Efficient Streaming Audio-Visual Target Speaker Extraction for Real-World Acoustic Scenes

Wendi Sang[1,2,3],* Kai Li[2],* Yifan Li[1,3] Jianqiang Huang[1,3],† Xiaolin Hu[2,4],†

[1] School of Computer Technology and Application, Qinghai University, Xining, China
[2] Department of Computer Science and Technology, Institute for AI, BNRist, IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing, China
[3] Intelligent Computing and Application Laboratory of Qinghai Province, Qinghai University, Xining, China
[4] Chinese Institute for Brain Research (CIBR), Beijing, China

* Equal contribution. † Corresponding author.

NeurIPS 2026

Background

Lip movements help identify a target voice in crowded, noisy scenes. Streaming extraction must combine quality with low computation.

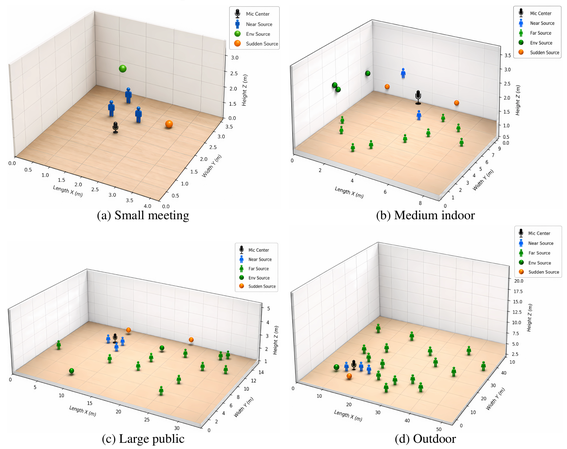
Contributions

1. RealSSA combines controllable acoustic scenes with held-out real recordings.
2. Falcon models four time-frequency scales in one encoder-decoder pass.
3. Speaker contrastive training reduces leakage with no added inference cost.

The RealSSA Dataset

RealSSA-Sim · approximately 153 hours

Four scene types; 1–5 near-field interferers;
10–15 far-field interferers where present.
Controlled reverberation, background and sudden noise.



(a) Small meeting (b) Medium indoor (c) Large public (d) Outdoor

RealSSA-Real · approximately 84 minutes

10 participants · 8 environments · held-out evaluation

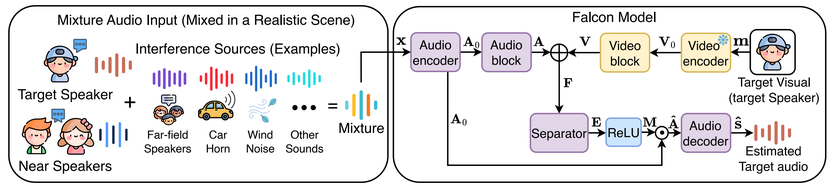
Benchmark coverage

Benchmark	NF speakers	FF speakers	Reverb. control	Env. noise	Sudden noise	Real eval.
NTCD-TIMIT+noise [2017]	2	✗	✗	✓	✗	✗
LRS2-2Mix [2018]	2	✗	✗	✗	✗	✗
LRS3-2Mix [2018]	2	✗	✗	✗	✗	✗
VoxCeleb2-2Mix [2018]	2	✗	✗	✗	✗	✗
LRS3+WHAM! [2023]	2	✗	✗	✓	✗	✗
RealSSA (ours)	**2–6**	✓	✓	✓	✓	✓

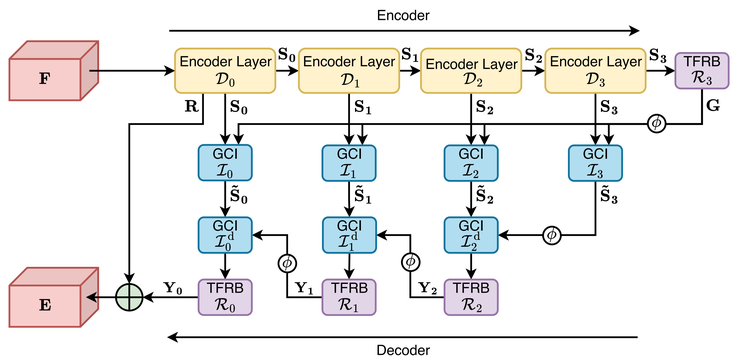
NF/FF: near-/far-field. NF counts include the target.
Env. noise: an external noise corpus is mixed in.
Real eval.: held-out real recordings.

Methods | Falcon

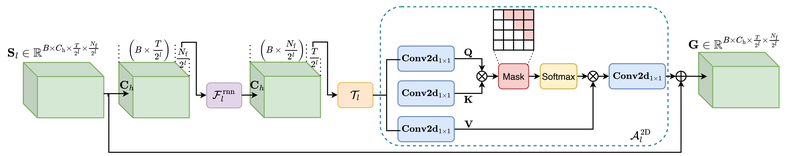
Audio-visual target speaker extraction



One encoder-decoder pass across four scales



Time-Frequency Refinement Block (TFRB)



Training-only speaker contrastive loss

Pull estimates toward the target and away from near-field interferers.

Reduces speaker leakage with no added inference cost.

Results

Compared with causal AV-TF-GridNet

0.5 M	10.6 G	51.3 ms
Parameters	**MACs**	**GPU time**
6.2 M → 0.5 M	213.8 G → 10.6 G	567.8 → 51.3 ms
91.9% fewer	**95.0% fewer**	**91.0% lower**

Separation quality · SI-SNRi (dB) ↑

Method	RealSSA-Sim	LRS2-2Mix	LRS3-2Mix	VoxCeleb2-2Mix
AV-TF-GridNet	7.6	12.7	14.3	5.8
AV-Mossformer2	6.0	13.2	16.2	7.7
Swift-Net-12	7.3	13.9	16.0	9.4
Dolphin	6.7	14.0	14.3	9.2
Falcon (Ours)	**7.7**	**14.1**	**16.4**	**10.5**

Selected causal systems. Falcon ranks first across all four simulated benchmarks in the full comparison.

LRS2 / LRS3 baseline results are from Swift-Net; the other baseline results use retrained models.

Speech content and target identity

	RealSSA-Sim		RealSSA-Real	
Method	**WER ↓**	**Similarity**	**WER ↓**	**Similarity**
AV-TF-GridNet	27.31	40.44	34.95	39.88
AV-Mossformer2	32.44	37.34	**32.00**	36.42
Swift-Net-12	28.59	41.79	33.72	40.04
Falcon (Ours)	**25.47**	**43.05**	**32.74**	**42.23**

WER (%) ↓; speaker similarity (cosine × 100) ↑.

Best speaker similarity on both sets; lowest simulated WER and second-lowest real WER.

Timing: 2 s audio + 50 video frames; batch size 1; RTX 3090.
GPU time measures the forward pass.

Contact: sangwendi@163.com
Falcon · Efficient streaming speech extraction

Code, data and demos
https://cslikai.cn/Falcon/

Code | Demo | Hugging Face Daily Papers