Parturition Hindi Speech Dataset for Automatic Speech Recognition

Vansh Bansal*, Thishyan Raj T†, Nagarathna Ravi‡, Shubham Korde§, Jaskaran Kalra¶, Sudha Murugesan‖, Ramkrishnan B**, Aboli Gore†† and Vipul Arora‡‡
*Department of Computer Science and Engineering, IIT Kanpur, India
†‡§¶ ‡‡Department of Electrical Engineering, IIT Kanpur, India
‖**††CARE India, India
Email: *vansh2002bansal@gmail.com, †thishyan20@iitk.ac.in, ‡rathna@iitk.ac.in, §skorde955@gmail.com, ¶mkalra2615@gmail.com, ‖smurugesan@careindia.org, **brkrishnan@careindia.org, ††aboli@careindia.org, ‡‡vipular@iitk.ac.in

***Abstract*—While automatic speech recognition (ASR) technologies have become mature, they are mostly being developed by industry for large scale commercial applications. There are many niche domains that can potentially benefit from ASR. These domains may need ASR for a specific limited vocabulary for a certain population with a distinct language and dialect. This paper details the complete procedure for developing an ASR solution for such an application. It presents Parturition Hindi Speech (PHS) dataset prepared for real-time ASR for a medical application in Bihar, India. The dataset is prepared for childbirth assistance with recordings done by nurses in situ. Finally, several ASR systems are developed for the PHS dataset and their performances are compared. The models are pre-trained on large datasets and are adapted to PHS dataset. In experiments, we find that end-to-end ASR models adapt more effectively as compared to GMM-HMM based models. Moreover, custom language models further boost the performance.**

***Index Terms*—Automatic speech recognition, speech dataset, limited vocabulary, digital healthcare**

I. Introduction

Automatic speech recognition (ASR) has been a problem of interest for many decades. Starting from the principles developed by linguists and phoneticians, empirical success has been brought about by computational models such as Hidden Markov models (HMM) [1]. Traditionally, HMMs consist of Gaussian Mixture Model (GMM) based acoustic models and a Finite State Transducer (FST) based language models [2]. With the rise of deep learning, hybrid Deep Neural Network (DNN)-HMM based models bring further performance improvements. Recently, End-to-End (E2E) deep learning-based models [3]–[5] have shown superior performance to the earlier models.

However, all the above technologies need a good amount of annotated speech data for training the models. Large datasets are commercially available. But speech technologies are expanding to different domains and ought to be used by diverse groups. They are finding a great value in the medical domain [6] [7] [8]. ASR technologies have a great potential for documentation and human-machine interaction, especially in places where hands-free mode is desirable, e.g., inside an operation theatre where the hands are busy. However, speech datasets in the medical domain are meager.

Likewise, it is desirable to have datasets for diverse languages. Different languages are likely to have different phonetic structures. There have been ongoing efforts to generate language resources for Indian languages and their dialects [9].

To contribute to the above two directions, this paper presents Parturition Hindi Speech (PHS) dataset. It consists of medical data uttered by nurses inside an operation theatre. The dataset is available at https://github.com/madhavlab/2023_NCC_parturitionASR.

The PHS dataset consists of 2000 annotated audio files with a length of approximately 10 sec each. This totals 5.6 hours of audio. We provide the transcriptions in two formats: one is in Devanagri script and the other is the Romanized transliteration.

The rest of the paper is organized as follows. Section II presents the motivation for building this dataset and the associated ASR models. Section III provides an overview of the available open-source datasets and highlights their domain-specific differences. We elaborate on the data collection procedure in Section IV. Section V highlights the properties of the PHS dataset. We describe our baselines, namely, GMM-HMM and E2E ASR models, built on the PHS dataset, along with the evaluation results, in section VI. Finally, we conclude the paper with pointers to further research.

II. Motivation

The PHS dataset is prepared with an aim to improve intrapartum care in hospitals in rural Bihar. During delivery, the nurses note down various parameters to track the progress of delivery and to ensure safe delivery. It is immensely tedious for the nurses to record these details

This project has been funded by Google AI for social good award 2021.
†‡Equal Contribution

manually. In delivery wards, the nurses are often busy with multiple patients, and their hands are covered with blood. Hence, there is a need for a hands-free solution [10]. Here, ASR has great potential. ASR model can transcribe the speech of the nurses and automatically fill in the observations in the record sheet of the patients.

