VoxEffects: A Speech-Oriented Audio Effects Dataset and Benchmark

Zhe Zhang, **Yigitcan Özer**, **Junichi Yamagishi**
National Institute of Informatics, Tokyo, Japan *zhe@nii.ac.jp*

Paper

Data + Code

Interspeech 2026 · Sydney, Australia

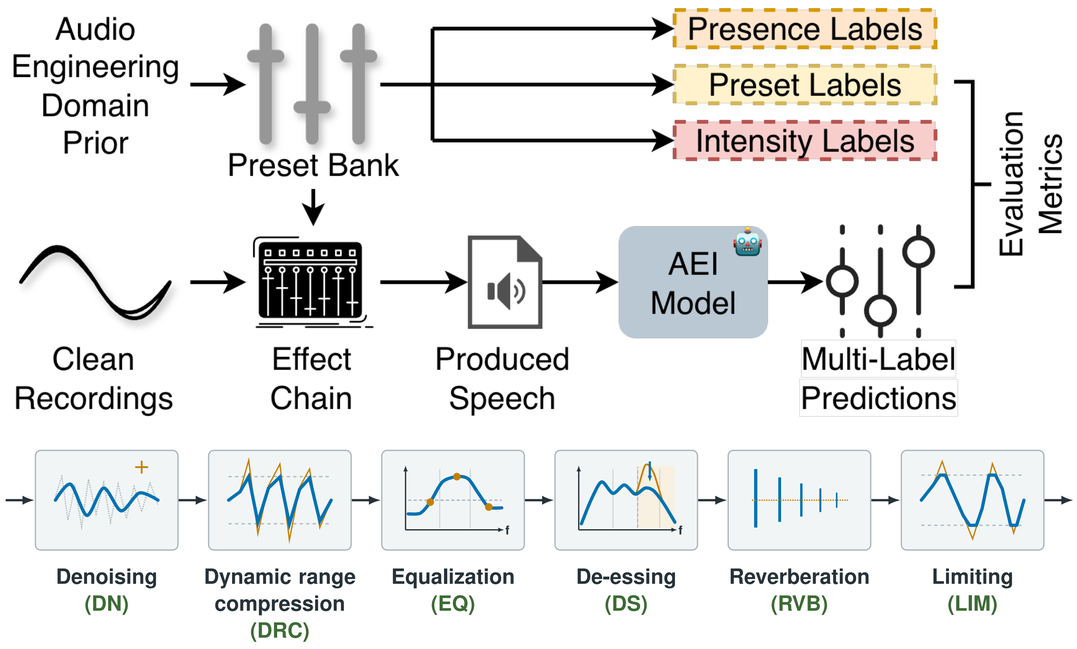
Why Speech Audio Effects

- Real-world speech is routinely shaped by audio effects.
- Few speech datasets expose exact effect and parameter labels.
- **Audio effect identification (AEI)** infers **which effects are present** and **how they are applied**.

What We Contribute

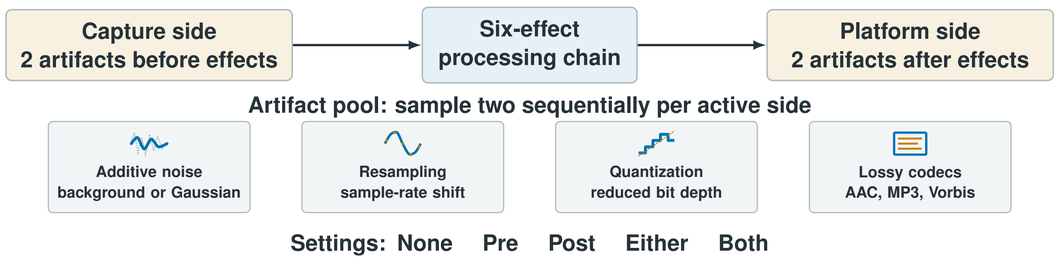
- **Open VoxEffects dataset**: processed speech with exact, multi-granularity effect labels, ready for research and education.
- **Reproducible renderer + benchmark**: four AEI tasks evaluated under domain shift.
- **AudioMAE-Fx**: a multi-task baseline with robustness, effect, and duration analyses.

