Mapudungun–Spanish Speech Translation: A Low-Resource End-to-End System for the IWSLT 2026 Shared Task

Diego Barriga[*1] **Amilkar Gazque**[1] **Mikel Segura Elizalde**[2]

Carlos Mena[3] **Ximena Gutierrez-Vasques**[4] **Ivan Meza**[5]

[1]Posgrado en Ciencias e Ingeniería de la Computación, Universidad Nacional Autónoma de México (UNAM)

[2]Facultad de Ciencias, Universidad Nacional Autónoma de México (UNAM)

[3]Barcelona Supercomputing Center (BSC)

[4]CEIICH, Universidad Nacional Autónoma de México (UNAM)

[5]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México (UNAM)

Abstract

We present an end-to-end speech translation system for Mapudungun–Spanish developed for the IWSLT 2026 low-resource task. Building on the Canary-1B-v2 model, we apply parameter-efficient fine-tuning with a lightweight adapter and leverage an English-centered configuration as a proxy to enable translation. Experiments show that the system captures key phonetic patterns despite limited data, though it exhibits biases toward repetitive Spanish outputs. Our results highlight both the feasibility and the challenges of adapting multilingual foundation models to low-resource Indigenous languages.[1]

1 Introduction

This work summarizes our entry for the Low-Resource Speech Translation shared task at the International Conference on Spoken Language Translation (IWSLT) 2026. This shared task aims to benchmark and advance speech translation for diverse dialects and low-resource languages, where limited parallel data constrain standard supervised approaches (Adelani et al., 2026). It emphasizes the need for novel methods that leverage heterogeneous resources to address data scarcity. The task comprises two tracks: (1) a traditional speech-to-text translation track, and (2) a data track focused on the creation and open release of speech translation datasets for under-resourced languages. The current edition explicitly prioritizes multilingual systems capable of handling a broad range of languages, encouraging participants to develop generalizable approaches and engage with multiple language pairs, while still allowing specialized systems for individual pairs.

Our entry focuses on Task 1 and on a single language pair. We build an end-to-end speech translation system from Mapudungun (arn) to Spanish (spa) using the Canary-1B-v2 model.

In the next section 2, we describe the main characteristics of the Mapudungun language and corpus. Section 3 presents our approach, Section 4 reports our results, Section 5 provides our concluding remarks, and Section 6 discusses the limitations of our study.

2 Mapudungun language and corpus

Mapudungun (arn) is an Indigenous language of the Americas spoken by the Mapuche people in south-central Chile and west-central Argentina. It has an estimated native speaker population of around 250,000. Under the Agglomerated Endangerment Scale (AES), Mapudungun is classified as *shifting (downward)*, meaning that the language is undergoing a process of increasing endangerment (Hammarström et al., 2024).

Our entry uses the provided Mapudungun corpus (Duan et al., 2020). The corpus consists of 142 hours of spoken Mapudungun, collected between 2001 and 2005 as part of the AVENUE project (Levin et al., 2000), in collaboration among Carnegie Mellon University, the Chilean Ministry of Education, and the Universidad de La Frontera. The data, which originally spanned 170 hours, were cleaned and publicly released with transcriptions and Spanish translations. The corpus covers three closely related dialects—Nguluche (approximately 110 hours), Lafkenche (approximately 20 hours), and Pewenche (approximately 10 hours).

The recordings are domain-specific, focusing on primary and traditional healthcare practices, and were collected through ethnographically grounded, open-ended conversational interviews. The dataset reflects culturally authentic interactions among native speakers, with participants aged 16 to 100 and balanced by gender. Ethical considerations were addressed through informed consent, anonymization,

[*]dbarriga@ciencias.unam.mx

[1]https://github.com/umoqnier/iwstl-low-resource-experiments

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 232–237

and the removal of sensitive cultural knowledge.

From a linguistic perspective, the corpus uses a non-standard, supra-dialectal orthography developed at the time of collection, consisting of 28 graphemes designed to be compatible with Spanish keyboards. Although this differs from the modern Azumchefe standard, both the original and normalized versions are valuable. The dataset also includes rich annotations capturing disfluencies, non-verbal sounds, and code-switching (e.g., Spanish borrowings), making it a comprehensive and challenging resource for speech and language processing in a low-resource setting.