Several open-source ASR datasets provide hours of labeled Hindi speech recordings [11], [13]–[15]. There are two main limitations to them. First, they do not contain medical vocabulary of interest. Second, the accent of Hindi speakers in Bihar is different from the accents of speakers in those datasets. Moreover, the hospital environment is usually noisy. Particularly, the maternity wards have noises from other nurses and patients. We also observe significant code-switching, i.e., the nurses speak in both English and Hindi. It is impractical to set speech protocols in such tense environments.

Another reason for building a customized dataset is the critical nature of the application. The medical records transcribed by the ASR technology will be used for decision-making. Hence, it is imperative to take all measures to ensure high-fidelity transcription. Using a customized dataset is one such measure.

III. Related Works

We give highlights of the available open-source Hindi datasets in this section. We also briefly describe how the existing datasets are different from our dataset.

Mozilla Common Voice Hindi dataset [11] contains 17 hours of audio recordings and their transcripts. This was created with the help of volunteers who read the requested phrases [12].

ULCA-ASR-data-corpus [13] contains 2398.76 hours of labeled Hindi data and 2432.92 hours of unlabelled data. It contains speeches recorded from various platforms like Swamyam Prabha, News on Air, etc.

Open Speech and Language Resources platform has published Hindi datasets for MUCS 2021 challenge and 1111 Hours Hindi ASR Challenge [14] [15]. The MUCS 2021 challenge dataset contains approximately 100 hours of audio and their transcripts are taken from Hindi stories. The 1111 Hours Hindi ASR challenge dataset contains 105 hours of labeled speech data and 1000 hours of unlabelled speech data. This data was scraped from telephone recordings.

Despite several hours of open-source Hindi data already being available, the task at hand presents different challenges. Hence, we need a custom-built dataset for our application. Further, in this paper, we also provide the details of how the dataset has been built. The approach could be implemented and evolved by research and industry personnel to generate other balanced domain-specific datasets, as required for their applications.

IV. Data Collection

A. Corpus Synthesis

PHS is a read-speech corpus, uttered inside a functioning maternity ward or operation theatre to emulate acoustics during parturition.

A text corpus is synthesized first. Since the goal is to facilitate the nurses record various parameters to track the labour progress, we sample terms $\mathcal{T}$ randomly from the set, $\mathcal{T}$ ={"BP", "Dilatation", "Discharge", "Drug", "Effacement", "FHR", "Pulse", "Temperature"}. These terms are frequently used as per the medical protocols. Taking into account the importance and frequency of occurrence of the above terms, we define a probability mass function over them and sample the terms as $t \sim P_T(t)$. The probability mass function we used for the above parameters is $P_T(t) = [0.2, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.2]$ for $t \in \mathcal{T}$, respectively. 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. A typical utterance is "बी पी एक सौ तेरह बटा इकसठ टेम्परेचर सैंतीस पॉइंट एक पल्स नवासी ". In a sentence, we allow the same parameter to be repeated, e.g., "एफ एच आर वन हंड्रेड टेन एफ एच आर वन हंड्रेड बीस बी पी वन हंड्रेड नाइन बाए सिक्सटी नाइन ".

We define the parameters and the range and format of the values they can take as follows:

- **"BP"**: Blood Pressure of the mother
  The systolic and diastolic BP, measured in mm Hg, can take integral values and are separated by a "/", pronounced as **बटा, बटे, बाय** . The systolic BP is sampled from a truncated Gaussian distribution with mean 120 and standard deviation 25, i.e., $x \sim \mathcal{N}(x; 120, 25)$, truncated such that $x \in [80, 240]$. The diastolic BP is sampled from $x \sim \mathcal{N}(x; 80, 25)$, with $x \in [40, 140]$. Both the BP values are then rounded off to the nearest integer [20]. A typical occurrence is "बी पी नब्बे बटा सत्तर"
- **"Dilatation"**: Cervical dilatation of the mother
  Dilatation is again allowed to take integral values from the range $[1, 10]$, and is sampled from truncated $\mathcal{N}(6, 2)$, following which it is rounded off to the nearest integer. The unit centimetre is appended after this.
  A typical occurrence could be "डाइलेटेशन सात सेंटीमीटर "
- **"Discharge"**: Colour of the vaginal fluids discharged by the mother
  Discharge colour could be one of {"रेड ","ब्राउन ", " येल्लो " ,"वाइट", " ग्रीन", "क्लियर''} and is sampled uniformly from the above set.
  A typical occurrence is "डिस्चार्ज येल्लो".
- **"Drug"**: Drugs given to the mother before delivery
  The drug could be one of {"टेबलेट मिसोप्रोस्टोल ", "इन्जेक्शन ऑक्सीटॉसिन ", "लिग्नोकेने ", "एंटीबायोटिक "} and is sampled

from a probability mass function given by *[0.35, 0.35, 0.2, 0.1]* (respectively).
A typical occurrence could be "ड्रग इन्जेक्शन ऑक्सीटॉसिन".

