Model Architecture and Additional Evaluations for paper "Target Speaker Extraction through Comparing Noisy Positive and Negative Audio Enrollments"

April 1, 2025

Table 1: Comparison between our method and the baselines' SNR (↑), SNRi (↑), SI-SNR (↑), SI-SNRi (↑), PESQ (↑), STOI (↑), DNSMOS (↑), and WER (↓) on samples synthesized from the Librispeech dataset. For the binaural audio extraction task, since target is to predict the reverberant binaural audio of the target speaker, even a perfect prediction may score poorly due to the reverberation effect. We included the WER metric here for completeness, and omit the WER in the rest evaluations when the model is expected to predict reverberant audio.

Method	**Audio Type**	**Metric**	**2 Speakers**	**3 Speakers**	**4 Speakers**	**5 Speakers**	**6 Speakers**
Ours (Classical-monaural)	Mono	SNR	**7.40 ± 2.17**	**5.51 ± 2.23**	**3.68 ± 2.47**	**2.44 ± 2.43**	**1.12 ± 2.35**
		SNRi	**10.81 ± 2.43**	**10.96 ± 2.74**	**10.33 ± 2.75**	**10.24 ± 2.73**	**9.80 ± 2.74**
		SI-SNR	**6.48 ± 3.41**	**3.87 ± 4.01**	**0.71 ± 5.42**	**-1.98 ± 6.22**	**-5.45 ± 7.12**
		SI-SNRi	**9.89 ± 3.42**	**9.33 ± 3.99**	**7.37 ± 5.19**	**5.79 ± 5.88**	**3.21 ± 6.41**
		PESQ	**1.35 ± 0.16**	**1.20 ± 0.11**	**1.13 ± 0.07**	**1.09 ± 0.05**	**1.08 ± 0.06**
		STOI	**0.79 ± 0.10**	**0.70 ± 0.13**	**0.60 ± 0.15**	**0.50 ± 0.17**	**0.40 ± 0.18**
		DNSMOS	**2.98 ± 0.22**	**2.87 ± 0.20**	**2.81 ± 0.18**	**2.75 ± 0.16**	2.71 ± 0.16
		WER (%)	**34.3 ± 35.7**	**52.4 ± 29.7**	**69.5 ± 26.2**	**81.3 ± 32.4**	**87.2 ± 17.1**
Ours (Film Fusion)	Mono	SNR	6.59 ± 2.11	4.53 ± 2.38	3.27 ± 2.34	1.78 ± 2.39	0.82 ± 2.36
		SNRi	10.01 ± 2.55	9.98 ± 2.68	9.93 ± 2.69	9.55 ± 2.61	9.38 ± 2.50
		SI-SNR	5.46 ± 3.41	2.22 ± 4.71	0.03 ± 5.14	-3.05 ± 5.69	-5.51 ± 6.39
		SI-SNRi	8.88 ± 3.52	7.66 ± 4.50	6.67 ± 4.92	4.73 ± 4.94	3.03 ± 5.61
		PESQ	1.30 ± 0.14	1.17 ± 0.09	1.10 ± 0.06	1.08 ± 0.05	1.07 ± 0.04
		STOI	0.75 ± 0.11	0.65 ± 0.14	0.52 ± 0.16	0.43 ± 0.18	0.38 ± 0.17
		DNSMOS	2.93 ± 0.22	2.83 ± 0.20	2.76 ± 0.17	2.74 ± 0.15	**2.73 ± 0.15**
		WER (%)	37.8 ± 27.0	63.1 ± 38.7	73.9 ± 37.5	84.7 ± 18.8	89.7 ± 29.3
TCE	Mono	SNR	4.89 ± 2.44	2.68 ± 2.54	1.22 ± 2.36	-0.11 ± 2.30	-0.83 ± 2.08
		SNRi	8.46 ± 2.57	8.24 ± 2.58	8.19 ± 2.55	7.73 ± 2.10	7.75 ± 2.05
		SI-SNR	3.18 ± 3.82	-0.42 ± 4.57	-3.28 ± 5.10	-5.89 ± 4.84	-8.00 ± 4.88
		SI-SNRi	6.75 ± 3.66	5.16 ± 4.11	3.69 ± 4.39	1.94 ± 3.47	0.57 ± 3.80
		PESQ	1.21 ± 0.12	1.11 ± 0.08	1.08 ± 0.05	1.06 ± 0.05	1.05 ± 0.05
