Poster No. 7978

TIME-DOMAIN AUDIO-VISUAL SPEECH SEPARATION ON LOW QUALITY VIDEOS

Yifei Wu[1], Chenda Li[1], Jinfeng Bai[2], Zhongqin Wu[2] , Yanmin Qian[1]

[1] MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

[2] TAL Education Group, China

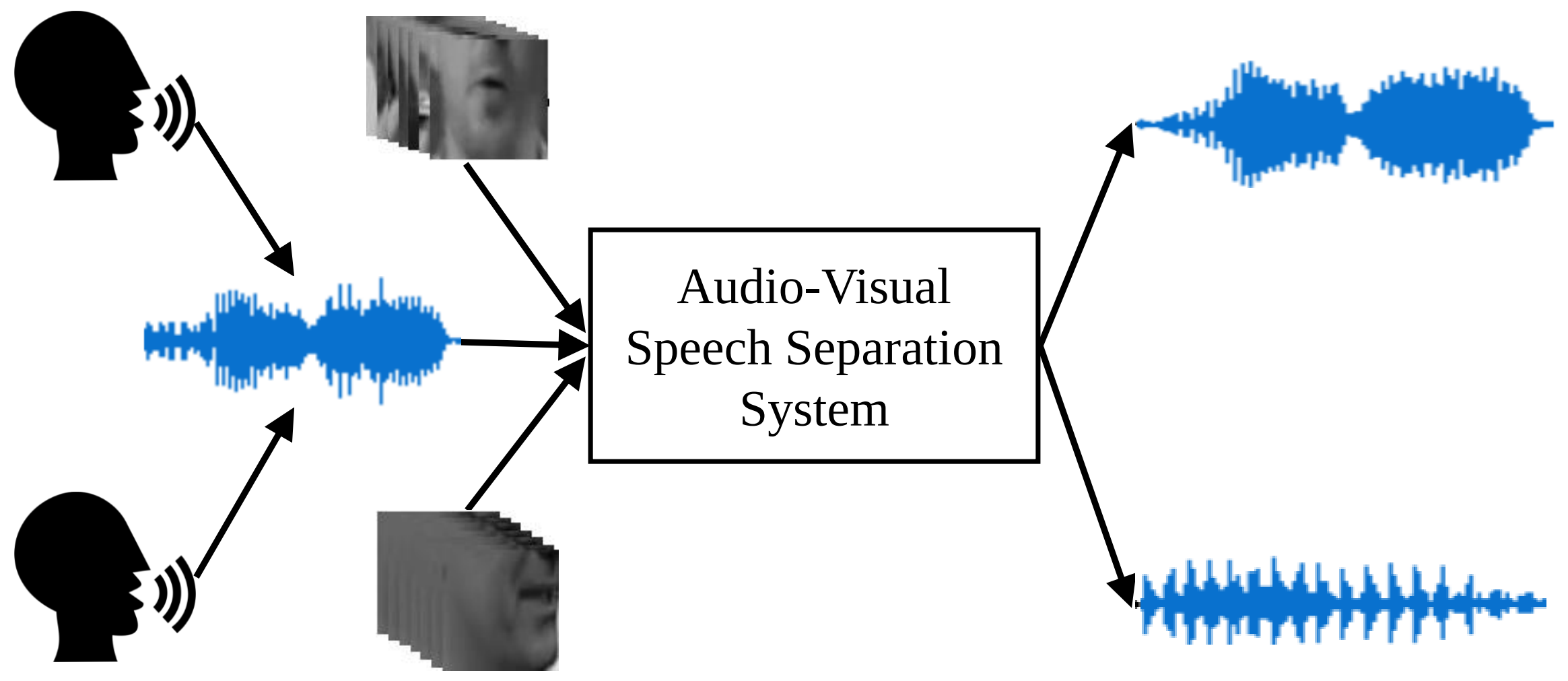
SJTU Cross Media
Language Intelligence Lab
上海交通大学跨媒体语言智能实验室

