Towards Inclusive Digital Health: An Architecture to Extract Health Information from Patients with Low-Resource Language

Prajat Paul[a], Mohamed Mehfoud Bouh[b] and Ashir Ahmed[c]
Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan

Keywords: Digital Health, Automatic Speech Recognition (ASR), Low Resource Language (LRL), Bangla, Health Data Extraction, Electronic Health Record (EHR).

Abstract: Collection of health information from the underserved community has been a challenge. Their health records are not digitized. The major population of the underserved community is text-illiterate but is not voice-illiterate. This article proposes a speech-based healthcare information collection system as an additional module to the traditional EHR system. Bangla is a language spoken widely across Bangladesh and Western parts of India by 210 million people, but it is still one of the LRLs when it comes to ASR resources. The existing research outcomes indicate the necessity of application-specific language resources for better performance. In addition, a system architecture for collecting speech data from doctor-patient conversations and an automated information retrieval system in the local language are put forward. The system also extends to extracting information that can provide assistance in operations like prescription prediction and creating new health records in digital medical history management systems.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Speech-to-text recognition technologies and their relevant application along with Machine Learning models and Artificial Intelligence has become significantly prominent in the last decade. Concurrently, research focusing on the development of technology to improve the current situation of digital healthcare services for the masses has gained a significant amount of attention. Generalizing EHR systems for the accumulation of health-related data and digitization has been introduced to several systems with dynamic measures (Hossain et al., 2022)(Ahmed et al., 2013). However, the form-based data collection process for these systems is mostly centered around high-resource languages (HRL). This dependency makes the accessibility of the technology difficult for countries that have one of the LRLs as their primary medium of communication and do not have an HRL as their second language. Languages that have not been studied extensively from the perspective of digital resources, significantly lack resources for research and development, and are less commonly used in speech-to-text technology compared to major languages are denoted as LRLs (Magueresse et al., 2020). In developing countries, the lack of adequate literacy tends to train the people to use technology based on muscle memory rather than understanding the operations because most of the application interfaces are not available adequately in their regional language. Filling up complex medical forms on a digital screen would be a farfetched expectation and not enough medical personnel is available to assist the mass number of patients. In a situation of this sort, speech-based data collection methodologies may pose as a handy tool. If the process only involves pressing a record button and speaking into a mobile device, that action is simple enough to associate with the muscle memory-based handling of technology done by general people.

A population of 210 million people across Bangladesh and Some parts of India speak Bangla as their first or second language (The Editors of Encyclopedia Britannica, 2023). Despite that, when it comes to applicable digital resources, the lack is quite noticeable. This situation leads to Bangla being one of the LRLs when it comes to ASR infrastructure. Some of the key aspects that make this language difficult to work with are the fact that it has a total of 28 diverse accents in use and only a mere 38.8% of the total

[a] https://orcid.org/0009-0002-2243-6078
[b] https://orcid.org/0000-0002-7716-7007
[c] https://orcid.org/0000-0002-8125-471X

Paul, P., Bouh, M. and Ahmed, A.
Towards Inclusive Digital Health: An Architecture to Extract Health Information from Patients with Low-Resource Language.
DOI: 10.5220/0012471500003657
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2024) - Volume 2*, pages 754-760
ISBN: 978-989-758-688-0; ISSN: 2184-4305
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

Table 1: Speech Recognition Tools and supported LRLs and HRLs. Source was respective websites and documentation.