		STOI	0.68 ± 0.12	0.55 ± 0.15	0.45 ± 0.14	0.36 ± 0.15	0.31 ± 0.13
		DNSMOS	2.68 ± 0.18	2.65 ± 0.15	2.65 ± 0.14	2.65 ± 0.13	2.63 ± 0.14
		WER (%)	72.6 ± 24.2	87.6 ± 17.4	92.8 ± 12.2	95.7 ± 8.9	97.1 ± 12.7
SpeakerBeam	Mono	SNR	-3.72 ± 1.34	-3.77 ± 1.29	-3.74 ± 1.39	-3.76 ± 1.32	-3.68 ± 1.34
		SNRi	-0.42 ± 2.15	1.51 ± 2.41	3.10 ± 2.52	3.85 ± 2.65	4.91 ± 2.72
		SI-SNR	-6.32 ± 6.22	-8.18 ± 5.33	-9.64 ± 4.44	-10.37 ± 4.28	-11.65 ± 4.15
		SI-SNRi	-3.01 ± 5.98	-2.90 ± 5.08	-2.81 ± 3.96	-2.75 ± 3.68	-3.03 ± 3.38
		PESQ	1.15 ± 0.14	1.07 ± 0.11	1.05 ± 0.04	1.05 ± 0.03	1.04 ± 0.02
		STOI	0.47 ± 0.28	0.40 ± 0.19	0.32 ± 0.17	0.28 ± 0.15	0.25 ± 0.11
		DNSMOS	2.57 ± 0.18	2.50 ± 0.18	2.43 ± 0.14	2.47 ± 0.13	2.40 ± 0.12
		WER (%)	92.1 ± 12.8	95.4 ± 9.9	95.8 ± 11.2	98.9 ± 8.3	97.4 ± 8.1
Ours (Classical-binaural)	Bi	SNR	**5.30 ± 1.98**	**3.34 ± 1.87**	**2.14 ± 1.51**	**1.40 ± 1.28**	**0.65 ± 0.78**
		SNRi	**8.97 ± 2.39**	**8.83 ± 2.64**	**9.04 ± 2.53**	**9.28 ± 2.51**	**10.06 ± 2.67**
		SI-SNR	**3.52 ± 3.27**	**-0.09 ± 4.40**	**-2.73 ± 4.14**	**-5.22 ± 4.41**	**-7.97 ± 3.62**
		SI-SNRi	**7.18 ± 3.25**	**5.42 ± 4.26**	**4.16 ± 3.82**	**2.66 ± 3.85**	**1.46 ± 3.22**
		PESQ	**1.40 ± 0.20**	**1.20 ± 0.13**	1.12 ± 0.09	1.10 ± 0.11	1.09 ± 0.15
		STOI	**0.73 ± 0.09**	**0.56 ± 0.15**	**0.45 ± 0.15**	**0.36 ± 0.14**	0.29 ± 0.13
		DNSMOS	**2.78 ± 0.20**	**2.68 ± 0.18**	**2.61 ± 0.16**	**2.56 ± 0.15**	2.52 ± 0.14
		WER (%)	**69.6 ± 26.4**	88.2 ± 18.0	**93.7 ± 10.7**	97.9 ± 53.7	**96.3 ± 7.3**
LookOnceToHear	Bi	SNR	3.58 ± 2.33	2.30 ± 2.66	1.51 ± 2.05	0.91 ± 1.92	0.12 ± 1.73
		SNRi	7.10 ± 3.16	7.72 ± 3.16	8.34 ± 2.75	8.62 ± 2.72	9.28 ± 2.61
		SI-SNR	0.48 ± 5.60	-2.55 ± 6.60	-4.54 ± 6.14	-6.53 ± 6.09	-9.29 ± 5.38
		SI-SNRi	4.05 ± 5.80	2.93 ± 6.45	2.32 ± 6.02	1.34 ± 5.71	0.09 ± 5.04
		PESQ	1.30 ± 0.19	1.18 ± 0.12	**1.14 ± 0.14**	**1.11 ± 0.15**	**1.11 ± 0.12**
		STOI	0.65 ± 0.16	0.52 ± 0.19	0.42 ± 0.18	0.36 ± 0.16	**0.30 ± 0.15**
		DNSMOS	2.72 ± 0.23	2.66 ± 0.20	2.60 ± 0.18	2.56 ± 0.17	**2.54 ± 0.15**
		WER (%)	70.8 ± 28.7	**87.8 ± 17.8**	95.2 ± 45.7	**95.1 ± 10.7**	96.7 ± 9.3

Table 2: Comparison of model parameter size, inference time, and memory usage when extracting 1-second audio on an Intel Xeon Silver 4314 @ 2.40GHz CPU. We report the average inference time for extraction in 50 repeated experiments. Since our model replaced the Film fusion method in the LookOnceToHear baseline with a cross-attention based fusion method, our model has significantly smaller parameter size.

