15	-10	-5	0	5	10	15	20	
Unproc.	0.04	0.08	0.15	0.24	0.35	0.47	0.58	0.67	0.74		0.04	0.08	0.14	0.23	0.34	0.46	0.57	0.66	0.73	
STSA-DM	0.10	0.17	0.27	0.39	0.51	0.61	0.68	0.73	0.77		0.10	0.16	0.25	0.37	0.50	0.60	0.68	0.72	0.76	
LSA-DM	0.11	0.17	0.26	0.38	0.51	0.62	0.69	0.74	0.78		0.11	0.17	0.25	0.36	0.49	0.60	0.68	0.73	0.76	
MSA-DM	0.11	0.18	0.28	0.40	0.52	0.61	0.68	0.73	0.77		0.11	0.17	0.26	0.38	0.50	0.60	0.68	0.72	0.76	
LMSA-DM	0.11	0.17	0.26	0.38	0.51	0.61	0.68	0.73	0.76		0.11	0.16	0.25	0.36	0.49	0.60	0.67	0.72	0.74	
PSSA-DM	0.10	0.17	0.27	0.40	0.53	0.62	0.69	0.74	0.77		0.10	0.16	0.25	0.38	0.51	0.61	0.68	0.73	0.76	
STSA-IM	0.11	0.18	0.28	0.40	0.52	0.62	0.69	0.74	0.78		0.10	0.17	0.26	0.38	0.50	0.60	0.68	0.73	0.77	
LSA-IM	0.09	0.15	0.24	0.35	0.47	0.58	0.67	0.73	0.78		0.09	0.15	0.23	0.34	0.46	0.57	0.66	0.72	0.77	
MSA-IM	0.11	0.18	0.28	0.41	0.53	0.63	0.69	0.74	0.78		0.11	0.18	0.27	0.39	0.51	0.61	0.69	0.73	0.77	
LMSA-IM	0.10	0.16	0.25	0.37	0.49	0.60	0.68	0.73	0.77		0.10	0.16	0.24	0.35	0.47	0.58	0.67	0.72	0.76	
PSSA-IM	0.10	0.17	0.27	0.40	0.52	0.62	0.70	0.75	0.78		0.10	0.16	0.25	0.38	0.50	0.61	0.69	0.73	0.77	
STSA-MA	0.11	0.18	0.27	0.39	0.51	0.62	0.70	0.75	0.79		0.11	0.17	0.26	0.37	0.50	0.61	0.69	0.74	0.78	
PSSA-MA	0.11	0.18	0.27	0.40	0.53	0.64	0.71	0.76	0.80		0.10	0.17	0.26	0.38	0.51	0.63	0.70	0.75	0.79	