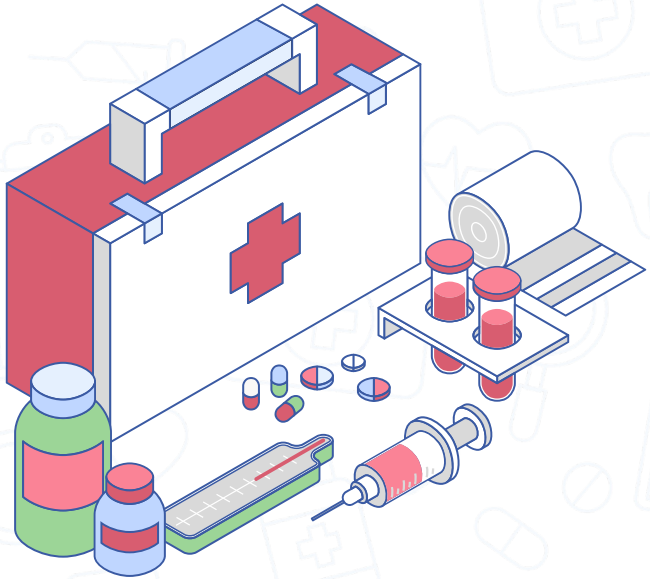


PARTURITION HINDI SPEECH DATASET FOR AUTOMATIC SPEECH RECOGNITION

VANSH BANSAL, THISHYAN RAJ, NAGARATHNA, SHUBAM KORDE, JASKARAN KALRA, SUDHA MURUGESAN, RAMKRISHNAN, ABOLI GORE, VIPUL ARORA

CONTENTS

01

INTRODUCTION

INTRODUCTION

- **Automatic Speech Recognition** has wide range of applications in varied sectors including the medical sector
- Traditionally ASR models were built using **Hidden Markov Models (HMM)** with **Gaussian Mixture Model (GMM)** based acoustic models and **Finite State Transducer (FST)** based language models
- **End-to-End (E2E)** models based on deep learning have provided superior results as compared to previous models
- It is desirable to have **domain specific datasets** for training ASR models for a specific task
- This paper presents the **Parturition Hindi Dataset (PHS)** which consists of the medical terms uttered by nurses in operation theatre

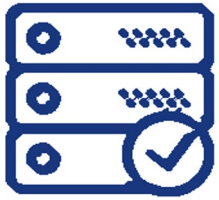
02

MOTIVATION

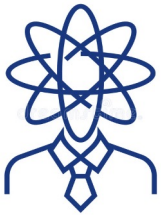
THE PROBLEMS

HEAVY PATIENT LOAD

LIMITED NURSES



COMPLICATED PROTOCOLS



TEDIOUS WORKFLOWS

AUTOMATIC SPEECH RECOGNITION

has a great potential to solve these problems allowing the nurses to automatically fill in data

RESEARCH CHALLENGES AND SOLUTION

HIGH ACCURACY REQUIREMENT

NOISY OPERATION ENVIRONMENT

CODE SWITCHING

LACK OF PROPER DATASETS

THEREFORE, CUSTOMIZED DOMAIN SPECIFIC DATASET IS NEEDED

PARTURITION HINDI DATASET

5.6

HOURS OF RECORDED DATA

2K

ANNOTATED AUDIO FILES

TRANSCRIPTIONS PROVIDED IN DEVANAGARI & ROMANIZED TRANSLITERATION

03

DATA COLLECTION

CORPUS SYNTHESIS

- PHS is a read-speech corpus, uttered inside a functioning maternity ward or operation theatre to emulate acoustics during parturition.
- Terms are sampled randomly from the set, ***T = {"BP", "Dilatation", "Discharge", "Drug", "Effacement", "FHR", "Pulse", "Temperature"}*** with an initial probability mass function ***$P_T(t)$ = [0.2, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.2]***
- Sample utterance: "बी पी एक सौ तेरह बटा इकसठ टेम्परेचर सैंतीस पॉइंट एक पल्स नवासी "
- Each sentence in the corpus is a sequence of three parameters randomly sampled according to the above distribution.
- The values of these parameters are also sampled from their specific probability distributions
- Numbers are sampled from Hindi as well as English numbers

