QUESPA Submission for the IWSLT 2025 Dialectal and Low-resource Speech Translation Task

John E. Ortega[1], **Rodolfo Zevallos**[2], **William Chen**[3], **Idris Abdulmumin**[4]
[1]Northeastern University, USA, [2]Barcelona Supercomputing Center, Spain
[3]Carnegie Mellon University, USA, [4]University of Pretoria
contact email: j.ortega@northeastern.edu

Abstract

This article describes the **QUESPA** team speech translation (ST) submissions for the Quechua to Spanish (QUE–SPA) track featured in the Evaluation Campaign of IWSLT 2025: dialectal and low-resource speech translation. This year, there is one main submission type supported in the campaign: *unconstrained*. This is our third year submitting our ST systems to the IWSLT shared task and we feel that we have achieved novel performance, surpassing last year's submission. This year we submit three total *unconstrained-only* systems of which our best (contrastive 2) system uses last year's best performing pre-trained language (PLM) model for ST (without cascading) and the inclusion of additional *Quechua–Collao* speech transcriptions found online. Fine-tuning of Microsoft's SpeechT5 model in a ST setting along with the addition of new data and a data augmentation technique allowed us to achieve 26.7 BLEU. In this article, we present the three submissions along with a detailed description of the updated machine translation system where a comparison is done between synthetic, unconstrained, and other data for fine-tuning.

1 Introduction

In this article, we describe three systems that were submitted to the IWSLT 2025 Low-Resource Track for Speech Translation (ST). The IWSLT task is particularly challenging for low-resource languages (LRLs) due to the lack of data needed to create, or even fine-tune, a pre-trained language model (PLM). While many problems are solvable with APIs provided by large corporations such as ChatGPT or Gemini, it is still the case that for LRLs, zero-to-few shot approaches are needed where corporate-level APIs do not contain enough data either. Here, we describe three main approaches that extend previous approaches submitted in the past three iterations of IWSLT (Ahmad et al., 2024; Agarwal et al., 2023; Anastasopoulos et al., 2022) where the best score gotten for ST until this publishing based on BLEU (Papineni et al., 2002) for the Quechua to Spanish task was: 19.7, submittted by this same team **QUESPA**.

Quechua is an indigenous language spoken by more than 8 million people in South America. It is mainly spoken in Peru, Ecuador, and Bolivia where the official high-resource language (HRLs) is Spanish. It is a highly inflective language based on its suffixes which agglutinate and found to be similar to other languages like Finnish. It is worthwhile to note that previous work (Ortega and Pillaipakkamnatt, 2018; Ortega et al., 2020) has been somewhat successful in identifying the inflectional properties of Quechua such as agglutination where another HRLs, namely Finnish, can aid for translation purposes achieving nearly 20 BLEU on religious-based (text-only) tasks. The average number of morphemes per word (synthesis) is about two times larger than English. English typically has around 1.5 morphemes per word and Quechua has about 3 morphemes per word. There are two main region divisions of Quechua known as Quechua I and Quechua II. This data set consists of two main types of Quechua spoken in Ayacucho, Peru (Quechua Chanka ISO:quy) and Cusco, Peru (Quechua Collao ISO:quz) which are both part of Quechua II and, thus, considered a "southern" languages. We label the data set with que - the ISO norm for Quechua II mixtures.

The **QUESPA** team this year consists of four organizers from four different institutions: Northeastern University, Pompeu Fabra University, Carnegie Melon University and University of Pretoria. A new organizer has been introduced this year who has expertise in machine translation (MT) of African languages. All of the IWSLT 2024 organizers have continued to work on the project with exception of one. Three of the four organizers have had experience with the QUE–SPA language

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 260–268
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

pair in the past and submitted have already submitted three times to IWSLT, making this article the fourth submission with an increase of BLEU score for each year's submission. We report the QUESPA consortium submission for the IWSLT 2025 and once again focus on the low-resource task at hand by combining *all* the two dialects *Quechua I and II* into one. However, we specifically make use of the Quechua II variant in Collao (ISO:quz), given the discovery of a new corpus.

