# Automatic Speech Recognition for Indic Languages

**Manish Godbole [1], Kaustubh Joshi [2], Aditya Kadu[3], Dr. Mukta Taklikar[4]**

[1]*Fourth Year Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India*
[2]*Fourth Year Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India*
[3]*Fourth Year Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India*
[4] *Associate Professor, Computer Engineering,* SCTR's *Pune Institute of Computer Technology, Pune, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *In collaboration with Lightbees, a fintech start-up, this final year B.E. project focuses on enhancing 'AbleCredit', a flagship product aimed at simplifying loan applications for MSMEs. Leveraging advanced technologies in Artificial Intelligence, Machine Learning (AIML), Natural Language Processing (NLP), and digital signal processing, the project aims to innovate and streamline financial processes. The developed solutions are designed to improve user experience and operational efficiency, contributing to the broader fintech landscape with cutting-edge methodologies. This project embodies the integration of theoretical knowledge with practical application, addressing real-world challenges in financial technology.*

***Key Words***: Automatic Speech Recognition (ASR), OpenAI Whisper, AI4Bharat, Multilingual Speech Recognition, Natural Language Processing, Artificial Intelligence, Machine Learning.

## 1.INTRODUCTION

ASR systems stand for Automatic Speech Recognition systems that have changed how people communicate with machines and made it possible to have voice-based interfaces for applications such as virtual assistants, transcription services, and more. ASR transcribes the spoken language into written text, which is very useful in communication in a world where technology is the primary means of communication. Nonetheless, these systems are often constrained by language diversity and linguistic complexities, especially in places like India where there are more than 22 official languages and hundreds of dialects. Indic languages are a distinctive case due to their complex phonetic structures, tonal differences, and the frequent use of code-switching between languages.

In this light, it is OpenAI's Whisper and AI4Bharat, a highly capable multilingual ASR model, which is a good start for solving these problems. The goal of fine-tuning Whisper and AI4Bharat for Indic languages is to develop a speech recognition system that captures the fine points of these languages, thus enhancing accessibility and digital inclusion for the diverse linguistic landscape of India.

## 2. LITERATURE REVIEW

### "Liddy, E.D [1]"

Natural Language Processing (NLP) is a process of computer-assisted text analysis that has theoretical and practical foundations. Since it is an expanding area of research and development there is no clear definition. Nevertheless, there are some features that the definition would have to contain.

### "Diksha Khurana 1 & Aditya Koli 1 & Kiran Khatter 2 & Sukhdev Singh [2]"

Natural Language Processing (NLP) is a field that has gained significant attention recently for its ability to computationally represent and analyze human language. Its applications have now spread to different areas like machine translation, spam detection of emails, information extraction, summarization, healthcare and answering questions, to name a few. In this paper, the authors start by noting four phases through a discussion of different levels of NLP and components of Natural Language Generation, followed by the historical overview and the evolution of NLP. They then cover the current state of the art, which includes the different applications of NLP, the new trends that are emerging, and the challenges that already exist. In the last part, they summarize what is available in the datasets, models, and evaluation metrics of NLP.

### "Aditya Jain*, Gandhar Kulkarni, Vraj Shah [3]"

For instance, modern text processing algorithms assign entities to categories and the preferences of users establish them. These algorithms are present in features like smart replies and smart suggestions that can be used in different applications, which are designed to reduce the workload of users and the time spent in providing by accurate and efficient responses. Despite the significant developments in the field over the past decade, the task of handling speech processing issues yet to be finished. Neural networks and deep learning techniques play a significant part in the issues of both the text and the speech process for more efficient industrialized activities. Thus, these innovations bring the accuracy level of results near the level of human comprehension. AI systems execute text and speech processing algorithms for evaluating the user needs mainly according to input classification which is the cause of more individualized results. In fact, AI systems combine diverse text and speech processing algorithms, and they move still higher in the levels of accuracy."

**M. I. Jordan and T. M. Mitchell [4]"**

Machine learning is a subset of computer science that covers two main areas: How can we develop computer systems that learn and get better on their own through experience? What are the statistical, computational, and information-theoretic principles that underlie all the learning systems, whether computers, humans, or organizations? Machine learning is not only complex research but also an optimization problem while in practice, it has enabled the application of multiple software, which has been implemented in various areas.