CORPUS SYNTHESIS

Parameter	Range/Set	Prior
Blood Pressure	[80, 240]	Truncated Gaussian
Dilatation	[1, 10]	Truncated Gaussian then rounded to nearest integer
Discharge	{“रेड ”,“ब्राउन ”, “येल्लो” ,“वाइट”, “ ग्रीन”, “क्लियर"}	Uniformly sampled from the set
Drugs	{“टेबलेट मिसोप्रोस्टोल”, “इन्जेक्शन ऑक्सीटॉसिन”, “लिग्नोकेने”, “एंटीबायोटिक ”}	Customized probability mass function
Effacement	{“मोटा ”, “पतला ”}	Uniformly sampled from the set
Fetal Heart Rate	[80, 200]	Truncated Gaussian
Pulse	[50, 130]	Truncated Gaussian then rounded to nearest integer
Temperature	[35, 40]	Truncated Gamma distribution with location, shape, and scale parameters as 35, 4, and 0.55 respectively

AUDIO RECORDING ENVIRONMENT

- To emulate this environment accurately, the audio has been recorded in the labour rooms as well as triaging rooms of hospital facilities established in rural Bihar.
- They are asked to read the sentences, from the corpus described above, and record them into microphones while they **are not performing a delivery**
- Labour rooms have **a low signal-to-noise ratio (*)**, because of the cries of mothers and the cross-talks between nurses.
- So, the nurses are provided with **close-talking microphones** to record their speech.

(*) C. C. Chiu et al. "Speech recognition for medical conversations arXiv preprint arXiv:1711.07274, 2017.

ETHICS

TRANSPARENCY

FAIRNESS

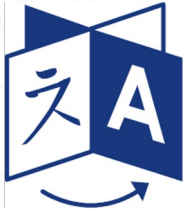
ACCURACY

PRIVACY

ACCOUNTABILITY

DATASET CHARACTERISTICS

This dataset consists of only a small set of words, totaling around 180.



LIMITED VOCABULARY

ACCENT

The nurses speak predominantly with a Bihar accent.

There are no restrictions as to whether the numbers are pronounced in Hindi or English.

BI-LINGUAL

MEDICAL SETUP

This dataset is created to have words in a medical setup.

AUDIO RECORDING ENVIRONMENT

- To emulate this environment accurately, the audio has been recorded in the labour rooms as well as triaging rooms of hospital facilities established in rural Bihar.
- They are asked to read the sentences, from the corpus described above, and record them into microphones while they **are not performing a delivery**
- Labour rooms have **a low signal-to-noise ratio (*)**, because of the cries of mothers and the cross-talks between nurses.
- To enhance the speech, the nurses are provided with **close-talking microphones** to record their speech.

(*) C. C. Chiu et al. "Speech recognition for medical conversations arXiv preprint arXiv:1711.07274, 2017.

DATA ANNOTATION

- There been a sufficient scope of natural variations such as **numerals, abbreviations, and word mispronunciations.**
- Therefore, a team of **25 annotators** re-transcribe the audio files manually.
- To ensure the correctness of the transcripts**, inter-annotator validation** has been performed.
- After training the model with the dataset, we feed in the audio to transcribe them using the ASR model. A comparison of the model-generated transcriptions with the ground truth helps in verifying the accuracy of the transcription.
- The data is collected with help of nurses from different backgrounds, with an **anonymized speaker index** with each audio file for possible use in **automatic speaker identification**