HIGHLIGHTS

- Time-domain audio-visual speech separation
- Attention-based feature fusion
- Robust to low-quality video inputs, including:
  - Low resolution
  - Lip concealment
  - Out of synchronization

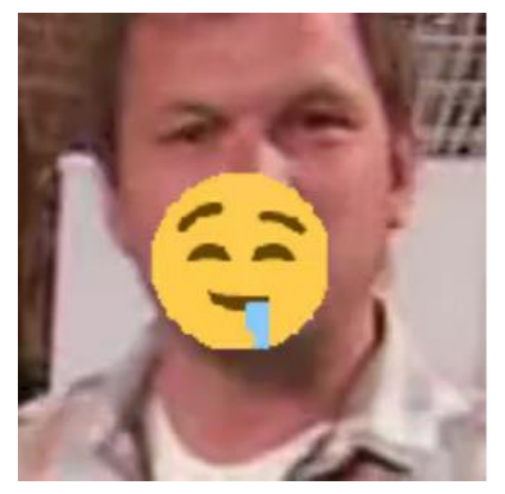
I. TASK DEFINITION

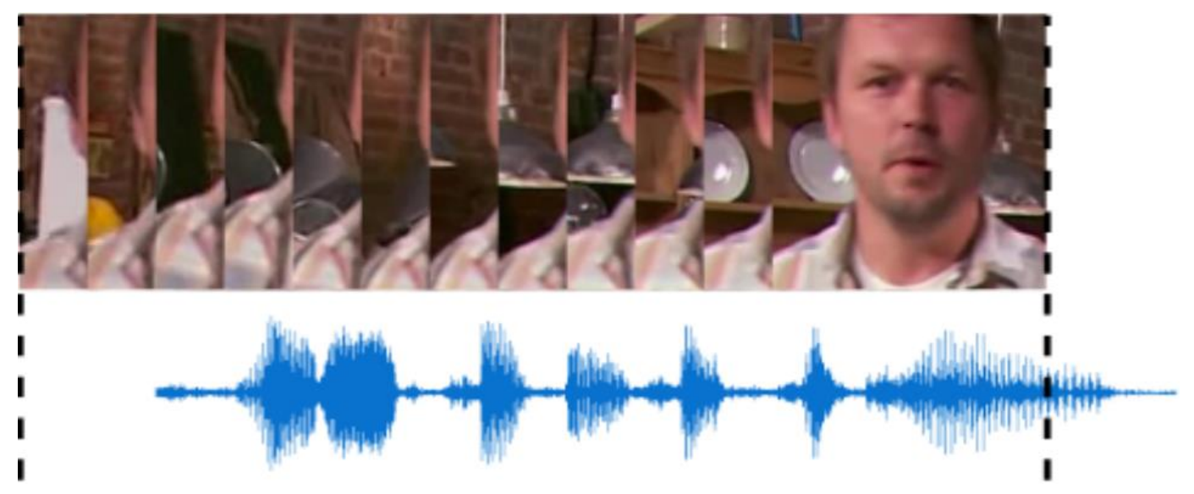
Audio-visual speech separation task:



Categories of low-quality video to be addressed:

