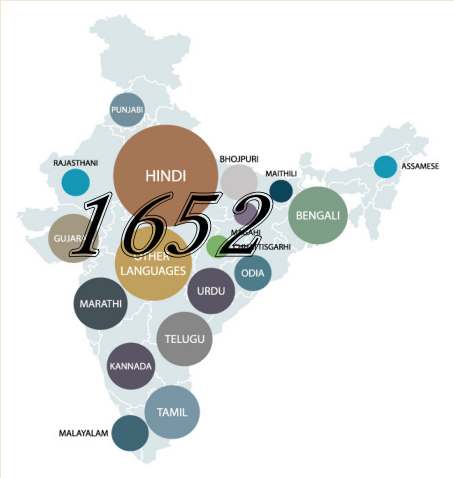
MUCS 2021: MUltilingual and Code-Switching ASR challenges for low resource Indian languages

Anuj Diwan[1], Rakesh Vaideeswaran[2], Sanket Shah[3], Ankita Singh[1], Srinivasa Raghavan[4], Shreya Khare[5], Vinit Unni[1], Saurabh Vyas[4], Akash Rajpuria[4], **Chiranjeevi Yarra**[6], Ashish Mittal[5], Prasanta Kumar Ghosh[2], Preethi Jyothi[1], Kalika Bali[3], Vivek Seshadri[3], Sunayana Sitaram[3], Samarth Bharadwaj[5], Jai Nanavati[4], Raoul Nanavati[4], Karthik Sankaranarayanan[5], Tejaswi Seeram[7], Basil Abraham[7]

[1]**Computer Science & Engineering, Indian Institute of Technology (IIT), Bombay, India**
[2]**Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India**
[3]**Microsoft Research India, Hyderabad, India**
[4]**Navana Tech India Private Limited, Bangalore, India**
[5]**IBM Research India, Bangalore, India**
[6]**Language Technologies Research Center (LTRC), IIIT Hyderabad, India**
[7]**Microsoft Corporation, Bangalore, India**

Introduction

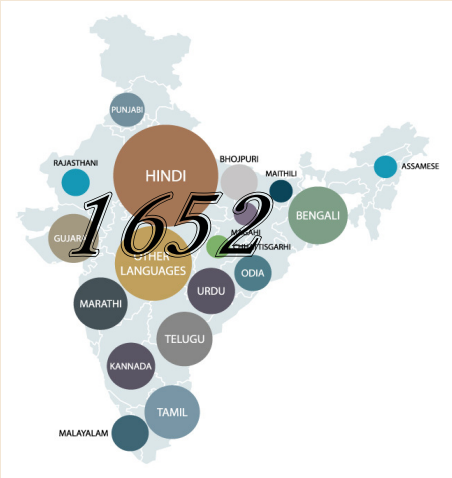
Language diversity



- Most of the languages are low-resource.
- Code-switching naturally happens between Indian languages and English.

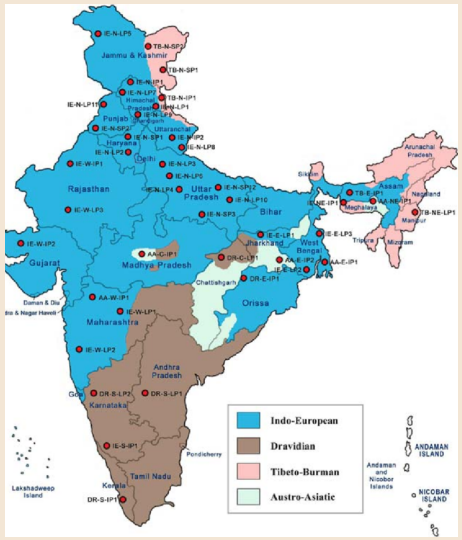
Introduction

Language diversity



- Most of the languages are low-resource.
- Code-switching naturally happens between Indian languages and English.

