Online Audio-Visual Target Speaker Extraction with Viseme-Guided Lightweight Visual Pretraining

Zixuan Li[1], Xueliang Zhang[1], Lei Miao[2], Zhipeng Yan[2], Sun Ying[2], Zhu Chong[2]

[1]**College of Computer Science, Inner Mongolia University, China**
[2]Lenovo, China



Motivation: visual cues help, deployment still hurts

1. Why audio-visual TSE?

- Lip motion adds articulatory evidence when speakers overlap.
- Visual observations remain informative at low acoustic SNR.
- Cameras are already available on phones, PCs, and meeting devices.

2. What blocks deployment?

- Many systems use non-causal designs.
- High computational cost in both audio part and visual part.
  - Audio branch: speaker separation network.
  - Visual branch: VSR-pretrained backbones.

Our Tagert: Online inference on an edge device with low latency and low computational cost.

Motivation: visual cues help, deployment still hurts

2. What blocks deployment?

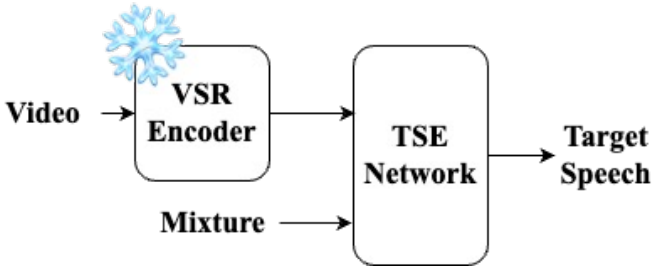
- Many systems use non-causal designs. → Online-AV-Crossnet explored fully causal design, but they still have unacceptable MACs for on-devices processing.
- High computational cost in both audio part and visual part.
  - Audio branch: speaker separation network. → Can be easily mitigate by using some efficient speaker separation networks, such as Voicefilter-lite, SKiM.
  - Visual branch: VSR-pretrained backbones. → **Less attention has been paid on.**

Motivated by these observations, we rethink the design of visual part in AV-TSE system.

How AV-TSE systems use visual information?

1. VSR/Lip-Reading Cue

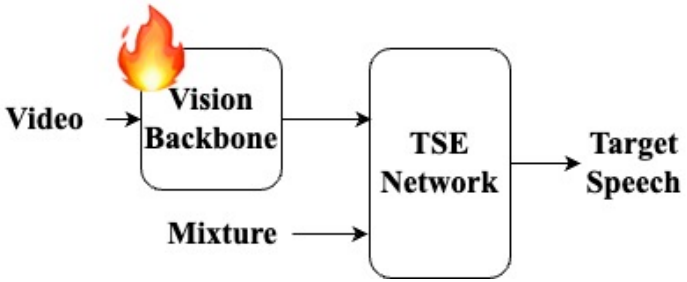
- The most common way
- Adopting a visual encoder pretrained for lip-reading



☹️ Because VSR is challenging, such encoders often have many parameters and high computational cost.

2. Mouth Cue

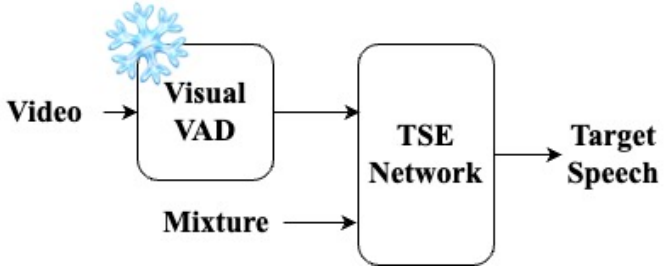
- A vision backbone such as Resnet is trained jointly with the separator.



😑 Mainly focus on network architecture and doesn't re-examine the design of visual cues

3. Visual VAD Cue

- Adopting visual information to obtain VAD cues for target speaker identification.

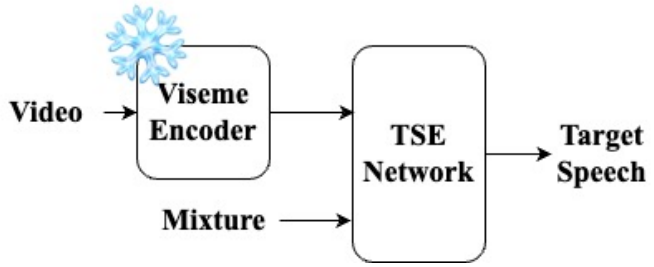


😨 Primarily designed for sparse speaker overlap; does not work in fully overlapping cases.