Speech Recognition Tool	**Supported HRL**	**Supported LRL (with insufficient resources & less usability)**
Google Cloud Speech-to-Text	Arabic, Danish, German, Greek, English, Spanish, Finnish, French, Hebrew, Japanese, Mandarin, Korean, Dutch, Russian, Italian, Hindi	Bengali, Burmese, Catalan, Hungarian, Kannada, Kazakh, Malay, Malayalam, Marathi, Nepali, Burmese, Nepali, Punjabi, Somali, Urdu, Vietnamese, Swahili, Zulu
IBM Watson Speech-to-Text	Arabic, Mandarin, Dutch, English Japanese, Korean, Hindi	Portuguese, Czech, Swedish
Amazon Transcribe and Amazon Transcribe Medical	Arabic, Chinese, English in multiple accents, French, Japanese, Korean, Italian	Malay, Portuguese, Swedish, Tamil, Telegu, Thai, Turkish, Vietnamese
Wit.Ai	Arabic, Chinese, English, Dutch, Finnish, French, German, Hindi, Italian, Japanese, Spanish	Bengali, Indonesian, Kannada, Malay, Malayalam, Marathi, Polish, Portuguese, Sinhalese, Swedish, Tagalog, Tamil, Thai, Turkish, Urdu, Vietnamese
Microsoft Azure Speech Service	Arabic, Danish, German, Greek, English, Spanish, Finnish, French, Hebrew, Japanese, Mandarin, Korean, Dutch	Afrikaans, Bengali, Bosnian, Catalan, Czech, Welsh, Estonia, Persian, Filipino, Gujarati, Hungarian, Kannada, Malayalam, Mongolian, Marathi, Malay, Burmese, Nepali, Punjabi, Somali, Urdu, Vietnamese
Nuance Dragon	English, French, German, Japanese, Italian, Spanish and Dutch	None
iSpeech	English, Spanish, Mandarin, Japanese, Korean, Dutch, Italian, German, Russian, Arabic (Male)	Cantonese, Hungarian, Catalan, Czech, Polish, Swedish (Female)
Yandex SpeechKit	German, English, Spanish, Finnish, French, Hebrew, Italian, Dutch, Russian	Kazakh, Polish, Portuguese, Swedish, Turkish, Uzbek
Speechmatics	Arabic, Dutch, English, Finnish, French, German, Greek, Hindi, Italian, Japanese, Korean, Mandarin, Russian, Spanish	Bashkir, Basque, Bulgarian, Cantonese, Croatian, Czech, Hungarian, Latvian, Malay, Marathi, Norwegian, Tamil, Thai, Welsh

speakers use the conventional accent for communication (Alam et al., 2022). In addition, this language is equipped with highly complex inflectional morphology, phonetic complexity, cultural nuance, termination, and multifarious orthography (Bhattacharya et al., 2005). All of this leads to a hardship concerning relevant research and the development of useful resources for this language.

Another noticeable aspect of the linguistic community of this language is that they mostly belong to a part of the world where literacy and adequate healthcare facilities are not available to their full potential for the masses. A very high doctor-patient ratio and limitation of interaction time often affect the efficiency of the system and do not have a structured flow of operations. In such a situation, an efficient speech-based process of patient health data collection is less time-consuming and divides the burden of work from the healthcare personnel to all who are involved in the process, namely the patients and their helping hands.

This paper aims to propose the research aspects of developing a speech-based data collection system for the LRL, Bangla, and explore methodologies required in building and optimizing resources for an efficient ASR system in digital healthcare. In addition to that, listening to doctor-patient conversations or automated systems conducts guided conversations for health data accumulation while maintaining a high level of ac-

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

curacy. Finally, it delves into the utilization of the extracted medical information for operations such as assistive prescription prediction and the creation and update of health records of individual patient profiles in EHR systems.

The following parts of the article are organized as follows. Section 2 talks about the existing research done on LRLs and the Bangla language in developing ASR architectures for speech recognition along with their limitations. Section 3 consists of the motivating factors behind this research and the research questions that determine the key factors of this concept. Following that, the proposed architecture of the technology and its process of accumulating speech data for extraction of health information is discussed in Section 4. Section 5 depicts the necessary initiatives that can contribute to the improvement of speech data processing for our LRL in focus, the Bangla language.

