IWSLT 2025 Indic Track System Description Paper: Speech-to-Text Translation from Low-Resource Indian Languages (Bengali and Tamil) to English

Sayan Das[1], **Soham Chaudhuri**[2], **Dipanjan Saha**[3], **Dipankar Das**[4], **Sivaji Bandyopadhyay**[5]
[1,3,4,5]Dept. of CSE, Jadavpur University, Kolkata, India
[2]Dept. of EE, Jadavpur University, Kolkata, India
{sayan.das200216, sohamchaudhuri.12.a.38, sahadipanjan6, dipankar.dipnil2005, sivaji.cse.ju}@gmail.com

Abstract

Multi-language Speech-to-Text Translation (ST) plays a crucial role in breaking linguistic barriers, particularly in multilingual regions like India. This paper focuses on building a robust ST system for low-resource Indian languages, with a special emphasis on Bengali and Tamil. These languages represent the Indo-Aryan and Dravidian families, respectively. The dataset used in this work comprises spoken content from TED Talks and conferences, paired with transcriptions in English and their translations in Bengali and Tamil. Our work specifically addresses the translation of Bengali and Tamil speech to English text, a critical area given the scarcity of annotated speech data. To enhance translation quality and model robustness, we leverage cross-lingual resources and word-level translation strategies. The ultimate goal is to develop an end-to-end ST model capable of real-world deployment for underrepresented languages.

1 Introduction

Speech-to-Text Translation (ST) has seen significant progress in recent years, driven by advancements in deep learning and large-scale multilingual datasets. However, the benefits of these advancements have not equally reached low-resource languages. Many Indian languages, despite being spoken by millions, lack sufficient parallel speech-text corpora to train high-performing supervised models. This paper addresses this gap by focusing on ST systems for Bengali and Tamil speech to English text, two major Indian languages that are often underrepresented in current ST research.

India's linguistic diversity presents both a challenge and an opportunity for Speech Translation research. In multilingual communities, there is a growing demand for ST systems that can facilitate communication across different linguistic groups, especially in education, healthcare, and public services. However, the shortage of human translators and limited digital resources for many Indian languages hamper efforts to build such systems. This work is motivated by the need to create language-inclusive ST models that cater to low resource Indian languages such as Bengali and Tamil. These languages are widely spoken but lack the large-scale annotated datasets available for high-resource languages like English, German, or Mandarin. By focusing on Bengali and Tamil audio-to-English text translation, this paper aims to fill a critical gap in current ST research. Additionally, the paper explores methods to overcome data scarcity, such as leveraging related language resources, incorporating multilingual pretraining, and utilizing word-level translation dictionaries. The broader goal is to build a scalable and adaptable ST pipeline that not only improves translation accuracy but also supports the integration of more languages in the future.

The work leverages a curated dataset of public speeches, including TED Talks and conference recordings, to build a system that can handle the unique linguistic and acoustic features of these languages. Our approach explores both end-to-end and cascaded architectures, aiming to strike a balance between performance and scalability. By addressing the linguistic diversity and resource limitations of these languages, we aim to contribute towards more inclusive language technology in India and beyond.

Speech-to-Text Translation (ST) research has traditionally been concentrated on high-resource languages, with significant advancements in languages such as English, German, and Mandarin. However, languages such as Bengali and Tamil, spoken by millions in multilingual regions like India, remain underrepresented in research due to the scarcity of parallel speech text corpora. The absence of large-scale annotated datasets for these languages presents a significant barrier to training high-performance models.

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 245–251
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

In recent years, there has been growing interest in leveraging existing resources from related languages and innovative techniques to overcome these challenges. For instance, the IWSLT shared tasks have provided valuable insights into the effectiveness of both cascaded and end-to-end (E2E) models for Speech-to-Text (ST) tasks. A cascaded architecture typically involves a two-step process: first, automatic speech recognition (ASR) converts speech to text, and then machine translation (MT) translates the recognized text from the source language into the target language. This approach has been shown to perform effectively in the handling of complex speech data.

In the context of low-resource languages, the cascaded model approach has demonstrated robustness, especially when training data is limited. Recent work has explored combining ASR models like OpenAI's Whisper (Radford et al., 2022), trained on multilingual data, with neural MT systems such as Helsinki-NLP/opus-mt-bn-en. This combination has proven effective in addressing the challenges posed by limited resources, especially for languages such as Bengali and Tamil. These models leverage the strengths of both components to improve translation accuracy and ensure scalability in real-world applications.

