ADAPT–MTU HAI at IWSLT2026: Robust Cascaded Speech Translation for Bhojpuri–Hindi and Irish–English

Pournima Sonawane
Munster Technological University
Cork, Ireland
pournima.sonawane@mymtu.ie

Haithem Afli
Munster Technological University
Cork, Ireland
haithem.afli@mtu.ie

Abstract

Low-resource speech translation remains challenging due to limited data, weak ASR support, and error propagation in cascaded systems. We present the ADAPT–MTU HAI submission to the IWSLT 2026 Low-Resource Speech Translation task, a robust cascaded framework combining Whisper-based ASR and NLLB-200 multilingual translation for Bhojpuri→Hindi and Irish→English language pairs.

We evaluate multiple ASR models and routing strategies, including direct and pivot-based translation. For Bhojpuri→Hindi, the best configuration (Whisper-large-v3 and direct NLLB) achieves BLEU 25.59, chrF++ 42.48, and TER 63.83 on the full development set, outperforming pivot and copy baselines. For Irish→English, replacing Whisper with a language-specific Wav2Vec2 ASR model improves ASR coverage from 94.8% to 100% on the test set while maintaining low repetition rates.

Our findings highlight the critical role of ASR quality in downstream translation performance, the conditional benefits of pivot translation, and the effectiveness of modular cascaded architectures for low-resource speech translation.

1 Introduction

Speech translation plays a critical role in enabling communication across linguistically diverse communities. While recent advances in neural speech and language technologies have led to strong performance for high–resource language pairs, progress for low-resource languages remains limited. Languages such as Bhojpuri and Irish are underrepresented in available datasets and benchmarks and are characterized by limited annotated data, high speech variability, and insufficient support in existing automatic speech recognition (ASR) and machine translation (MT) systems (Vistatec, 2025). These challenges significantly hinder the development of reliable and generalizable speech translation pipelines.

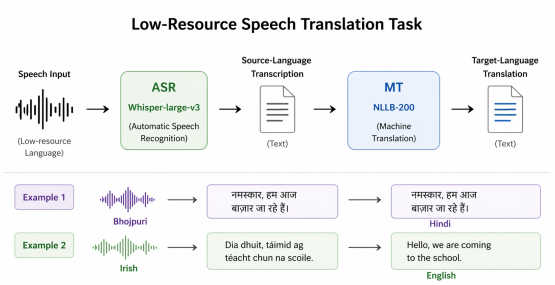

Figure 1: Overview of the low-resource speech translation task. Speech input in a source language (e.g., Bhojpuri or Irish) is first transcribed using an ASR system and then translated into the target language using a machine translation model.

The IWSLT 2026 Low-Resource Speech Translation task[1] provides an important benchmark addressing these limitations, focusing on realistic multilingual scenarios where data scarcity, robustness, and system adaptability are central concerns. Although end-to-end speech translation approaches have attracted increasing attention due to their conceptual simplicity, they typically require large amounts of parallel speech data and remain unstable in low-resource conditions (Cheng et al., 2021). In contrast, cascaded architectures, which decompose the task into ASR followed by MT, continue to offer a practical and effective alternative, particularly when leveraging strong pretrained multilingual models (Romney Robinson et al., 2024).

As illustrated in Figure 1, the cascaded speech translation pipeline consists of a task that can be formulated as a two-stage pipeline in which speech input is first transcribed into source-language text using an ASR model and subsequently translated into the target language using an MT system. This modular formulation allows each component to be

[1] See https://iwslt.org/2026/low-resource for details.

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 58–67
July 3-4, 2026 ©2026 Association for Computational Linguistics

optimised independently and provides flexibility in selecting models and routing strategies, which is particularly beneficial in low-resource settings.

In this paper, we present the ADAPT–MTU HAI submission to the IWSLT 2026 (Adelani et al., 2026) shared task, a robust cascaded speech translation system targeting two low-resource language pairs: Bhojpuri→Hindi and Irish→English. Our approach combines Whisper-based ASR with NLLB-200 multilingual translation, enabling multiple flexible routing strategies, including both direct and pivot-based translation. We systematically investigate the impact of ASR model selection, translation routing, and lightweight post-processing on end-to-end system performance.

Our findings highlight the central role of ASR quality in determining downstream translation performance. For Bhojpuri→Hindi, improvements in ASR enable direct translation to outperform pivot-based strategies. For Irish→English, the use of a language-specific ASR model substantially improves coverage and robustness. Overall, this work demonstrates that carefully designed cascaded pipelines, supported by strong pretrained models and minimal post-processing, provide a reliable and efficient solution for low-resource speech translation.