A substantial cleaning and normalization process was applied to make the corpus suitable for modern computational use. This includes fixing inconsistencies in utterance boundaries and speaker labels, resolving missing or mismatched files, converting text encodings to Unicode, standardizing formatting, and removing duplicates. Audio was normalized in power, and turn boundaries were adjusted to control leading and trailing silence. Additionally, forced alignment techniques were applied to improve synchronization between audio and transcripts, enabling the identification of well- and poorly aligned segments. Residual alignment issues are partly attributed to orthographic variation, which remains an area for future normalization.

3 Approach

Our approach focuses on fine-tuning NVIDIA's Canary-1B-v2 model for speech-to-text translation (ST) (Sekoyan et al., 2025), a multilingual/multitask model trained for various tasks, including automatic speech recognition (ASR). Its implementation is based on the Conformer architecture (Gulati et al., 2020); specifically, its more efficient variant known as FastConformer (Rekesh et al., 2023). The original model released by NVIDIA (NVIDIA, 2025) was trained on a 1.7-million-hour dataset covering 22 high-resource and 3 low-resource languages. It uses both pseudo-labeled and human-annotated data and is also exposed to non-speech audio associated with empty outputs to improve its robustness.

The Canary model supports speech translation through English-centered language pairs. In one direction, the source language is English and the target can be any of 24 languages ($eng \rightarrow X$). In the second configuration, the direction is reversed, from any of the 24 languages to English ($X \rightarrow eng$). As we show below, this design becomes a challenge because English is not part of our chosen language pair (arn, esp) for the shared task.

We chose the Canary model based on the following advantages:

- **Accessible codebase:** Canary uses the NeMo NVIDIA framework (Kuchaiev et al., 2019), which makes the training code readily accessible.
- **Spanish coverage:** Spanish is one of the languages covered by the model, so the system is already capable of "listening" and "generating" Spanish.
- **Hardware requirements:** The implementation is efficient, and it supports parameter-efficient fine-tuning (PEFT) that significantly reduces the amount of computer power; in our case, we perform experiments with two GPUS with 24GB of VRAM (12GB on each GPU) and a single GPU with 32GB of VRAM.
- **Preliminary testing:** Our initial experiments allowed us to easily prototype the scripts needed to train and evaluate the system.

3.1 Data preprocesing

The speech signal was augmented using an on-the-fly waveform augmentation pipeline (Jordal, 2023) applied directly during training, eliminating the need to store additional preprocessed data. Prior to batch collation and zero-padding, the raw audio signals were stochastically augmented through a combination of background noise injection, Gaussian noise, aliasing, room impulse responses, band-pass filtering, and pitch shifting. Additionally, in one experimental run, we applied a duration-based filtering step that discarded recordings longer than 15 seconds. The latter was done to reduce instabilities during training that would otherwise halt the process and to maintain more uniform sequence lengths within batches.

The Spanish translations were also preprocessed by eliminating capitalization and removing punctuation.

3.2 Architecture

The NeMo framework allows us to fine-tune the Canary-1B-v2 model with considerable flexibility. In particular, we adopt a parameter-efficient fine-tuning (PEFT) strategy (Houlsby et al., 2019) based

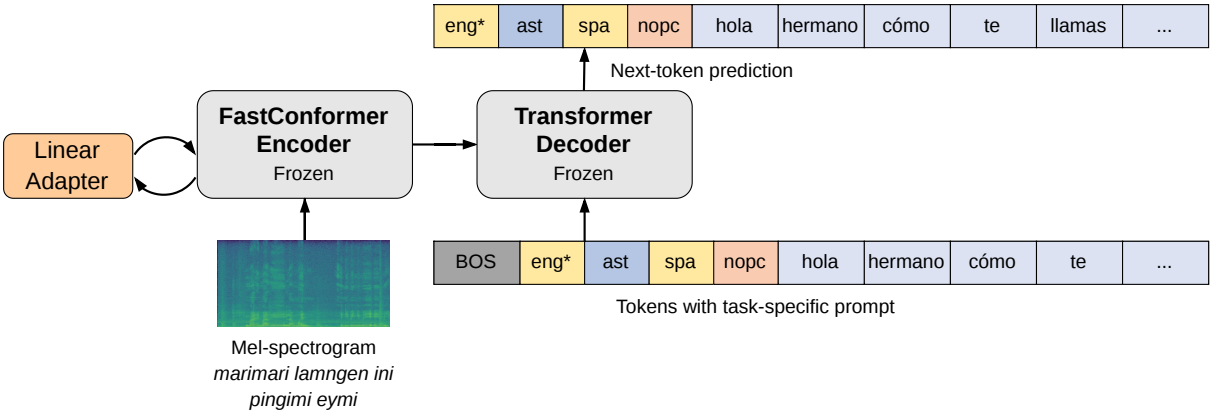

Figure 1: Modified Canary model for AST adopting a PEFT strategy. A Linear Adapter is integrated into the frozen backbone to enable Mapudungun translation. The asterisk in `eng*` denotes the use of English as a placeholder for Mapudungun as the source language.

