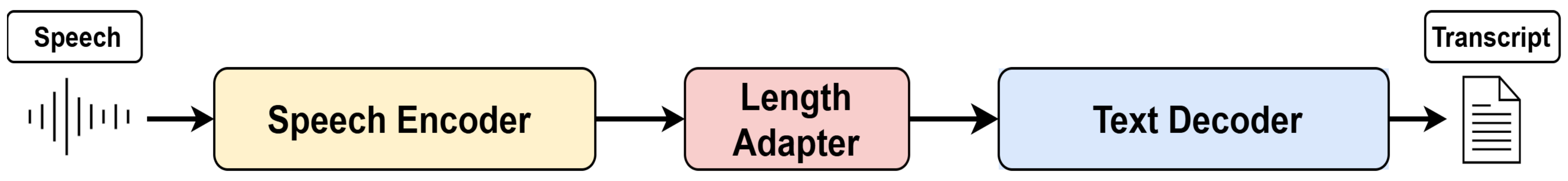
Parameter-efficient Adaptation of Multilingual Multimodal Models for Low-resource ASR

Abhishek Gupta* Amruta Parulekar* Sameep Chattopadhyay* Preethi Jyothi
Indian Institute of Technology Bombay, Mumbai, India
* Equal Contribution

Aim: To efficiently utilize textual and cross-lingual speech data in a computationally efficient manner to enhance the ASR performance of multilingual multimodal models for low-resource languages.

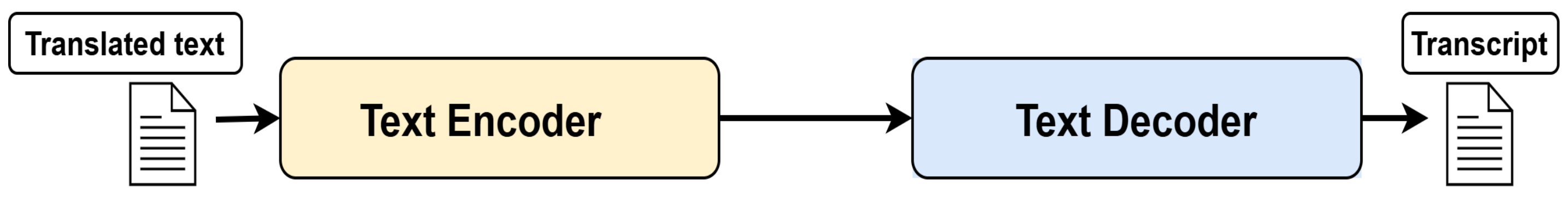
PEFT For ASR



Components fine-tuned	Learnable Parameters	Maitihli WER	Maitihli CER	Malayalam WER	Malayalam CER	Kannada WER	Kannada CER	Gujarati WER	Gujarati CER	Odia WER	Odia CER	Bengali WER	Bengali CER
None	-	82.20	43.39	56.15	20.65	69.29	29.11	41.03	24.50	42.81	17.38	37.70	18.44
Length adapter	46M	54.97	26.10	52.82	18.14	55.48	20.38	33.91	16.40	35.48	13.75	35.90	17.08
Text Decoder	201M	54.56	26.21	54.04	19.28	54.3	20.57	33.62	17.12	35.14	13.48	36.14	17.95
Speech Encoder	311M	43.87	17.79	46.99	13.45	47.91	14.93	27.79	11.58	29.82	9.24	29.07	12.09

Length adaptor with just **5 hours** of labeled speech provides significant improvement in ASR performance using < **10 %** of the total parameters.