VoxEffects at a Glance



- Reproducible rendering provides exact multi-granularity labels across **2,520 chains** from preset counts **3, 5, 7, 3, 4, 2**.
- **DAPS, EARS, TSP** supply 8:1:1 splits; **VCTK** is OOD test-only.
- Offline/on-the-fly rendering plus noise, resampling, quantization, and codec degradations at capture- and platform-side stages.

Robustness Augmentation



Tasks and AudioMAE-Fx Baseline

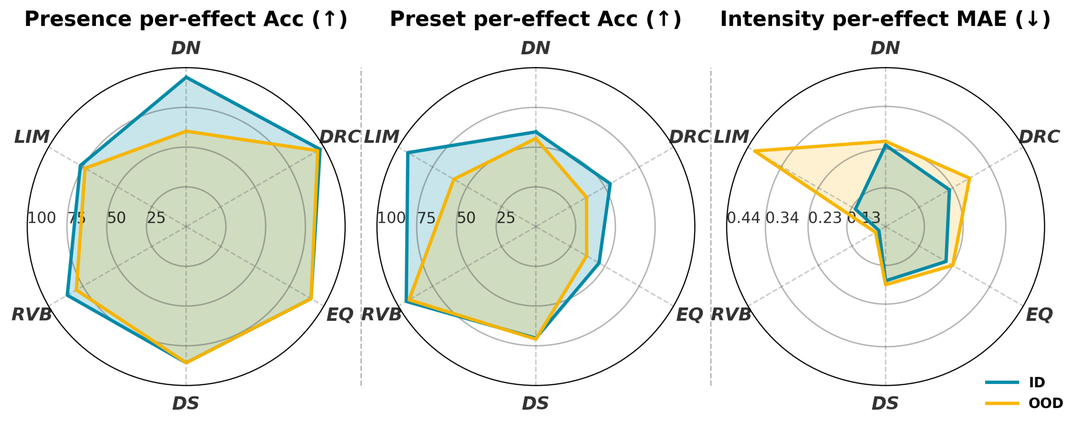
Task	Prediction	Metric
Presence (*Which*)	Six independent effect labels	$\text{Acc}_{\text{macro}}$, EMR
Preset (*How*)	One of 2,520 configurations	Top-1 / Top-5
# active (*How many*)	Number of active effects, 0–6	Accuracy
Intensity (*How much*)	Overall scalar + six-effect vector	MAE

- 16-kHz waveform → log-mel features → pretrained AudioMAE backbone → five lightweight prediction heads

Robustness Fine-Tuning Pays Off

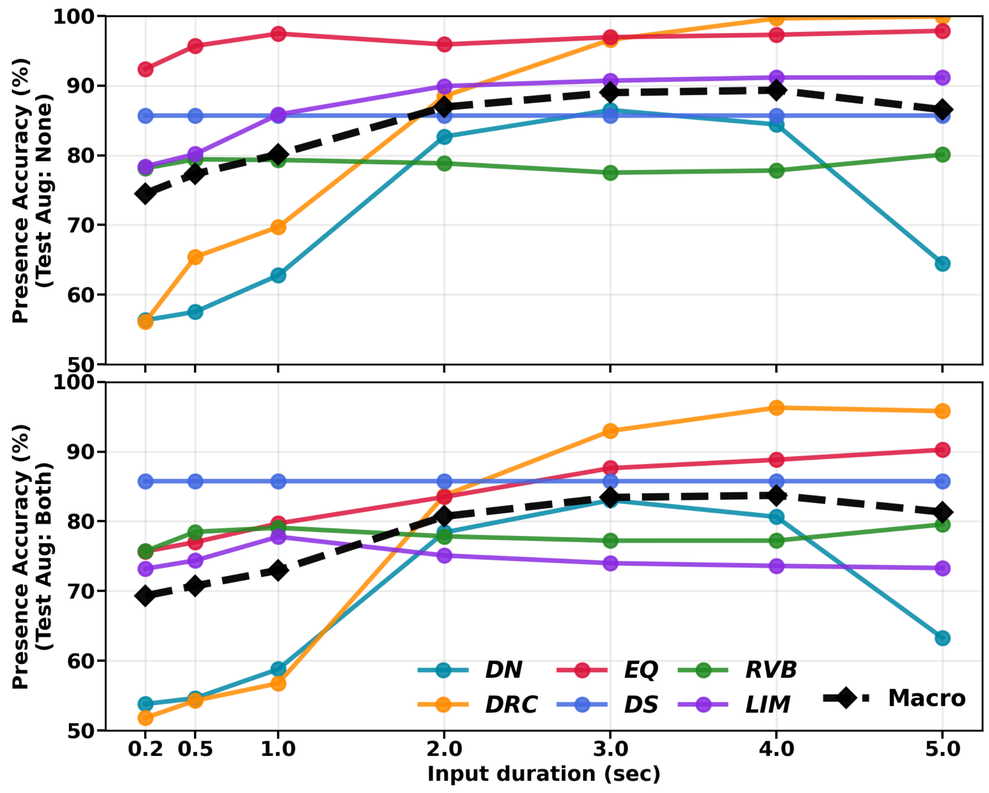
- **Stage 1**: clean fine-tuning; **Stage 2**: 50,000 robust steps. Evaluate 60 ID + 60 OOD utterances across all 2,520 tuples.
- Train Both improves every presence/preset condition; full ID/OOD results appear below.
- The largest gains appear under Post/Both shifts, while 2,520-way exact preset prediction remains hard.