2 EXISTING RESEARCH ON LRL SPEECH RECOGNITION AND THEIR LIMITATIONS

Research on diverse dynamics is noticed when it comes to both the implementation of ASR in digital healthcare and exploring ways of mitigating the shortcomings of LRLs. In the case of Bangla, ASR algorithms and corpora development have seen a significant amount of work done. A combination of fine-tuned deep learning algorithms and large language models (LLM) has shown performance that has been evaluated with various metrics of evaluation such as Word Error Rate (WER), Character Error Rate (CER), Levenshtein Distance Score (LDS), Accuracy, Precision and Recall.

The exploitation of labeled data from HRLs with the intent of improving the LRLs using Hidden Markov Models (HMM) (Schultz and Waibel, 2001) to advanced neural network systems (Tong et al., 2017)(Toshniwal et al., 2018)(Conneau et al., 2020) has been a field of research interest for quite some time. Khare *et al.* adapted a mapping process of scripts by transliterating HRL (English) resources to target LRL (six different languages), where the HRL and LRL are belonging to dissimilar language families (Khare et al., 2021). It was evaluated on wav2vec2.0 and transformer-based ASR architectures, resulting in a reduction of WER by 8.2%. One of the six LRLs was Bangla which had a WER of 88.9% for the wav2vec2.0 model. The wav2vec2.0 being a deep learning ASR architecture with the incorporation of the transformer architecture has made it quite a popular choice of algorithm in the recent works done on Bangla language. Training this self-supervised model on the Bengali Common Speech Dataset led to an LDS value of 6.234 and a WER of 0.2524 after running 71 epochs (Shahgir et al., 2022). A similar approach with the integration of post-processing using an n-gram language model and hyperparameter tuning achieved a CER of 1.54%, an LDS value of 1.65, and a WER of 4.66% (Rakib et al., 2023b). A modification of wav2vec2.0 using IndicWav2Vec designed by AI4Bharat had an LDS value of 3.819 (Showrav, 2022). Deb *et al.* combined a wav2vec2.0 ASR architecture with a Marian-NMT translation model for the Bangla language. This multimodal perspective resulted in a precision of 0.94 and a recall of 0.91 (Deb et al., 2023). Fine-tuned Convolutional Neural Network(CNN) -Recurrent Neural Network (RNN) based models have also shown decent performance (Islam et al., 2019)(Mandal et al., 2020).

Simultaneously, over the years of research on Bangla ASR, the development of language corpus has had progressive work done. A campaign using Bengali.AI with the motive of creating a collection of 5000 hours of audio data led to the collection of 400 hours of voice data (Alam et al., 2022). Kibria *et al.* used the RNN model to develop a language corpus of 299 hours of speech data with the contribution of 61 volunteers from all across Bangladesh (Kibria et al., 2022). A significantly large out-of-distribution (OOD) database, titled OOD-Speech, is claimed to be the largest open-source resource for benchmarking (Rakib et al., 2023a). It consists of crowd-sourced data in a controlled environment and accent diversity was maintained for better dynamics. Murtoza *et al.* made a collection of 977 sentences with coverage of 77.56% of bi-phones and 5.06% of tri-phones (Murtoza et al., 2011).

However, despite all the work that is done, the resources still lack in being efficient when it comes to domain-specific applications. Digital healthcare is a field relevant to medical operations. Use of medical terms, the conversation being multilingual and mispronunciation are some of the key difficulties faced in the case of LRLs that have not been addressed so far. ASR tools that are available commercially have a very limited range of supported languages. In most cases, LRLs are not prioritized, and even if they do have a language model for them, the results are not satisfactory. Table 1 displays some of the commonly used tools for speech recognition and their supported HRLs and LRLs. Even though these tools claim to support LRLs, the language models are not formidable