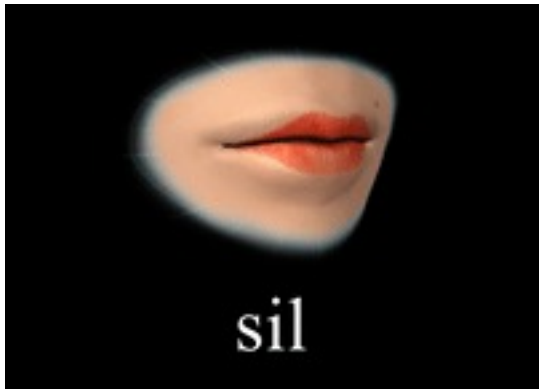
How AV-TSE systems use visual information?

4. Viseme Cue (Proposed)

- Employing viseme representation as the visual guidance signal in AV-TSE system.



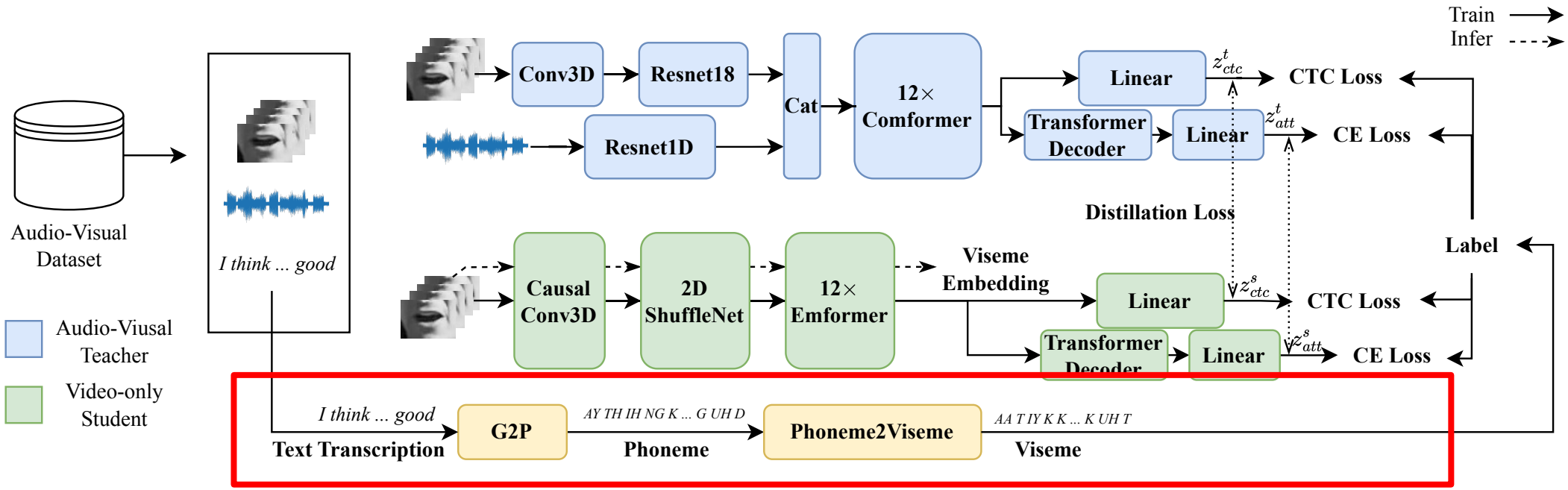
A viseme is a group of phonemes that appear visually similar. Such as /p/, /b/, and /m/.



Viseme Class	Phoneme Set
P	B, P, M
T	D, T, S, Z, TH, DH
K	G, K, N, NG, L, Y, HH
CH	JH, CH, SH, ZH
F	F, V
W	R, W
IY	IY, IH
EY	EH, EY, AE
AA	AA, AW, AY
AH	AH
AO	AO, OY, OW
UH	UH, UW
ER	ER

- ✓ **VS VSR/Lip-Reading Cue:** The label space is reduced from thousands of words to just 13 viseme classes, simplifying the task and enabling a lighter visual neural network.
- ✓ **VS Mouth Cue:** Provides task-aligned supervision and a more interpretable representation than jointly learned mouth cues.
- ✓ **VS Visual VAD Cue:** Viseme-level supervision keeps mouth movement details beyond speech activity, giving finer guidance when speakers fully overlap.