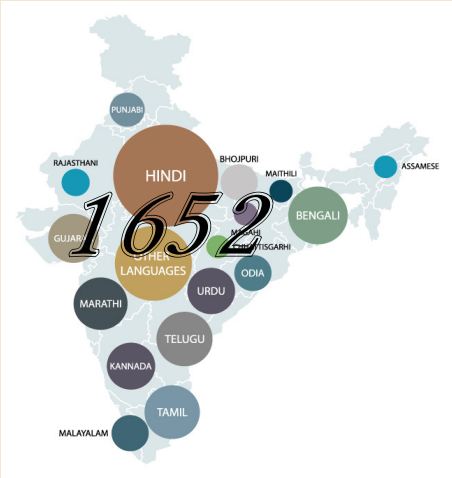
Language characteristics



- Evolved from four language families.
- Similar based on demographic regions.

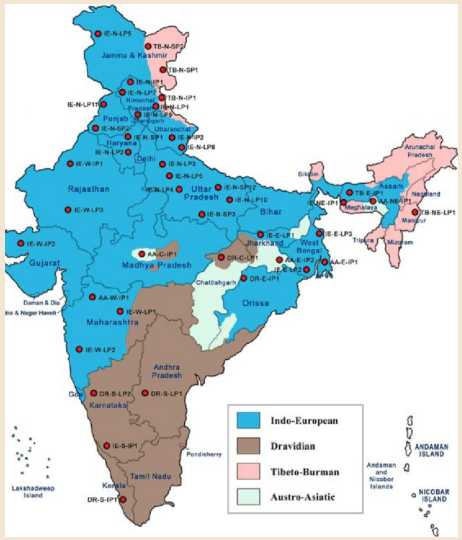
Introduction

Language diversity



- Most of the languages are low-resource.
- Code-switching naturally happens between Indian languages and English.

Language characteristics



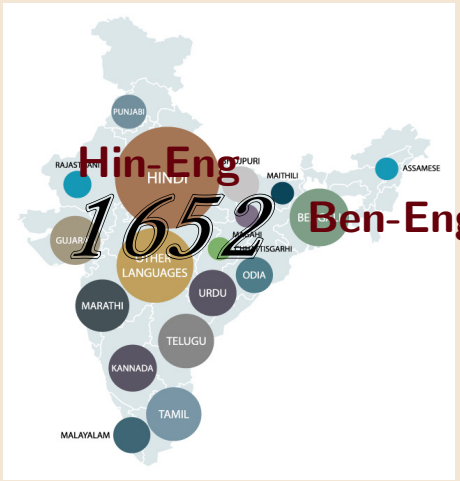
- Evolved from four language families.
- Similar based on demographic regions.

Objectives

1. **Multilingual ASR**: Exploring the common acoustic properties for better ASR model.
2. **Code-switching ASR**: Exploit code-switching patterns to help in modelling ASR.

Introduction

Language diversity



- Most of the languages are low-resource.
- Code-switching naturally happens between Indian languages and English.