- **"Effacement"**: Cervical effacement of the mother
In lieu of percentage, nurses usually report effacement as "मोटा " or "पतला ". We uniformly sample effacement from the set {"मोटा ", "पतला "}.
A typical occurrence could look like: "एफेसमेन्ट मोटा ".
- **"FHR"**: Fetal Heart Rate
FHR can take integral values from the range $[80, 200]$, and is sampled from truncated $\mathcal{N}(130, 25)$ following which it is rounded off to the nearest integer [21].
A typical occurrence could look like: "एफ एच आर एक सौ बीस ".
- **"Pulse"**: Pulse rate of the mother
Pulse can take integral values from the range $[50, 130]$, and is sampled from $\mathcal{N}(80, 25)$ following which it is rounded off to the nearest integer.
A typical occurrence is "पल्स एक सौ बीस ".
- **"Temperature"**: A body temperature of the mother
Temperature is a decimal number (rounded up to one decimal digit) and could be in $^\circ C$ or $^\circ F$. The temperature in $^\circ C$ is allowed to take values in $[35, 40]$ and is sampled from a truncated Gamma distribution with location, shape, and scale parameters as 35, 4, and 0.55 respectively, i.e $Gamma(35, 4, 0.55)$. On the other hand, temperature in $^\circ F$ is sampled from $Gamma(95, 4, 0.55)$ [22].
A typical occurrence could look like: "टेम्परेचर अट्ठानबे दशमलव एक" or "टेम्परेचर थर्टी फाइव पॉइंट सेवन".

B. 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. Nurses have been recruited with their prior consent. 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 [7] , 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. Ethics of Data Collection

The data has been collected with due approval from the Institute Ethics committee. Approved consent forms printed in both Hindi and English languages have been signed by the participating nurses. We ensure that the following ethics have been practised during data collection:

- **Transparency**: The nurses have been well informed about the purpose of data collection, and the intended uses of the data, i.e., training ASR systems for medical assistance.
- **Fairness and Diversity**: The data is collected by various nurses at different hospital facilities, both at district and block level facilities, irrespective of their socio-economic backgrounds.
- **Accuracy**: The data collection process takes into account various factors that could affect its quality. While simulating training data and recording the audio, we focus on the relative frequency of usage of particular medical terms, the range, and distribution of the parameters, the environment where audios are recorded, etc. The recording instructions do not put any conditions, other than the customary medical protocols, on the nurses concerning language and format of numerals and abbreviations. Even common mispronunciations are accounted for.
- **Privacy**: Nurses do not speak anything other than the artificially generated data, which does not contain any personal information.

D. Data Annotation

Although it is a read-out speech, there has been a sufficient scope of natural variations such as numerals, abbreviations, and word mispronunciations. A team of 25 annotators re-transcribe the audio files manually. To ensure the correctness of the transcripts, inter-annotator validation has been performed.

Another level of validation is performed with the help of the ASR model. 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 samples with any discrepancy between the ground truth and predictions are verified by the annotators again.

Since the data is collected with help of nurses from different backgrounds, we have added an anonymized speaker index with each audio file for possible use in automatic speaker identification.

V. Properties of Dataset

A. Limited Vocabulary

This dataset consists of only a small set of words, totalling around 180. Eight of them are the parameters used in the medical protocols, namely, {बी पी, डाइलेटेशन, डिस्चार्ज, ड्रग, एफेसमेन्ट, एफ एच आर, पल्स, टेम्परेचर}. There are words, such as बटा, बटे, बाय, पॉइंट, दशमलव, पतला, मोटा, क्लीयर, पीला, ग्रीन, ब्राउन, येलो, रेड, वाईट, which are non-numerical terms. The rest all are numbers spoken in Hindi and English.

B. Accent

The nurses speak predominantly with a Bihar accent. The pronunciation varies across nurses. Hence, we made sure to collect audio recordings from different nurses to help the model learn the variations in speech.