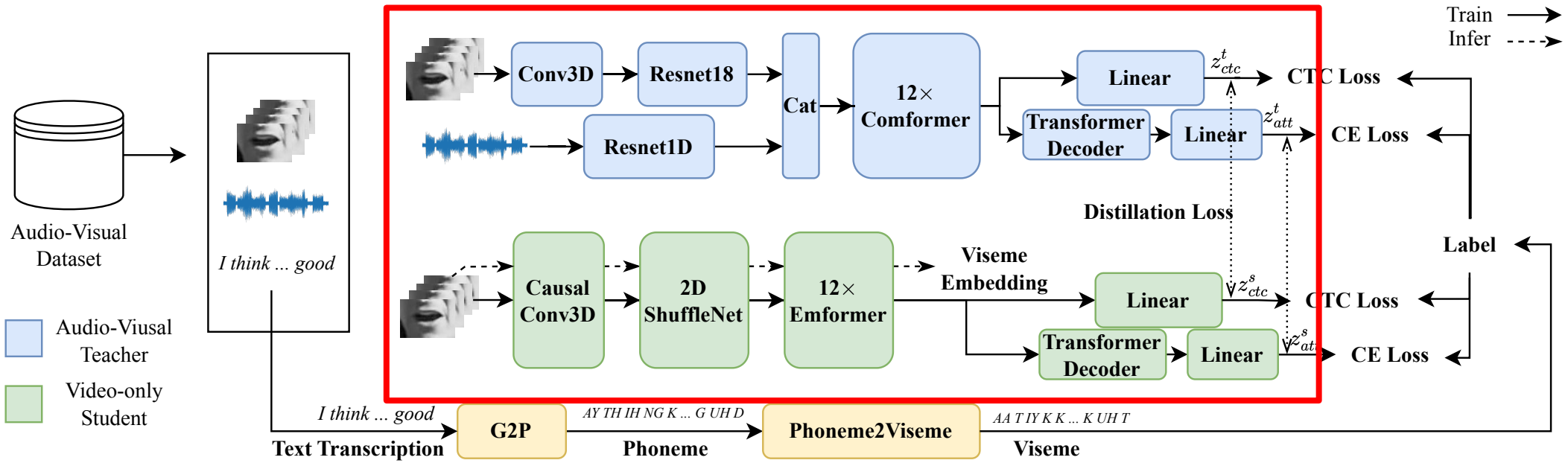
Online viseme recognition via cross-modal distillation



How to get the Viseme label:

- The text transcription in audio-visual dataset is transformed to phoneme sequence by G2P toolkit
- The phoneme sequence is mapped to viseme sequence as the label

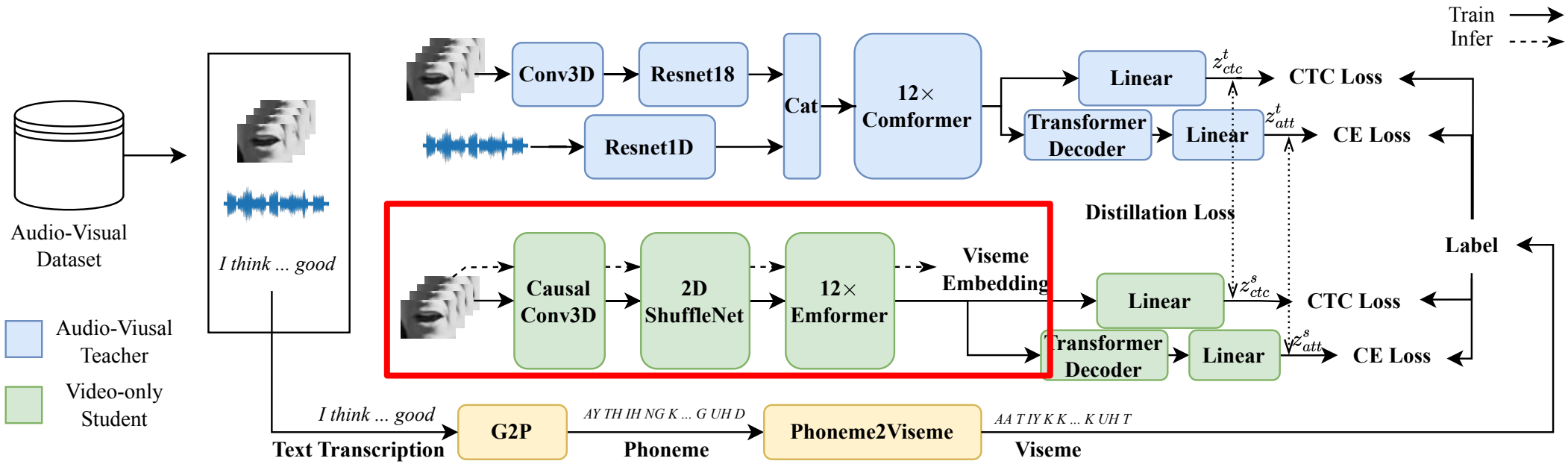
Online viseme recognition via cross-modal distillation