2 Background and Related Work

Low-resource multilingual speech translation has become an important research area in natural language processing and speech technologies, driven by the need to enable communication across underserved linguistic communities. While modern systems achieve strong performance for high–resource language pairs such as English→French or Spanish→English, many real-world languages remain underrepresented due to limited annotated data, scarce linguistic resources, and insufficient tool support. These limitations pose significant challenges for both automatic speech recognition (ASR) and machine translation (MT).

A fundamental challenge in low-resource speech translation lies in the interaction between ASR and MT within cascaded pipelines. Errors introduced during speech recognition propagate directly into the translation stage, making overall system performance highly sensitive to transcription quality. Despite advances in multilingual modeling, only a small fraction of the world's languages are adequately supported by current systems (Vistatec, 2025). Traditional cascaded approaches often rely on large amounts of transcribed and parallel data, which are typically unavailable in low-resource settings, motivating research into alternative modeling strategies (Bansal et al., 2018).

Recent advances in large-scale pretrained models have significantly improved the feasibility of low-resource speech translation. Models such as Whisper leverage large-scale weak supervision to provide robust multilingual ASR, while NLLB-200 enables many-to-many translation across a wide range of languages, including low-resource ones. These models facilitate transfer learning, reduce the need for language-specific training, and support flexible strategies such as pivot-based translation. However, challenges remain, including pronunciation variability, dialectal diversity, limited corpora, and the lack of standardised evaluation benchmarks.

The broader development of AI systems for low-resource languages is also shaped by systemic challenges. The Stanford HAI whitepaper highlights persistent issues such as data scarcity, limited community-driven datasets, and imbalances in research and development priorities (Pava et al., 2025). These factors emphasise the importance of designing modular, adaptable, and resource-efficient systems capable of operating under constrained conditions.

Within the speech translation literature, a key distinction exists between cascaded and end-to-end approaches. End-to-end models aim to directly map speech in one language to text (or speech) in another, offering conceptual simplicity and reduced latency. However, they typically require substantial training data and remain unstable in low-resource scenarios (Cheng et al., 2021). In contrast, cascaded systems, which decompose the task into ASR followed by MT, continue to demonstrate strong and reliable performance, particularly when combined with pretrained multilingual models (Romney Robinson et al., 2024).

Empirical evidence from recent IWSLT shared tasks supports this observation. Systems combining Whisper-based ASR with NLLB-based translation frequently outperform standalone or end-to-end approaches when fine-tuning resources are limited. Similarly, (Meng and Anastasopoulos, 2025) show that end-to-end models such as SeamlessM4T perform well primarily when their pretraining data closely matches the target language, limiting their effectiveness in truly low-resource settings. The

Whisper model itself has demonstrated strong multilingual ASR capabilities, making it a reliable front-end for cascaded pipelines (NVIDIA, 2024; OpenAI, 2023).

A growing body of work explores strategies for mitigating data scarcity. For example, (Bhogale et al., 2023) propose methods for constructing large-scale aligned speech-text datasets, while (Moslem, 2024) demonstrate the effectiveness of synthetic audio data and back-translation for improving Irish-to-English speech translation. Complementary approaches focus on mining parallel data from weakly aligned or comparable sources. (Afli et al., 2013) introduce a multimodal framework for extracting parallel phrases from comparable corpora by leveraging ASR-transcribed audio and information retrieval techniques, showing that useful parallel data can be constructed even in the absence of explicitly aligned resources.

On the machine translation side, multilingual pretrained models such as NLLB-200 have demonstrated strong adaptability across diverse language pairs. Studies on Telugu–English and Indonesian–Javanese translation show that fine-tuning and data filtering can significantly improve translation quality even when parallel data is limited (Govil et al., 2026; Yuliawati et al., 2025). In the specific context of Irish, (Lankford et al., 2022) provide a detailed human evaluation of Transformer-based English–Irish neural machine translation systems, showing that architectural choices and subword modeling significantly impact both fluency and adequacy. Their findings highlight the importance of fine-grained evaluation beyond automatic metrics, particularly for low-resource languages where subtle linguistic phenomena may not be fully captured by BLEU or related scores.