The 2023 IWSLT Evaluation Campaign, for example, evaluated offline SLT systems for translating speech from English to German, Japanese, and Chinese, using both cascaded and E2E models. The campaign highlighted the importance of combining ASR and MT components, as well as the performance improvements that could be achieved by integrating large-scale language models, data augmentation, and ensemble methods (Agarwal et al., 2023). Although E2E models are more direct, cascaded systems are particularly well-suited for low-resource languages, as they allow for leveraging pre-existing, powerful models for both ASR and MT tasks.

For Bengali and Tamil, this cascaded approach has shown promise by using pre-trained ASR models (such as Whisper) followed by fine-tuned MT models (like Helsinki-NLP/opus-mt-bn-en) to handle the translation from these languages to English. By employing this two-step architecture, the system benefits from the specific strengths of both ASR and MT, improving overall translation accuracy and ensuring adaptability to the challenges presented by these languages.

The work conducted in the IWSLT2024 Indic Track system description paper (Showrav, 2022) focuses on speech-to-text translation for multiple Indian languages, including Bengali and Tamil, and follows a similar cascaded approach to tackle the challenges of low-resource languages. This aligns closely with the goal of our paper, which aims to build a robust and scalable ST system that can effectively handle the translation of Bengali and Tamil speech to English text.

By incorporating these strategies, our work contributes to advancing Speech-to-Text Translation for low-resource languages, filling a critical gap in the research landscape, and offering a path forward for scalable models that can support the diverse linguistic landscape of India.

2 Dataset Description

The IWSLT 2025 Indic Track (Abdulmumin et al., 2025) focuses on Speech-to-Text (ST) translation between English and three low-resource Indian languages: Hindi (hi), Bengali (bn), and Tamil (ta) and vice-versa. These languages belong to two major language families—Indo-Aryan (Hindi and Bengali) and Dravidian (Tamil)—and are widely spoken across South Asia.

The dataset for this task specifically supports Speech-to-Text translation from Indic languages to English, where the source is audio in a low-resource Indian language, and the target is English text. It includes:

- Bengali and Tamil speech recordings as the source audio.
- English text transcriptions serving as the target translations.
- YAML metadata files that define audio segmentation with information like file name, offset, duration, and speaker ID.

Each language dataset is carefully aligned. Every English transcript line has a corresponding line in the target language (Hindi, Bengali, or Tamil), along with metadata in YAML format. This metadata provides information such as file name, offset, duration, and speaker ID.

The corpus is divided into training, validation, and test subsets. Each audio file corresponds to a talk by a single speaker, contributing to diverse speaking styles and accents. While the number of segments is consistent across the aligned files, token counts may differ across languages due to linguistic variations.

In our work, we focused specifically on the Bengali-to-English and Tamil-to-English translation directions. We used Bengali and Tamil audio files aligned with their English translations. The actual dataset contains significantly more than 50,000 samples for each language pair, providing a rich and diverse resource for training, evaluation, and fine-tuning.

This well-structured, multilingual dataset serves as a strong foundation for building effective Speech-to-Text translation systems for low-resource Indian languages like Bengali and Tamil.

3 System Overview

The system integrates advanced speech-to-text (ASR) and machine translation (MT) models to transcribe Bengali audio files and translate the transcriptions into English. The architecture consists of several interconnected modules, each playing a crucial role in ensuring accuracy, efficiency, and robustness. The key components of the system include an input module that accepts audio files in WAV format, which is a standard format for audio processing due to its lossless nature. Along with the audio file, a YAML metadata file is provided ,which contains the following information:

- **Offset**: The time point in the audio from which the transcription should begin.
- **Duration**: The duration of the audio clip to be processed.
- **Speaker ID**: Used to identify the speaker in case of multiple speakers in the audio file.

The ***data validation and preprocessing*** module validates the provided metadata, ensuring that the offset, duration, and speaker ID align correctly with the audio file. Audio segmentation is then performed based on the provided offset and duration to ensure precise transcription.

The ***audio processor and transcription*** module consists of two sub-modules, namely, the *Audio Chunk Extraction* and the *model integration*. In the *Audio Chunk Extraction* module, ***Librosa*** (McFee et al., 2015) and ***SoundFile*** libraries were used to extract precise segments from the original audio file based on the metadata. These libraries are efficient in processing and manipulating audio data. The ***Whisper-small model*** is loaded via the ***Hugging Face Transformers*** library (Wolf et al., 2020). Whisper is a robust, multilingual ASR model that can handle diverse languages and dialects with zero-shot capabilities (Radford et al., 2023).

