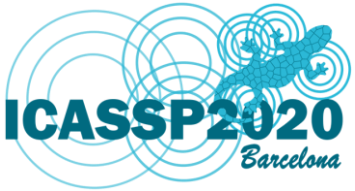


A DATASET FOR MEASURING READING LEVELS IN INDIA AT SCALE

Dolly Agarwal[1], Jayant Gupchup[2], Nishant Baghel[1]

[1]Pratham Education Foundation, [2]Pratham Volunteer*

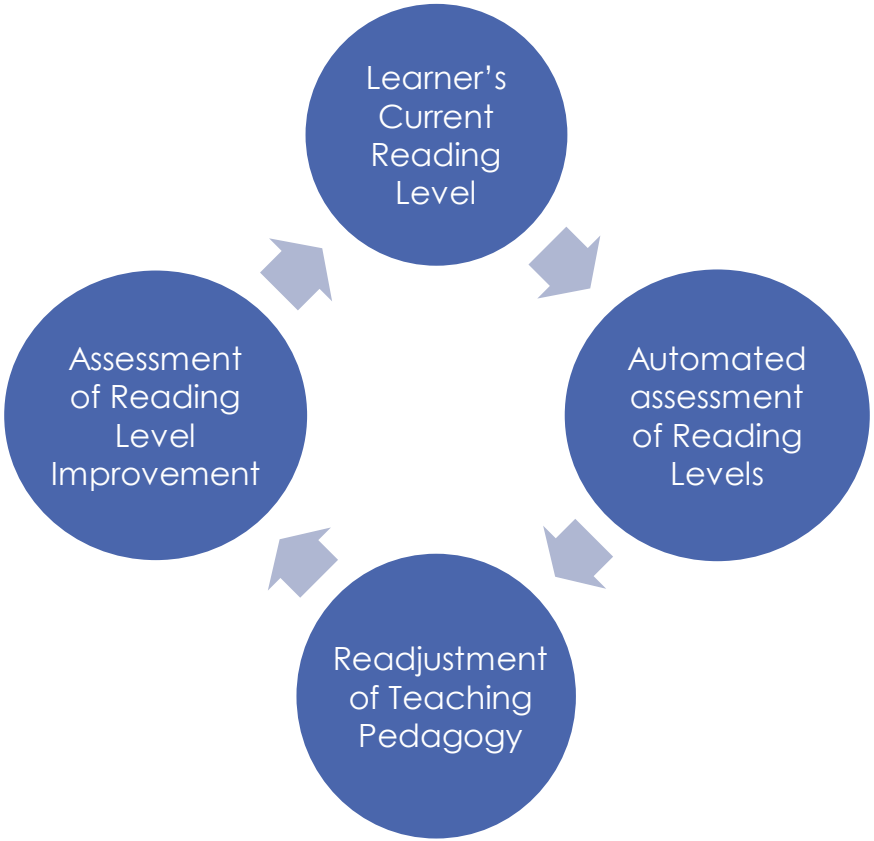
[1]{dolly.agarwal, nishant.baghel}@pratham.org, [2]gupchup@gmail.com

One out of four children in India are leaving grade eight without basic reading skills.

The Problem

Measuring reading level is a manual, laborious, expensive and error-prone process in high pupil teacher ratio environment

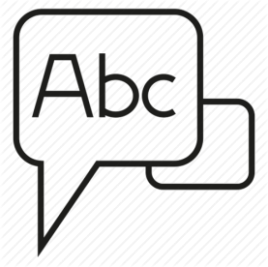
The Solution



Recent advances in deep learning and speech recognition technology opens up the possibility of automating this task.

The Challenges

Unlike adult speech corpora, there are less speech data of children



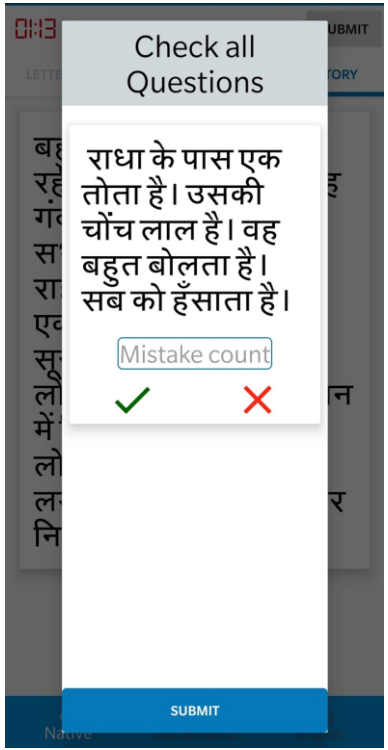
Most children's speech databases which are available are either in English or some European languages.