Hybrid systems combining pretrained ASR encoders and MT decoders have also shown promise; (Avila and Crego, 2025) present a compact pipeline integrating Whisper and NLLB with minimal fine-tuning, achieving competitive performance under resource constraints.

Overall, the literature highlights three key insights relevant to this work: (i) the critical role of ASR quality in cascaded speech translation pipelines, (ii) the effectiveness of large-scale multilingual pretrained models for addressing data scarcity, and (iii) the trade-offs between cascaded and end-to-end architectures in low-resource scenarios. Building on these insights, the present work investigates how ASR selection, translation routing strategies, and lightweight post-processing can be combined to improve robustness and performance for Bhojpuri→Hindi and Irish→English speech translation.

Language Pair	Hours	Segments
Bhojpuri→Hindi	≈23h	252
Irish→English	≈206h	722

Table 1: Overview of IWSLT 2026 test datasets.

3 Dataset and Data Preparation

Experiments were conducted using the official datasets provided for the IWSLT 2026 Low-Resource Speech Translation task. We focus on two language pairs: Bhojpuri→Hindi and Irish→English, which represent distinct low-resource scenarios in terms of data availability, linguistic characteristics, and ASR support.

Table 1 summarises the evaluation datasets. The Bhojpuri→Hindi dataset consists of approximately 23 hours of speech, primarily drawn from news and conversational domains. The official test set includes 252 audio segments, each paired with a Hindi reference translation. This setting represents an extremely low-resource scenario, as Bhojpuri lacks large-scale annotated corpora and well-established ASR models (Upadhyay, 2026a).

In contrast, the Irish→English dataset is substantially larger, comprising approximately 206 hours of speech collected from both natural and synthetic sources, including Wikimedia Speech, Tatoeba, and EUbookshop. The test set contains 722 audio segments spanning multiple domains such as news, public speeches, and narrative recordings. Despite the larger data volume, Irish remains challenging due to phonetic variability, dialectal diversity, and limited representation in mainstream ASR systems (Upadhyay, 2026b).

The datasets are provided as segmented audio files accompanied by timestamp metadata in TSV format. However, the raw `stamped.tsv` files required additional preprocessing due to inconsistencies such as missing headers and irregular filename formatting (e.g., absent zero-padding in segment identifiers). To address these issues, we developed a custom manifest construction pipeline that parses TSV entries, reconstructs valid file paths, and verifies the integrity of all audio samples.

The final manifests were stored in a structured

CSV format, including fields such as `audio_path`, `sample_id`, `src_lang`, `tgt_lang`, `split`, and timing metadata. After preprocessing, both datasets achieved full coverage, with 252 valid samples for Bhojpuri→Hindi and 722 valid samples for Irish→English, ensuring consistent and reliable input for downstream ASR and translation experiments.

4 System Architecture

4.1 Overview

We propose **WhiNN-ST**, a cascaded speech translation framework in which speech input is first transcribed using an automatic speech recognition (ASR) model and subsequently translated using a multilingual machine translation (MT) model. This modular design enables independent optimization of each component while providing flexibility in model selection and translation routing, which is particularly beneficial in low-resource settings.

As illustrated in Figure 2, audio input is processed by Whisper-large-v3 to produce a source-language transcription, which is then passed to the NLLB-200 model for translation. The MT component supports both direct and pivot-based routing, enabling multiple translation pathways within a unified framework. Specifically, we evaluate the following configurations:

- **BHO→HI** (direct)
- **BHO→EN→HI** (pivot via English)
- **GA→EN** (direct)
- **GA→HI→EN** (pivot via Hindi)

4.2 ASR Component

The ASR component is based on the Whisper model family, which provides strong multilingual transcription capabilities through large-scale weak supervision. We evaluate three variants: Whisper-small, Whisper-medium, and Whisper-large-v3.

Prior to inference, all audio inputs are standardised using `librosa`, including resampling to 16 kHz and conversion to mono, ensuring consistent input quality across datasets.

Whisper-small offers computational efficiency but produces unstable and error-prone transcriptions in low-resource settings, particularly for Irish. Whisper-medium improves fluency and grammatical structure but occasionally sacrifices faithfulness to the source utterance. Whisper-large-v3 provides the best overall performance, achieving higher transcription accuracy and consistency across both language pairs, and is therefore adopted as the primary ASR component in the final system.

4.3 MT Component

For the translation stage, we employ the NLLB-200 distilled 600M multilingual model, which supports many-to-many translation across a broad range of languages. This model is particularly suited to low-resource scenarios due to its strong cross-lingual generalization capabilities (Lankford, 2024).