04

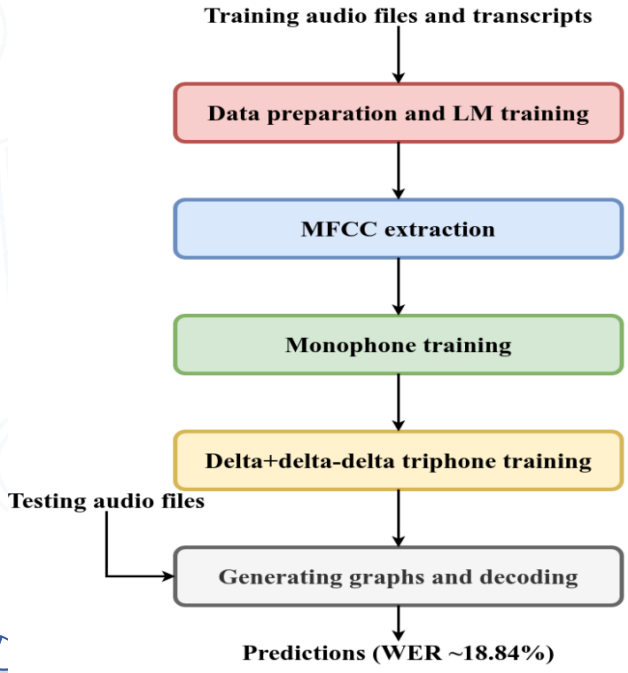
ASR MODELS

ASR MODELS

- The main intention of building this dataset is to evolve a model that can identify the readings spelt out by medical practitioners during intensive procedures.
- We use this data to train and evaluate two kinds of ASR systems – a traditional model and an E2E model.

Model	Word Error Rate (WER)
GMM-HMM (Devanagari Script)	18.84%
GMM-HMM (Romanized Transliteration)	21.77%
E2E (without language model)(Romanized Transliteration)	12.01%
E2E (with language model)(Romanized Transliteration)	2.7%

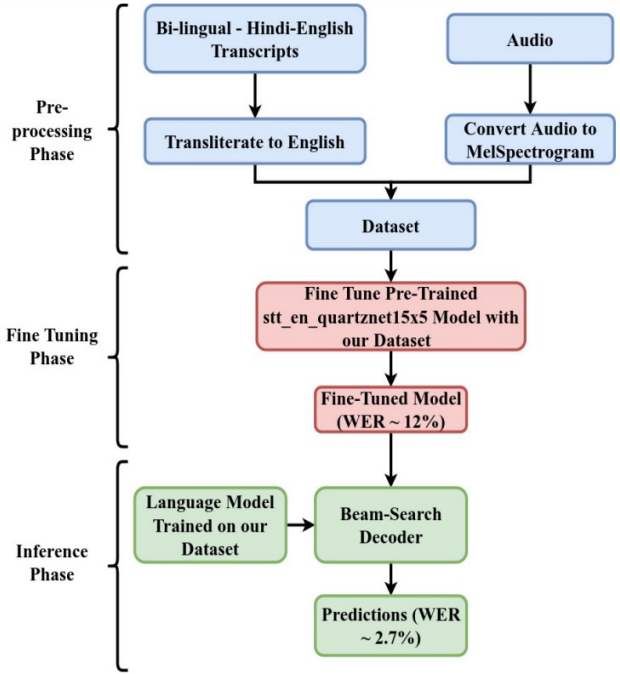
TRADITIONAL TRAINING WORKFLOW



- A Gaussian Mixture Model–Hidden Markov Model (GMM-HMM) based system built using the Kaldi toolkit, is used as the traditional model.
- We use GMMs to model the phoneme probability densities, which the HMMs use as the emission probability densities.