The differences in languages, pronunciation and accent across India pose many challenges in the development of accurate evaluation tools

ASER Dataset

- ASER is a household-based standardized survey of children's schooling and learning status
- An external evaluation of ASER conducted in 2013-14 calculated that the ASER survey costs a little over Rs 100 per child (approximately U.S. $1.40).
- This cost albeit lower compared to other large scale learning assessments is notable because ASER involves assessing about 700,000 children every year

Data Collection Methodology



ASER Sample Collection (Early Access)
https://play.google.com/store/apps/details?id=info.pratham.asersample

Complete Vocabulary details is available at:
https://github.com/PrathamOrg/ASER-Dataset.git

The Dataset

5,301 CHILDREN

81,330 AUDIO CLIPS

123 HRS OF SPEECH

HINDI, MARATHI & ENGLISH

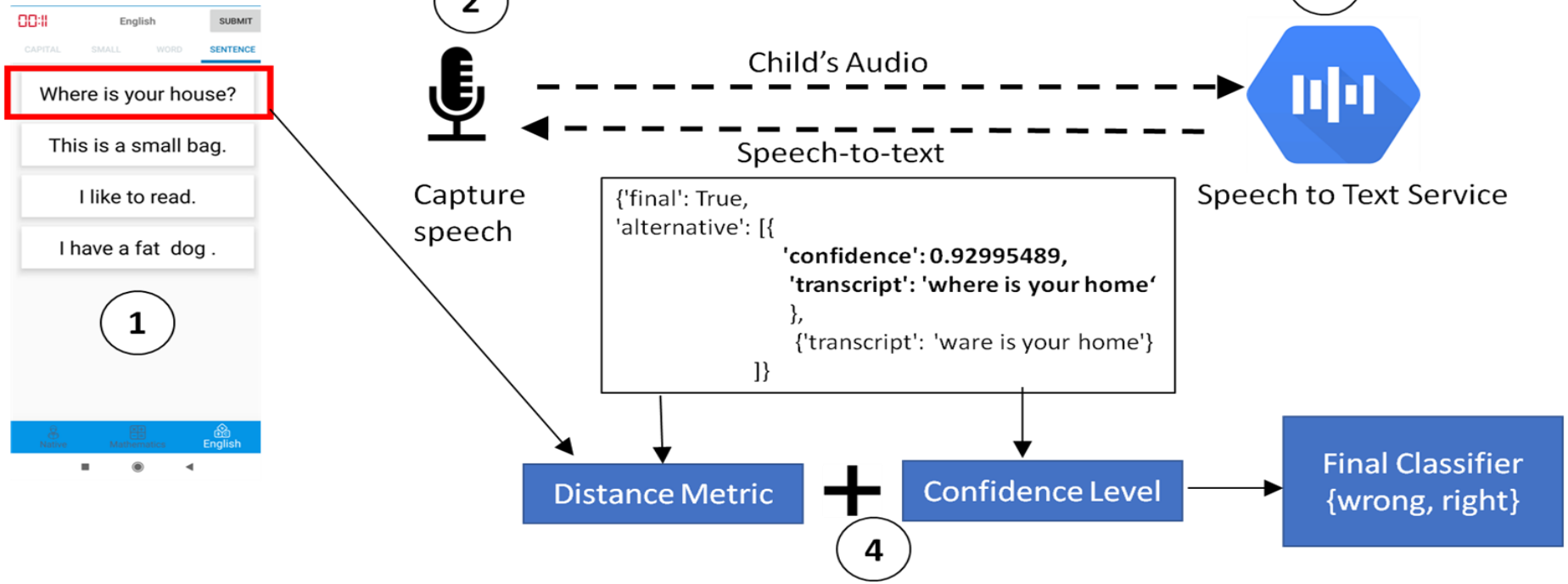
The Dataset is available at:
https://github.com/PrathamOrg/ASER-Dataset.git

Hindi (3959 subjects)					
	Story	Para	Word	Letter	Total
No. of distinct Samples	4	8	41	17	70
No. of audio clips	2244	3290	4669	5667	15870
Duration (hrs)	28.30	14.81	5.65	4.92	53.68