The MT component is configured to support both direct and pivot-based translation strategies. This enables flexible routing across Bhojpuri, Hindi, Irish, and English, allowing us to systematically evaluate the impact of intermediate pivot languages on translation quality.

4.4 Post-processing

We introduce a lightweight post-processing step to address script inconsistencies observed in ASR outputs. Specifically, Bhojpuri transcriptions produced by Whisper-large-v3 occasionally contain a mixture of Devanagari and Urdu script. To ensure consistency, we apply a Unicode-based normalization step that removes non-Devanagari characters prior to translation.

Although this step does not alter the semantic content, it improves the stability and reliability of downstream translation, particularly for direct Bhojpuri→Hindi configurations.

4.5 Alternative Models

In addition to the primary Whisper and NLLB pipeline, we evaluate several alternative architectures and ASR front ends.

SeamlessM4T-v2-large is explored as an end-to-end speech translation model; however, it fails to produce usable outputs in our setup (0% non-empty outputs on the Irish test set) and is therefore excluded from further analysis.

For Irish ASR, the MCV Fleurs Combined Irish ASR model produces complete outputs but exhibits higher noise levels compared to Whisper-large-v3 (average repetition score: 0.051 vs. 0.027). Another model (Eimhin03) fails to run due to missing configuration files, highlighting practical limitations in model deployment.

We further evaluate a Wav2Vec2 large XLSR-53 Irish ASR model, a CTC-based architecture specif-

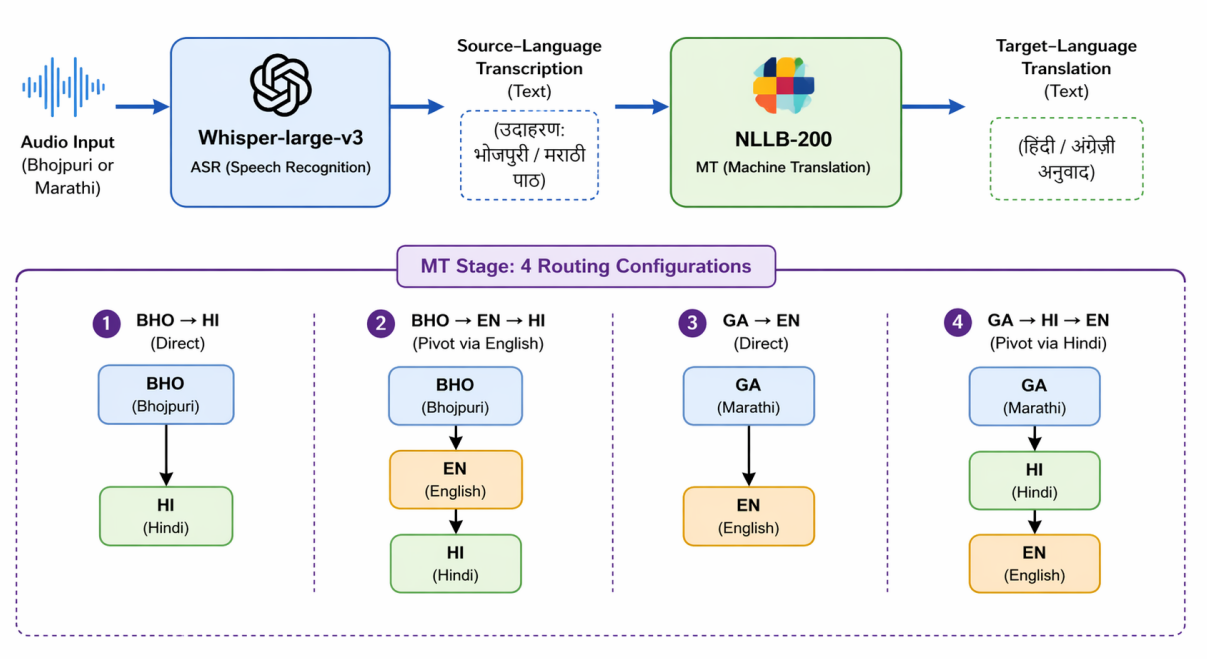

Figure 2: WhiNN-ST cascaded speech translation pipeline. Audio input is transcribed using Whisper-large-v3 and subsequently translated using NLLB-200. Both direct and pivot-based translation routes are supported for Bhojpuri–Hindi and Irish–English.

ically tailored for Irish. This model achieves 100% ASR coverage on all 722 test segments and maintains low repetition rates (0.039 for both direct and pivot configurations), demonstrating strong robustness in the Irish setting.