Parameter	Value
Adapter Type	Linear
Input Dimension	1024
Output Dimension	8
Activation	Swish (default)
Normalization Position	Pre-norm (default)
Dropout Rate	0.0 (default)

Table 1: Configuration and hyperparameters of the PEFT adapter layers.

on a linear adapter in the encoder, as shown in Figure 1, which substantially reduces the number of trainable parameters (from 1.7 billion to 589k). Table 1 summarizes the main adapter configuration. We set the translation direction to $eng \rightarrow spa$ even though our actual source language is Mapudungun ($arn \rightarrow spa$). This strategy effectively hijacks one of the predefined languages by using English as a placeholder for Mapudungun. We chose this setting for technical reasons. First, adding a new language to the model would require fine-tuning the entire model. Second, we hypothesized that the speech signal would have a stronger effect on the task because the model already knows how to generate Spanish. In this sense, our main hypothesis is that we can transfer the model's knowledge of English acoustic patterns to Mapudungun while preserving its ability to translate into Spanish.

Training was conducted using bfloat16 mixed precision and dynamic duration-based audio bucketing via Lhotse (Żelasko et al., 2023) to efficiently process variable-length utterances.

3.3 Submitted runs

We developed three system runs:

- **Augmented (Main):** Incorporates transcription pre-processing and on-the-fly speech data augmentation. The model was trained on a single GPU for 19 epochs, utilizing early stopping.
- **Vanilla:** Achieved the best performance during our validation and internal testing phases. It shares the training configuration of the `augmented` run (19 epochs, single GPU, early stopping) and includes transcription preprocessing, but strictly omits speech data augmentation.
- **Filtered:** Follows the `vanilla` configuration, with an additional data curation step that discards audio recordings exceeding 15 seconds in duration. This model was trained across two GPUs.

Additionally, after submission, we prepared a second version of the third run, which we call **Filtered***. We did this since we had lost a development evaluation during experimentation.

4 Results

The training logs indicate a progressive decrease in Word Error Rate (WER) as expected. Interestingly, the augmented model initially outperformed the vanilla variant; however, by the end of training, the vanilla model surpassed the augmented version by a marginal 0.1% WER. This suggests that data augmentation may offer limited benefit in this specific case, as the source corpus already contains

significant ambient noise. Both models exhibited a steady decline in loss and were terminated according to our early stopping policy. Final performance metrics are summarized in Table 2.

Qualitative analysis indicates that our models capture Mapudungun speech phenomena to some extent. For instance, the model successfully transcribes terms like `huinca` (Example 1) and `mapuche` (Example 3), which are likely underrepresented in the pre-training corpus due to their specific regional context and the model's high-resource language bias. Conversely, as seen in Example 2, the model occasionally falls back on repetitive, basic Spanish structures. This pathological decoding results in higher WER and lower BLEU scores, likely due to the model's bias toward the higher-resource language.

1. **Accurate Rare-Word Transcription**

   Reference: sí pero yo creo que hay eso sí que no sé pero en las cosas del **huinca** parece que hay una tableta
   Predicted: sí pero yo creo que hay eso sí que no sé pero en el **huinca** parece que hay una tableta

2. **Pathological Decoding (Repetition)**

   Reference: en el monte en la quebrada esta en la montaña
   Predicted: en la tierra de la tierra de la tierra

3. **Partial Rare-Word Capture**

   Reference: como nosotros somos **mapuches** pero estamo aparte entonce
   Predicted: el **mapuche** es muy bueno pues

Overall, the loss and error metrics exhibited stable convergence. We observe that both WER and BLEU scores reflect the significant challenge of this low-resource task. The high WER and low BLEU scores are expected given the extreme scarcity of training data and language-specific phenomena such as code-switching in Mapudungun. Despite these challenges, the *Filtered* run shows a noticeable improvement over the baseline, suggesting that constraining utterance duration is a critical factor in reducing the model's tendency toward pathological repetitions.