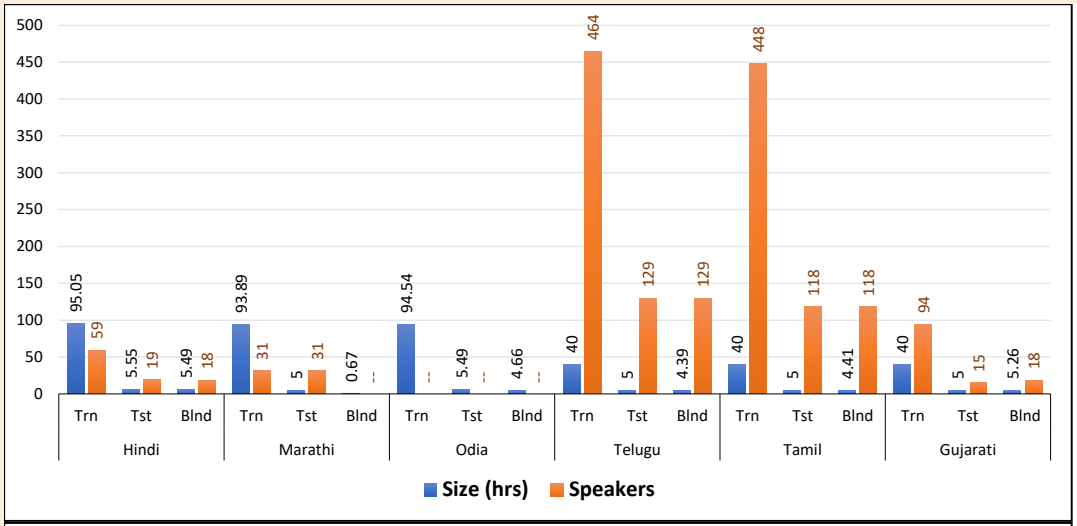
Language characteristics

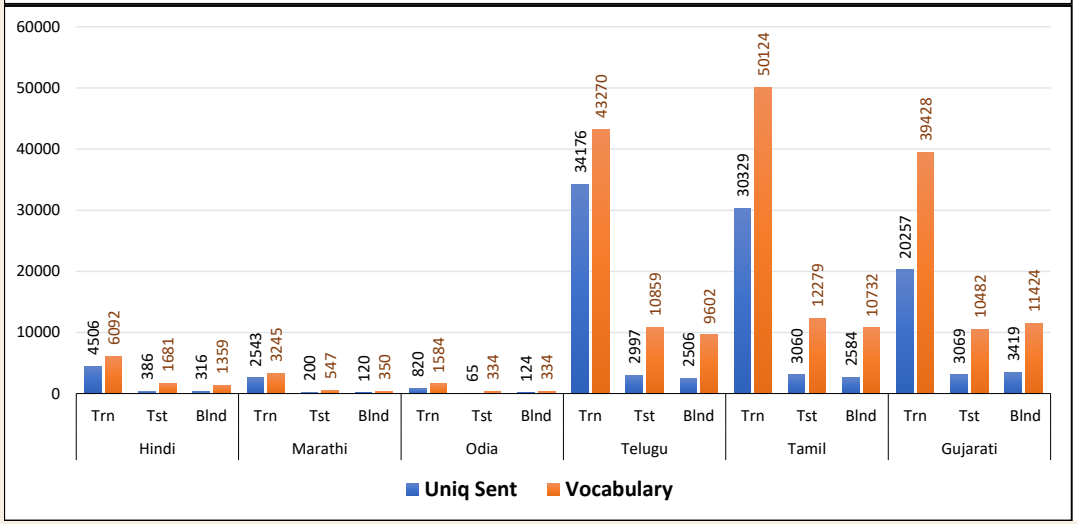
- Evolved from four language families.
- Similar based on demographic regions.

Objectives

1. **Multilingual ASR**: Exploring the common acoustic properties for better ASR model.
2. **Code-switching ASR**: Exploit code-switching patterns to help in modelling ASR.

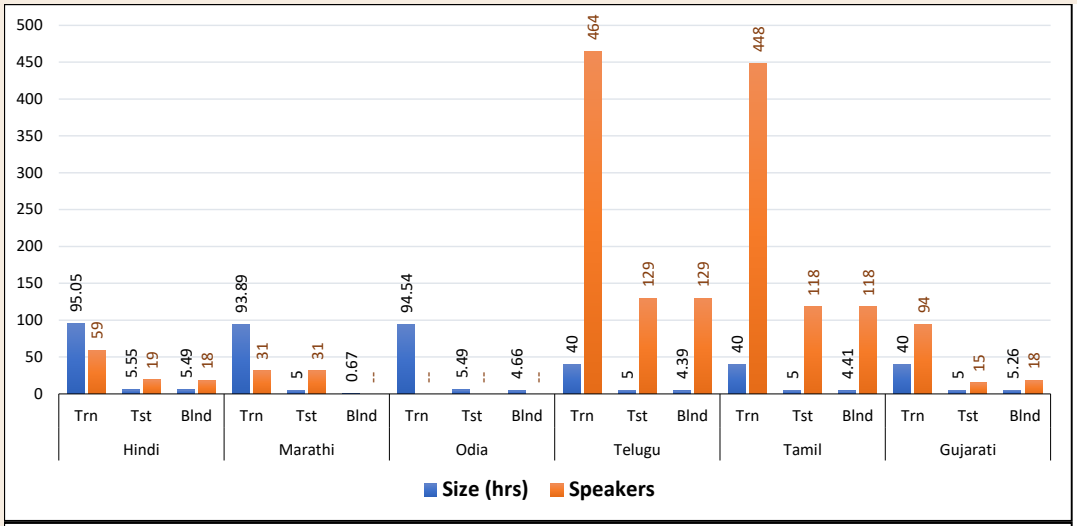
Data details: Multilingual ASR[1]