C. Medical Terms

This dataset is created to have words in a medical setup. So the words in the vocabulary are not commonly used in day-to-day speech. These terms are specific to the medical domain, thus training the model in the relevant application domain becomes convenient.

D. Bi-lingual

There are no restrictions as to whether the numbers are pronounced in Hindi or English. This is because each person will have their individual preference to pronounce numbers in either of the languages. It would be biased and unrealistic to restrict the nurses by settings rules or protocols, impacting the effectiveness of real-world deployments.

VI. Baseline ASR Methods

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. Hence, we cannot expect them to follow a protocol in spelling out the readings. For instance, nurses may tend to utter the reading 97.1 in one of the following ways - {"ninety seven point one", "सत्तानबे point one", "सत्तानबे दशमालव एक", "सत्तानबे point एक ", "nine seven point one", "नौ सा point एक", "नौ सात दशमालव एक", "नौ सात point one"}. The nurses work in a rural setup, where the accents are different from Hindi spoken by the vast majority of people. Also, the nurses make pronunciation errors while spelling some of the words. For example, the term 'dilatation' is sometimes mispronounced as 'dilation', 'misoprostol' is mispronounced as 'misoprost', etc. So the model should be robust enough to capture these variations. To ensure this, we split the dataset randomly into train and validation sets.

The audio recordings are bilingual, with some of the utterances being made in Hindi or English. It would be infeasible to train a unified model to recognize both Hindi and English words, especially since the amount of available recordings is limited. Hence, our annotators prepare two different versions of the data set - the first containing only Hindi, with the English words transliterated to Hindi (Devanagari script), and the second containing only English (Romanized transliteration), with all the Hindi words transliterated to English.

We use this data to train and evaluate two kinds of ASR systems – a traditional model and an E2E model. Table I gives the Word Error Rate (WER) of the baseline models.

Figure 1 gives an overview of the traditional training model. We use a Gaussian Mixture Model–Hidden Markov Model (GMM-HMM) based system built using the Kaldi toolkit, as the traditional model [2]. We use GMMs to model the phoneme probability densities, which the HMMs use as the emission probability densities. We then utilize these density values and the state transition probabilities of the HMMs to infer the state transitions using the Viterbi algorithm and obtain the most probable state sequence. To decode utterances while evaluating, we utilize more statistical information that is available from the data. This comprises context information of phonemes, a pronunciation lexicon, and the n-gram model.

Table I: WER (%)

Model	**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%

Table II: Sample Ground-truth and Predictions - GMM-HMM ASR - Hindi-only Dataset

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 एफेसमेन्ट मोटा टेम्परेचर नाइंटी सैवन पॉइंट सिक्स	बी *** पल्स फोर्टी सैवन बटे हंड्रेड नाइन एफेसमेन्ट मोटा टेम्परेचर नाइंटी सैवन *** पल्स

The Hindi lexicon is built by splitting the individual characters of every word using a space. The transliterated lexicon is built by transliterating Hindi words to English and then using CMUs online dictionary [17] to convert these words into their ARPABET representations [18]. These are all individually built using Weighted Finite State Transducers (WFSTs) and then compiled together using the composition, minimization, and determinization operations. We use the joint WFST to decode and obtain the transcriptions for the evaluation data. We do this for the Devanagari data as well as the Romanized data. Table II gives sample ground truth and predictions, that we see when the model is trained using Hindi transcripts. Table III shows the ground truth and predictions of the model when trained with Romanized transliterated transcripts. In tables II and III, '***' in the ground truth indicates a word was inserted in that position by the model in the predictions. '***' in the prediction indicates that a word that was there in the ground truth is missing in the prediction. The best WER achieved by the tri-phone model trained in Hindi is **18.84%**. The best WER achieved by the tri-phone model trained with Romanized data is **21.77%**.