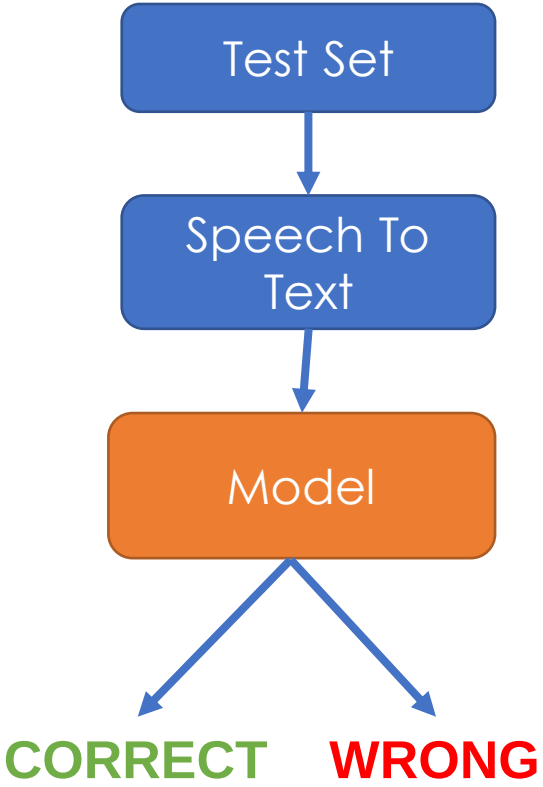
Marathi (1342 subjects)					
	Story	Para	Word	Letter	Total
No. of distinct Samples	4	8	36	23	71
No. of audio clips	860	1225	1307	1106	4498
Duration (hrs)	14.68	5.40	1.45	0.99	22.52

English (5047 subjects)					
	Sentence	Word	CL	SL	Total
No. of distinct Samples	22	28	24	23	97
No. of audio clips	6224	14611	22186	17941	60962
Duration (hrs)	9.80	13.00	14.14	10.34	47.28

Baseline Solution: Classifier Design for Sentence Level



Baseline Solution: Evaluation



Performance of our Data Model

Accuracy	Precision	Recall
86.0%	90.0%	80.0%

The desired accuracy level for an automated solution is greater than 95% for efficient utilization of resources.

The Reach and Impact of the Solution

- An automated solution for assessing the reading levels in India will result in saving ~**1 million dollars** and **10 minutes** time per assessment.
- Help teachers to adopt **"Teaching at the Right Level"** (TaRL) approach.
- Same logic and models can be extended to **TaRL** inspired countries - Botswana, Côte d'Ivoire, Ghana, India, Kenya, Madagascar, Mozambique, Niger, Nigeria, Uganda, Zambia by generating relevant dataset. This will extend the reach to over **60 million children**.

The Opportunities

Computer assisted language learning

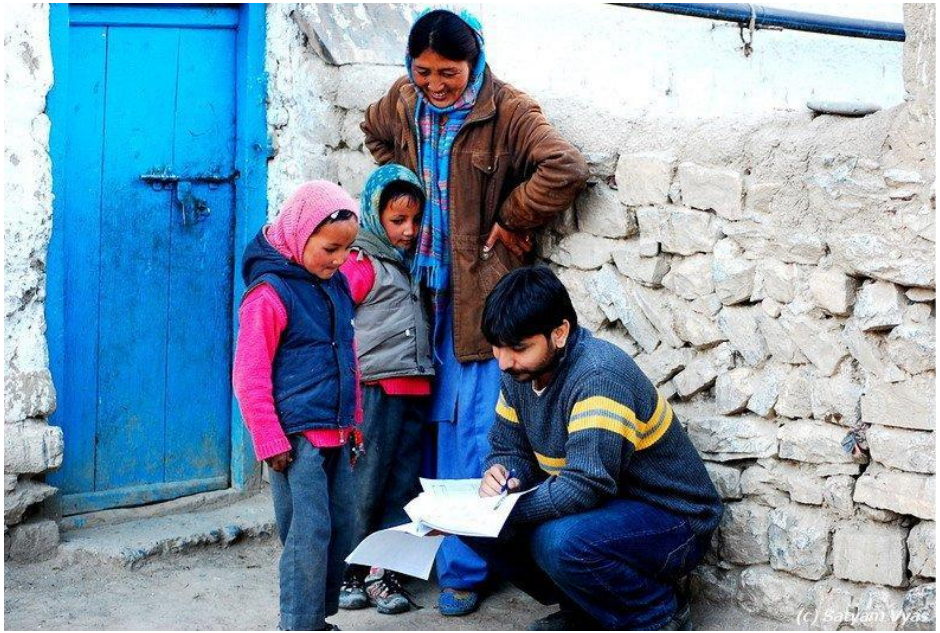
Foreign language learning

Improve ASR for children

Benefit the ecosystem of academia & researchers

Adaptive learning tools/products

Next Steps



- Collect half a million samples
- 11 regional languages
- Kaggle competition
- Innovative solutions


Thank You