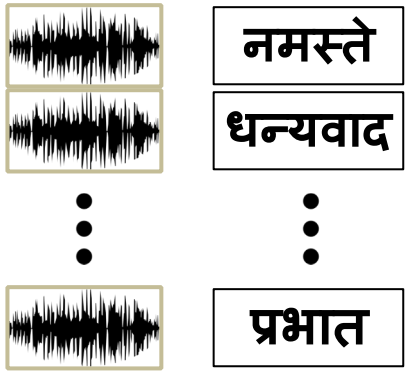
Automatic Speech Recognition system for Indian Languages “IndicWav2Vec”

AI4Bharat Speech Team

What is ASR

Challenges for Indian ASR Systems

Manual Transcription *is a time-consuming process causing delays*



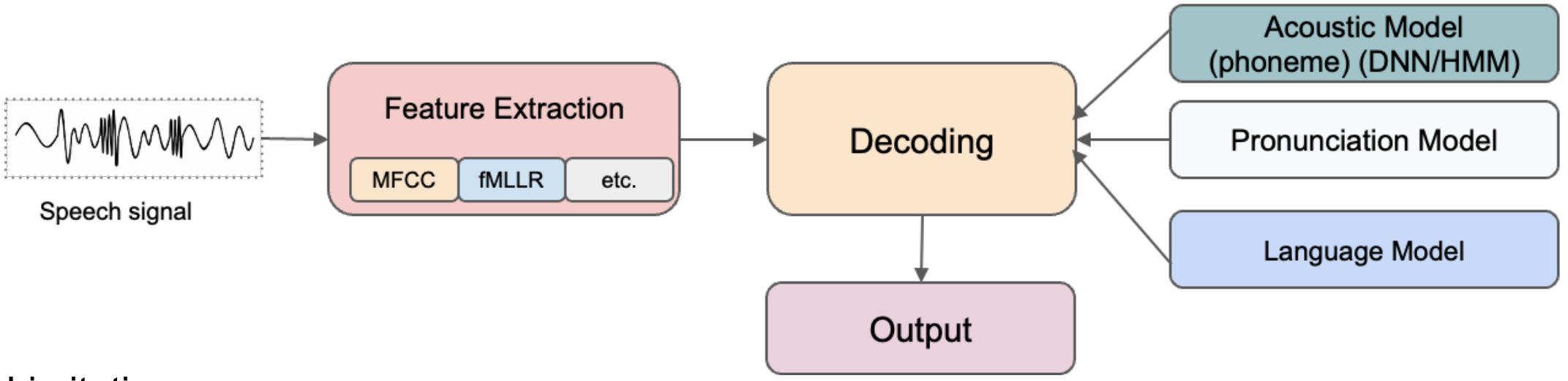
Modern ASR models rely on ***large amounts of labeled data*** *for each language.*
It is -

- ***Expensive***
- ***Not scalable***

Accurate, free, open-source *ASR systems are* ***not*** *available for all the Indic languages of interest*

Complex inflectional systems *leads to* ***larger vocabulary*** *sizes posing challenges to incorporate* ***language models***

Traditional ASR Modelling Systems

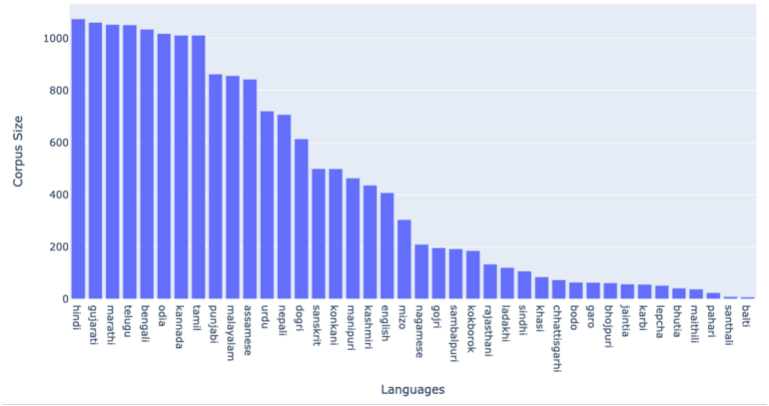


Limitations:

