International Journal of Information Science and Computing
Citation: *IJISC*: **12**(02): 189-201, December 2025
DOI: 10.30954/2348-7437.2.2025.6

INFORMATION SCIENCE AND TECHNOLOGY FOUNDATION

RESEARCH PAPER

Whispering in Sylheti Language

Swapnajeet Das*, Amit Kumar Roy, Bijay Kumar Singh and Bipul Syam Purkayastha

Department of Computer Science, Assam University, Silchar, India

**Corresponding author: swapnajeet.das@zohomail.in*

Received: 16 Sept., 2025 **Revised:** 12 Nov., 2025 **Accepted:** 24 Nov., 2025

ABSTRACT

Automatic Speech Recognition (ASR) converts the spoken words/sentence to text, which allows voice activation features, accessibility, and real-time translation to all. In low resource languages, ASR opens the door to digital inclusion, heritage conservation, and community-based data collection in areas where there were previously limited resources. We consider an ASR system of the Sylheti language that was created on the Whisper model of Open AI. Sylheti is a language with low resources, used in India and Bangladesh, but there are no significant digital and linguistic resources, which makes it challenging to develop the ASR. A hybrid dataset was developed to fill this gap by the addition of a custom Sylheti corpus and Common-Voice dataset Bengali corpus. The audio information was pre-processed by removing noise, cutting, and matching the audio with written transcriptions. The small model of the Whisper was also optimized with the help of the combined dataset, as its model is based on the transformer-based encoder decoder architecture, which was trained on the task of multilingual transcription. Word Error Rate (WER) metrics is used to evaluate trained model. The system got a WER of 66.8% which proved to be a good system in transcription accuracy of Sylheti speech. The system was implemented via an interactive Gradio interface of interactive transcription. Its results confirm Whisper to be flexible towards low-resource languages and the possibility of facilitating linguistic inclusivity, cultural continuity, and accessibility.

Keywords: ASR, Whisper Model, Sylheti Language, WER, Gradio

The first appeal was published in 1956, Artificial Intelligence (AI) has been evolving at an accelerated pace and providing strength to various spheres of contemporary life[1]. further emphasise that, between nearly 200,000 and more than 496,000 AI-related publications have grown in 2010 and 2021, respectively, indicating the increased interest of scientists and industries in this area. Automatic Speech Recognition (ASR) is an actively developing branch of AI that allows computers to translate speech into text[2]. ASR systems have been integrated into everyday use in voice searching the internet, smart assistance, biometric authentication, and controlling IoT devices and communication between humans[3].

How to cite this article: Das, S., Roy, A.K., Singh, B.K. and Purkayastha, B.S. (2025). Whispering in Sylheti Language. *IJISC.,* **12**(02): 189-201.

Source of Support: None; **Conflict of Interest:** None

The efficiency and performance of ASR research have led to the shift system to End-to-End (E2E) systems instead of the traditional hybrid pipelines[4]. Recurrent Neural Network Transducer (RNN-T), Attention-based Encoder-Decoder (AED) and Connectionist Temporal Classification (CTC) are all successful E2E frameworks[5]. Some of the latest models that use such methods include Whisper - a Transformer-based encoder-decoder ASR system created by the OpenAI[6]. Whisper is trained on 6,80k+ hours of multi-lingual and multi-task information on language identification, speech recognition and translation of speech .

Nevertheless, the performance of Whisper remains poor in those cases when it works with the languages that are not represented in digital corpora, which are often referred to as low-resource languages[6]. The problem of such languages is that they lack adequate training data, thus having low recognition accuracy in terms of Word Error Rate (WER) or Character Error Rate (CER). The existing literature indicates that Whisper has a problem with zero-shot recognition of underrepresented languages[7]. These challenges are seen in such languages as Javanese and other regional languages in South Asia[8].

Other low-resource languages that are given little consideration in the development of speech technology are Sylheti, a language used by more than 10 million people in Northeast India (Southern Assam India, Tripura and others) and some parts of Bangladesh. Though it bears certain linguistic resemblance with Bengali, the phonetics and accent styles and vocabulary used require a specific ASR system, instead of a transfer-assumption to Bengali.

Fine-tuning is also known as an effective solution to the problem of the small amount of data to solve the problem and make Whisper more efficient in downstream tasks[9]. Nevertheless, it is normally expensive to train large models such as Whisper. To solve this, Low-Rank Adaptation (LoRA) , a type of Parameter-Efficient Fine-Tuning (PEFT) method, can be used to minimize memory needs and trainable parameters without compromising performance[10].