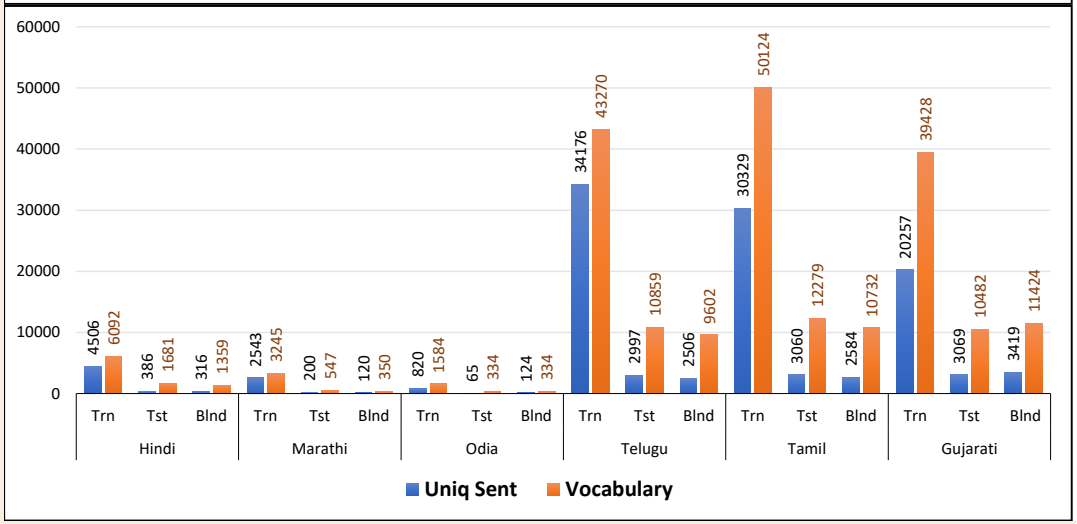




[1], "Indian Language Speech sound Label set (ILSL12)", Indian Language TTS Consortium and ASR Consortium

Data details: Multilingual ASR[1]