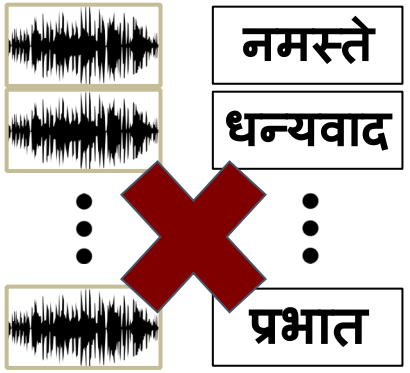
- Complexity
- Chained Inefficiencies
- Reliance on labelled datasets
- Deployment

Can we do better?

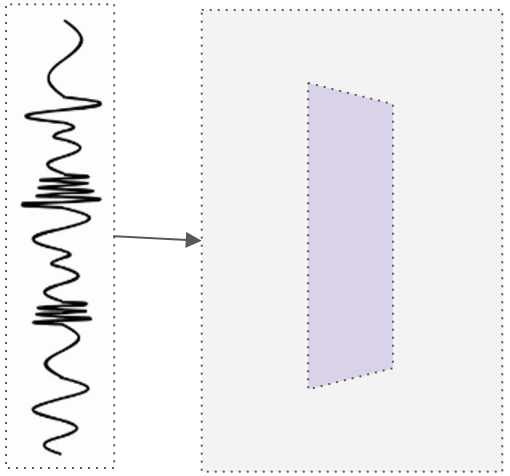
How can we do better?



Utilize large amounts of ***freely available unlabeled data*** *on Youtube, NewsOnAir to perform* ***Self-Supervised Pretraining***