Ground Truth	Predictions
डिस्चार्ज रेड एफेसमेन्ट मोटा डिस्चार्ज रेड	डिस्चार्ज रेड एफेसमेन्ट मोटा डिस्चार्ज रेड
डिस्चार्ज ब्राउन टेम्परेचर सैंतीस पॉइंट चार बी पी एक सौ चालीस बटा छियासी	डिस्चार्ज ब्राउन टेम्परेचर सैंतीस पॉइंट चार बी पी एक सौ चालीस बटा छियासी
टेम्परेचर *** सत्तानवे एफ एच आर एक सौ सैंतीस बी पी एक सौ बीस बटा एक सौ नौ	टेम्परेचर सोलह सत्तानवे एफ एच आर एक सौ सैंतीस *** पी एक सौ बीस बटा एक सौ ***
बी पी वन फोर्टी सैवन बटे हंड्रेड नाइन एफेस-मेन्ट मोटा टेम्परेचर नाइंटी सैवन पॉइंट सिक्स	बी *** पल्स फोर्टी सैवन बटे हंड्रेड नाइन एफेसमेन्ट मोटा टेम्परेचर नाइंटी सैवन *** पल्स

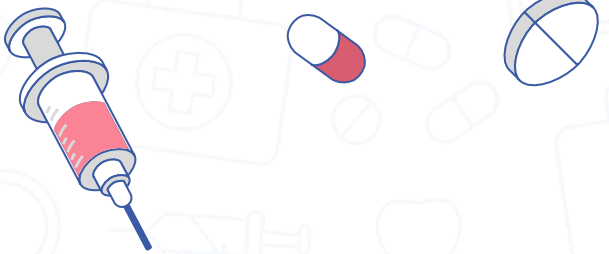
E2E TRAINING WORKFLOW



Ground Truth	Predictions
b p one hundred forty nine by one hundred four effacement patla temperature ninety five point nine	b p one hundred forty nine by one hundred four effacement patla temperature ninety five point nine
temperature thirty seven temperature ninety six point five discharge red	temperature thirty seven temperature ninety six point five discharge red
temperature ninety seven point five discharge green pulse sixty	temperature ninety seven point five discharge green pulse sixty
temperature chhattees point zero temperature untaalees point do temperature saintees point one	temperature chhattees point zero temperature untaalees point saat temperature saintees point one

The **NeMo toolkit** is used to build the E2E ASR model
During inference, we add a language model **(KenLM model*)** trained on our application domain to improve prediction accuracy.

* K. Heafield, "KenLM: Faster and smaller language model queries," In Proceedings workshop on statistical machine translation, pp. 187-197. 2011.



05

CONCLUSION

CONCLUSION

- This paper presents a complete method for developing **automatic speech recognition (ASR)** solutions for limited vocabulary tasks, especially for local languages.
- This includes dataset preparation as well as developing ASR models. In particular, the paper offers a new Hindi dataset for **healthcare applications**.
- The experimental results validate the efficacy of the presented method.
- In future work, we plan to employ the presented method to prepare more datasets from **regional dialects** and other regional languages.
- We also plan to use the PHS dataset for diverse applications, such as to develop models which can detect **out-of-vocabulary words**.
- We will also develop models that can detect **half-spelled words**.

DEMO

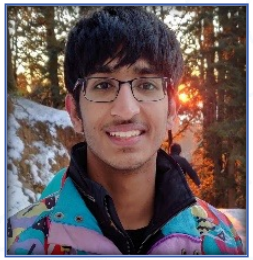
Demo of Digital Parturition Assistant

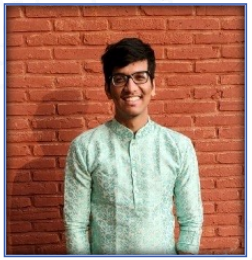
https://youtu.be/qNk5xuop27Y

AUTHORS

IIT Kanpur


Prof. Vipul Arora


Jaskaran Kalra


Shubham Korde


Vansh Bansal


Thishyan Raj


Nagarathna

Care India


Dr. Aboli Gore


B Ramakrishnan


Sudha Murugesan

Thanks to **Google AI for Social Good** Grant for supporting this work