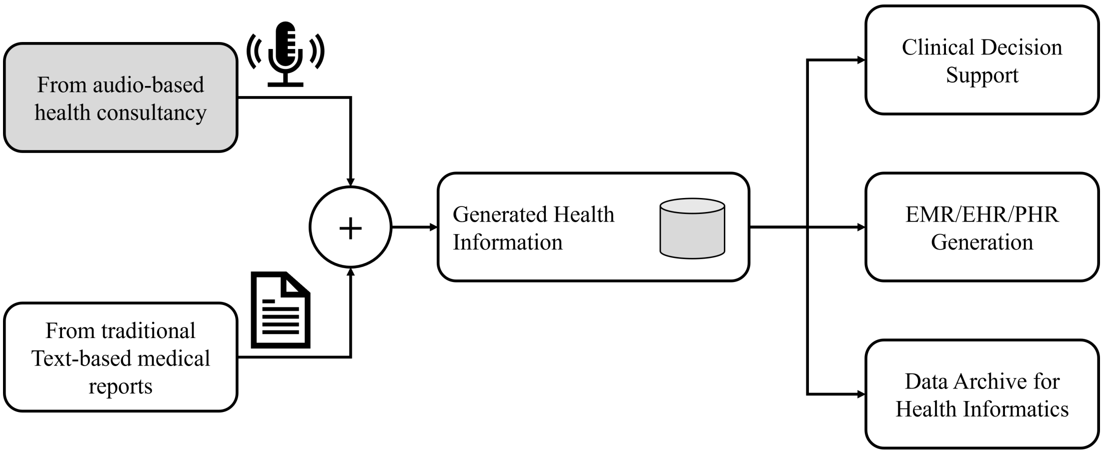
From audio-based health consultancy
From traditional Text-based medical reports
+
Generated Health Information
Clinical Decision Support
EMR/EHR/PHR Generation
Data Archive for Health Informatics

Figure 1: Overview of the system to generate and increase volume of health information.

enough for application. As for the Bangla language, the problems are very similar due to the complexity of the language along with the lack of application-guided development. A system that can transcribe a medically relevant conversation in Bangla consisting of one or multiple speakers with high precision and convert that transcription into usable information for EHR systems is yet to be developed. From table 1 Google Speech-to-text, Wit.Ai, and Microsoft Azure Speech Service have a claim of supporting Bangla. However, the lack of resources and semantics leads to a very poor quality of transcription.

3 MOTIVATING FACTORS AND THE RESEARCH QUESTIONS

This section outlines the motivating factors behind this study and itemizes the research questions.

3.1 Motivating Factors

The motivation for this research is sourced from the intention of creating a system for LRLs and ensuring the inclusiveness of its users in using speech-based technology for healthcare services. Figure 1 displays an overview of the health information collection methodology and the role of voice-based data collection in it. The following are the key aspects denoting the importance of this research:

- Primarily, Digital systems that are implemented in the healthcare domain have a text-based input system for convenience. It puts people who cannot read or type information due to illiteracy or some physical disability at an unavoidable disadvantage. EHR systems require a speech-based data collection system in LRLs, which in this case is Bangla. This will contribute immensely to the aspect of inclusive participation of the masses in the use of EHR systems.
- In addition, The efficient use of EHR systems requires a large amount of patient data. A majority of the population speaking Bangla unfortunately belongs to the underprivileged crowd. Given that the geographical region of this population is densely populated, the use of convenient systems like speech-based information extraction methodology can contribute greatly to EHR database enrichment.
- The conversation between a patient and a doctor usually leads to the formation of a chief complaint about a specific health event that denotes the patient's current health issue. It also includes basic inquiries from the doctor's end for a better understanding of the situation. Overall, the speech data consists of information that can create a new health record, update previous health events, or provide assistance to the doctor in the prescription generation process to provide more efficiency within the available time constraint.
- Automated systems with advanced voice synthesis models that can conduct guided question and answer sessions with the patient for the initial gathering of information, problem diagnosis, and filtering out important elements for the doctor's observation can, in turn, reduce the time required in individual patient assessment. This, along with the integration of patient medical history, can provide a structured version of the required medical particulars that medical personnel need to know to provide healthcare services.
- The Bangla language requires the application-domain-specific development of speech recognition resources. Even though there are corpora