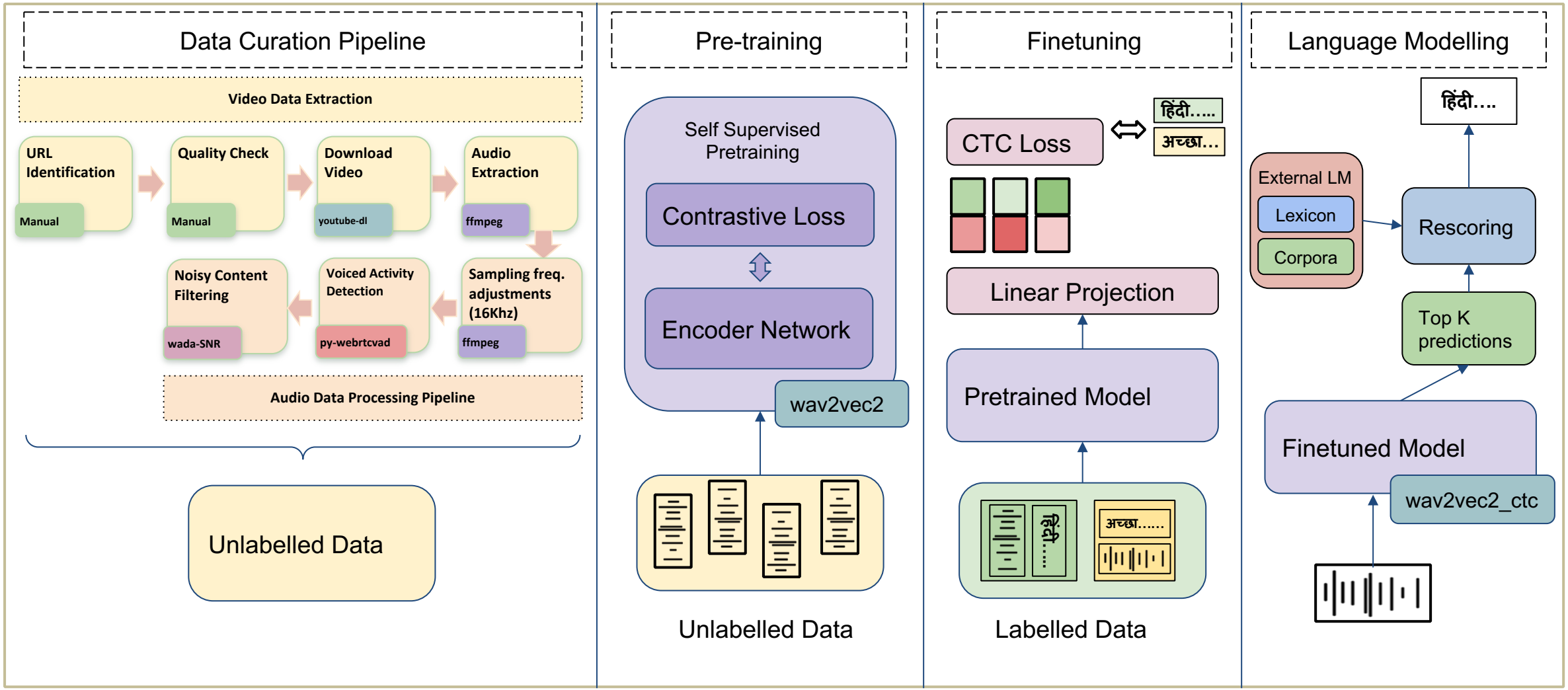


Reduce *the requirement for labeled fine-tuning data.*

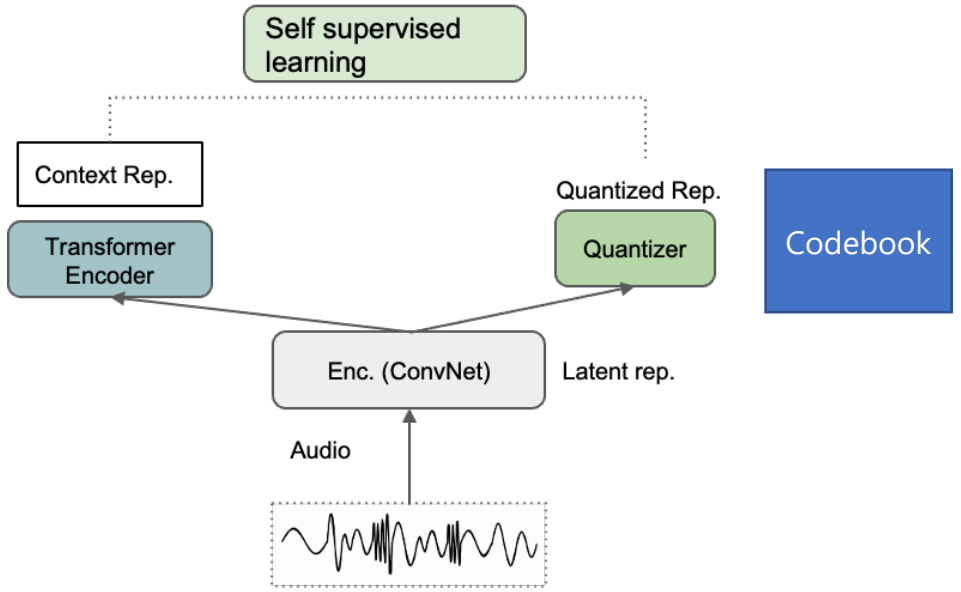