On the Bhojpuri side, a Wav2Vec2 Bhojpuri ASR model (BHO-M 60) introduces tokenization artifacts that significantly degrade downstream translation quality, as confirmed by evaluation metrics. This further reinforces the importance of ASR quality in cascaded speech translation pipelines.

5 Experiments

5.1 Experimental Setup

All experiments were conducted in Python 3 using Google Colab with a T4 GPU accelerator. Initial pilot runs were performed on CPU, but inference times extended to several hours, making large-scale experimentation impractical. The workflow was therefore migrated to GPU execution, reducing ASR runtime for the Irish test set to approximately 47 minutes and translation runtime to approximately 12 minutes.

To accommodate GPU memory constraints and improve fault tolerance, the pipeline was split into two sequential stages: ASR and MT. Intermediate outputs were stored as CSV files, allowing experiments to resume from the last completed stage without requiring full recomputation. This setup was particularly useful for long test-time runs and for systematically comparing multiple ASR and routing configurations.

5.2 Evaluation

The official IWSLT 2026 test references are blind and therefore not publicly available. As a result, blind inference was used to generate submission files in the required format for the shared task.

For Bhojpuri→Hindi, local evaluation was carried out on the combined 2024–2025 development datasets, comprising 1056 segments with aligned Hindi reference translations. This enabled the computation of standard MT metrics, including BLEU, chrF++, TER, and COMET. Word error rate (WER) was not computed because aligned source-language transcript references were unavailable for the ASR outputs.

For Irish→English, no equivalent public development set with reference translations was available. Evaluation for this language pair therefore relied on proxy diagnostics, including non-empty ASR coverage, repetition scores, and output length ratios, together with comparisons across alternative ASR front ends and routing configurations.

ASR	MT Route	BLEU	chrF++	TER	COMET
Wh.-med	Direct	0.00	0.00	100.00	0.1999
Wh.-med	Pivot (EN)	0.00	3.56	92.08	0.2668

Table 2: Phase 1: Bhojpuri→Hindi development evaluation with Whisper-medium and NLLB-200-distilled-600M.

MT Route	BLEU	chrF++	TER
Copy baseline	9.91	27.74	82.21
Direct NLLB	**25.59**	**42.48**	**63.83**
Pivot via EN	21.62	37.27	70.72

Table 3: Phase 2: Bhojpuri→Hindi development evaluation (full 1056 segments) with Whisper-large-v3 and NLLB-200. Best results are shown in **bold**.

5.3 Experiment Phases and Results

5.3.1 Phase 1: Baseline (Whisper-medium and NLLB-200)

The earliest working system combined Whisper-medium with the NLLB-200 distilled 600M model. Results on the Bhojpuri→Hindi development set is shown in Table 2, which effectively shows weak baseline results. The direct route failed to produce useful translations, yielding BLEU 0.00 and TER 100.00. In contrast, pivoting through English produced a modest improvement, with chrF++ of 3.56, TER of 92.08, and COMET of 0.2668. Although performance remained very weak overall, this result suggested that pivot-based translation could partially compensate for deficiencies in the direct route under a weaker ASR front end and supports the motivation for improvements.

5.3.2 Phase 2: Improved Pipeline (Whisper-large-v3 and NLLB-200)

The next phase upgraded the ASR front end to Whisper-large-v3 and replaced the conservative copy baseline with a true direct NLLB translation route. As shown in Table 3, this produced substantial improvements across all metrics. The direct NLLB route achieved BLEU 25.59, chrF++ 42.48, and TER 63.83 on the full 1056-segment development set, outperforming both the copy baseline and the pivot route. This is one of the most important tables in this study, clearly demonstrating that direct translation outperforms pivot in terms of ASR.

This phase highlights one of the key findings: in Phase 1, pivoting was preferable because the direct path was too weak to produce usable translations. In Phase 2, however, once ASR quality improved sufficiently, direct translation became clearly superior. This shows that pivot translation is not inherently superior; rather, it serves primarily as a compensatory strategy when the direct route is unreliable.