Thus, the goal of this study is to improve Whisper for the Sylheti language and assess its effectiveness using the WER metric. The primary objective is to improve the quality of ASR for Sylheti and support continued initiatives to increase low-resource language populations' access to technology.

LITERATURE REVIEW

The last few decades have witnessed a substantial development of speech and language technologies due to the use of Artificial Intelligence (AI) as noted by Zhang, C. and Lu, Y.[1], there has been a rapid development of AI-driven systems, especially natural language processing and speech applications. Automatic Speech Recognition (ASR) no longer involves the use of traditional modular pipelines that combine acoustic and language models Levinson, S.E. and Roe, D.B.[2] to deep learning based end-to-end (E2E) designs. Contemporary ASR systems, which are divided into CTC, Attention-based Encoder-Decoder (AED), and RNN-T systems by Alharbi, S. *et al.*[5], transformer architecture proposed by Vaswani, Ashish[11] has helped gain a great benefit in terms of efficient modeling of long-range dependencies with the help of self-attention..

One of the recent advances in multilingual ASR is Whisper by Radford, Alec *et al.*[6], which was trained on multilingual speech data (680,000+ hours). Despite the high zero-shot capabilities of Whisper, it has been shown that it performs weakly on studies by Rouditchenko, A. *et al.*[7] report underper-formance of greatly underrepresented low-resource languages. The lack of data, dialect, and phonetic complexity are still vexing. It is partly solved by crowd sourced datasets, which are discussed in Butryna, A. *et al.*[8], and

partly cross-lingual approaches Novitasari, S. *et al.*[12] and Bengali ASR studies Al Amin, M.A. *et al.*[13] show the power of related language transfer. Nevertheless, this transfer is frequently unable to reproduce dialect-specific variation, which results in large Word Error Rates (WER).

In order to decrease computational cost, Parameter-Efficient Fine-Tuning (PEFT) models like LoRA Hu, E.J. *et al.*[4] and quantization Dettmers, T. *et al.*[14] are used to adapt large models efficiently. However, according to Pratama, R.S.A. and Amrullah, A.[15], zero-shot Whisper is still inadequate with low-resource languages without fine-tuning, which is why it is necessary to consider language-specific strategies to adapt it.

RESEARCH GAP

Despite the excellent generalization of multilingual transformer-based ASR systems like the OpenAI Whisper, there are still major gaps in the low-resource language research. Majority of the studies done are on well-resourced or moderately represented languages, and therefore the regional and dialectal languages remain underexplored. Although some positive results have been obtained in relation to the cross-lingual transfer between related languages such as Bengali, little of the studies have been done to assess the appropriateness of cross-lingual transfer to reflect dialect-based phonetic and lexical features. Moreover, even though Parameter-Efficient Fine-Tuning (PEFT) methods, including LoRA and quantization, decrease the computational expenses, their performance as adapters of large multilingual ASR models to particular low-resource dialects has not yet been well studied.

The language of millions of people in Northeast India and Bangladesh, Sylheti has received little regard in the ASR research. No effort has been done to systematically fine-tune Whisper on Sylheti with hybrid corpora and measure performance in terms of Word Error Rate (WER). This paper fills that gap by doing specific fine-tuning and evaluation, which can be added to the development of inclusive speech technology

METHODOLOGY

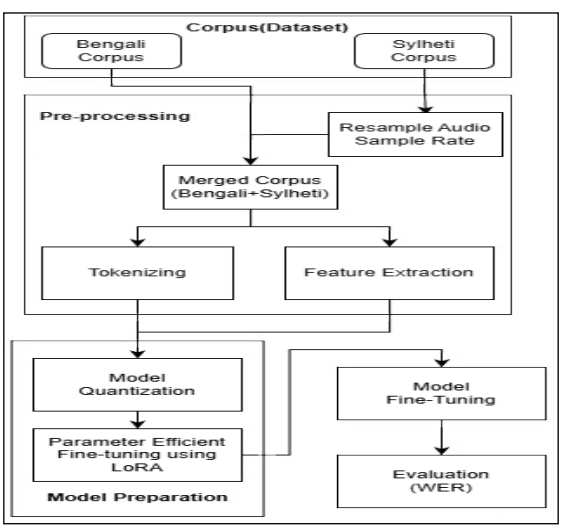


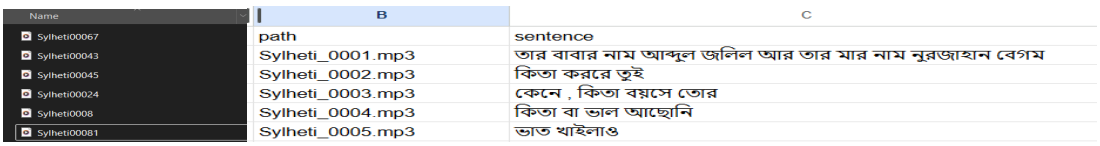
Fig. 1: Flow Chart of the proposed model