The rest of this article is organized as follows. Section 2 presents the related work. Since we would like to highlight the addition of our MT comparisons and systems by a new author, we present a section dedicated to the MT delivery in Section 3.1. Afterwards, we present experiments for the for QUE–SPA low-resource track are presented in Section 3 and present their results in Section 4 provides.

2 Related Work

In this section, we first cover the different approaches used in previous speech processing shared tasks for Quechua (Section 2.1). We then discuss prior work that used a similar strategy to our primary submission to the unconstrained track (Section 2.2).

2.1 Quechua Speech Processing

The previous iteration of IWSLT (Agarwal et al., 2023) was the first time that Quechua–Spanish was featured in the low-resource ST track. Due to the small amount of available paired data, the participants focused on exploiting PLMs for speech and/or text in the unconstrained track. The teams all converged on using XLS-R 128 (Babu et al., 2021) as the pre-trained speech encoder, while NLLB 200 (NLLB Team et al., 2022) was the most popular text PLM. However, the teams used the PLMs in very different manners. QUESPA (E. Ortega et al., 2023) separated the PLMs into distinct systems for an ASR+MT cascade, GMU (Mbuya and Anastasopoulos, 2023) performed full fine-tuning on XLS-R for direct ST, and NLE (Gow-Smith et al., 2023) combined the two PLMs via adapter fine-tuning. By using PLMs for both the input and output modalities, NLE and QUESPA obtained the best performances at 15.7 and 15.4 BLEU respectively. For the constrained track, developing a usable system was far more difficult to achieve. In this setup, the best performing model was a direct ST system by GMU that achieved 1.46 BLEU. The QUESPA team adopted a near-identical strategy to achieve 1.25 BLEU.

Quechua–Spanish ST was also featured as part of a similar competition in the 2022 edition of AmericasNLP (Ebrahimi et al., 2022). Similar to IWSLT 2023, participants experimented with different ways of leveraging PLMs. XLS-R and NLLB were popular choices, but some teams also experimented with DeltaLM (Ma et al., 2021) and Whisper (Radford et al., 2023).

Quechua was most recently part of the 2023 ML-SUPERB Challenge (Shi et al., 2023), which tasked participants on evaluating different self-supervised (SSL) speech encoders on long-tail languages. Chen et al. (2023a) found that XLS-R 128 outperformed all other SSL encoders on Quechua, further validating its popularity in the other competitions.

2.2 Multilingual Speech Processing

Multilingual training is a common strategy to facilitate cross-lingual transfer learning, with the goal of boosting performance on LRLs. While this is generally done by pairing HRLs with low-resource ones, it can also be beneficial in settings where only LRLs are available. Chen et al. (2023b) trained multilingual ASR systems on 102 languages, each in a low-resource setting, and obtained state-of-the-art (SOTA) results on the FLEURS benchmark (Conneau et al., 2023). Radford et al. (2023) and Peng et al. (2023) then combined multilingual ASR and ST at scale, developing SOTA models through supervised training on hundreds of thousands of audio samples. Our strategy for the unconstrained track can be viewed as a combination of these two methods, enhancing performance on Quechua–Spanish using multilingual ST training with other LRLs.

3 Quechua-Spanish

In this section we present our experiments for the QUE–SPA dataset provided in the low-resource ST track at IWSLT 2025[1], identical to the dataset from IWSLT 2024. As a reminder, the audio consists of contains 1 hour and 40 minutes of *unconstrained* speech along with its corresponding translations and nearly 48 hours of ASR data (with transcriptions) from the Siminichik (Cardenas et al., 2018)

[1] https://github.com/Llamacha/IWSLT2025_Quechua_data

corpus. Additionally, an MT dataset is offered from previous neural MT work (Ortega et al., 2020). The audio and corresponding transcriptions along with their translations are mostly made of radio broadcasting from the mountainous region in the Andes, Peru. This dataset has been used in other tasks but not in its entirety (Ebrahimi et al., 2023, 2022; Zevallos et al., 2022a). This year there has been a new addition to the dataset provided by the task which is a machine-translated and post-edited text of the Huqariq corpus (Zevallos et al., 2022b) that was used last year by this team (Ortega et al., 2024) for augmentation of the best performing T5 model (Raffel et al., 2020).