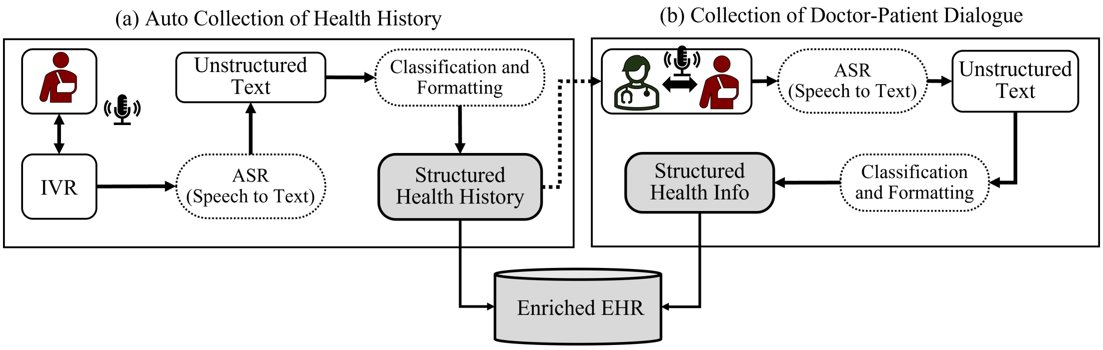
(a) Auto Collection of Health History
Unstructured Text
Classification and Formatting
IVR
ASR (Speech to Text)
Structured Health History
(b) Collection of Doctor-Patient Dialogue
ASR (Speech to Text)
Unstructured Text
Structured Health Info
Classification and Formatting
Enriched EHR

Figure 2: Extraction of Health Information from two audio sources: (a) Health history-taking from the patient-IVR conversation and (b) patient-doctor conversation.

available sourced from different media content, a health event-related conversation has medical terms and a mix of languages. This requires a dedicated corpus equipped with healthcare-related terminologies along with ASR models that can detect the use of words from other languages during a conversation.

3.2 Research Questions

Following are the Research Questions (RQ) that would guide this research in further studies and endeavors.
RQ 1: How can the health information of the population using LRLs be included in EHR systems?
RQ 2: How can the accuracy of Bangla ASR tools be improved and make it applicable for EHR generation?

Answers to these questions project the system that can facilitate the process of collecting health information for EHR systems from a patient who cannot operate text-based data collection systems and use LRL for communication. It also sheds light on the aspect of performance enhancement of existing ASR tools for the Bangla language for better accuracy of speech transcription to be applied for digital healthcare.

4 PROPOSED SYSTEM ARCHITECTURE & APPLICATIONS

Subsections 4.1 and 4.2 discuss the process of data extraction displayed in figure 2. The extracted information is classified and organized for integration with the patient health information database. Some of the multifarious applications of this retrieved information would be applicable in assisting in clinical decision-making, electronic medical record generation, and the enhancement of information archives for health informatics. Subsection 4.3 talks about the probable scopes of applications and their projected impact on the overall scenario.

4.1 Automated Information Retrieval System

The system architecture would involve the patient interacting with the system to log health data by vocally communicating with it. It is equipped with an Interactive Voice Response (IVR) module that generates Call Detail Records (CDR). The automated system can listen to the patient speaking in their local language and process the information to generate follow-up questions to determine the Chief Complaint (CC) and gather other necessary information. A voice synthesis system will also be asking the questions in the patient's local language. The generated CDRs can be transcribed into useable text data. With Filtering and analysis of the data, structured health information would be retrieved that can be added to the patient health history informatics along with assisting the doctor in the process of patient observation.

4.2 Health Info Extraction from Doctor-Patient Interaction

This step involves the doctor meeting the patient face-to-face for a direct consultation. The interaction is provided with gathered information by the previously mentioned automated system. It consists of relevant parts of health history and new lookouts for the current health event.

- Patient health data retrieved from the speech-based communication done with the system along with previous health information stored in the database, would have a summary of all the necessary information that can assist the doctor in further clinical decision-making. The conversation between the doctor and patient would also contain information related to diagnosis, key observations, the patient's current health condition, and medical advice from the doctor.