The objective is to use computation to fine-tune the Whisper ASR model for the Sylheti language. To maximise performance throughout the fine-tuning procedure, the audio sample rate of the dataset must first be converted from 48 kHz to 16 kHz[16]. Tokenisation and feature extraction are the two phases of preparing data. Following preprocessing, the model's trainable parameter was reduced by using PEFT Low-Rank Adaptation (LoRA) to compute minimal computing cost. After all the preparations are complete, then fine-tuning can begin. Upon the fine-tuning, the performance of the model will be assessed with the help of WER and compared to the model that will not be fine-tuned (the zero-shot model). The specifics of the research flow are presented in Fig. 1.

DATASET

There are two datasets are used Primary Dataset and Secondary Dataset, which will be merged together to perform fine tuning with whisper model for ASR.

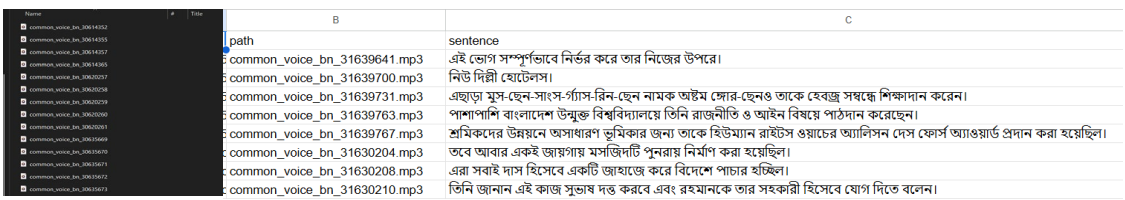
A specially created Sylheti Corpus as primary dataset, which includes spoken audio files with the extension ".mp3," file IDs, and transcriptions of each audio file, will be utilised. There are 3000 audio files total, each with an ID and transcription, and they are 2-3 hours long[15].



Name	B	C
Sylheti00067	path	sentence
Sylheti00043	Sylheti_0001.mp3	তার বাবার নাম আব্দুল জলিল আর তার মার নাম নুরজাহান বেগম
Sylheti00045	Sylheti_0002.mp3	কিতা কররে তুই
Sylheti00024	Sylheti_0003.mp3	কেনে , কিতা বয়সে তোর
Sylheti0008	Sylheti_0004.mp3	কিতা বা ভাল আছোনি
Sylheti00081	Sylheti_0005.mp3	ভাত খাইলাও

Fig. 2: Sylheti Corpus

The Common Voice (Bengali corpus v_19) secondary dataset, which includes spoken audio files with the extension ".mp3," file IDs, and transcriptions of each audio file, will be utilised. There are 37,000 audio files total, each with an ID and transcription, and they are all 54 hours long[13].



	B	C
	path	sentence
	common_voice_bn_31639641.mp3	এই ভোগ সম্পূর্ণভাবে নির্ভর করে তার নিজের উপরে।
	common_voice_bn_31639700.mp3	নিউ দিল্লী হোটেলস।
	common_voice_bn_31639731.mp3	এছাড়া মুস-ছেন-সাংস-র্গ্যাস-রিন-ছেন নামক অষ্টম ঙ্গোর-ছেনও তাকে হেবজ্র সম্বন্ধে শিক্ষাদান করেন।
	common_voice_bn_31639763.mp3	পাশাপাশি বাংলাদেশ উন্মুক্ত বিশ্ববিদ্যালয়ে তিনি রাজনীতি ও আইন বিষয়ে পাঠদান করেছেন।
	common_voice_bn_31639767.mp3	শ্রমিকদের উন্নয়নে অসাধারণ ভূমিকার জন্য তাকে হিউম্যান রাইটস ওয়াচের অ্যালিসন দেস ফোর্স অ্যাওয়ার্ড প্রদান করা হয়েছিল।
	common_voice_bn_31630204.mp3	তবে আবার একই জায়গায় মসজিদটি পুনরায় নির্মাণ করা হয়েছিল।
	common_voice_bn_31630208.mp3	এরা সবাই দাস হিসেবে একটি জাহাজে করে বিদেশে পাচার হচ্ছিল।
	common_voice_bn_31630210.mp3	তিনি জানান এই কাজ সুভাষ দত্ত করবে এবং রহমানকে তার সহকারী হিসেবে যোগ দিতে বলেন।

Fig. 3: Bengali Corpus

The Corpus will be split into two portions, one for model performance evaluation and the other for model training or fine-tuning. The ideal ratio for dividing datasets is 8:2[17], where 8 is the training corpus and 2 is the test corpus.