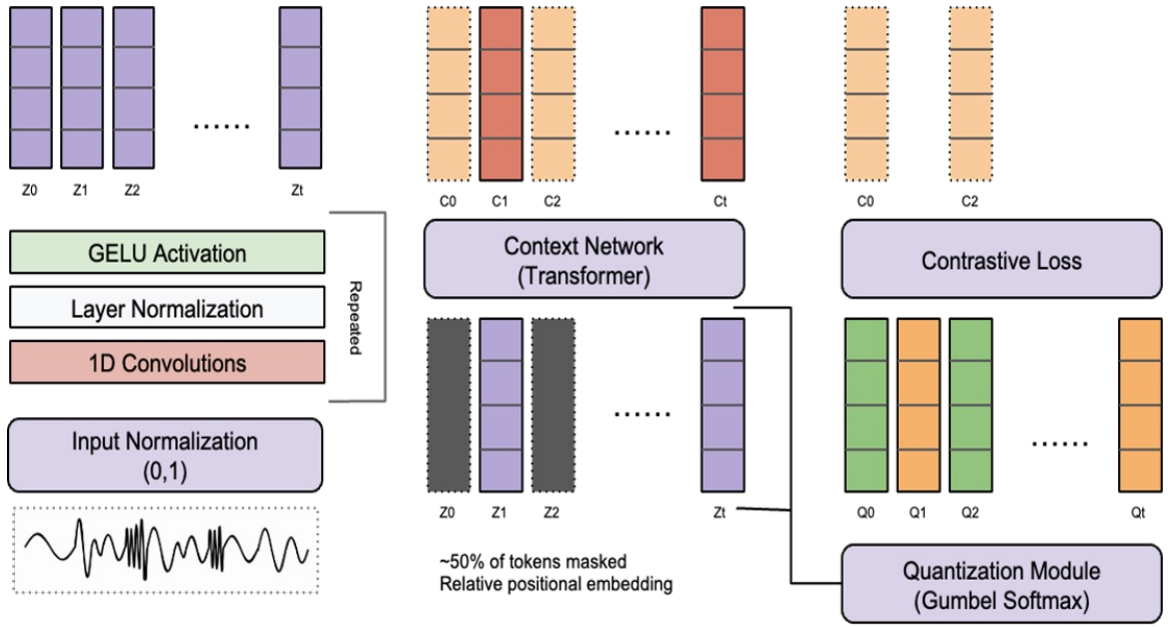
Build ***robust ASR systems*** *by pretraining a multilingual model across Indian languages*