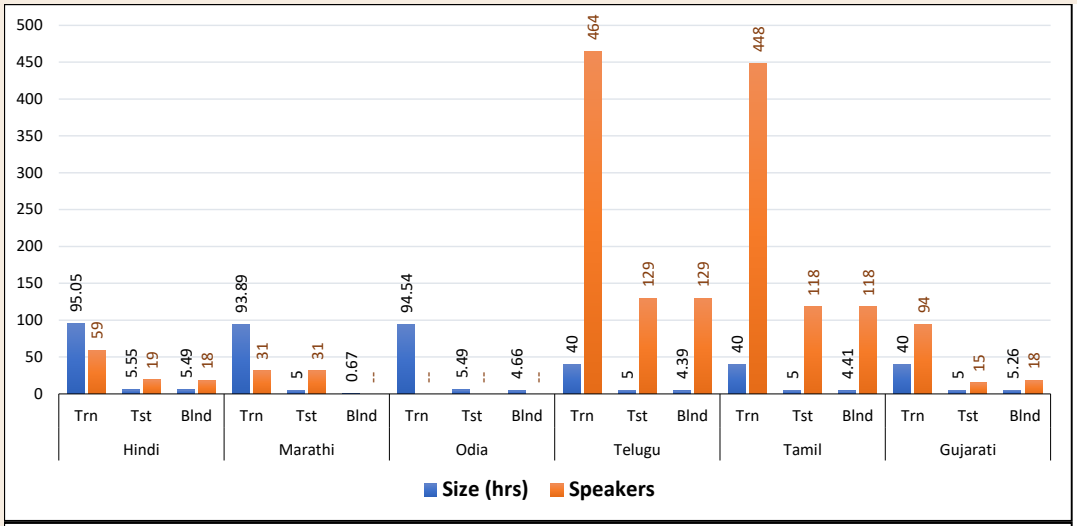


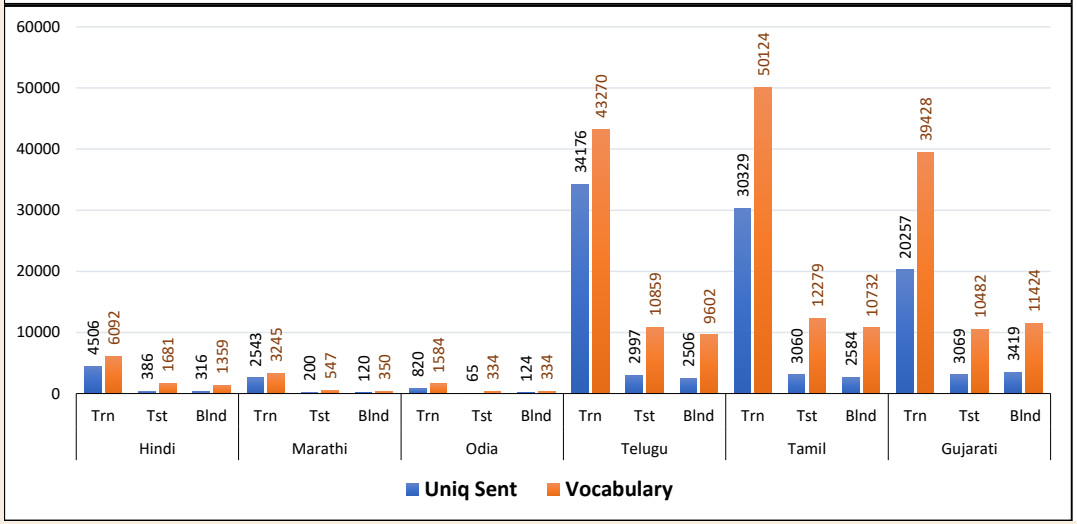


- A total of ∼450 hrs.

[1], "Indian Language Speech sound Label set (ILSL12)", Indian Language TTS Consortium and ASR Consortium

Data details: Multilingual ASR[1]