PREPROCESSING

(a) Speech Resampling

The dataset that will be utilised has an audio sample rate of 48 kHz. According to[6], Whisper functions best at a sampling rate of 16 kHz, particularly its feature extractor. Thus, audio must be resampled from

48 kHz to 16 kHz in order to maximise performance throughout the fine-tuning phase. Bandlimited interpolation was used to resample the audio using the Torch audio library[14].

(b) Tokenization

In low-resource languages, tokenisation has a considerably greater impact on language model performance. Sentences will be divided into tokens using the Whisper Model's Byte-level BPE tokenizer[6]. Bengali is one of the 96 languages that Whisper's built-in tokenizer recognises[18].

(c) Features Extraction

Whisper model's built-in feature extractor will be used to extract features. There are two ways to use this feature extractor. The audio stream is first padded to match the 30-second time using the feature extractor. The audio will be divided into two parts if it is longer than thirty seconds. The audio stream will either be silent for 30 seconds or padded with zero if it is shorter than the allocated 30 seconds. The audio is then converted into an 80-channel log mel-spectrogram image. First, the Fast Fourier Transform (FFT) will be used to measure the audio stream. The logarithm operation will be carried out once this measurement has been entered into the mel-filter bank[6].

This leads to log-mel spectrogram visualisation. Fig. 4 shows the workflow.

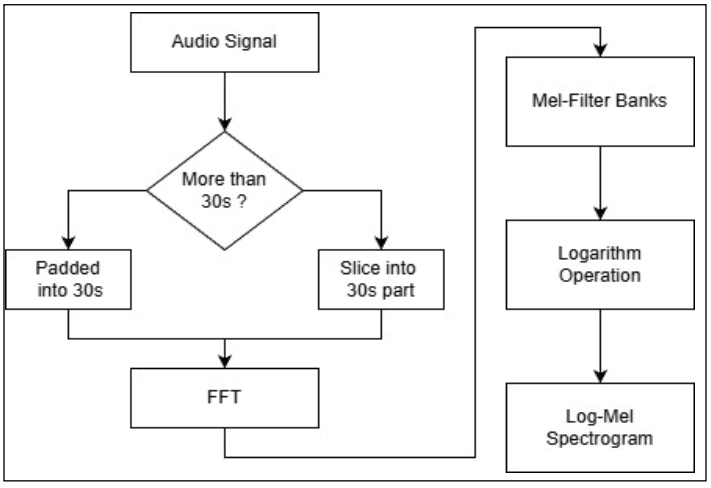


Fig. 4: Feature Extractor Process in Whisper Model

MODEL QUANTIZATION

Whisper is the model that uses the concept of an encoder-decoder transformer[11]. As it learns to encode the input sequence into word sequences, it can also be referred as a seq-2-seq model[6], Weak supervisory training over 6,80k+ hours of multilingual speech or audio have already been performed[6]. It will be the Whisper model where the log-mel extractor spectrogram of the feature extractor will be coded to give a sequence of the hidden state of the encoder as shown in Fig. 3. It will then be decoded using this sequence to autoregressively predict the text token using the encoder hidden state and the prior token condition.