Technical Description of Proposed Methodology



Self-Supervised Pretraining



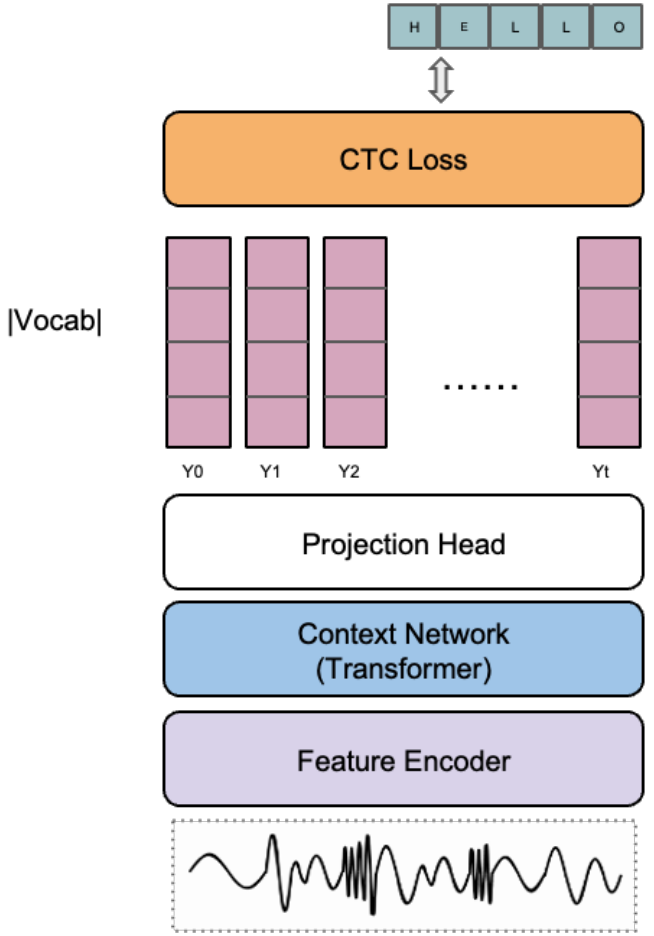


Self-supervision

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

Finetuning

Fine-tuning:
- Add a projection layer
- CTC loss

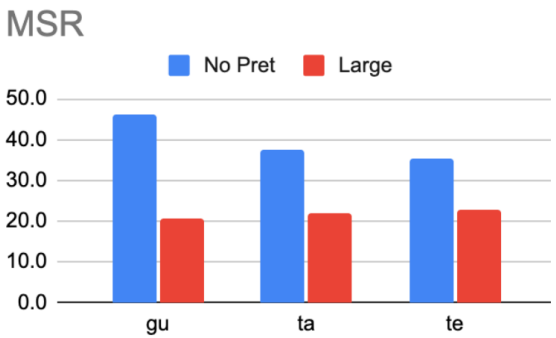


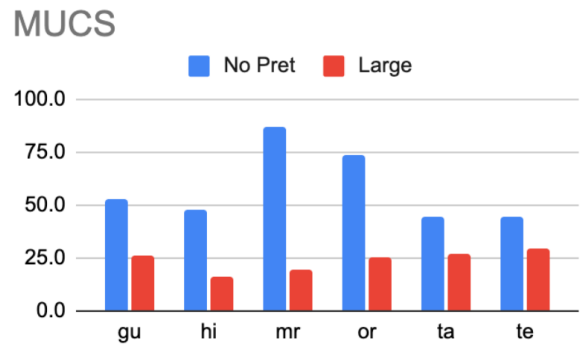
Results

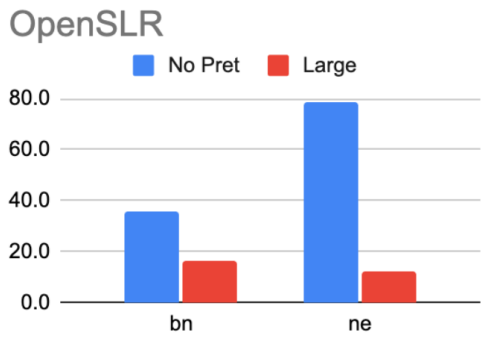
Pretraining Data source: **Dhwani**

Finetuning Data source: **MUCS, MSR, OpenSLR**

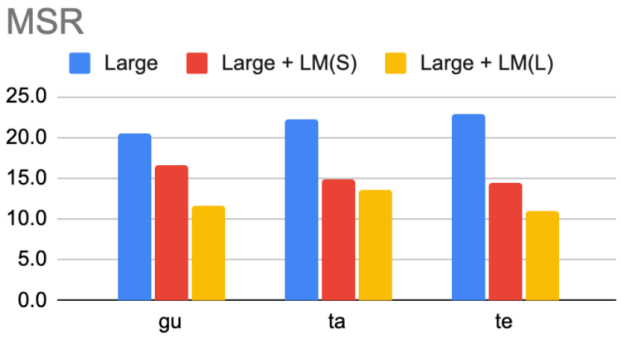
Role of Pretraining

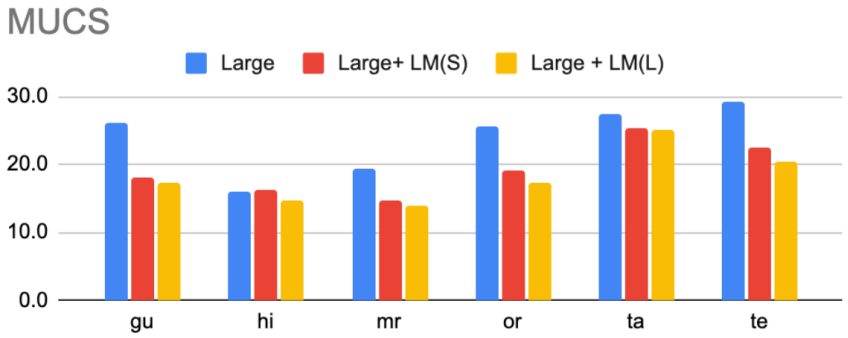






Role of Language Model





MODELS

Siri, phone mai 12 baje ki alarm lagado!