Run	WER ↓	BLEU ↑
Main	2.5	1.015
Vanilla	2.3	1.05
Filtered*	**1.56**	**3.7**

Table 2: Performance comparison on the development set. WER is reported as a percentage (where 1.0 represents 100% error), and BLEU scores are reported on a scale of 0–100. Lower WER and higher BLEU indicate better performance. The results indicate that the *Filtered* run (restricting audio duration to less than 15 seconds) significantly improves performance, suggesting that shorter utterances reduce decoding instabilities and repetitions common in low-resource settings.

The fact that training was terminated early suggests that our current architecture has not yet saturated its learning capacity. These observations indicate that performance could be further improved by scaling the data, fine-tuning training configurations, and unfreezing more adapter layers. Beyond configuration tuning, adopting a model that incorporates speech-input pre-training on low-resource languages would allow the architecture to better adapt to the phonetic nuances of the target language.

4.1 Official Test Set Evaluation

To validate the generalization capability of our models, we evaluated three configurations on the official test set. As shown in Table 3, the *Filtered* configuration outperformed both the Main and the Vanilla submission.

While the BLEU scores remain low, the chrF2++ results (reaching 14.31 for *Filtered*) provide a more nuanced view of the model's performance. The relative improvement in the *Filtered* run confirms that our strategy for refining the training data—specifically limiting utterance duration—effectively reduces decoding instabilities and improves generalization on unseen test data.

Submission	BLEU ↑	chrF2++ ↑
Main	0.73	12.70
Vanilla	0.34	6.92
Filtered	**0.82**	**14.31**

Table 3: Official test set results provided by the organizers. All submissions were lower-cased with punctuation removed. The results show that the *Filtered* configuration generalizes best to unseen data, achieving the highest scores in both BLEU and chrF2++ metrics.

5 Conclusions

We presented our approach to adapting a multilingual-multitask foundation model for speech-to-text translation from Mapudungun to Spanish. Despite resource constraints, our methodology successfully captured distinct Mapudungun phonetic features, demonstrating the viability of parameter-efficient fine-tuning for this task.

For future work, we intend to extend this framework to develop broader multilingual translation systems that include other indigenous languages spoken in Latin America, such as Quechua and Nahuatl. Ultimately, our research aims to understand the challenges that arise in these specific settings, helping to close the technological divide between high-resource and low-resource languages. This is one of the elements that may contribute to strengthening the representation of linguistic diversity in modern speech technologies.

6 Limitations

In developing our approach, we encountered the following notable constraints:

1. Given the available time and computational resources, we were not able to fine-tune the entire model. We addressed this challenge by using a PEFT training strategy.

2. The chosen model supports English-centric speech translation, which limited our ability to model the direction directly ($arn \rightarrow spa$) and forced us to use a placeholder language. Because we wanted to preserve the model's knowledge of generating Spanish, we had to rely on a strong English acoustic model.

3. We encountered training-stability issues, particularly when using two GPUs to accelerate training.

It is noteworthy that this work constitutes a first approach to tackling this complex task in a resource-constrained setting. Further investigation is needed before deploying a speech translation tool, along with close collaboration with the speaker communities.

7 Acknowledgemnts

The authors would like to thank PAPIIT-UNAM for its support through the projects *"Traducción de voz a voz para lenguas originarias de México"* (37-IN107224) and *"Fortaleciendo la diversidad en las tecnologías del lenguaje: procesamiento automático de las lenguas en México"* (TA100725). We would like to thank the project ILENIA with reference 2022/TL22/00215337 and the Government of Catalonia through the Aina project.

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.

Mingjun Duan, Carlos Fasola, Sai Krishna Rallabandi, Rodolfo Vega, Antonios Anastasopoulos, Lori Levin, and Alan W Black. 2020. A resource for computational experiments on Mapudungun. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2872–2877, Marseille, France. European Language Resources Association.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.2 - mapudungun. Accessed 2026-04-21.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Ivér Jordal. 2023. Audiomentations: A python library for audio data augmentation. https://iver56.github.io/audiomentations/. Accessed: 2026-04-23.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, and 1 others. 2019. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*.

Lorraine Levin, Rodolfo M Vega, Jaime G Carbonell, Ralf D Brown, Alon Lavie, Eliseo Cañulef, and Carolina Huenchullan. 2000. Data collection and language technologies for mapudungun.

NVIDIA. 2025. Canary-1b-v2: Multitask speech transcription and translation model. https://huggingface.co/nvidia/canary-1b-v2. Hugging Face model card.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *Preprint*, arXiv:2305.05084.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast. *Preprint*, arXiv:2509.14128.

Piotr Żelasko, Dominika Gajewska, and David Povey. 2023. Lhotse: a python toolkit for working with audio data for speech applications. In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 129–136.