TWO-STAGE AUDIO-VISUAL TARGET SPEAKER EXTRACTION SYSTEM FOR REAL-TIME PROCESSING ON EDGE DEVICE

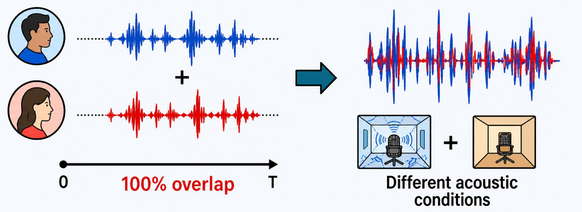
Zixuan Li[1], Xueliang Zhang[1*], Lei Miao[2], Zhipeng Yan[2], Ying Sun[2], Chong Zhu[2]
[1]College of Computer Science, Inner Mongolia University, China [2]Lenovo, China

1) Introduction / Motivation

Audio-Visual Target Speaker Extraction (AVTSE) aims to isolate a target speaker from multi-speaker mixtures by leveraging visual cues. Existing End-to-End methods

1 Unrealistic simulation in end-to-end AVTSE

- Popular datasets (LRS2-2mix, LRS3-2mix, VoxCelob2-2mix) are built from artificially mixed signals with 100% temporal overlap.
- Mixtures are often created by simply summing signals recorded under different acoustic conditions.
- Models may learn spurious acoustic shortcuts rather than robust target-speaker cues, causing poor real-world generalization



2 Poor generalization due to data scarcity

- End-to-end systems depend heavily on large-scale synchronized audio-visual corpora.
- Compared with audio-only or visual-only data, synchronized audio-visual data is much harder and more expensive to collect.
- Therefore, improving generalization merely by scaling synchronized corpora is impractical.

Audio-only	Visual-only	Audio-Visual (synchronised)
Abundant	Abundant	Scarce & Expensive

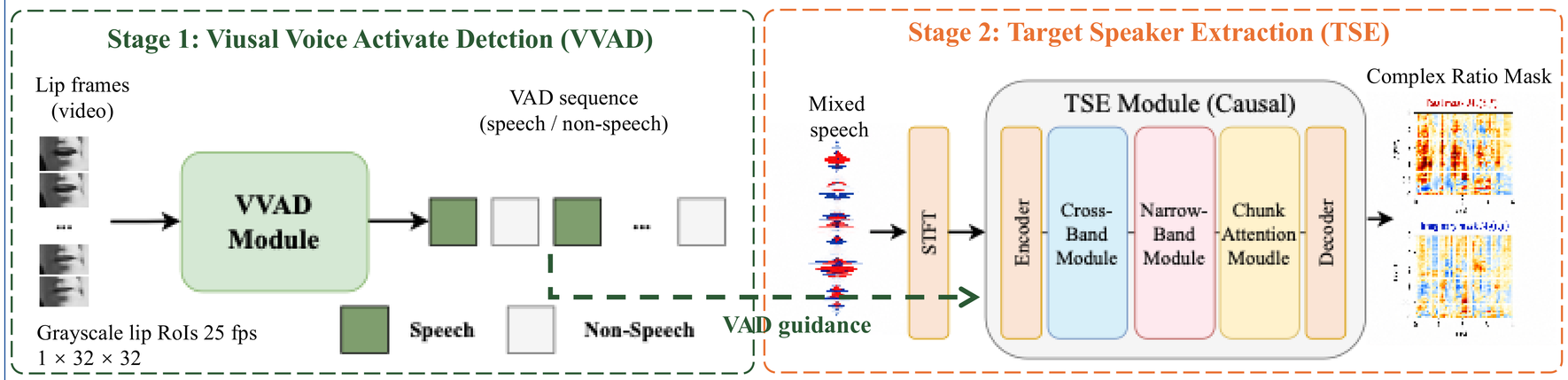
3 High complexity and non-causal design

- Many existing AVTSE systems are computationally heavy and non-causal.
- This limits real-time processing and deployment on resource-constrained edge devices.