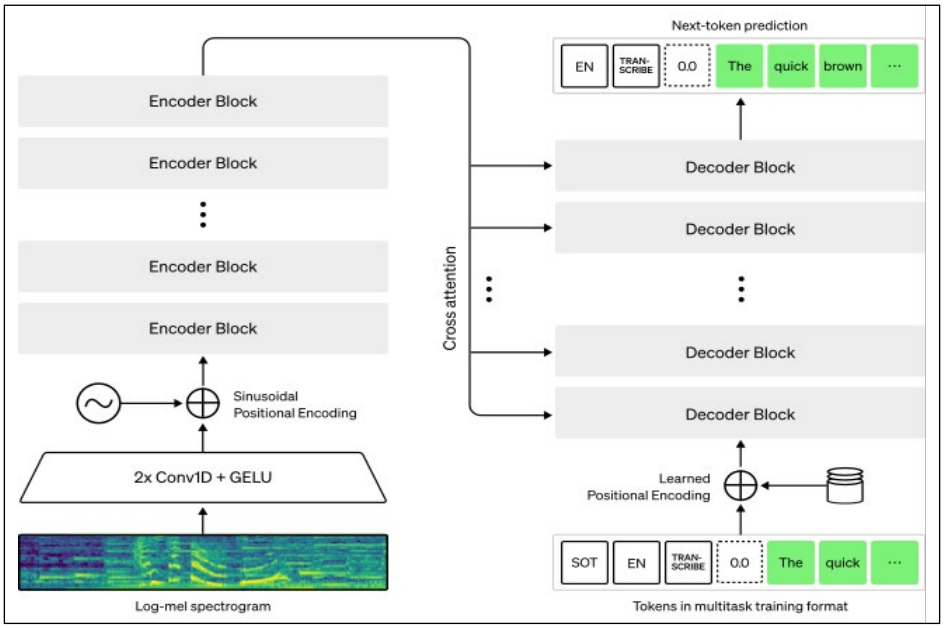


Fig. 5: Architecture of the Whisper Model

The whisper variation model can be contemplated in the table below. It includes the Whisper Base model, the Small model, the Medium model, and the Large model.

Table 1: Comparisons of Different Whisper-based Models

Model	No. of Layer	Width	No. of Head	Size
Tiny	4.0	384.0	6.0	39 M
Base	6.0	512.0	8.0	74 M
Small	12.0	768.0	12.0	244 M
Medium	24.0	1024.0	16.0	769 M
Large	32.0	1280.0	20.0	1500 M

Models with parameters that are larger than fine-tuning models (more than 200M) are prohibitively expensive to compute. In order to solve this problem, model quantization is suggested. 8-bit quantization is the original proposal of model quantization by[14]. The entire model is operating at full accuracy with the type of data of fp32. The model data type will be loaded in quarter precision or an 8-bit data type to save memory and computing cost, and it was demonstrated that the loss of accuracy on the model was just the loss of the full precision model, which was about 5% loss.

PARAMETER-EFFICIENT FINE-TUNING

Parameter-Efficient Fine-Tuning (PEFT) is suggested as a way to further reduce memory and computing requirements. Only necessary parameters associated with certain tasks will be taught since PEFT operates by freezing model parameters[9]. The LoRA approach makes this reduction achievable. One of the methods

that was proposed by[4] is the Low-rank Adaptation (LoRA), which minimizes the number of parameters to be trained and, consequently, minimizes memory and computational expenses by introducing an additional decomposition matrix (Update matrix) to represent weights and only new weights are trained.

MODEL FINE-TUNING

32GB of RAM, a 12.288GB RTX 3060 GPU, and an i7 12thGen CPU were used to fine-tune the whisper model in an Ubuntu system. The fine-tuning hyperparameter is displayed in Table 2.

Table 2: Hyperparameter used for training

Hyper Parameter	Values
Rate of Learning	1e-05
Train batch Size	16
Evaluation batch Size	16
No. of Seeds	42
Optimizer	Use adamw_torch with betas = (00.90,0.9990)
Epsilon	1e-08
Optimizer arguments	No additional optimizer arguments
Learning Scheduler type	constant_with_warmup
Learning Scheduler_warmup steps	500
Training steps	5000

EVALUATION

The model is assessed using the Word Error Rate (WER) evaluation metric. The Character Error Rate (CER) and WER calculations are identical. However, because the error rate is assessed using words rather than characters, WER more properly depicts ASR performance[19]. As a result, it may be concluded that ASR has a larger error rate when assessed using WER than when assessed using CER. The ASR model was evaluated using CER in a prior study[12]. Because the language has some characters that are not part of the regular alphabet, this approach was selected. However, the standard alphabet can be used to write it in Sylheti. Therefore, a more accurate assessment of ASR performance in identifying Sylheti speech will be provided by the WER evaluation. The following equation can be used to determine WER mathematically: “S” stands for substitution, “D” for deletion, “I” for insertion, and “C” for correct[15].

$$WER = \frac{(S + D + I)}{(S + D + C)}$$

RESULTS AND ANALYSIS

The result demonstrates a consistent decrease in the evaluation loss and WER as training progresses, indicating improved accuracy in speech recognition. There is a steady increase in samples per second,

suggesting enhanced efficiency in model inference over time.

Table 3: Results Summary

Metric	Initial	Final
Training Loss	1.1084	0.1148
Validation Loss	1.09012	0.2100
WER	205.33%	66.80%

The programs used to train the Whisper model to fine-tune it to the Sylheti speech were used to train and validate it during the training period. The main evaluation measure to be employed in this study is Word Error Rate (WER), and the trends on training and validation loss as a measure of convergence.

