# SPEECH RECOGNITION FOR INDIAN LANGUAGES

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra


## SUMMARY

- Curated 17000 hours of raw speech corpus for 40 Indian languages.
- Mined 6000 hours of data for 12 Indian languages.
- Created Multilingual Benchmark for 12 Indian Languages
- Trained Large-scale Language Models from IndicCorp to improve ASR performance.
- Trained and deployed efficient models for on-device ASR.


## What is Automatic Speech Recognition?

... सारे जहाँ से अच्छा ...

## What are the Challenges for Indian Languages?

नमस्ते
धन्यवाद
प्रभात

Lack of *labelled data*

Complex *Inflectional Systems*

पढ़ती पढी पढ़ पढा पढ़ता पढना पढाई

*On-device* models are necessity

Lack of robust *benchmarks*

Lack of existing *pretrained* models

## Our Approach

1. Exploiting raw audio data
2. Mine labelled data from Govt. sources
3. Create a robust multilingual benchmark
4. Train large scale LMs to handle inflections
5. Train efficient models for on-device ASR

## 1. Exploiting use of Raw Audio Data

NewsOnAir *11000 hours*

**17000 hours** of Audio Data

YouTube *6000 hours*

## 2. Mining Labelled data from Govt. Sources

NEWS SERVICES DIVISION ALL INDIA RADIO

News Bulletin *(Long Audio)*

प्रदेश में नवरात्रि पर्व की शुरूआत श्रद्धालु माँ दुर्गा की पूजा कर रहे हैं दुर्गा पंडलों में साजसज्जा की गई

**Full Transcript**

Audio-Text Alignment

प्रदेश में नवरात्रि पर्व की शुरूआत

श्रद्धालु माँ दुर्गा की पूजा कर रहे हैं

दुर्गा पंडलों में सा जसज्जा की गई

**Sentence-level** *Aligned*

**6000 hours** of Mined Data

## 3. Create a Robust Multilingual Benchmark

Automatic Speech Recognition

Speaker Identification (Monolingual /Multilingual)


Language Identification

Automatic Speaker Verification

Query By Example

**6 tasks** across 12 Indian Languages

## 4. Large Scale LMs to handle inflections

ASR model

पर्देश में नवराती परव

प्रदेश में नवरात्रि पर्व

परदेश में नवरात्रि पर्व

Multiple Hypothesis

Language Model

पर्देश में नवराती परव

प्रदेश में नवरात्रि पर्व

परदेश में नवरात्रि पर्व

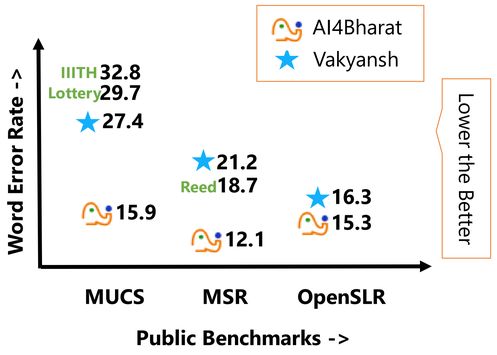
Select *Best One*

## 5. Efficient Models for on-device ASR

Train on Cloud → Quantization → TorchScripting + Mobile Conversion → Infer on Mobile

Streaming Mode ... प्रदेश में नवरात्रि पर्व की ...

## Results




## OUR PLAN AHEAD

**Support all 22 Indian languages**

**Adapt the trained models for specific domains**

**Collect data and train models on conversational speech**

**Train smaller models, reduce APK sizes for on-device ASR**


## ACKNOWLEDGEMENTS

We would like to thank the Nilekani Philanthropies for their generous grant which helped in setting up the "Nilekani Centre at AI4Bharat, IIT Madras" to support our students and research staff, as well as data and computational requirements. We would like to thank The Ministry of Electronics and Information Technology for its grant to support the creation of datasets and models for Indian languages under its ambitious Digital India Bhashini project. We would also like to thank the Centre for Development of Advanced Computing, India (C-DAC) for providing access to the Param Siddhi supercomputer for training our models. Lastly, we would like to thank Microsoft for its grant to create datasets and tools for Indian languages.


**The focus of AI4Bharat, an initiative of IIT Madras, is on building open-source language AI for Indian languages, including datasets, models, and applications.**

https://ai4bharat.iitm.ac.in/speech-recognition
https://github.com/AI4Bharat/IndicWav2Vec
Contact: Tahir Javed, Kaushal Bhogale, Abhigyan Raman