Personal Smart Assistants

Differently abled persons can hear through text

Stepping stone for building S2S Translation systems

ASR MT TTS

High quality education content in native language

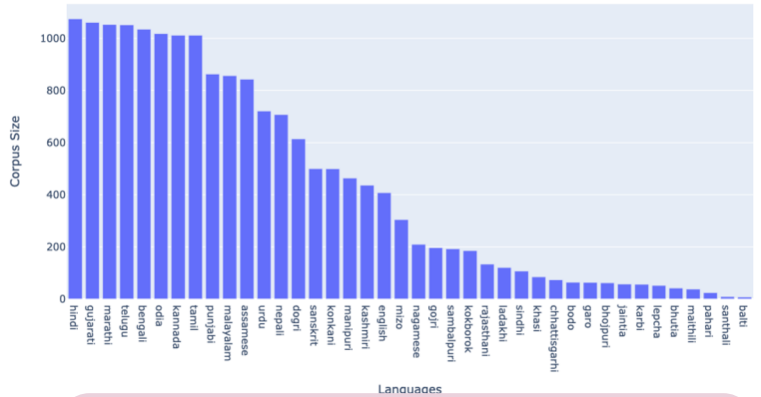
Realtime Translations

ASR Systems

अच्छा.....

SOTA ASR models for 9 languages** across 3 datasets*.

IndicWav2Vec



Dhwani Dataset
17000 hours of Pretraining Data across 40 Languages

Open sourced Pretrained model which can be used in building speech applications

Extensive ablations and insights about building ASR Systems, informing LM Choice, Lexicon Choice, Size of Pretraining data.

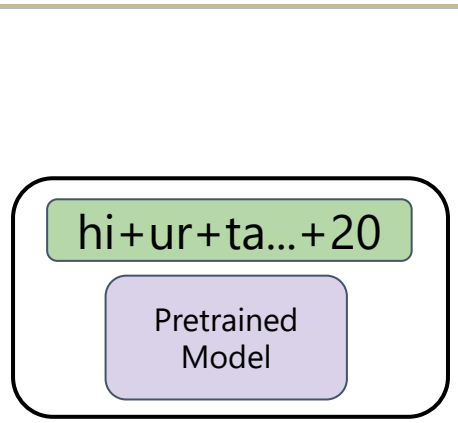
Pretraining Significantly helps

Choice of Lexicon is crucial

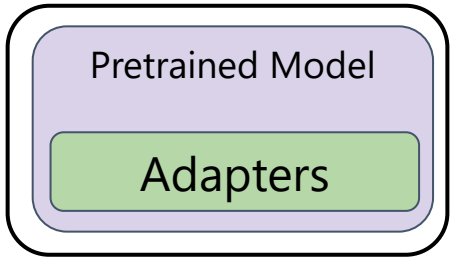
Moving to bigger LM Helps

**hi, ta, te, gu, mr, or, ne, bn, si *MUCS, MSR, OpenSLR
Results on pretraining effects and comparison to SOTA approaches on benchmark datasets are provided in Appendix

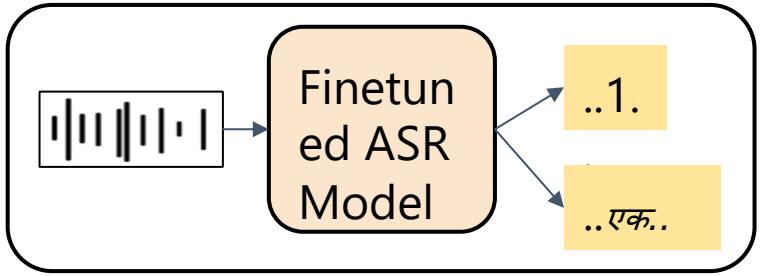
Future Directions



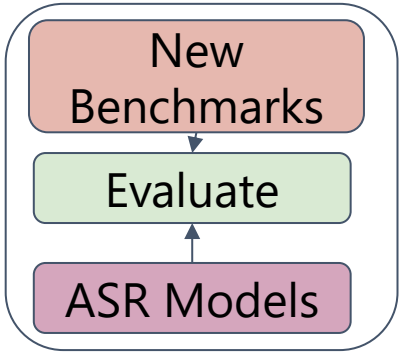
Completely Multilingual setup



Using domain specific adapters



Standardization of the output by using ITN (1 and एक causes confusion even if model is correct in both cases)



Building Benchmark to better assess model Performance

Questions?