Low Resolution

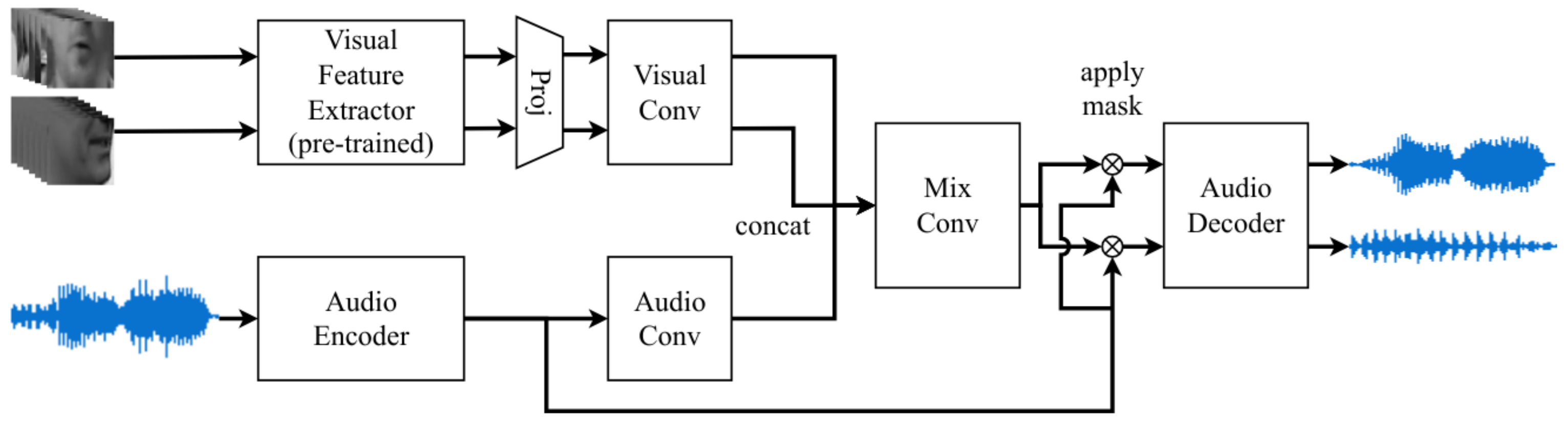

Lip Concealment


Out of synchronization

II. BASELINE MODEL

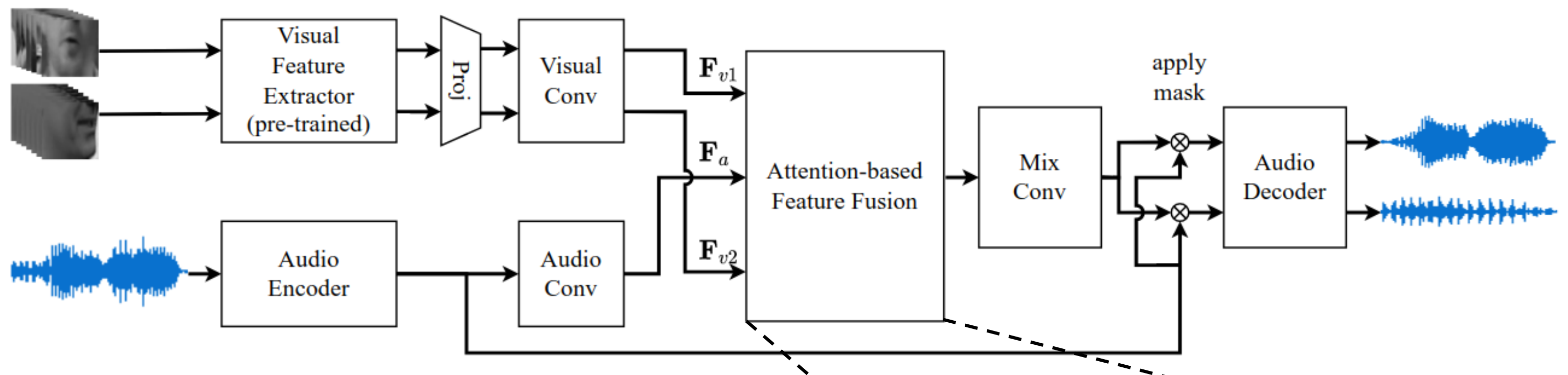
Backbone: Conv-TasNet

Visual feature extractor: pre-trained TCN [1]