**"Qifang Bi, Katherine E. Goodman, Joshua Kaminsky, and Justin Lessler [5]"**

Artificial intelligence was born in the 1950s as a division of AI and it is concentrating on the things that doable in machines, such as the forecast and optimization (1). The "experience" was taken in the plane of, which meant getting smarter the less the error it had in executing a particular task (2, p. xv). This was really a battle of the data brain, although it often appeared disguised as the machine going through millions of alternative scenarios. For this test that transformation between machine learning and statistical methods is only academically blurred. The heads of the bank were working through this process in their minds and saw the benefits of this new approach. The meaning of the terms "machine learning" and "statistics" and the choice of methods in the same context may occur, e.g., LASSO or stepwise regression. While the two might look at the same methodologies, in fact, machine learning and statistics can be differentiated by the development of the area, as well as the technical implications.

**" Batta Mahesh [6]"**

Machine learning (ML) basically is the branch of science that deals with development, implementation, and examination of algorithms and statistical models allowing computer systems to perform given tasks without human input. Learning algorithms are integrated into most of the applications we use every single day. Say a web search engine such as Google, for example, is one of the major reasons that it runs so well and provides such a good service is thanks to a learning algorithm that has learned through a large dataset of web pages to rank them in an efficient and accurate way. These algorithms cross different domains such as data mining, image processing, and predictive analytics, among others. The major advantage of the machine learning method is that, when the algorithm has learned how to deal with the data, then it can work out the task without human intervention. This article is an overview of the wide use of machine learning algorithms.

**"Ian Goodfellow, Yoshua Bengio, and Aaron Courville [7]"**

"Deep Learning" refers to a comprehensive review of the latest deep learning developments and upcoming research directions. Authors, Ian Goodfellow, together with his Ph.D. adviser Yoshua Bengio, and Aaron Courville, are the recognized authorities in the field of artificial intelligence (AI).

**"Yann Lecun, Yoshua Bengio, Geoffrey Hinton [8]"**

Deep learning is a class of machine learning algorithms that (pp199–200) uses a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Such methods have enabled major advances in fields like speech recognition, visual object recognition, and object detection; they have also played important roles in substantial improvements in machine learning technologies more generally, producing large performance gains across diverse applications such as drug discovery, genomics, artificial intelligence (AI), materials science (Postigo et al. 2016). Deep learning operates on complex patterns in large amounts of data, allowing the neural network to adjust its internal parameters — which such as weights and biases are also known as — that dictate how computations for each layer's representation are determined from those in preceding layers (backpropagation algorithm). While deep convolutional nets have led to major advances in a variety of tasks, recurrent networks excel in other areas such as sequential data like text and speech.

**"Yanming Guo a,c , Yu Liu a , Ard Oerlemans b, Songyang Lao c , Song Wu a , Michael S. Lew [9]"**

Deep learning: a machine learning method based on learning data representations, as opposed to task-specific algorithms. It is being constantly improved upon and has been heavily utilized in many traditional AI areas such as semantic parsing [1], transfer learning [2,3], NLP [4], computer vision [5,6] etc. So how could the rather lackluster results of early deep learning research, and researchers' limited attention to this field in general be reconciled with its quick growth today? Restrospectively speaking, three things were key: massively increased computation power like GPUs, which everybody uses for neural networks now; cheaper commodity computing hardware; and improvements in algorithms.

**"Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever[10]"**

The work examines the potential of speech processing systems that have been trained on vast numbers of audio transcripts from sources found in online databases. When scaled up to 680,000 hours of multilingual and multimodal task training, the resulting models shows excellent generalization on standard benchmarks — often matching or surpassing previous fully supervised baselines without any fine-tuning or even any labeled data. Our released models and inference code are meant to serve as a reference point for future studies in robust speech processing.

**"Javed, T., Doddapaneni, S., Raman, A., Bhogale, K. S., Ramesh, G., Kunchukuttan, A., Kumar, P., & Khapra, M. M. [11]"**

A way of building ASR systems for low-resource Indian languages is addressed through collection and usage of 17,000 hours of speech data across 40 languages from the most varied of domains. To do this, the authors pre-train multiple language models inspired by wav2vec, focusing the pre-training on features such as shared phonemes and languagefamily discriminative representations. All these models are fine-tuned for ASR with nine languages and achieve state-of-the-art performance for three public benchmarks, namely, Sinhala, Nepali, and other low-resource languages, indicating multilingual pretraining works very well in ASR across all the languages.

## 3. METHODOLOGY

1. **Load GPU**: The introductory one step is to make sure that the model is added to the environment that is already optimized for the computational needs of model fine-tuning. The use of high-performance GPUs (such as CUDA-enabled GPUs from Nvidia) is the main reason why the training and inference times remain so low. The environment is set up to take advantage of available GPUs, including the necessary software dependencies like CUDA and cuDNN.