We present the three submissions for *unconstrained* task ony as this year the constrained task has been abandoned:

1. a **primary unconstrained** system consisting of a Mamba ASR model (Zhang et al., 2024) fine-tuned with unconstrained data and cascaded the best performing NLLB MT system from our case study;

2. a **contrastive 1 unconstrained** system consisting of a Whisper (Radford et al., 2023) ASR model fine-tuned with the unconstrained data and cascaded with the best performing NLLB MT system from our case study;

3. a **contrastive 2 unconstrained** system consisting of a SpeechT5 model (Ao et al., 2021) fine-tuned for speech translation with two data augmentation techniques and an additional newly introduced corpus based on Quechua Collao (iso: quz) (Paccotacya-Yanque et al., 2022).

We present the experimental settings and results for unconstrained systems starting off with the MT case studies in Section 3.1. Then, we describe the task further in Section 3.2. Primary, Contrastive 1 and Constrastive 2 descriptions are found in Sections Sections 3.3, 3.4 and 3.5, respectively. Afterwards, we offer results and discussion in Section 4.

3.1 Machine Translation

Our MT systems were all trained by fine-tuning the 1.3B parameter version[2] of the NLLB_200 (NLLB Team et al., 2022). For fine-tuning, we set maximum token lengths of 128 for both inputs and outputs. Each model was trained for 10 epochs with a batch size of 8 for both training and evaluation, using 5 beams during generation. We saved model checkpoints every 10,000 steps and set a random seed of 65 to ensure reproducibility.

We trained four models, with each model using a different training dataset. The first three models were trained strictly on datasets provided in the shared task. The first model was fine-tuned on the unconstrained data (U; Cardenas et al. 2018). We then increased the training data using the provided `additional_mt_text` dataset (A; Ortega et al. 2020) to train the second model. This data consists of texts from JW300 (Agić and Vulić, 2019) and Hinantin websites. For the third model, we further expanded the training data by incorporating the provided `synthetic data` (S; Zevallos et al. 2022b) dataset. The sizes of the training data for the three models are 573, 15,857, and 17,265 sentences, respectively.

The fourth model was trained on the largest available dataset. In this setting, we used additional resources (AR) including SMOL (Caswell et al., 2025), GATITOS (Jones et al., 2023), spanish-to-quechua,[3] and cuzco-quechua-translation-spanish[4]. The SMOL and GATITOS datasets consist of 863 and 3,717 sentences, respectively. The two latter datasets each contain over 100k sentences (103k and 106k), though we observed overlap between them. To address this, we deduplicated the Quechua sentences after merging the datasets. After merging all available datasets, including those provided in the shared task, and performing deduplication, the total number of training sentences amounted to 167,052.

For each of the four models, we experimented with two different validation datasets. The first was the 125 parallel sentences provided for validation in the shared task. In the second, we expanded this set by adding the 2,500-sentence JW300 validation dataset, also provided in the shared task. In the latter setup, our goal was to ensure more generalizable models. However, we identified several issues in the JW300 validation data that required preprocessing, including instances where the source and target sentences were identical. After preprocessing and cleaning, the expanded validation set consisted of

[2] https://huggingface.co/facebook/nllb-200-1.3B

[3] https://huggingface.co/datasets/somosnlp-hackathon-2022/spanish-to-quechua

[4] https://huggingface.co/datasets/pollitoconpapass/cuzco-quechua-translation-spanish