We use a **non-causal audio-visual teacher** to distill a **causal visual-only student**:

- The teacher model follows the AVSR framework.
- In the student model, ResNet is replaced with ShuffleNet to reduce computation.
- In the student model, Conformer is replaced with Emformer to ensure causality.

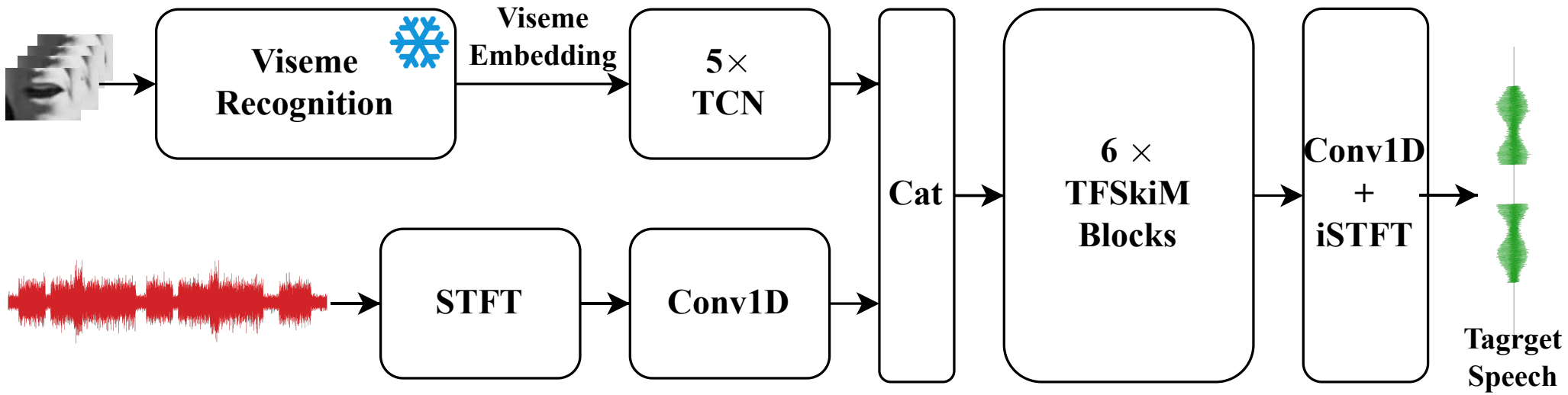
Online viseme recognition via cross-modal distillation



Inference:

- During inference, we use only the visual-only student.
- The output of the final Emformer layer is used as the viseme embedding for downstream TSE.

Viseme-guided target speaker extraction



We use an adaption version of TFSkiM-Net as TSE backbone:

- **TF-SKiMNet** is a **SOTA speech enhancement network** with about **2G MACs**.
- We replace the **Local-T Module** in TFSKiM-Net with a **narrow-band LSTM** to better fit the TSE task.
- $\mathcal{L}_{\text{sep}} = \left\| |\hat{S}| - |S| \right\|_1 - \text{SI-SNR}(\hat{s}, s).$

Experimental setup and Baselines

Viseme Recognition：

- LRS2-pretrain, LRS2-train for train
- LRS2-test for evaluation

Exp.	Visual Cue	Network	Visual Frontend	Separator	Total
Baseline-1	Mouth	SkiM	2.1 G	5.8 G	7.9 G
Baseline-2	Viseme	SkiM	3.3 G	5.8 G	9.1 G
Baseline-3	VSR	TF-SkiMNet	7.9 G	3.4 G	11.3 G
proposed	Viseme	TF-SkiMNet	3.3 G	3.4 G	7.7 G

Table 1:Computational cost (MACs) of different components. Total equals Visual Frontend + Separator.