5.3.3 Phase 3: Irish→English Submission

For Irish→English, an initial full test-set run using Whisper-large-v3 produced non-empty ASR outputs for 685 of the 722 segments, corresponding to 94.8% coverage. To improve robustness, we replaced the ASR front end with `cpierse/wav2vec2-large-xlsr-53-irish`.[2], a Wav2Vec2 CTC model specifically trained for Irish at 16 kHz. This alternative achieved 100% non-empty ASR coverage across the full test set.

Both direct and pivot-based translation routes produced complete 722-line submissions, with repetition scores of 0.039 for both routes (Table 5). Compared with Whisper-large-v3, the Wav2Vec2 model produced shorter outputs on average (11.00 words direct, 10.14 words pivot), suggesting a more compact transcription style. Given its full ASR coverage and stable diagnostic profile, the Wav2Vec2-based system was selected as the primary Irish submission.

5.3.4 Phase 4: Optional Model Benchmarks

To further assess robustness, we evaluated several alternative models for both language pairs. For Bhojpuri→Hindi, Table 4 reports benchmark results on a 50-sample subset. The custom Bhojpuri ASR model Vakyansh.[3], performed dramatically worse than Whisper-large-v3 across all metrics, with BLEU of 0.33 and TER of 170.47 for the direct route. These results confirm that the well-integrated Whisper-large-v3 and NLLB pipeline substantially outperforms the more specialized Bhojpuri-specific alternative in our experimental setting.

For Irish→English, Table 5 compares ASR front ends on the full 722-segment test set. The Wav2Vec2 Irish model achieved 100% non-empty coverage, compared with 94.8% for Whisper-large-v3, while maintaining competitive repetition rates. SeamlessM4T-v2 failed to generate usable outputs and was therefore excluded from further consider-

[2]https://huggingface.co/cpierse/wav2vec2-large-xlsr-53-irish

[3]https://huggingface.co/Harveenchadha/vakyansh-wav2vec2-bhojpuri-bhom-60

ation. It provides a useful diagnostic comparison with strong evidence for ASR improvements (coverage from 94.8% - 100%).

Finally, Table 6 presents a unified diagnostic comparison summary across both language pairs and all datasets. The results confirm consistent 100% ASR coverage in the final selected systems, low repetition rates, and well-controlled output length ratios throughout the pipeline.

5.4 Submission Summary

The final submission package follows the IWSLT 2026 shared-task format. For Bhojpuri→Hindi, the primary system uses Whisper-large-v3 with direct NLLB translation and Devanagari post-processing, while the contrastive system uses the pivot route via English. For Irish→English, the primary system uses `cpierse/wav2vec2-large-xlsr-53-irish` with direct NLLB translation, and the contrastive system uses the pivot route via Hindi. All submission files were verified to contain the correct number of lines: 252 for Bhojpuri→Hindi and 722 for Irish→English.

6 Results and Discussion

The experimental results highlight the importance of ASR quality in determining the behaviour of cascaded speech translation systems. In Phase 1, pivoting through English yielded better performance than the direct route under the weaker Whisper-medium baseline, particularly in COMET and chrF++. This suggests that an intermediate high–resource language can stabilise translation when the direct path is insufficiently robust. However, once the ASR front end was strengthened with Whisper-large-v3 and the direct route was implemented as a true NLLB-based translation path rather than a conservative copy baseline, direct translation became the best-performing configuration across all surface metrics. This reversal is one of the key findings of the study that pivot translation is not inherently better but instead serves as a compensatory strategy when the direct path is weak.

The results also show that relatively simple engineering choices can meaningfully improve system robustness. Audio preprocessing with `librosa`, including resampling to 16 kHz mono, contributed to more stable ASR behaviour, particularly for Irish, where the source audio exhibited greater variability in sampling rates. Similarly, the Devanagari normalization step applied to Bhojpuri ASR outputs addressed a practical multilingual ASR issue: Whisper-large-v3 occasionally mixed Urdu and Devanagari characters in the transcription. Removing non-Devanagari characters improved script consistency and downstream translation reliability without requiring any additional model training.

For Irish→English, the main bottleneck was not the MT stage but ASR coverage. Whisper-large-v3 produced non-empty outputs for 94.8% of the test set, whereas the Irish-specific Wav2Vec2 model successfully processed all 722 segments, yielding 100% non-empty coverage with low repetition rates for both direct and pivot routes. The unified diagnostic summary in Table 6 confirms that the final pipeline achieved full coverage for both language pairs, with bad-repetition rates of at most 0.004 and length ratios within a stable range of 1.01–1.14. These results indicate that the selected Systems produce consistently well-formed outputs with limited degeneration.