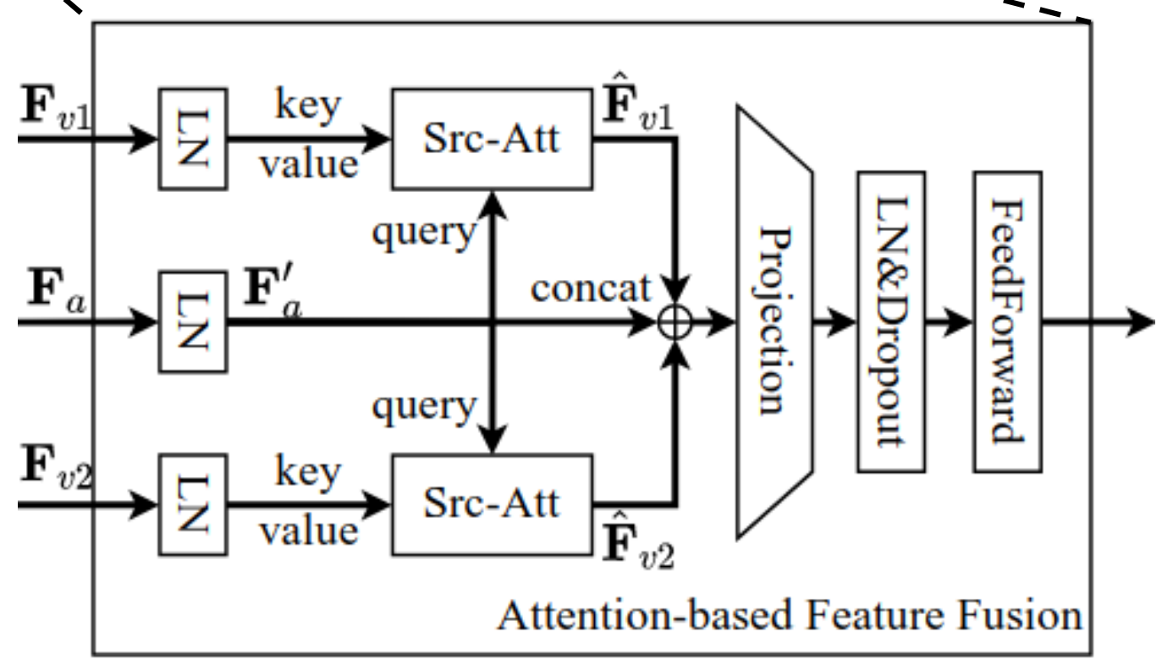


[1] Petridis, Stavros, et al. "End-to-end audiovisual speech recognition." *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.

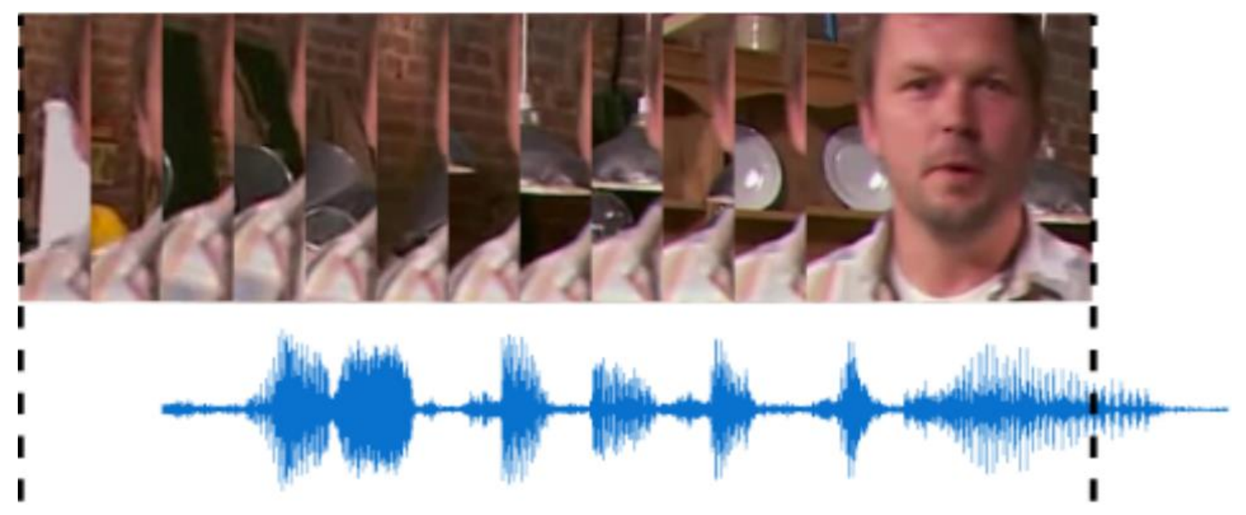
III. METHODS



- Audio-visual feature fusion module: Query Vision
- For adapting to long input sequences in time-domain model,
  - Self-attention is removed
  - Local attention is introduced