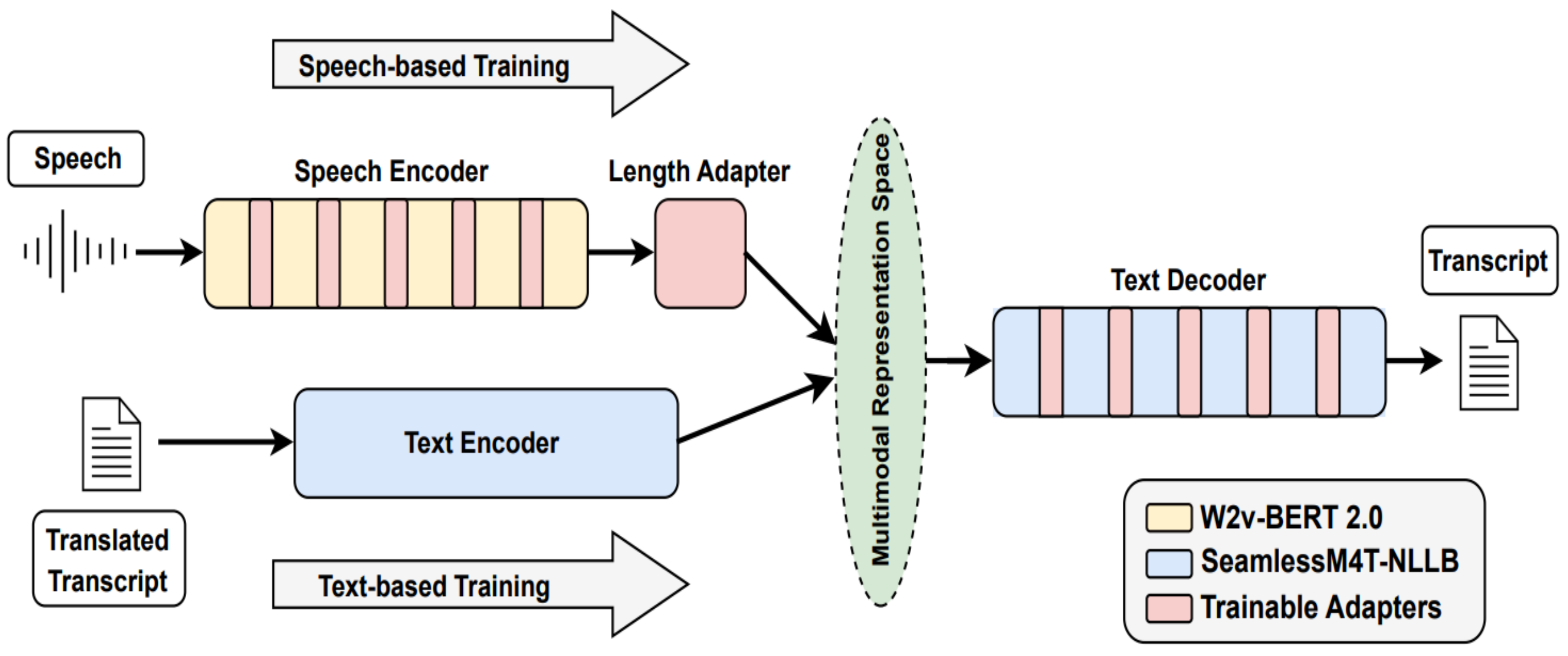
Text-only Adaptation



Text-only Adaptation	Learnable Parameters	Maithili WER	Maithili CER	Malayalam WER	Malayalam CER	Kannada WER	Kannada CER	Gujarati WER	Gujarati CER	Odia WER	Odia CER	Bengali WER	Bengali CER
None	-	82.20	43.39	56.15	20.65	69.29	29.11	41.03	24.50	42.81	17.38	37.70	18.44
5hr Transcript	6M	71.32	37.92	**53.96**	**18.94**	70.52	32.54	35.67	19.19	38.77	**14.84**	**35.28**	**16.77**
Full Transcript	6M	**68.24**	**36.84**	55.30	20.43	**68.13**	**26.91**	**35.45**	**18.66**	**38.39**	16.22	35.44	17.73

Adapting the **Text Decoder** with **translated transcript pairs** improves the ASR performance of SeamlessM4T, achieving both data and parameter efficiency.

Combining Both Techniques



A multimodal model like SeamlessM4T can be fine-tuned in a parameter-efficient manner with either speech or text data by inserting adapters in the pretrained base model.

We test our adaptation techniques using the primarily conversational speech data from **IndicVoices** Dataset.