The model is fine-tuned using a ***Quantized Low-Rank Adaptation (QLoRA)*** (Dettmers et al., 2023) technique on a custom dataset of $10,000$ Bengali-English audio pairs. QLoRA is a parameter-efficient fine-tuning technique that allows the model to adapt to new tasks with minimal computational overhead while retaining its generalization ability and quantization greatly reduces memory usage by reducing precision of floating points. The extracted audio chunks are then passed to the model, which transcribes the Bengali audio to Bengali text or Tamil audio to Tamil text . We have used the Kaggle free resources for all our task which provides us with 2 *Tesla P100* GPUs due to which we faced computational constraints and used QLoRA as an alternative.

In the ***Translation Module***, the system uses the **Helsinki-NLP/opus-mt-bn-en** model, a state-of-the-art model pre-trained for Bengali-to-English translation tasks. The model is fine-tuned using **CSV-aligned Bengali-English pairs**. This alignment ensures the model learns the appropriate context, improving translation accuracy. The model is also trained with the ***Seq2SeqTrainer*** framework, which is highly effective for sequence-to-sequence tasks such as translation. This method optimizes the model for better handling Bengali syntax and semantics complexities during translation.

After the completion of the *transcription* and *translation* phase, the system merges the **filename**, **transcription**, and **translated output** into a unified output. The results are stored in both CSV (for structured data) and TXT (for easy reading and further processing) formats. This allows for easy extraction and post-processing of results. To evaluate the translation quality, our system uses ***SacreBLEU*** and ***chrF++*** metrics, which are standard in machine translation tasks. We have achieved a BLEU score of 8.6945 and a chrF++ score of 35.5653 for the Bengali-English pair. These scores suggest that the system provides a reasonably high-quality translation, with strong character-level accuracy.

In addition to the Bengali-English translation system, the architecture supports **Tamil-to-English (ta-en) translation** using the **facebook / nllb-200-**

distilled-600M model (Team et al., 2022). This is a multilingual, distilled version of the NLLB-200 model by Meta AI, designed to handle translations across 200 languages with enhanced efficiency. For Tamil (language code *ta_Taml*) to English (language code *eng_Latn*), the system takes Tamil transcriptions (e.g., from speech recognition output or manually curated corpora), tokenizes them using the NLLB tokenizer, and then applies the sequence-to-sequence model for translation. The translation is done using forced BOS (beginning-of-sentence) tokens to ensure the output is directed towards English.

This translation pipeline is implemented using Hugging Face Transformers and evaluated using standard machine translation metrics. The system achieves a BLEU and chrF++ score of 13.3904 and 39.0237 on the Tamil-English test set. These results reflect a strong translation performance, especially given the morphological richness of Tamil. Like the Bengali-English pipeline, the Tamil-English system operates in an **unconstrained** setting, where no limitations are placed on the type of data or preprocessing methods used. This allows maximum flexibility in improving performance through data augmentation, custom preprocessing, or enhanced model assembling techniques.

The logic of fine-tuning the Transformers are mentioned below:

3.1 Whisper Fine-Tuning

For the bengali to english task we used *bangla-speech-processing/BanglaASR*[1] (Islam, 2023) model which is a Whisper-small fine-tuned on Bangla Mozilla Common Voice dataset and used QLoRA to fine-tune it efficiently on shared task development dataset. Before creating the Dataloader for training and validation set of the shared task, the audio files were preprocessed with Librosa and Pyloudnorm to generalize all audio files and normalize loudness. Then Dataloader was created to efficiently load data for training in a memory efficient way.

We trained our model using the ***Seq2SeqTrainingArguments*** class with a batch size of 4 per device and a ***gradient accumulation*** of 8 steps. The learning rate was set to 1×10^{-4}, and training was conducted for 3 epochs. We enabled mixed-precision training using FP16. For parameter-efficient fine-tuning, we used LoRA with a rank of 8, scaling factor (***lora_alpha***) of 32, and a dropout rate of 0.1. LoRA was applied specifically to the attention layers, targeting the ***q_proj*** and ***v_proj*** modules.

Similarly, for the Tamil to English task we used the *vasista22/whisper-tamil-small*[2] model, which is also a Whisper-small fine-tuned on multiple publicly available Tamil dataset. The parameters were kept the same as while fine-tuning the whisper-bangla model.