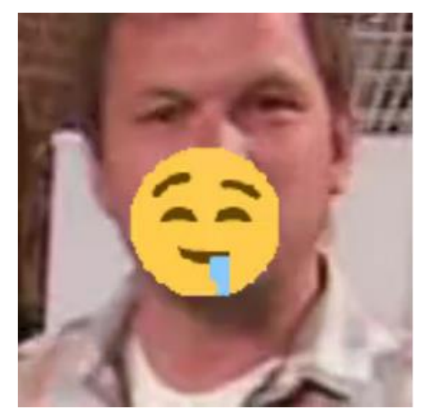
- Data augmentation methods:
  - Low resolution: Down-sample and up-sample each frame
  - Lip concealment: Conceal the lip region in some consecutive frames with a noise square
  - Random audio-video offset



Loss function: Scale-invariant SNR (no permutation-invariant training)

IV. DATASET

- Dataset: LRS2
  - Mixtures are generated with SNR in [-10, 10] dB
- Data augmentation
  - Low resolution: 80x80, 40x40, 20x20
  - Lip concealment: 25%, 50% or 75% of duration
  - Random offset: maximum 5 frames
- Low-quality test sets
  - LR10: 10x10 low resolution
  - LE75: 75% of duration is patched with an emoji
  - RO10: maximum 10 frames random offset



V. RESULTS

Q	Data Augmentation	Model	SDR(dB)			
			Normal	LR10	LE75	RO10
0	None	Audio-only	12.47	-	-	-
		Baseline	13.45	12.54/9.67	12.89/11.59	10.54/6.10
		Proposed	**14.66**	**14.09/11.55**	**14.06/12.57**	**11.89/6.94**
1	Low Resolution	Baseline	13.64	12.97/11.02	13.28/12.50	11.28/7.27
		Proposed	**14.86**	**14.53/13.29**	**14.47/13.65**	**12.69/8.23**
	Lip Concealment	Baseline	13.75	13.53/12.35	13.56/12.95	11.40/7.52
		Proposed	**14.77**	**14.48/13.30**	**14.48/13.92**	**12.63/8.36**
	Max. 5 Frames Async.	Baseline	13.08	12.85/11.80	12.77/12.04	12.76/**10.26**
		Proposed	**14.17**	**13.91/13.11**	**13.87/13.35**	**12.89**/10.10
	All	Baseline	12.87	12.74/12.28	12.73/12.36	12.14/9.92
		Proposed	**14.34**	**14.16/13.64**	**14.20/13.90**	**13.03/10.53**
2	Low Resolution	Baseline	13.27	13.24/12.73	13.12/12.72	10.88/7.40
		Proposed	**14.81**	**14.56/13.72**	**14.48/13.93**	**12.74/8.87**
	Lip Concealment	Baseline	13.59	13.44/12.85	13.42/13.07	11.57/8.71
		Proposed	**14.67**	**14.45/13.75**	**14.48/14.11**	**12.80/9.47**
	Max. 5 Frames Async.	Baseline	13.10	12.88/12.41	12.67/12.33	12.31/11.60
		Proposed	**13.53**	**13.33/13.06**	**13.26/13.01**	**12.81/11.86**
	All	Baseline	12.51	12.36/11.98	12.35/12.13	10.33/7.61
		Proposed	**14.00**	**13.86/13.42**	**13.85/13.55**	**12.80/10.33**

Q: Number of augmented input visual streams.
For LR10, LE75 and RO10 test sets, each column contains SDRs evaluated with one/two low quality visual streams.

VI. SUMMARY

- Explore the attention-based multi-modal fusion method to build a robust time-domain audio-visual speech separation system.
- To force the model to adapt to the low quality video inputs, 3 types of data augmentation are introduced.
- The proposed methods outperforms the concatenation-based baseline on all the 3 types of low quality video inputs, and is robust to low quality training dataset.