Target Speaker Extraction：

- LRS2-Mix (11-hour training data)
- LRS3-Mix (28-hour training data)
- Vox2-Mix (56-hour training data)
- SIR: [-5, 5]dB

- **Baseline-1:** AV-SkiM, SOTA online AV-TSE system
- **Baseline-2:** Replace mouth cue in Baseline-1 with viseme cue
- **Baseline-3:** VSR cue with our modifiedTF-SkiMNet TSE Module
- **Proposed:** Viseme cue with our modifiedTF-SkiMNet TSE Module, **achieving the lowest total MACs in the paper.**

Results and Comparison

Viseme Recognition:

Model	Input	Causal	VER (%) ↓	MACs
Teacher	Audio + Video	×	9.60	19.82 G
Student (distilled)	Video	✓	28.90	3.33 G
Student w/o distillation	Video	✓	33.93	3.33 G

Table 2: Viseme recognition error-rate results on the LRS2 test set. Lower VER is better.

- VER means Viseme Error Rate:

$$\mathrm{VER}(\%) = 100 \times \frac{S + D + I}{N},$$

It compute in the same way as word error rate (WER):

- S: Substitutions
- D: Deletions
- I: Insertions

Distilled causal student achieves 28.9% VPR:

- It requires **less computation** than the non-causal audio-visual teacher.
- It outperforms the version without distillation.

Results and Comparison

Viseme Recognition:

Model	Input	Causal	VER (%) ↓	MACs
Teacher	Audio + Video	×	9.60	19.82 G
Student (distilled)	Video	✓	28.90	3.33 G
Student w/o distillation	Video	✓	33.93	3.33 G

Table 2: Viseme recognition error-rate results on the LRS2 test set. Lower VER is better.

- VER means Viseme Error Rate:

$$\mathrm{VER}(\%) = 100 \times \frac{S + D + I}{N},$$

It compute in the same way as word error rate (WER):

- S: Substitutions
- D: Deletions
- I: Insertions

Distilled causal student achieves 28.9% VPR:

- It requires **less computation** than the non-causal audio-visual teacher.
- It outperforms the version without distillation.

Results and Comparison

Viseme Guided TSE: In-Domain Test

Exp.	Cue	Network	LRS2-Mix			LRS3-Mix			Vox2-Mix		
			SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI
Mixture	-	-	0.00	1.18	0.52	0.00	1.21	0.52	0.00	1.25	0.50
Baseline-1	Mouth	SkiM	7.17	1.61	0.71	9.88	1.89	0.78	**5.45**	1.61	0.62
Baseline-2	Viseme	SkiM	7.84	1.67	0.74	9.65	1.86	0.77	5.38	1.62	0.62
Baseline-3	VSR	TF-SkiMNet	8.88	1.91	0.77	10.03	2.08	0.80	5.14	1.70	0.63
Ablation	Viseme	TF-SkiMNet*	9.39	1.96	0.79	-	-	-	-	-	-
Proposed	Viseme	TF-SkiMNet	**10.07**	**2.07**	**0.81**	**10.78**	**2.18**	**0.82**	5.32	**1.70**	**0.64**

Table 3: In-domain results on LRS2-Mix, LRS3-Mix, and Vox2-Mix. Best results in each dataset/metric are in bold.

Viseme Cue VS Mouth Cue:

- Viseme cues improve performance on LRS2.
- On LRS3 and VoxCeleb2, viseme cues achieve comparable performance to mouth cues, with only a slight degradation.

  This gap is likely due to **domain mismatch**: the viseme encoder is trained only on LRS2, making LRS3 and VoxCeleb2 cross-domain test sets for viseme cues, whereas they remain in-domain for mouth cues.