We use the NeMo toolkit to build the E2E ASR model [5]. Figure 2 gives an overview of the training workflow. Initially, we process the audio files and their transcripts to convert them into the format expected by the model. Mel-Spectrograms are extracted out of the audio files

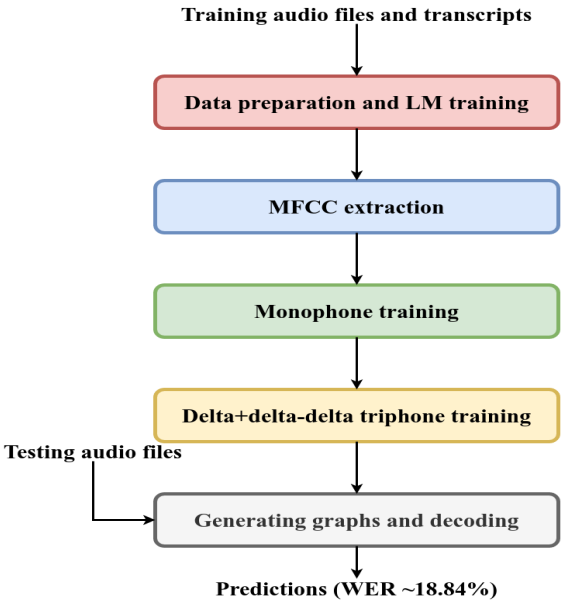

Figure 1: Traditional Training Workflow

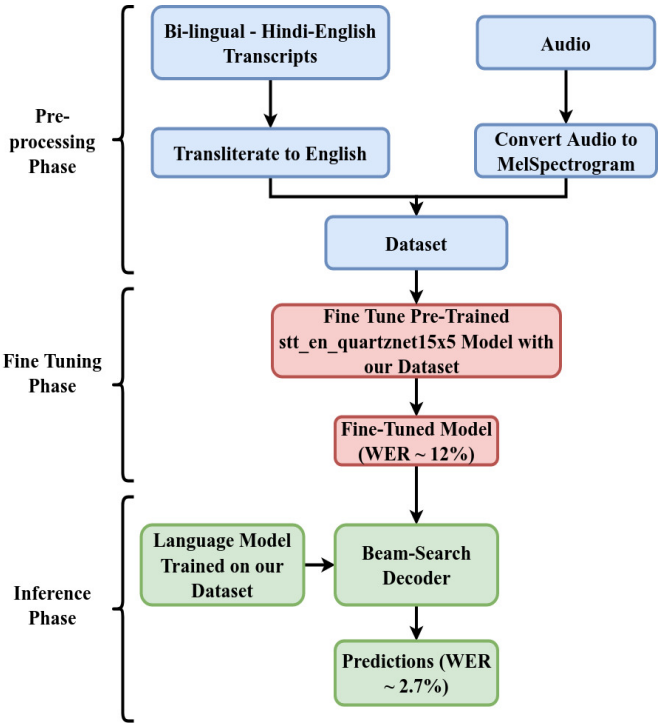

Figure 2: E2E ASR Training Workflow

Table III: Sample Ground-truth and Predictions - GMM-HMM - Romanized Transliterated Dataset

Ground Truth	Predictions
DRUG TABLET MISOPROSTOL PULSE EIGHTY FIVE B P ONE HUNDRED SIXTEEN BY SEVENTY	DRUG TABLET MISOPROSTOL PULSE EIGHTY FIVE B P ONE HUNDRED SIXTEEN BY SEVENTY
PULSE EK SAU SATRAH PULSE THIRAANAVE B P EK SAU PAINTEES BATA CHHIYAALEES	PULSE EK SAU SATRAH PULSE THIRAANAVE B P EK SAU PAINTEES BATA CHHIYAALEES
TEMPERATURE NINETY SEVEN POINT TWO DISCHARGE WHITE F H R ONE HUNDRED *** NINETEEN	TEMPERATURE NINETY SEVEN POINT TWO DISCHARGE WHITE F H R ONE HUNDRED NINE NINETY
TEMPERATURE NINETY EIGHT POINT EIGHT TEMPERATURE THIRTY SIX POINT FOUR TEMPERATURE NINETY NINE POINT FIVE	TEMPERATURE NINETY POINT P EIGHT TEMPERATURE THIRTY SIX POINT FOUR TEMPERATURE NINETY NINE POINT FIVE
TEMPERATURE THIRTY FIVE POINT SEVEN B P EK SAU SOLAH BY EK SAU PANDRAH B P EK SAU PAINTEES BY ATHATTAR	TEMPERATURE THIRTEEN FIVE POINT SEVEN B P EK SAU SOLAH BY EK SAU PANDRAAH B P EK SAU PAINTEES BY ***

(with audio files down-sampled to 16kHz). We use the Romanized transliterated transcripts to train the model. Since the size of our training data set is small, we prefer to use a pre-trained network that is trained on thousands of hours of ASR data to increase the model convergence rate. We use the stt_en_quartznet15x5 model as the pre-trained model [19], which is trained using 7,057 hours of English speech. We freeze the encoder layers and train only the final layer to adapt the ASR model to the domain of interest. With this, the model exhibits **12.01%** WER. During inference, we add a language model (KenLM model [23]) trained on our application domain to improve prediction accuracy. We use a beam search decoder along with the trained language model to obtain the transcriptions. This dropped the WER to **2.7%**. We give sample ground truth and predictions of the E2E ASR model in Table IV.