Model	Param Size	Inference Time	Inference Memory Usage
Ours	1.88 M	0.37s	2.37 GB
TCE	2.54 M	0.11s	1.94 GB
LookOnceToHear	4.41 M	0.35s	2.38 GB

Table 3: Affect of audio reverberation on model performance. We fine-tune our model on reverberant monaural audio and test its performance to extract reverberant monaural audio. In comparison to our model's performance on non-reverberant monaural audio (Shown in Table 1), the fine-tuned (Reverb-monaural) model shows lower performance, demonstrating the significant difficulty in extracting reverberant audio.

Method	**Audio Type**	**Metric**	**2 Speakers**	**3 Speakers**	**4 Speakers**	**5 Speakers**	**6 Speakers**
Ours (Reverb-monaural)	Mono	SNR	3.35 ± 1.97	1.56 ± 2.29	0.09 ± 2.28	-0.80 ± 2.43	-1.68 ± 2.31
		SNRi	7.25 ± 2.43	7.49 ± 2.33	7.39 ± 2.13	7.68 ± 1.93	7.69 ± 1.80
		SI-SNR	0.29 ± 4.00	-3.23 ± 5.06	-6.26 ± 5.16	-8.12 ± 5.26	-10.05 ± 5.10
		SI-SNRi	4.18 ± 3.64	2.69 ± 4.14	1.05 ± 4.06	0.34 ± 3.96	-0.69 ± 3.70
		PESQ	1.19 ± 0.12	1.12 ± 0.07	1.09 ± 0.05	1.08 ± 0.06	1.07 ± 0.05
		STOI	0.56 ± 0.11	0.44 ± 0.13	0.34 ± 0.12	0.29 ± 0.12	0.24 ± 0.11
		DNSMOS	2.65 ± 0.18	2.61 ± 0.17	2.60 ± 0.16	2.60 ± 0.16	2.56 ± 0.15

Table 4: Comparison between our monaural model and the baseline model TCE's SNR (↑), SNRi (↑), SI-SNR (↑), SI-SNRi (↑), PESQ (↑), STOI (↑), DNSMOS (↑), and WER (↓) performance on samples synthesized from the WSJ0 dataset.

Method	**Audio Type**	**Metric**	**2 Speakers**	**3 Speakers**	**4 Speakers**	**5 Speakers**	**6 Speakers**
Ours (Classical-monaural)	Mono	SNR	**5.59 ± 2.72**	**3.42 ± 2.63**	**1.77 ± 2.72**	**0.79 ± 2.37**	**0.10 ± 2.10**
		SNRi	**9.00 ± 2.99**	**8.47 ± 2.80**	**8.40 ± 2.54**	**8.34 ± 2.27**	**8.30 ± 1.97**
		SI-SNR	**3.23 ± 6.49**	**-0.78 ± 7.03**	**-4.20 ± 7.55**	**-6.54 ± 7.25**	**-7.81 ± 6.38**
		SI-SNRi	**6.63 ± 6.58**	**4.31 ± 6.93**	**2.42 ± 7.07**	**1.03 ± 6.83**	**0.40 ± 5.81**
		PESQ	**1.31 ± 0.14**	**1.20 ± 0.11**	**1.13 ± 0.08**	**1.10 ± 0.05**	**1.09 ± 0.06**
		STOI	**0.70 ± 0.16**	**0.57 ± 0.19**	**0.46 ± 0.18**	**0.37 ± 0.19**	**0.32 ± 0.16**
		DNSMOS	**2.75 ± 0.27**	**2.66 ± 0.22**	**2.59 ± 0.20**	**2.56 ± 0.18**	2.51 ± 0.16
TCE	Mono	SNR	3.45 ± 2.23	2.08 ± 2.02	0.75 ± 1.99	0.08 ± 1.89	-0.83 ± 1.77
		SNRi	6.65 ± 2.60	6.96 ± 2.10	7.28 ± 1.97	7.15 ± 1.49	7.39 ± 1.24
		SI-SNR	-0.19 ± 6.10	-3.12 ± 5.73	-6.45 ± 5.98	-8.01 ± 6.00	-9.94 ± 5.38
		SI-SNRi	3.02 ± 6.15	1.75 ± 5.31	0.10 ± 5.43	-0.90 ± 4.88	-1.70 ± 4.26
		PESQ	1.19 ± 0.11	1.13 ± 0.08	1.08 ± 0.06	1.08 ± 0.05	1.07 ± 0.08
		STOI	0.56 ± 0.17	0.46 ± 0.17	0.36 ± 0.16	0.31 ± 0.14	0.26 ± 0.12
		DNSMOS	2.59 ± 0.17	2.59 ± 0.15	2.58 ± 0.13	2.56 ± 0.14	**2.57 ± 0.13**

Table 5: Affect of enrollment audio quality on model performance. We compare ours and the baseline methods' performance when noisy or clean target speaker enrollment is used.