Results and Comparison

Viseme Guided TSE: In-Domain Test

Exp.	Cue	Network	LRS2-Mix			LRS3-Mix			Vox2-Mix		
			SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI
Mixture	-	-	0.00	1.18	0.52	0.00	1.21	0.52	0.00	1.25	0.50
Baseline-1	Mouth	SkiM	7.17	1.61	0.71	9.88	1.89	0.78	**5.45**	1.61	0.62
Baseline-2	Viseme	SkiM	7.84	1.67	0.74	9.65	1.86	0.77	5.38	1.62	0.62
Baseline-3	VSR	TF-SkiMNet	8.88	1.91	0.77	10.03	2.08	0.80	5.14	1.70	0.63
Ablation	Viseme	TF-SkiMNet*	9.39	1.96	0.79	-	-	-	-	-	-
Proposed	Viseme	TF-SkiMNet	**10.07**	**2.07**	**0.81**	**10.78**	**2.18**	**0.82**	5.32	**1.70**	**0.64**

Table 3: In-domain results on LRS2-Mix, LRS3-Mix, and Vox2-Mix. Best results in each dataset/metric are in bold.

Viseme Cue VS VSR Cue:

- Viseme cue outperform VSR cue across all three datasets.
- Viseme cue require only about half the computational cost of VSR cues.

Exp.	Visual Cue	Network	Visual Frontend	Separator	Total
Baseline-1	Mouth	SkiM	2.1 G	5.8 G	7.9 G
Baseline-2	Viseme	SkiM	3.3 G	5.8 G	9.1 G
Baseline-3	VSR	TF-SkiMNet	7.9 G	3.4 G	11.3 G
proposed	Viseme	TF-SkiMNet	3.3 G	3.4 G	7.7 G

Viseme pretraining provides a more task-aligned cue for AV-TSE.

Results and Comparison

Viseme Guided TSE: Cross-Domain Test

Exp.	LRS2-Mix			LRS3-Mix		
	SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI
Mixture	0.00	1.18	0.52	-0.00	1.21	0.52
Baseline-1	6.44	1.57	0.69	7.90	1.70	0.72
Baseline-2	7.04	1.61	0.70	**8.06**	1.72	0.73
Baseline-3	6.35	1.69	0.69	7.08	1.78	0.71
Proposed	**7.41**	**1.76**	**0.72**	7.88	**1.84**	**0.74**

Table 4: Cross-domain results on LRS2-Mix and LRS3-Mix.

- All models are trained on Vox2-Mix and evaluated on LRS2-Mix and LRS3-Mix.
- On LRS2-Mix, the proposed system achieves the best performance across all three metrics.
- On LRS3-Mix, it achieves the highest PESQ and ESTOI scores.
- These results demonstrate competitive performance under cross-domain conditions.

Results and Comparison

Ablation on TSE Module Modification

Exp.	Cue	Network	LRS2-Mix			LRS3-Mix			Vox2-Mix		
			SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI
Mixture	-	-	0.00	1.18	0.52	0.00	1.21	0.52	0.00	1.25	0.50
Baseline-1	Mouth	SkiM	7.17	1.61	0.71	9.88	1.89	0.78	**5.45**	1.61	0.62
Baseline-2	Viseme	SkiM	7.84	1.67	0.74	9.65	1.86	0.77	5.38	1.62	0.62
Baseline-3	VSR	TF-SkiMNet	8.88	1.91	0.77	10.03	2.08	0.80	5.14	1.70	0.63
Ablation	Viseme	TF-SkiMNet*	9.39	1.96	0.79	-	-	-	-	-	-
Proposed	Viseme	TF-SkiMNet	**10.07**	**2.07**	**0.81**	**10.78**	**2.18**	**0.82**	5.32	**1.70**	**0.64**

Table 3: In-domain results on LRS2-Mix, LRS3-Mix, and Vox2-Mix. Best results in each dataset/metric are in bold.

- TF-SkiMNet∗ denotes the original TF-SkiMNet configuration.
- Our modified version achieve better performance.

Conclusion

- Viseme pretraining provides a more interpretable and task-aligned cue for AV-TSE.
- We modify TF-SKiMNet, originally developed for speech enhancement, to make it more suitable for TSE.
- The proposed online system reaches leading in-domain results at the lowest reported total MACs.
- Cross-domain results show strong robustness, especially on LRS2-Mix and perceptual metrics on LRS3-Mix.

Visemes make the visual cue smaller, clearer, and more useful for online AV-TSE.

Thank you

Questions?

Contact me at: cslzx@mail.imu.edu.cn