The optional model comparisons further reinforce a broader engineering lesson: more specialized or more ambitious models do not necessarily outperform a well-integrated baseline. SeamlessM4T-v2-large, despite its appeal as an end-to-end architecture failed to produce usable outputs in our experimental setup, likely due to implementation or compatibility issues in the Colab environment. Likewise, the custom Bhojpuri ASR model introduced severe tokenization artifacts that made downstream translation effectively unusable. Taken together, these experiments validate the final system design: Whisper-large-v3 and NLLB-200 provide the most reliable and effective combination for the two language pairs considered in this work. More broadly, the observed improvements are consistent with previous findings showing that transformer-based multilingual models offer strong advantages for low-resource translation settings (Lankford, 2024).

7 Conclusion

This paper presented **WhiNN-ST**, the ADAPT–MTU HAI submission to the IWSLT 2026 Low-Resource Speech Translation task. The proposed system adopts a cascaded architecture that combines Whisper-based ASR with NLLB-200 multilingual machine translation, together with lightweight audio preprocessing and script normal-

ASR Model	Route	BLEU	chrF++	TER
Whisper-large-v3	Direct	**22.82**	**43.74**	**63.13**
Whisper-large-v3	Pivot	18.22	37.65	76.17
Custom BHO ASR	Direct	0.33	11.65	170.47
Custom BHO ASR	Pivot	0.28	11.08	171.09

Table 4: Phase 4: Bhojpuri→Hindi benchmark on a 50-sample subset comparing Whisper-large-v3 and Vakyansh Bhojpuri ASR. Best results are shown in **bold**.

ASR Front-end	Route	N	ASR%	Rep.↓	Len.R
Whisper-large-v3	Direct	722	94.8	0.027	—
Whisper-large-v3	Pivot	722	94.8	0.029	—
Wav2Vec2 Irish	Direct	722	**100.0**	**0.039**	1.144
Wav2Vec2 Irish	Pivot	722	**100.0**	0.039	1.050
SeamlessM4T-v2	Direct	—	0.0	—	—

Table 5: Phase 4: Irish→English optional model comparison on 722 test segments. Rep. = average repetition rate; Len.R = average length ratio (output/input words); ASR% = non-empty ASR coverage.

ization.

Our experiments demonstrate that ASR quality is a decisive factor in end-to-end speech translation performance. For Bhojpuri→Hindi, improving the ASR front end enabled direct translation to outperform both pivot-based and copy-based alternatives. For Irish→English, replacing Whisper-large-v3 with an Irish-specific Wav2Vec2 model improved ASR coverage from 94.8% to 100%, resulting in a more robust final submission. The study also shows that lightweight preprocessing and post-processing steps can provide practical gains without requiring additional model training.

Overall, the results confirm that carefully designed cascaded pipelines remain a robust and effective solution for low-resource speech translation. The best Bhojpuri→Hindi configuration achieved BLEU 25.59, chrF++ 42.48, and TER 63.83 on the full development set, while the selected Irish→English system achieved full ASR coverage on the complete 722-segment test set. These findings provide further evidence that these approaches are effective in low-resource settings.

Limitations

This study has several limitations. First, evaluation of Irish→English was constrained by the absence of a publicly available development set with reference translations. As a result, the Irish experiments relied on proxy diagnostics such as ASR coverage, repetition rate, and output length ratio rather than standard MT evaluation metrics. Second, WER could not be computed for either language pair because aligned source-language transcript references were not available for the ASR outputs.

Third, some alternative systems could not be fully evaluated. In particular, SeamlessM4T-v2-large encountered runtime and deployment issues that prevented a complete comparison under the same experimental conditions. Similarly, the Bhojpuri normalization step was implemented as a heuristic Unicode filter and was not validated against a gold-standard normalised reference.

Finally, the findings reported here are specific to the two language pairs and datasets examined in the IWSLT 2026 task. Although the results provide useful evidence for the effectiveness of cascaded pipelines in low-resource settings, further experiments are required before drawing stronger conclusions about generalisability to other languages, domains, or speech conditions.

Acknowledgments

We acknowledge the use of AI, such as Anthropic's Claude Code, OpenAI's ChatGPT, and Google's Gemini for assisted coding and writing, e.g., for improving the language of our paper.

This research was partially supported by the Horizon Europe project GenDAI (Grant Agreement ID: 101182801) and by the ADAPT Research Centre at Munster Technological University. ADAPT is funded by Taighde Éireann – Research Ireland through the Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) via Grant 13/RC/2106_P2.