Method	MACs	Params	Causality
AV-Speformer	105.23G	29.83M	✗
CTCNet	92.56G	18.34M	✗
AV-CrossNet	208.48G	22.32M	✗
2S-AVTSE	1.89G	1.36M	✓

Demos: http://www.cslzx.cn/2S-AVTSE/

2) Method Overview

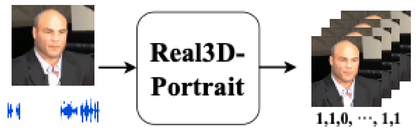


VVAD Training:

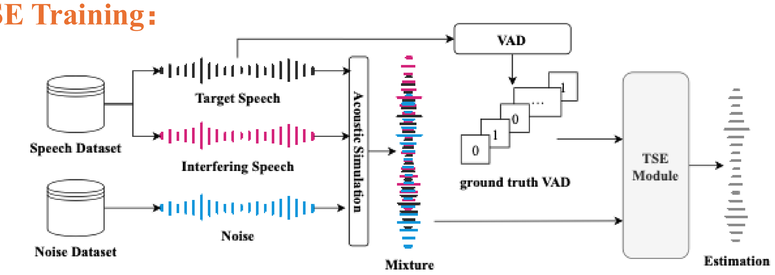
- Pre-train on **VAD-LRS3 dataset**
- Most frames in **VoxCeleb2** are speaking, esulting in an **imbalance** between positive and negative examples.
- Synthesizing **frame level training data** Using 3D Portrait Synthesis technology.

Segement level data

The entire video segment is labeled as 0/1

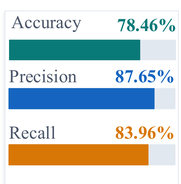


TSE Training:



3) Experimental Results

A. VVAD test



B. In-domain test on LRS2-2mix

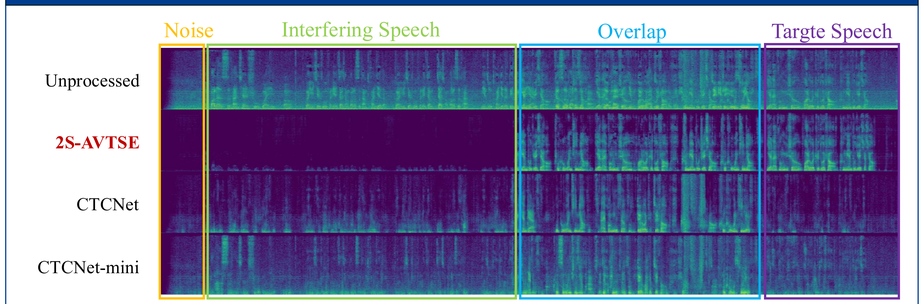
Method	Causal	MACs (G)	Params (M)	SI-SNR (dB)	STOI	PESQ
Unprocessed	-	-	-	0	0.66	1.53
CTCNet	✗	92.56	18.34	13.72	0.92	3.07
CTCNet-mini	✗	2.26	0.54	8.92	0.60	2.17
2S-AVTSE	✓	1.89	1.36	6.97	0.84	2.03

C. Cross-domain test on FaceStar-Mix

Method	SI-SNR	STOI	PESQ
Unprocessed	1.06	0.71	2.05
Audio Only(clean anchor)	5.48	0.75	2.28
Audio Only(noisy anchor)	4.51	0.73	2.17
CTCNet	5.90	0.77	2.20
CTCNet-mini	-0.59	0.60	1.54
Upper Bound*	8.69	0.81	2.54
2S-AVTSE	7.09	0.78	2.41

* Upper Bound used oracle VAD as input during test

4) Real-World Example



On a real laptop-recorded sample, 2S-AVTSE suppresses both background noise and interfering speech while preserving the target speech structure.

5) Conclusion

- 2S-AVTSE is a lightweight and causal two-stage AVTSE system.
- It generalizes better to realistic and cross-domain scenarios than conventional end-to-end baselines.
- The system is suitable for real-time processing on edge devices.