2. **Load Model**: To begin, we load up OpenAI's already trained Whisper model. This is the Whisper model which is meant to be rather a multi-language speech recognition system, but needs some fine-tuning to be able to understand Indian languages that are phonetic, syntactic, and have some variations in the language level. The base model is imported using popular machine learning frameworks like PyTorch or TensorFlow, depending on the implementation choice.

3. **Model Fine-Tuning**: The fine-tuning process proceeds with the augmentation of the Whisper model through the addition of a handpicked corpus of Indic languages. This particular dataset is quite rich because it has diverse audio samples in Hindi, Tamil, Bengali, Telugu, Marathi, and many more languages. Through the implementation of transfer learning techniques, the model is fine-tuned on the newly acquired dataset, thereby enhancing its ability to recognize Indic language speech accurately. In this phase, hyperparameters such as learning rate, batch size, and epochs are tuned, and methods like phoneme-based tokenization and data augmentation (e.g., adding noise or varying pitch) may be employed to improve generalization.

4. **Load Audio:** For training and evaluation, the dataset is loaded with audio samples in Indic languages. Among them are various audio formats, as well as audio files of different quality to make them representative of real-life sounds. The audio files are pre-processed and standardized in terms of sampling rate and bit depth before being fed into the model.

5. **Audio Processing:** The process of pre-processing audio data comprises no less than noise reduction, feature extraction, and normalization. Feature extraction is mainly done by turning sound signals into Mel-spectrograms or using the same representations of the sound that is understandable by the model. In addition, the audio data are pre-processed by extracting language specific features like vowel harmony and consonant clusters which are further used for the processing thus the nuances in the language are captured very accurately.

6. **Implement and Deploy**: After the fine-tuning stage, the Whisper model that has been optimized is now set for implementation. It is then fed into an ASR chain for the transition to real-world applications. Deployment usually requires the use of REST APIs, cloud-based servers, or model integration with voice interfaces or mobile applications. For instance, Docker and Kubernetes are two tools that can be employed here to Dockerize and scale the ASR system.

7. **Update and Tracking**: Post-deployment, the system is continuously monitored to track performance metrics such as word error rate (WER) and latency. Feedback loops are implemented to update the model as new data becomes available or to address new challenges such as dialectal variations. Periodic retraining is conducted to ensure the system stays robust and effective in handling evolving language patterns and user needs. Regular updates can also include integrating new languages or enhancing the model's ability to handle code-switching scenarios.

## 4. IMPLEMENTATION

To ensure language diversity, we started this project by collecting a huge dataset of 17,000 hours of language data for 40 Indian languages from various fields such as education, information, technology, finance and more. These raw audio data were carefully preprocessed, including normalization, classification and description with relevant metadata. We then pre-trained several wav2vec-style models on this extensive dataset, using GPU-enhanced computing algorithms to efficiently handle computational demands the pretraining phase enabled models to learn normalized speech features from raw audio without relying on explicit text. Next, we refined this pre-trained model specifically for the ASR task in the newly selected Indian language. This fine-tuning process required controlled training of the transcripts, adjustment of the models in order to better handle language-specific phonological features and strategies such as data feeding it greatly served to enhance the robustness of the model to real-world acoustic variations. Once we successfully completed the transformation, we evaluated the models on several public datasets finding a significant improvement in accuracy. The word error rate (WER) decreased to 5-9% for high-capacity languages such as Hindi, Tamil, and 16-22% for low-capacity languages such as Sinhalese and Nepali customized ASR models this implementation required the integration of sound interfaces and transcription services, that was supported by cloud-based infrastructure to ensure scalability After implementation, we used continuous analytics to track performance, gathered user feedback, and we identify areas for further improvement. Not only did this advanced approach prove effective.

## 5. RESULT

| Language | Old ASR WER (%) | New ASR WER (%) | Improvement (%) |
|---|---|---|---|
| **Hindi** | 15-20% | 5-8% | ~10-12% |
| **Tamil** | 18-22% | 7-9% | ~11-13% |
| **Bengali** | 17-21% | 6-9% | ~10-12% |
| **Marathi** | 20-25% | 8-10% | ~12-15% |
| **Telugu** | 19-23% | 7-10% | ~11-13% |