LP	Set	Rt	N	ASR%	Rep.↓	Bad↓	L.R
BHO→HI	Dev	Dir	1056	100	0.047	0.001	1.015
BHO→HI	Dev	Pvt	1056	100	0.046	0.000	1.038
BHO→HI	Test	Dir	252	100	0.045	0.004	1.032
BHO→HI	Test	Pvt	252	100	0.038	0.000	1.053
GA→EN	Test	Dir	722	100	0.039	0.000	1.144
GA→EN	Test	Pvt	722	100	0.039	0.000	1.050

Table 6: Unified diagnostic summary across all language pairs and datasets (Whisper-large-v3 for Bhojpuri, Wav2Vec2 Irish for Irish). LP = language pair; Rt = route (Dir/Pvt); Rep. = average repetition rate; Bad = bad-repetition rate; L.R = average length ratio.

Pair	System	ASR	Route
BHO→HI	Primary	Whisper-large-v3	Direct (NLLB)
BHO→HI	Contrastive	Whisper-large-v3	Pivot via EN
GA→EN	Primary	Wav2Vec2 Irish	Direct (NLLB)
GA→EN	Contrastive	Wav2Vec2 Irish	Pivot via HI

Table 7: Final submission configurations.

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Haithem Afli, Loïc Barrault, and Holger Schwenk. 2013. Multimodal comparable corpora as resources for extracting parallel data: Parallel phrases extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*.

Marko Avila and Josep Crego. 2025. SYSTRAN @ IWSLT 2025 low-resource track. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 324–332, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Low-resource speech-to-text translation. In *Proceedings of Interspeech 2018*, pages 1298–1302.

Kaushal Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023. Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yao-Fei Cheng, Hung-Shin Lee, and Hsin-min Wang. 2021. Allost: Low-resource speech translation without source transcription. In *Proceedings of Interspeech 2021*, pages 2252–2256.

Akshat Govil, C Ravindra Reddy, Haren A, and Manju Venugopalan. 2026. Fine-tuning meta's nllb-200 model for telugu–english neural machine translation and real-time wikihow deployment via android app. In *2026 International Conference on Computing, Communication, Control and Cyber-Physical Systems (I5CPS)*, pages 1–7.

Séamus Lankford. 2024. Enhancing neural machine translation of low-resource languages: Corpus development, human evaluation and explainable ai architectures. *Preprint*, arXiv:2403.01580.

Séamus Lankford, Haithem Afli, and Andy Way. 2022. Human evaluation of english–irish transformer-based neural machine translation. *Information*, 13(7):309.

Chutong Meng and Antonios Anastasopoulos. 2025. GMU systems for the IWSLT 2025 low-resource speech translation shared task. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 289–300, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Yasmin Moslem. 2024. Leveraging synthetic audio data for end-to-end low-resource speech translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 265–273, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

NVIDIA. 2024. Whisper large v3 - nvidia ai foundation models. `https://build.nvidia.com/openai/whisper-large-v3`. Accessed: 2026.

OpenAI. 2023. Whisper large v3 model. https://huggingface.co/openai/whisper-large-v3. Accessed: 2026.

Juan N. Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Elena Cryst, Vukosi Marivate, and Sanmi Koyejo. 2025. Mind the language gap: Mapping the challenges of llm development in low-resource language contexts.

Nathaniel Romney Robinson, Kaiser Sun, Cihan Xiao, Niyati Bafna, Weiting Tan, Haoran Xu, Henry Li Xinyuan, Ankur Kejriwal, Sanjeev Khudanpur, Kenton Murray, and Paul McNamee. 2024. JHU IWSLT 2024 dialectal and low-resource system description. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 140–153, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Shashwat Upadhyay. 2026a. Iwslt 2026 bhojpuri-hindi dataset repository. https://github.com/shashwatup9k/iwslt2026_bho-hi.

Shashwat Upadhyay. 2026b. Iwslt 2026 irish-english dataset repository. https://github.com/shashwatup9k/iwslt2026_ga-eng.

Vistatec. 2025. How to overcome the need for data for low-resource languages.

Arlisa Yuliawati, Ika Alfina, and Indra Budi. 2025. From corpus to benchmark: Evaluating pretrained language models for indonesian-javanese krama translation. In *2025 International Conference on Asian Language Processing (IALP)*, pages 199–204.