Effect-wise ID/OOD Generalization



- **DN** has the largest presence gap from corpus-specific noise cues.
- **RVB** transfers well; **DRC**/**LIM** need effect-specific augmentation.

How Much Audio Is Enough



- Longer context lifts OOD macro accuracy and EMR.
- Degradations flatten gains; **DN** remains duration-sensitive.

Takeaways and Next Steps

Open dataset + Renderer + Presets
Scan the **Data + Code** QR to download and reproduce all **2,520 chains**.

- Robust training stabilizes AEI under domain shift.
- Next: broaden to more engines, chain orders, continuous controls, and real speech production.

VoxEffects Benchmark Results (ID/OOD)

Test Aug.	Train Aug.	Effect Presence $\text{Acc}_{\text{macro}}$ ↑	Effect Presence EMR ↑	Effect Preset Top-1 Acc. ↑	Effect Preset Top-5 Acc. ↑	#Active Acc_{num} ↑	Intensity Reg. MAE_{mean} ↓	Intensity Reg. $\text{MAE}_{\text{overall}}$ ↓
None	None	91.59 / 82.81	58.96 / 30.86	21.52 / 5.76	47.59 / 18.01	61.11 / 45.81	0.14 / 0.22	0.16 / 0.14
	Both	**95.58** / **86.15**	**76.48** / **39.22**	**36.78** / **12.19**	**75.98** / **32.97**	**77.24** / **47.36**	**0.10** / **0.19**	0.16 / 0.17
Pre	None	84.31 / 78.06	39.94 / 22.56	11.11 / 3.80	27.73 / 11.76	53.07 / 41.65	0.20 / 0.26	0.16 / 0.14
	Both	**91.58** / **83.11**	**60.71** / **32.38**	**23.19** / **8.81**	**55.11** / **25.50**	**65.53** / **42.79**	**0.13** / **0.21**	0.16 / 0.17
Post	None	78.29 / 72.64	28.45 / 16.64	8.01 / 2.56	19.95 / 8.07	41.12 / 41.18	0.24 / 0.30	0.18 / 0.15
	Both	**90.25** / **82.38**	**56.05** / **30.96**	**18.06** / **7.01**	**44.54** / **21.08**	**60.74** / **42.64**	**0.15** / **0.22**	**0.16** / 0.16
Either	None	81.31 / 75.43	34.37 / 19.84	9.49 / 3.11	23.78 / 9.97	47.23 / 41.57	0.22 / 0.28	0.17 / 0.15
	Both	**90.95** / **82.77**	**58.57** / **31.75**	**20.58** / **7.82**	**49.88** / **23.20**	**63.23** / **42.63**	**0.14** / **0.22**	**0.16** / 0.16
Both	None	75.42 / 71.13	21.68 / 13.85	4.54 / 1.76	12.84 / 5.83	40.72 / 39.85	0.27 / 0.31	0.17 / 0.15
	Both	**88.48** / **80.87**	**49.77** / **27.58**	**12.57** / **5.48**	**35.20** / **17.47**	**56.57** / **39.78**	**0.17** / **0.23**	**0.16** / 0.16