(a) "Word Error Rate" (WER) Trend

The results of the WER decrease during 5000 training steps are shown in Fig. 6. The WER is also extremely high (205.33% at step 500) because the model is not very adaptable to Sylheti phonetics and linguistic structure. As training advances however, WER falls off sharply so that at step 5000, the WER is 66.80 percent.

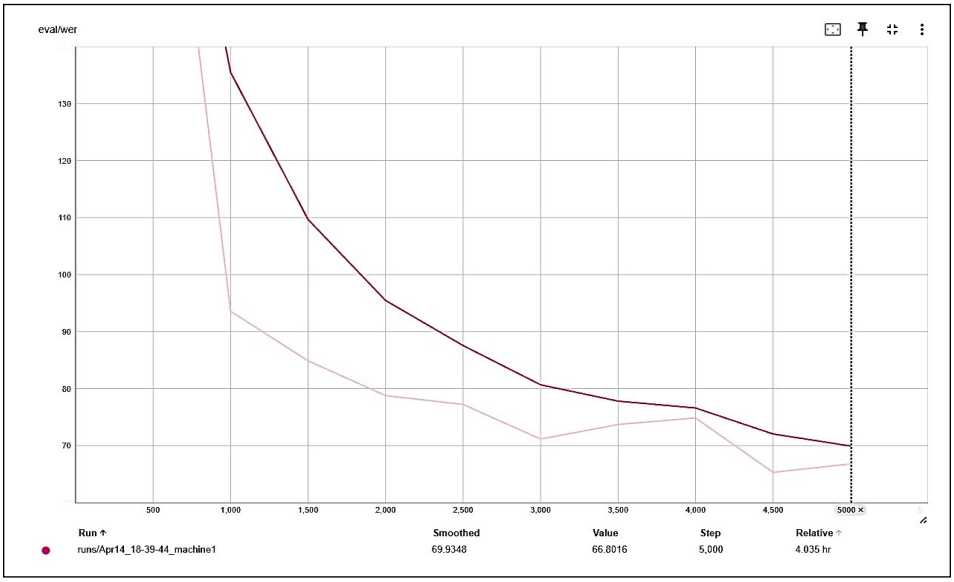


Fig. 6: Evaluation WER Graph

This significant change in WER demonstrates that the model is more and more able to properly identify Sylheti words, which proves that fine-tuning works.

(b) Training Loss Behavior

The training loss reduction vs. the steps is shown in Fig. 7. The training loss takes a nosedive at the initial stage (until step ~800) before levelling off to convergence. The last loss in training was 0.1148, which showed successful optimization.

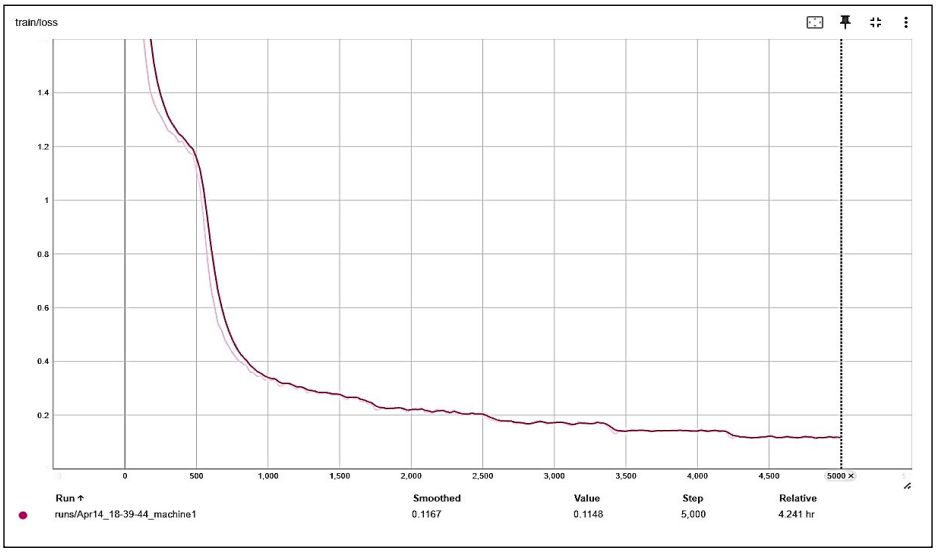


Fig. 7: Training/Loss Graph

The continuous and monotonically decreasing curve demonstrates that the model does not have any unstable moments when training.

(c) Training Results Summary

The table below summarises key performance values at major checkpoints.