- The filtering and analysis module can go through the data and retrieve useful health information in a structured manner to contribute to the process of populating the integrated patient health info database.

4.3 Applications and Expected Impact

The IVR-guided automatic system of communicating with the patient is an additional module to the current healthcare-providing methodology. It adds a dynamic of gathering initial information from the patient that can populate the health history database, along with creating assistive notes for the doctor to work with. The categories of gathered information will be personal health history, family health history, currently administered drug history, and initial complaint.

The doctor-patient interaction usually consists of a conversation that includes detailed discussion and observation of symptoms, chief complaint generation, prescribed medication, lifestyle adjustments, and medical investigation tests. The system will filter and analyze this information from the speech data and format it in a structured manner to include it in the EHR system.

This framework of data collection would ultimately contribute to the development of a structured medical history of the individual patient that can be accessed for future operations and data visualization for efficient observation.

5 PERFORMANCE OF SPEECH TRANSCRIPTION

One of the fundamental components that has a major role to play in keeping the mentioned system of information accumulation operational is the ASR tool for the LRL, which in this case is Bangla. The transcription and translation accuracy would greatly affect the smooth conveyance of the information. The semantics of the speech data would be greatly required to understand the context of the information. Following are some of the initiatives that can improve the current situation of lack of resources.

- ASR error correction for medical conversations to mitigate the misinterpretation of medical terminologies has been observed for the English language (Mani et al., 2020). A similar approach can be adapted to detect the use of medical terms in a Bangla language conversation or speech and classify it correctly.
- Transformer-based acoustic models have displayed formidable performance in WER reduction for HRLs (Wang et al., 2021). Bangla language has a complex acoustic structure with multifarious di-phones and tri-phones. Implementation of such systems for Bangla language corpus designing can improve the overall quality of transcription.
- Use of Transcription and Translation to HRL (English) using Transformer Models such as Generative Pre-trained Transformer (GPT) or Bidirectional Encoder Representations from Transformers (BERT) can mitigate transcription error and keep the semantics of the speech somewhat unaffected. This results in a better quality of information retrieval.
- Use of wav2vec2.0 architecture has gained popularity in speech recognition due to its self-learning algorithm along with an integrated transformer architecture. Fine-tuning this system on systemically designed Bangla language corpora can result in higher accuracy.
- For the scenario of a doctor-patient conversation, the system will require a speaker recognition system to determine the context of the exchange of information.

6 CONCLUSION

This article introduces the use of a speech-based health information extraction system that can formidably support LRL as a medium of communication. The LRL in consideration here is the Bangla language. Despite having a decent-sized population primarily depending on this language for communication, the digital resources, and ASR tools are not sufficient to provide for the above-mentioned system. System structures and ASR development initiatives were proposed as a future prospect for this research, aiming for an efficient data collection system for digital health informatics along with the inclusion of the underserved community in the use of modern healthcare technology. The gathered information shows the scope of serving applications such as support in clinical decision-making, data archive enrichment, and electronic record management. Future endeavors would be focused on the prototyping and performance observation of the proposed IVR-guided

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

system in diverse scenarios of varying language complexity. In addition, a combination of ASR algorithms and generative AI will be evaluated and trained on the unstructured format of speech data from doctor-patient interactions. Moreover, data collection to mitigate the requirement for linguistic resources will be conducted with the inclusion of medical terminologies and their Bangla synonyms.

REFERENCES

Ahmed, A., Inoue, S., Kai, E., Nakashima, N., and Nohara, Y. (2013). Portable health clinic: A pervasive way to serve the unreached community for preventive healthcare. In *Distributed, Ambient, and Pervasive Interactions: First International Conference, DAPI 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013. Proceedings 1*, pages 265–274. Springer.

Alam, S., Sushmit, A., Abdullah, Z., Nakkhatra, S., Ansary, M., Hossen, S. M., Mehnaz, S. M., Reasat, T., and Humayun, A. I. (2022). Bengali common voice speech dataset for automatic speech recognition. *arXiv preprint arXiv:2206.14053*.