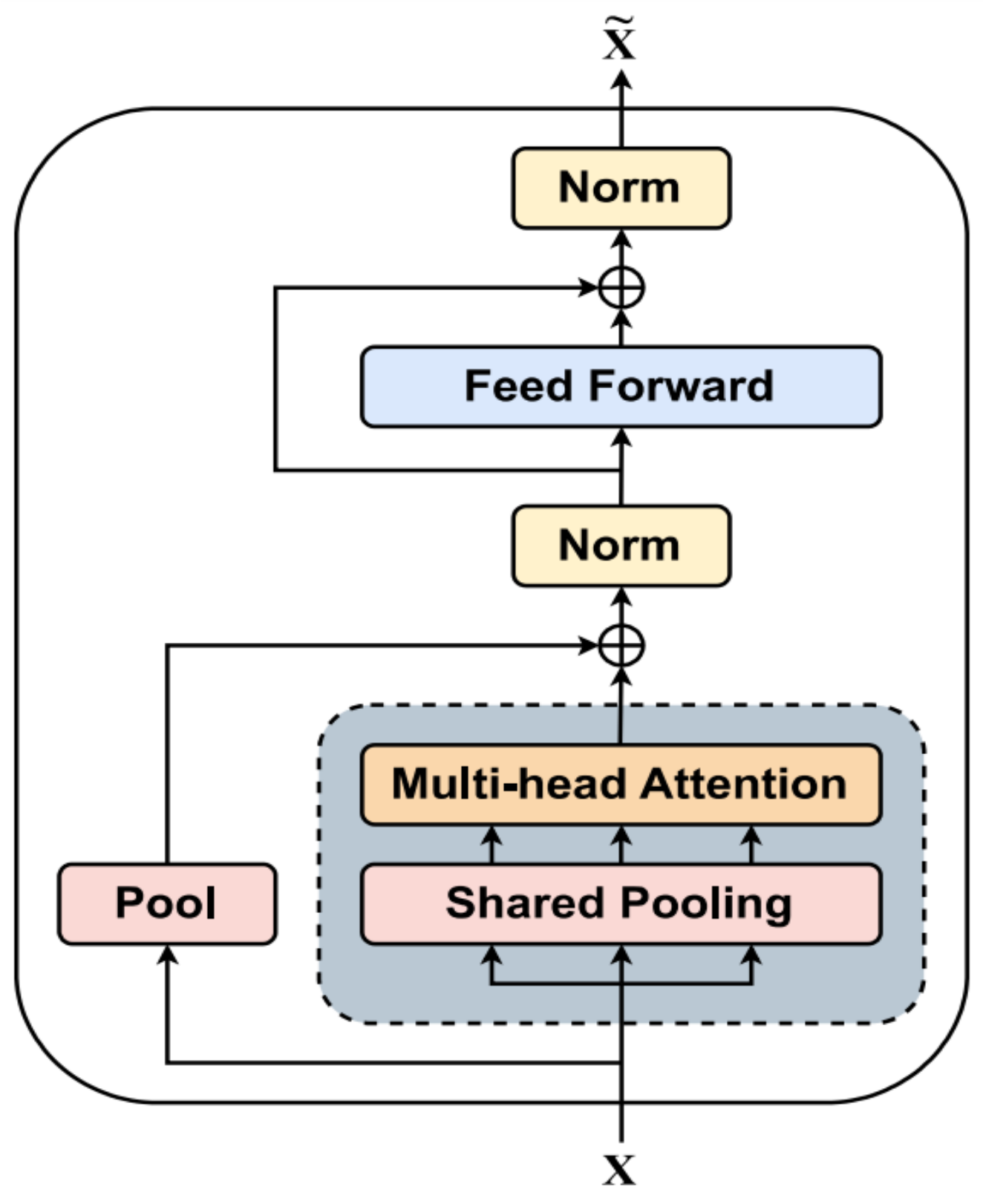
Language	Component Fine-tuned	None		Length Adapter		Encoder Adapter		Decoder Adapter		Len+Enc Adapter		Encoder Adapter (L)		All Components	
	Learnable Parameters	-		46 M		6 M		6 M		52 M		50 M		571 M	
	System	A	T-A	A	T-A	A	T-A	A	T-A	A	T-A	A	T-A	A	T-A
Maithili	WER	82.20	68.24	54.97	54.74	52.95	48.14	63.52	58.39	47.92	45.98	46.08	**44.60**	42.58	46.54
	CER	43.39	36.84	26.10	27.10	22.86	21.58	31.60	29.70	20.56	20.47	19.20	**19.52**	17.14	20.78
Malayalam	WER	56.15	55.3	52.82	52.51	49.71	50.14	56.03	53.71	48.22	48.19	47.81	**47.75**	47.38	45.9
	CER	20.65	20.43	18.14	18.87	15.34	16.35	20.21	20.00	14.76	15.46	14.12	**14.92**	13.86	13.38
Kannada	WER	69.29	68.13	55.48	53.83	52.54	53.29	62.88	58.71	49.36	48.24	49.14	**47.75**	45.48	43.5
	CER	29.11	26.91	20.38	20.94	16.95	18.84	23.76	23.44	15.63	16.51	15.26	**14.92**	14.06	14.18
Gujarati	WER	41.03	35.45	33.91	34.41	29.20	**27.72**	38.88	35.53	28.03	27.73	28.09	27.90	25.56	26.31
	CER	24.50	18.66	16.40	17.41	11.96	**12.05**	19.28	17.80	12.63	12.35	12.00	12.50	11.28	11.67
Odia	WER	42.81	38.39	35.48	34.99	32.03	32.97	38.55	36.24	30.09	31.18	30.04	**28.92**	30.54	30.17
	CER	17.38	16.22	13.75	14.62	10.57	11.25	14.50	14.57	10.11	11.32	10.01	**9.92**	10.37	10.30
Bengali	WER	37.70	35.44	35.90	35.09	29.65	28.77	38.10	35.60	29.96	**28.50**	29.30	31.92	28.12	27.62
	CER	18.44	17.73	17.08	17.22	12.76	12.58	18.59	17.72	13.06	**12.38**	12.52	14.63	12.12	11.91