Table 4: Training Results

Training results

Training Loss	Epoch	Step	Validation Loss	Wer
1.1084	0.5959	500	1.0912	205.3312
0.3323	1.1919	1000	0.3413	93.5779
0.2757	1.7878	1500	0.2803	84.8822
0.2238	2.3838	2000	0.2543	78.7731
0.202	2.9797	2500	0.2323	77.2269
0.172	3.5757	3000	0.2224	71.1526
0.1384	4.1716	3500	0.2171	73.7138
0.14	4.7676	4000	0.2103	74.8553
0.1273	5.3635	4500	0.2117	65.3154
0.1148	5.9595	5000	0.2100	66.8016

Both training and validation loss decrease progressively, and the gap between them remains small, indicating no over-fitting.

(d) Real-Time Inference Demonstration Using Gradio

A Gradio-based speech recognition interface was created to test the practical potential of the fine-tuned Whisper model, which allowed the use of unseen Sylheti speech data to test the model live. It can be seen that the interface enables users to record speech or upload audio, and the model provides them with the corresponding transcription (Fig. 8).

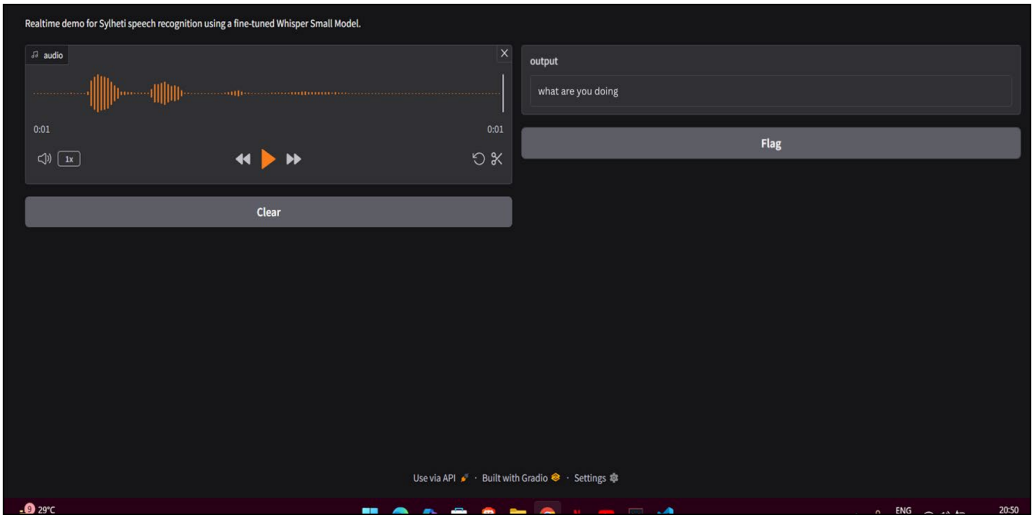


Fig. 8: Gradio Interface

The model demonstrates good recognition performance for short and simple Sylheti sentences, successfully converting them into meaningful text (*"What are you doing"*). Since Word Error rates (WER) is at 66.80%, the system has difficulty with longer sentences or complex speech patterns, and therefore the words are usually not translated correctly. The limitation is also affected by the fact that the experiments were done on the Whisper Small model with limited dataset, which although, in terms of computation, is an efficient model, has lower representational capacity than Whisper variants of large sizes.

The general impression of the demo is that the model is workable and holds potential of use in real life situations, particularly in short speech exchanges. This is likely to improve performance over time with longer fine-tuning and bigger model with bigger dataset configurations which will help minimize WER and enhance generalization on different speech of Sylheti.

FUTURE DIRECTION

Future research can extend this work in several important directions:

- **Bigger and more heterogeneous Sylheti corpus**: Bigger and more heterogeneous Sylheti corpus: By increasing the data size with speakers of various ages, gender, accents and recording conditions, it will be possible to enhance the generalization of models.

- **Investigations in Whisper scaled up variants:** Fine-tuning Whisper Medium or Large models can substantially decrease WER because it is capable of having much more representations.
- **State of art PEFT approaches:** There are various more computationally efficient approaches that can be investigated, such as QLoRA, adapters, and prompt-tuning.
- **Language model integration:** For longer sentences, decoding accuracy may be increased by using shallow fusion or external language models.
- **Noise-robust training:** Techniques for augmentation (such as noise injection and SpecAugment) can improve robustness in the real world.
- **Adoption in the real world:** Integration with educational resources, voice assistants, and Sylheti-speaking accessibility systems.
- **Cross-lingual transfer learning:** Utilizing multilingual training techniques to take use of closely related languages (like Bengali).