Method	**Audio Type**	**Metric**	**2 Speakers (Clean Enroll.)**	**2 Speakers (Noisy Enroll.)**	**3 Speakers (Clean Enroll.)**	**3 Speakers (Noisy Enroll.)**
Ours (Classical-monaural)	Mono	SNR	5.99 ± 2.90	**7.40 ± 2.17**	3.72 ± 2.81	**5.59 ± 2.23**
		SNRi	9.27 ± 2.87	**10.81 ± 2.43**	9.01 ± 2.65	**10.90 ± 2.68**
		SI-SNR	4.19 ± 4.89	**6.49 ± 3.41**	0.72 ± 5.38	**3.99 ± 3.95**
		SI-SNRi	7.47 ± 4.65	**9.90 ± 3.42**	6.01 ± 4.92	**9.31 ± 3.99**
		PESQ	1.26 ± 0.16	**1.35 ± 0.16**	1.14 ± 0.09	**1.20 ± 0.11**
		STOI	0.73 ± 0.14	**0.79 ± 0.10**	0.61 ± 0.16	**0.70 ± 0.13**
		DNSMOS	**2.94 ± 0.23**	**2.98 ± 0.22**	**2.85 ± 0.18**	**2.87 ± 0.20**
		WER (%)	60.7 ± 36.2	**34.3 ± 35.7**	72.9 ± 29.9	**52.4 ± 29.7**
SpeakerBeam	Mono	SNR	**9.29 ± 3.21**	-3.72 ± 1.34	**4.97 ± 3.85**	-3.77 ± 1.29
		SNRi	**12.87 ± 3.63**	-0.42 ± 2.15	**10.36 ± 3.02**	1.51 ± 2.41
		SI-SNR	**8.28 ± 6.20**	-6.32 ± 6.22	**2.97 ± 6.38**	-8.18 ± 5.33
		SI-SNRi	**11.85 ± 6.36**	-3.01 ± 5.98	**8.37 ± 5.27**	-2.90 ± 5.08
		PESQ	**1.52 ± 0.24**	1.15 ± 0.14	**1.23 ± 0.17**	1.07 ± 0.11
		STOI	**0.83 ± 0.14**	0.47 ± 0.28	**0.67 ± 0.17**	0.40 ± 0.19
		DNSMOS	2.64 ± 0.18	2.57 ± 0.18	2.57 ± 0.18	2.50 ± 0.18
		WER (%)	**30.5 ± 21.9**	101.4 ± 7.3	**39.8 ± 15.7**	101.6 ± 8.7
Ours (Classical-binaural)	Bi	SNR	**4.28 ± 2.75**	**5.30 ± 1.98**	2.09 ± 2.42	**3.34 ± 1.87**
		SNRi	**7.94 ± 2.91**	**8.97 ± 2.39**	7.62 ± 2.65	**8.83 ± 2.64**
		SI-SNR	**1.88 ± 4.73**	**3.52 ± 3.27**	-1.93 ± 5.02	**-0.09 ± 4.40**
		SI-SNRi	**5.55 ± 4.51**	**7.18 ± 3.25**	3.61 ± 4.70	**5.42 ± 4.26**
		PESQ	1.31 ± 0.22	**1.40 ± 0.20**	1.16 ± 0.13	**1.20 ± 0.13**
		STOI	0.66 ± 0.15	**0.73 ± 0.09**	0.51 ± 0.17	**0.56 ± 0.15**
		DNSMOS	**2.80 ± 0.18**	**2.78 ± 0.20**	**2.74 ± 0.16**	**2.68 ± 0.18**
LookOnceToHear	Bi	SNR	4.01 ± 2.02	3.58 ± 2.33	**3.00 ± 1.93**	2.30 ± 2.66
		SNRi	7.69 ± 3.01	7.10 ± 3.16	**8.38 ± 2.74**	7.72 ± 3.16
		SI-SNR	1.71 ± 4.00	0.48 ± 5.60	**-0.53 ± 4.53**	-2.55 ± 6.60
		SI-SNRi	5.43 ± 4.44	4.05 ± 5.80	**4.86 ± 4.60**	2.93 ± 6.45
		PESQ	**1.32 ± 0.19**	1.30 ± 0.19	**1.21 ± 0.14**	1.18 ± 0.12
		STOI	**0.68 ± 0.12**	0.65 ± 0.16	**0.57 ± 0.15**	0.52 ± 0.19
		DNSMOS	2.75 ± 0.22	2.72 ± 0.23	2.65 ± 0.19	2.66 ± 0.20