VII. 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

Table IV: Sample Ground-truth and Predictions - E2E ASR - Romanized Transliterated Dataset

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

out-of-vocabulary words. We will also develop models that can detect half-spelled words.

Acknowledgment

The authors acknowledge the help of all the stakeholders of the project and those who helped in data collection and annotation.

References

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[2] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and J. Silovsky, "The Kaldi speech recognition toolkit," in IEEE 2011 workshop on automatic speech recognition and understanding, IEEE Signal Processing Society, 2011.

[3] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Davison, "Huggingface's transformers: State-of-the-art natural language processing," arXiv preprint arXiv:1910.03771, 2019.

[4] A. Graves, F. Santiago, G. Faustino, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd international conference on Machine learning, pp. 369-376. 2006.

[5] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, and P. Castonguay, "Nemo: a toolkit for building AI applications using neural modules," arXiv preprint arXiv:1909.09577, 2019.

[6] A. Schulte, R. Suarez-Ibarrola, D. Wegen, P. F. Pohlmann, E. Petersen, and A. Miernik, "Automatic speech recognition in the operating room–An essential contemporary tool or a redundant gadget? A survey evaluation among physicians in form of a qualitative study," Annals of Medicine and Surgery, vol. 59, pp. 81-85, 2020.

[7] C. C. Chiu, A. Tripathi, K. Chou, C. Co, N. Jaitly, D. Jaunzeikare, A. Kannan, P. Nguyen, H. Sak, A. Sankar, J. Tansuwan, N. Wan, Y. Wu, X. Zhang, "Speech recognition for medical conversations," arXiv preprint arXiv:1711.07274, 2017.

[8] G. P. Finley, E. Edwards, A. Robinson, N. Sadoughi, J. Fone, M. Miller, D. Suendermann-Oeft, M. Brenndoerfer, and N. Axtmann, "An Automated Assistant for Medical Scribes," In INTERSPEECH, pp. 3212-3213, 2018.

[9] https://bhashini.gov.in/en

[10] A. Rahman, T. Begum, F. Ashraf, S. Akhter, D. M. E. Hoque, T. K. Ghosh, M. Rahman, J. Stekelenburg, S. K. Das, P. Fatima, and I. Anwar, "Feasibility and effectiveness of electronic vs. paper partograph on improving birth outcomes: A prospective crossover study design," PloS one, vol. 14, no. 10, p.e0222314, 2019.

[11] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," arXiv preprint arXiv:1912.06670, 2019.

[12] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," arXiv preprint arXiv:1804.03209, 2018.

[13] "Country Report — India," 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), 2021, pp. 1-6.

[14] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, and A. Mittal, "Multilingual and code-switching ASR challenges for low resource Indian languages," arXiv preprint arXiv:2104.00235, 2021.

[15] https://www.openslr.org/118/

[16] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," in Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2, pp. 957-964, 2000.

[17] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[18] https://en.wikipedia.org/wiki/ARPABET

[19] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_quartznet15x5

[20] https://utswmed.org/medblog/fetal-heart-rate-monitor/

[21] J. Cohen, D. Vaiman, B.M. Sibai, B. Haddad. Blood pressure changes during the first stage of labor and for the prediction of early postpartum preeclampsia: a prospective study. Eur J Obstet Gynecol Reprod Biol., 184:103-7, 2015.

[22] E. Ashwal, L. Salman, Y. Tzur, A. Aviram, T. B.-M. Bashi, Y. Yogev, and L. Hiersch. "Intrapartum fever and the risk for perinatal complications–the effect of fever duration and positive cultures." The Journal of Maternal-Fetal & Neonatal Medicine, vol. 31, no. 11, pp. 1418-1425, 2018.

[23] K. Heafield, "KenLM: Faster and smaller language model queries," In Proceedings of the sixth workshop on statistical machine translation, pp. 187-197. 2011.