Bhattacharya, S., Choudhury, M., Sarkar, S., and Basu, A. (2005). Inflectional morphology synthesis for bengali noun, pronoun and verb systems. In *Proc. of the National Conference on Computer Processing of Bangla (NCCPB 05)*, pages 34–43.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Deb, A., Nag, S., Mahapatra, A., Chattopadhyay, S., Marik, A., Gayen, P. K., Sanyal, S., Banerjee, A., and Karmakar, S. (2023). Beats: Bengali speech acts recognition using multimodal attention fusion. *arXiv preprint arXiv:2306.02680*.

Hossain, F., Islam, R., Ahmed, M. T., and Ahmed, A. (2022). Technical requirements to design a personal medical history visualization tool for doctors. In *Proceedings of the 8th International Conference on Human Interaction and Emerging Technologies. IHIET, https://ihiet. org*.

Islam, J., Mubassira, M., Islam, M. R., and Das, A. K. (2019). A speech recognition system for bengali language using recurrent neural network. In *2019 IEEE 4th international conference on computer and communication systems (ICCCS)*, pages 73–76. IEEE.

Khare, S., Mittal, A. R., Diwan, A., Sarawagi, S., Jyothi, P., and Bharadwaj, S. (2021). Low resource asr: The surprising effectiveness of high resource transliteration. In *Interspeech*, pages 1529–1533.

Kibria, S., Samin, A. M., Kobir, M. H., Rahman, M. S., Selim, M. R., and Iqbal, M. Z. (2022). Bangladeshi bangla speech corpus for automatic speech recognition research. *Speech Communication*, 136:84–97.

Magueresse, A., Carles, V., and Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Mandal, S., Yadav, S., and Rai, A. (2020). End-to-end bengali speech recognition. *arXiv preprint arXiv:2009.09615*.

Mani, A., Palaskar, S., and Konam, S. (2020). Towards understanding asr error correction for medical conversations. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 7–11.

Murtoza, S., Alam, F., Sultana, R., Chowdhur, S., and Khan, M. (2011). Phonetically balanced bangla speech corpus. In *Proc. Conference on Human Language Technology for Development*, volume 2011, pages 87–93.

Rakib, F. R., Dip, S. S., Alam, S., Tasnim, N., Shihab, M. I. H., Ansary, M. N., Hossen, S. M., Meghla, M. H., Mamun, M., Sadeque, F., et al. (2023a). Ood-speech: A large bengali speech recognition dataset for out-of-distribution benchmarking. *arXiv preprint arXiv:2305.09688*.

Rakib, M., Hossain, M. I., Mohammed, N., and Rahman, F. (2023b). Bangla-wave: Improving bangla automatic speech recognition utilizing n-gram language models. In *Proceedings of the 2023 12th International Conference on Software and Computer Applications*, pages 297–301.

Schultz, T. and Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1-2):31–51.

Shahgir, H., Sayeed, K. S., and Zaman, T. A. (2022). Applying wav2vec2 for speech recognition on bengali common voices dataset. *arXiv preprint arXiv:2209.06581*.

Showrav, T. T. (2022). An automatic speech recognition system for bengali language based on wav2vec2 and transfer learning. *arXiv preprint arXiv:2209.08119*.

The Editors of Encyclopedia Britannica (2023). *Bengali language*.

Tong, S., Garner, P. N., and Bourlard, H. (2017). Multilingual training and cross-lingual adaptation on ctc-based acoustic model. *arXiv preprint arXiv:1711.10025*.

Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., and Rao, K. (2018). Multilingual speech recognition with a single end-to-end model. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4904–4908. IEEE.

Wang, Y., Shi, Y., Zhang, F., Wu, C., Chan, J., Yeh, C.-F., and Xiao, A. (2021). Transformer in action: a comparative study of transformer-based acoustic models for large scale speech recognition applications. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6778–6782. IEEE.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS