AMPS: ASR with Multimodal Paraphrase Supervision

Abhishek Gupta*, Amruta Parulekar*, Sameep Chattopadhyay and Preethi Jyothi

Indian Institute of Technology Bombay

SCAN ME

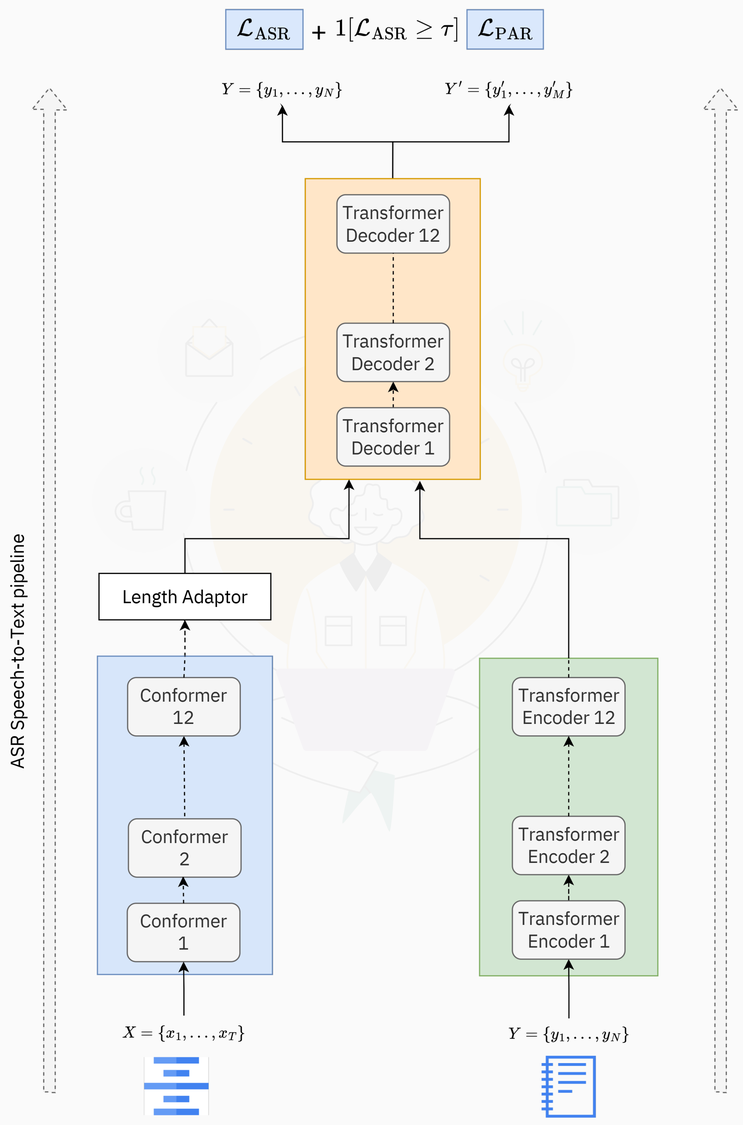
REVISITING ASR: IS FAITHFUL TRANSCRIPTION NECESSARY IF THE MEANING IS PRESERVED?

1 Central Idea

Speech recognition systems struggle with conversational speech, leading to high word error rates (WERs) in real-world settings

For ASR, use paraphrase as additional supervision to multimodal models that jointly model speech and text in joint embedding spaces

2 What are we proposing?



3 About the Dataset

IndicVoices

- Hindi (50 hour)
- Marathi (50 hour)
- Kannada (44 hour)
- Malayalam(50 hour)

ZambeziVoice

- Nyanja (5 hour)

4 Key Results

ASR Results

LANGUAGE	EVALUATION	INFERENCE	ASR	AMPS	AMPS$_\tau$
MARATHI	WER ↓	38.65	21.18	21.58	**20.20**
	BERTScore ↑	81.01	90.40	**92.31**	91.92
HINDI	WER ↓	29.16	20.63	20.83	**20.12**
	BERTScore ↑	88.65	93.60	93.65	**93.76**
MALAYALAM	WER ↓	56.15	42.06	42.09	**39.97**
	BERTScore ↑	84.35	91.50	91.56	**92.02**
KANNADA	WER ↓	69.26	41.41	40.10	**39.50**
	BERTScore ↑	76.65	89.84	90.21	**90.41**

LANGUAGE	EVALUATION	INFERENCE	ASR	AMPS	AMPS$_\tau$
NYANJA	WER ↓	42.34	22.16	21.90	**21.59**
	METEOR ↑	66.71	79.25	79.30	**80.10**

Human Evaluations

LANGUAGE	ASR	AMPS	AMPS$_\tau$
MARATHI	4.199	4.271	**4.314**
HINDI	3.608	3.625	**3.689**
MALAYALAM	3.635	3.688	**3.902**
KANNADA	3.433	3.542	**3.597**

5 Paraphrasing Techniques

Uske charge lagenge sir, satrah rupaye per kilometer ke hisaab se dekh lijiye
(There will be charges for that, sir. Check the rate of it's seventeen rupees per kilometer)

LLM → **Uske charge honge sattar-ek rupaye prati kilometer ki dar se**
(The charges for that will be seventeen rupees per kilometer)

Round trip translation → **Iska shulk satrah rupaye prati kilometer liya jayega**
(A charge of seventeen rupees per kilometer will be taken)

LANGUAGE	PARAPHRASE TYPE		BEAM SEARCH RT		LLM-PARA		RT	
		ASR	AMPS	AMPS$_\tau$	AMPS	AMPS$_\tau$	AMPS	AMPS$_\tau$
HINDI	WER ↓	23.14	23.14	**22.80**	22.35	**22.20**	**22.58**	22.81
	BERTScore ↑	92.60	92.59	**92.78**	92.89	**92.90**	**92.63**	92.62

6 Key Highlights

- Paraphrase-based **supervision** of multimodal models
- Predicts **semantically** similar words in **spontaneous** speech
- **Validated** using **Human evaluation**
- **Code** is available at the **QR** code above

1 LEARNABLE THRESHOLD

Future work will aim at making the threshold τ learnable

2 EVALUATION METRIC

Devise an improved metric to reliably measure the quality of our predictions

Poster by: **Abhishek Gupta**

Email: **abhishekgupta@cse.iitb.ac.in**

NAACL 2025