CONCLUSION

This project developed an automatic speech recognition model for Sylheti using the Whisper architecture. Achieving a final word error rate of 66.8016 % further validates the effectiveness of our approach, contributing to the broader field of speech technology and emphasizing the need for continuous evaluation and adaptation to achieve the best results in ASR system development. Future work will focus on enhancing the model's robustness and exploring broader applications, contributing to broader efforts in promoting language inclusivity and accessibility in the digital domain. Future work will focus on:

- Expanding the Sylheti corpus with dialect diversity
- Fine-tuning larger Whisper variants
- Exploring PEFT (LoRA, QLoRA) for computational efficiency
- Deploying Sylheti ASR models for real-world applications (education, accessibility, voice assistants)

REFERENCES

1. Zhang, C. and Lu, Y. 2021. "Study on artificial intelligence: The state of the art and future prospects," *Journal of Industrial Information Integration*, **23**: 100224.
2. Levinson, S.E. and Roe, D.B. 1990. "A perspective on speech recognition," *IEEE Communications Magazine*, **28**(1): 28–34.
3. Zhang, X., Peng, Y. and Xu, X. 2019. "An Overview of Speech Recognition Technology," in *2019 4th International Conference on Control, Robotics and Cybernetics (CRC)*, pp. 81–85. doi: 10.1109/CRC.2019.00025.
4. Hu, E.J. *et al.* 2021. "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv*: arXiv:2106.09685.

5. Alharbi, S. *et al.* 2021. "Automatic Speech Recognition: Systematic Literature Review," *IEEE Access*, vol. 9, pp. 131858–131876, doi: 10.1109/ACCESS.2021.3112535.

6. Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C. and Sutskever, I. 2025. "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Jul. 2023, pp. 28492–28518. Accessed: Dec. 15, 2025. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

7. Rouditchenko, A. *et al.* 2023. "Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages," *arXiv*: arXiv:2305.12606. doi: 10.48550/arXiv.2305.12606.

8. Butryna, A. *et al.* 2020. "Google Crowdsourced Speech Corpora and Related Open-Source Resources for Low-Resource Languages and Dialects: An Overview," *arXiv*: arXiv:2010.06778. doi: 10.48550/arXiv.2010.06778.

9. Liu, H. *et al.* 2022. "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning," *Advances in Neural Information Processing Systems*, **35**: 1950–1965.

10. "On the Effectiveness of Parameter-Efficient Fine-Tuning | Proceedings of the AAAI Conference on Artificial Intelligence." Accessed: Dec. 15, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26505

11. Vaswani, A. 2025. "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30. p. I, 2017. Accessed: Dec. 15, 2025. [Online]. Available: https://cir.nii.ac.jp/crid/1370849946232757637

12. Novitasari, S., Tjandra, A., Sakti, S. and Nakamura, S. 2020. "Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis," *arXiv*: arXiv:2011.02128. doi: 10.48550/arXiv.2011.02128.

13. Md. Amin, A.A., Md. Islam, T., Kibria, S. and Rahman, M.S. 2019. "Continuous Bengali Speech Recognition Based On Deep Neural Network," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–6. doi: 10.1109/ECACE.2019.8679341.

14. Dettmers, T., Lewis, M., Belkada, Y. and Zettlemoyer, L. 2022. "GPT3.int8(): 8-bit Matrix Multiplication for Transformers at Scale," *Advances in Neural Information Processing Systems*, **35**: 30318–30332.

15. "Analysis of Whisper Automatic Speech Recognition Performance on Low Resource Language | Jurnal Pilar Nusa Mandiri." Accessed: Dec. 16, 2025. [Online]. Available: https://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/4633

16. Pham, N.-Q., Nguyen, T.-S., Niehues, J., Müller, M., Stüker, S. and Waibel, A. 2025. "Very Deep Self-Attention Networks for End-to-End Speech Recognition," *arXiv*: arXiv:1904.13377. doi: 10.48550/arXiv.1904.13377.

17. "Optimal ratio for data splitting - Joseph - 2022 - Statistical Analysis and Data Mining: The ASA Data Science Journal - Wiley Online Library." Accessed: Dec. 15, 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/sam.11583

18. Toraman, C., Yilmaz, E.H., Şahinuç, F. and Ozcelik, O. 2023. "Impact of Tokenization on Language Models: An Analysis for Turkish," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, **22**(4): 116:1-116:21.

19. Wei, J. *et al.* 2022. "Finetuned Language Models Are Zero-Shot Learners," Feb. 08, 2022, *arXiv*: arXiv:2109.01652. doi: 10.48550/arXiv.2109.01652.