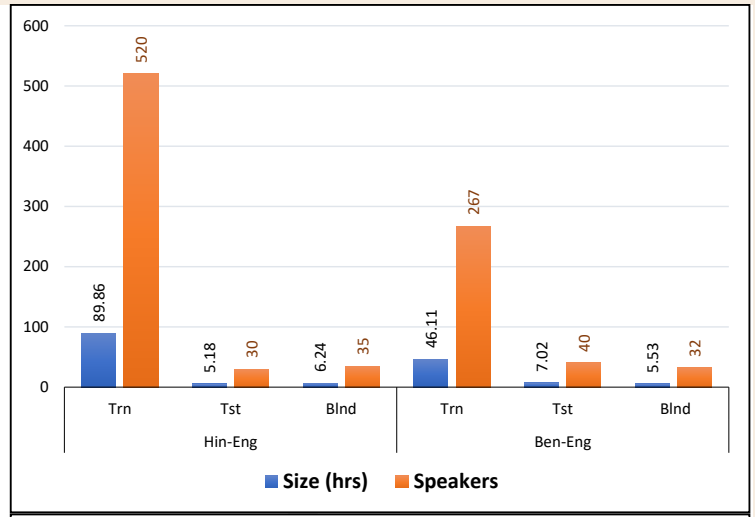


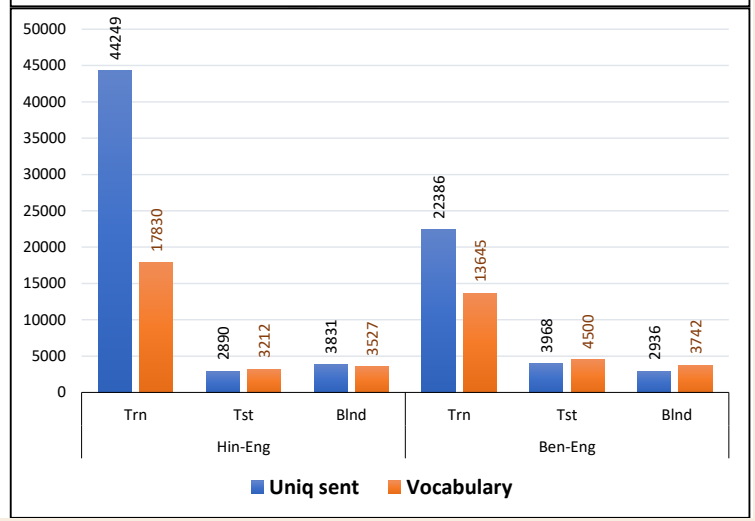


- A total of ∼450 hrs.
- Grapheme set:
  - Indian language speech sound label set (ILSL12) standard.
  - Size: 69, 61, 68, 64, 50 and 65.

[1], "Indian Language Speech sound Label set (ILSL12)", Indian Language TTS Consortium and ASR Consortium

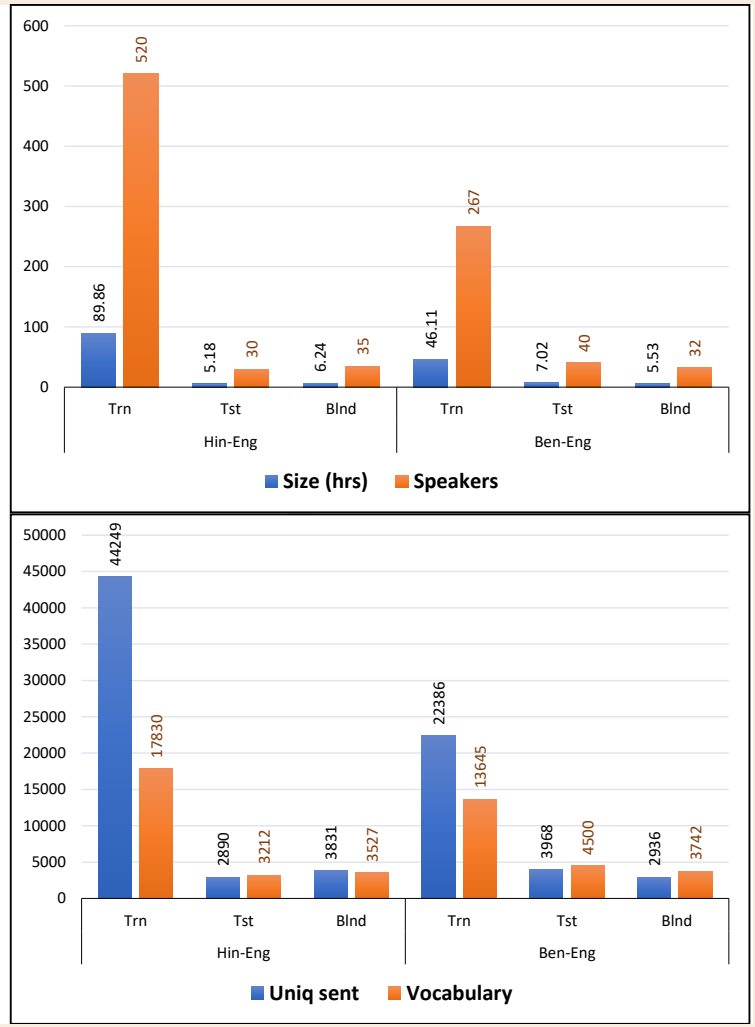
Data details: Code-switching ASR





Data details: Code-switching ASR

- A total of ∼150 hrs.



Data details: Code-switching ASR

- A total of ∼150 hrs.
- Technical content of lectures
  - Sentence time-stamps available
  - Alignment is used to obtain audio files for each sentence segment.