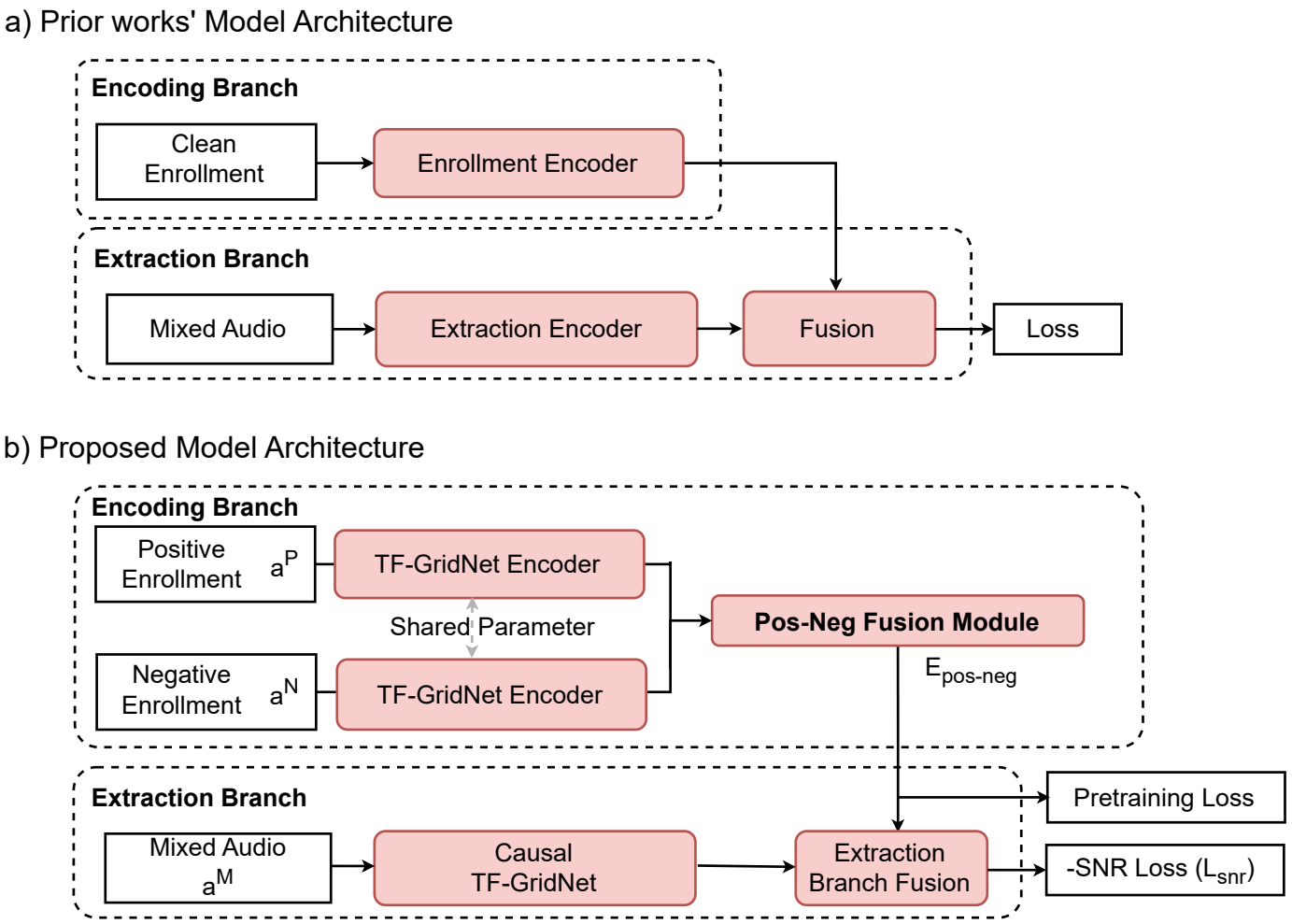

Figure 1: Comparison of the proposed model architecture with the commonly used architecture in prior works (e.g. LookOnceToHear, SpeakerBeam). We encode the Positive and the Negative Enrollments separately using two encoders with shared parameters, and fuse their embeddings using an attention based Pos-Neg Fusion Module.