We analyze the adaptation strategies for **6** Indic languages.

System A: Only ASR finetuning (ASR FT)

System T-A: Text-only adaptation followed by ASR FT

Cross-lingual Transfer Learning



Hypothesis

A length adapter (left) can capture the prosodic features of a language without overfitting on its syntax.

Can fine-tuning the length adaptor with speech from a related language, combined with target language text adaptation, improve the ASR quality in an extremely low-resource setting without any available speech data?

Language 1 (Target)	Language 2 (ASR Fine-tuning)	Genetic Distance	Text- only Adaptation	ASR fine-tuned Component	Number of Parameters	WER	CER
Maithili	None	-	No	None	-	82.2	43.39
	Bengali	0.625	No	Length Adapter	46M	79.77	40.04
			No	Encoder Adapter	50M	81.81	41.61
			No	Len. + Enc. Adapter	52M	80.81	40.44
			Yes	Length Adapter	6M+46M	**72.52**	**39.31**
	Kannada	1.000	No	Length Adapter	46M	80.29	38.37
			No	Encoder Adapter	50M	85.25	41.58
Odia	None	-	No	None	-	42.81	17.38
	Bengali	0.375	No	Length adapter	46M	41.05	15.07
			No	Encoder Adapter	50M	43.67	16.03
			No	Len. + Enc. Adapter	52M	42.4	15.27
			Yes	Length Adapter	6M+46M	**35.45**	**13.92**
	Kannada	1.000	No	Length Adapter	46M	41.21	14.08
			No	Encoder Adapter	50M	44.01	14.59

Target Languages: Maithili & Odia

High-resource Pivots: Bengali (Related) & Kannada (Unrelated)

Key Result: 17% reduction in relative WER in a zero-shot setting without any labeled speech from the target language.

The authors acknowledge the support of the Amazon IITB AI ML Initiative and the consortium project on "Speech Technologies in Indian Languages" under National Language Translation Mission (NLTM), MeitY, Government of India.