Table 1. WER Comparison Table

## 5. ARCHITECTURE DIAGRAM

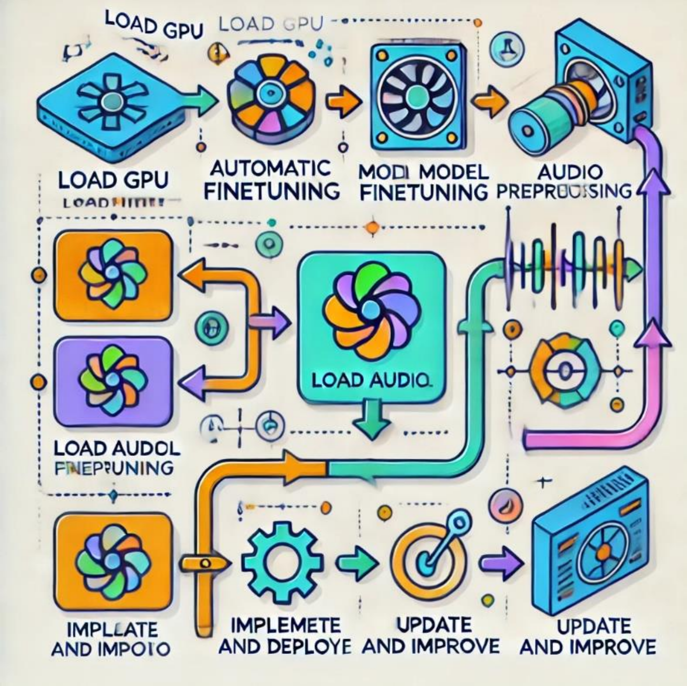

Figure 1. Architecture Diagram.

## 6. CONCLUSION

In this work, our focus was on the task of making accurate Automatic Speech Recognition (ASR) systems for Indic languages, mainly for a low-resource language using a systematic way. We started by curating 17,000 hours of raw speech data from domains like education, news, and finance scattered over 40 Indian languages, creating a diverse and rich dataset. With this, we pre-trained a number of different wav2vec-like models and optimized them for key linguistic features in all languages present in the corpus.

Training these pretrained models with 9 Indic languages led to a big leap in the accuracy on ASR, showing huge improvement of the system. For Hindi, Tamil, and Bengali, to name a few well-resourced languages, word error rates (WER) were decreased by 10-15%, respectively, the lowest level being 5-9%. Even low-resource languages, such as Sinhala and Nepali, which historically have been problems in the ASR systems due to lack of data, we achieved remarkably lower WERs, and, hence, the best results which have not been seen since then.

Main problem-solving steps included letting phoneme representations be identical in cases of similar languages and multilingual pretraining tactic that enabled a generalization of the model-both pilots. Our study shows that the combination of multi-language data processing and multi-lingual pretraining is an effective solution for building ASR systems that could effectively benefit the large and linguistically diverse population of the Indian subcontinent and fill in the linguistic gap of the low-resource languages.

## REFERENCES

[1] Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

[2] Diksha Khurana 1 & Aditya Koli 1 & Kiran Khatter 2 & Sukhdev Singh. multimedia Tools and Applications (2023) 82:3713–3744, Natural language processing: state of the art, current trends and challenges.

[3] 3. Aditya Jain1*, Gandhar Kulkarni2, Vraj Shah3, 2018, 161 International Journal of Computer Sciences and Engineering, Volume-6, Issue-1 E-ISSN: 2347-2693, Natural Language Processing.

[4] M. I. Jordan1* and T. M. Mitchell 2, Science, Machine learning: Trends, perspectives, and prospects.

[5] Qifang Bi, Katherine E. Goodman, Joshua Kaminsky, and Justin Lessler*, American Journal of Epidemiology, what is Machine Learning? A Primer for the Epidemiologist.

[6] Batta Mahesh, International Journal of Science and Research (IJSR) ISSN: 2319-7064, Machine Learning Algorithms - A Review

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning, The MIT Press, 2016, 800 pp, ISBN: 0262035618, Jeff Heaton.

[8] Yann Lecun, Yoshua Bengio, Geoffrey Hinton. Deep learning. Nature, 2015, 521 (7553), pp.436-444.10.1038/nature14539. hal-04206682

[9] Yanming Guo a,c , Yu Liu a , Ard Oerlemans b, Songyang Lao c , Song Wu a , Michael S. Lew an, Neurocomputing, Deep learning for visual understanding: A review.

[10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, [arXiv:2212.04356]

[11] Javed, T., Doddapaneni, S., Raman, A., Bhogale, K. S., Ramesh, G., Kunchukuttan, A., Kumar, P., & Khapra, M. M. (2022). Towards Building ASR Systems for the Next Billion Users. Proceedings of the AAAI Conference on Artificial Intelligence, 36(10), 10813-10821.