3.2 MarianMT Fine-Tuning

The *Helsinki-NLP/opus-mt-bn-en* model[3] from the MarianMT family (Junczys-Dowmunt et al., 2018) was fine-tuned to perform Bengali-to-English translation using shared task's development dataset of aligned sentence pairs in CSV format. The dataset was prepared by merging Bengali transcriptions and their corresponding English translations based on identical audio file names. This ensured accurate one-to-one alignment without the need for external alignment tools such as *fast_align* or *awesome-align*. The fine-tuning procedure (Li et al., 2021) employed a ***batch size*** of 8 and a ***learning rate*** of 3×10^{-5}. Training was conducted over 5 epochs. Preprocessing included tokenizing the source and target sentences with truncation and padding up to a maximum length of 128 tokens. The model was trained using the Hugging Face ***Seq2SeqTrainer*** framework. Evaluation was performed after every epoch, and the best checkpoint was selected based on validation loss. This approach helped the model learn both syntactic and semantic structures effectively, resulting in improved translation quality from Bengali to English.

3.3 NLLB Unconstrained Translation

The *facebook/nllb-200-distilled-600M*[4] (Team et al., 2022) model has been employed for Tamil-to-English translation without additional fine-tuning. This distilled multilingual model was pretrained on a large corpus covering over 200 languages, including Tamil, so while no further task-specific adaptation was performed, the translation is not strictly zero-shot. The translation pipeline starts by tokenizing the Tamil **(tam_Taml)** input sequences

[1] https://huggingface.co/bangla-speech-processing/BanglaASR

[2] https://huggingface.co/vasista22/whisper-tamil-small

[3] https://huggingface.co/Helsinki-NLP/opus-mt-bn-en

[4] https://huggingface.co/facebook/nllb-200-distilled-600M

and specifying English **(eng_Latn)** as the target language using the ***forced_bos_token_id*** parameter. The encoder-decoder architecture of the model generates English output directly. This approach reduces training overhead while leveraging the model's strong pretrained multilingual capabilities to produce effective Tamil-to-English translations without additional supervised training.

Model Used	**BLEU**	**chrF++**
Whisper + Helsinki-NLP/opus-mt-bn-en	8.69	35.56
Whisper + NLLB-200-distilled-600M for ta-en	13.39	39.02

Table 1: BLEU and chrF++ scores for Bengali-English and Tamil-English translation systems on test set

4 Workflow

Figure 1 illustrates the basic workflow of our system which we have named **SpeechSync**, consists of four key stages:

- **Input Processing**: Audio files are provided to the system through a predefined directory or batch input process. The Input Module then validates and preprocessed these files. It extracts metadata from associated YAML files, ensuring that the audio is correctly segmented and ready for transcription.

- **Transcription**: The Transcription Module leverages a fine-tuned Whisper model to convert the segmented audio into text. This model, known for its multilingual and zero-shot capabilities, ensures accurate and reliable transcription across both Bengali and Tamil languages.

- **Translation**: The translated output is generated using language-specific models such as Helsinki-NLP/opus-mt-bn-en or NLLB. These models are either fine-tuned on aligned data or used in an unconstrained setup to produce high-quality, contextually relevant English translations from the source text.

- **Output Delivery**: The system compiles the original filename, transcription, and translated output into structured TXT and CSV formats. These outputs are made available for download, enabling users to easily integrate the translated results into their workflows or downstream applications.

This streamlined and modular workflow enables efficient conversion from audio to translated text, supporting diverse use cases in multilingual environments and helping bridge communication gaps across languages.

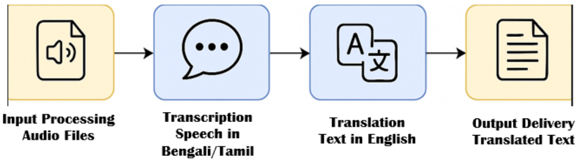

Figure 1: Basic Workflow of the SpeechSync System

5 Limitations

While we acknowledge the significant challenges ahead, such as the shortage of multilingual individuals and insufficient data for certain languages, we are determined to find innovative solutions. Some of the limitations of our current approach include:

- Language Support: The input module currently supports only one language at a time. If an audio file contains multiple languages (e.g., a conversation with code-switching between Tamil and English or Bengali and Hindi), the application processes only the primary language while ignoring others. This limitation restricts the system's effectiveness in multilingual environments and conversational scenarios common in many Indian contexts.

- Processing Time: The transcription and translation modules are computationally intensive. This is especially true when using models like Whisper and Helsinki-NLP/opus-mt-bn-en or NLLB, which require substantial processing resources. To address this, we are exploring model optimization strategies, such as quantization, reduced precision inference (e.g., FP16 or INT8), and parallel processing to enhance efficiency and throughput.

- Translation Performance for Low-Resource Pairs: While the system performs reasonably well for both Bengali-to-English and Tamil-to-English translation tasks, the Bengali-to-English translation still hovers around baseline performance. This is due to limited high-quality parallel data for Bengali, which impacts the model's ability to capture complex sentence structures and semantics. In contrast, Tamil-to-English translation demonstrates relatively improved performance, but further refinement is still necessary to handle domain-specific vocabulary and informal language constructs accurately.

Despite these limitations, we remain committed to enhancing system performance. Ongoing research focuses on expanding language support, improving inference speed, and increasing the quality of both transcription and translation outputs. These efforts are part of a broader goal to make **SpeechSync System** a reliable and efficient multilingual speech-to-text translation system, particularly for underrepresented Indian languages.

6 Future Work

While the current version of the **SpeechSync System** demonstrates strong performance in Bengali-to-English and Tamil-to-English speech translation, there remain several promising directions for future improvement and expansion.

One important enhancement is the incorporation of **speaker diarization and multi-speaker handling**. This would allow the system to differentiate between individual speakers in a single audio stream. This feature is essential for accurately processing meetings, interviews, or conversational datasets. By integrating diarization models, the system could associate transcription segments with specific speaker labels, improving readability and structure.

Another potential development is **real-time streaming transcription and translation**. This would significantly expand the system's usability in live scenarios such as conferences, classrooms, and emergency response settings. Achieving this would involve optimizing the current pipeline to minimize latency and memory usage, allowing for faster and more efficient processing.

Currently, the ASR outputs include only basic punctuation, which can hinder readability. To address this, future iterations will aim to integrate **advanced punctuation and formatting**. This includes accurate sentence boundaries, speaker turn indicators, and proper capitalization. These enhancements would make both transcriptions and translations more natural and easier to follow.

Further improvement could come from **multi-modal integration**, where additional visual cues such as lip movements or gestures are used to aid transcription accuracy, especially in noisy or acoustically challenging environments. This would position the system for use in richer, context-aware applications like video subtitling or assistive communication.

7 Conclusion

In summary, our key contributions lie in the rigorous experimentation conducted to identify effective models for speech translation, especially for low-resource languages like Bengali and Tamil. We perform extensive preprocessing of data to ensure quality and suitability for training. The proposed solution establishes a robust pipeline, including code development and workflow setup, allowing for efficient transcription and translation tasks. The training and experimentation were focused on Bengali to English translation for an in-depth analysis, which included fine-tuning Whisper for transcription tasks using LoRA and Helsinki-NLP/opus-mt-bn-en for translation. In addition, we extended our work to Tamil to English translation using the facebook/nllb-200-distilled-600M model, which was fine-tuned on Tamil-English parallel data to improve translation quality and generalization. This enabled the system to support multilingual speech-to-text translation more broadly. Close monitoring of performance metrics, including BLEU and chrF++ scores, was carried out to assess model performance and guide future improvements.

This paper is committed to advancing speech translation (ST) technology for low-resource languages. Through the creation of dedicated datasets and the development of robust models for both Bengali and Tamil, our aim is to facilitate seamless communication and accessibility across diverse linguistic communities, ultimately promoting inclusivity and empowerment.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Tsz Kin Lam, Danni Liu, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Salima Mdhaffar, Yasmin Moslem, Kenton Murray, Satoshi

Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John E. Ortega, Sara Papi, Pavel Pecina, Peter Polák, Piotr Połeć, Beatrice Savoldi, Nivedita Sethiya, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Patrick Wilken, Rodolfo Zevallos, Vilém Zouhar, and Maike Züfle. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austia (in-person and online). Association for Computational Linguistics. To appear.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Md Saiful Islam. 2023. Transformer based whisper bangla asr model.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Raymond Li, Wen Xiao, Lanjun Wang, Hyeju Jang, and Giuseppe Carenini. 2021. T3-vis: visual analytic for training and fine-tuning transformers in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 220–230, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *SciPy*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Tushar Talukder Showrav. 2022. An automatic speech recognition system for bengali language based on wav2vec2 and transfer learning. *arXiv preprint arXiv:2209.08119*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.