Model-[Data]	BLEU						CHRF					
	$\mathbf{V}_s$	$\mathbf{V}_l$	$\mathbf{T}_s$	$\mathbf{T}_l$	$\mathbf{Tr}_w$	$\mathbf{Tr}_m$	$\mathbf{V}_s$	$\mathbf{V}_l$	$\mathbf{T}_s$	$\mathbf{T}_l$	$\mathbf{Tr}_w$	$\mathbf{Tr}_m$
MT-[U]	18.5	18.5	17.3	17.3	11.8	11.4	53.9	53.8	54.1	54.1	46.5	46.0
MT-[U + A]	19.5	19.3	18.0	17.5	15.0	14.8	54.7	54.3	54.9	54.3	52.4	51.8
MT-[U + A + S]	14.6	14.3	13.3	13.6	12.3	11.6	48.0	47.7	48.4	48.8	46.9	47.4
MT-[U + A + S + AR]	15.1	14.2	13.2	13.3	12.4	12.5	48.4	48.5	48.1	48.2	46.9	47.3

Table 1: Performance of the four models on the validation and test sets. We also report results on transcripts generated from the test set, evaluated on models trained with the large validation set. **KEY:** T = test set, V = validation set, Tr = transcripts. *s* and *l* denote the small and large validation sets, respectively. *w* and *m* denote the Whisper and Mamba models, respectively. U = unconstrained, A = `additional_mt_text`, S = synthetic, AR = additional resources.

2,309 parallel sentences.

Table 1 presents the performance of the four machine translation models across different evaluation setups, measured using both BLEU and CHRF scores. Overall, the results indicate that while more data leads to better performances, the quality of the additional data matters. The first model, MT-[U], shows decent performance with a BLEU score of 18.5 on the large validation set and 17.3 on the test set, with strong CHRF scores ranging between 46 and 54. The second model, MT-[U + A], achieves better BLEU and CHRF scores, particularly on the transcript evaluations.

The third model, MT-[U + A + S], which incorporates synthetic data, shows a noticeable decline in both BLEU and chrF scores across all evaluation sets—most prominently on the test and validation sets. This drop suggests that the inclusion of synthetic data, if not carefully curated, can adversely affect model performance. The final model, MT-[U + A + S + AR], demonstrates a slight improvement over MT-[U + A + S] across the evaluation sets. However, it does not fully recover the performance lost when synthetic data was added to MT-[U + A]. This outcome highlights a crucial insight: although expanding training data with additional and diverse resources can enhance model generalization, introducing even a small amount of lower-quality data can undermine those gains. Careful data quality control is therefore essential when scaling datasets for low-resource machine translation.

3.2 Unconstrained Setting

Just like in IWSLT 2024, the organizers provided a total of 48 hours of audio along with their corresponding transcriptions. In addition, we translated the 48 hours of audio provided by the organizers into Spanish. Furthermore, we utilized a portion of the AmericasNLP[5] (ANLP) 2022 speech translation competition corpus, which consists of 19 minutes of Guarani and 29 minutes of Bribri, fully translated into Spanish. Although it is not a Quechua corpus, these languages have morphological similarities with Quechua, so we decided to experiment to see if that improves our models. As a new addition, we used the data set from previous work on Quechua Collao (Paccotacya-Yanque et al., 2022) which, much like the IWSLT 2025 corpus, is part of the Quechua II division. Finally, all the datasets described in this section allowed for further fine-tuning of the previously trained end-to-end speech translation model.

3.3 Primary System

The Primary System for the unconstrained setting consists of a cascaded architecture, where the output of an automatic speech recognition (ASR) model is passed as input to a machine translation (MT) model. For the ASR component, we employ ConMamba (Jiang et al., 2024), a recent extension of the Mamba architecture that integrates convolutional modules into its encoder blocks, inspired by Conformer (Gulati et al., 2020). This hybrid design enhances the model's ability to capture both global and local dependencies. The encoder architecture comprises a sequence of modules: an initial feedforward layer with residual connection, a bidirectional Mamba module (BiMamba) for long-range dependency modeling, a convolutional layer for local context enhancement, and final layer normalization and refinement through another feedforward module (Tang et al., 2024). This combination results in a balanced and efficient encoding mechanism for speech signals.

[5] https://turing.iimas.unam.mx/americasnlp/2022_st.html

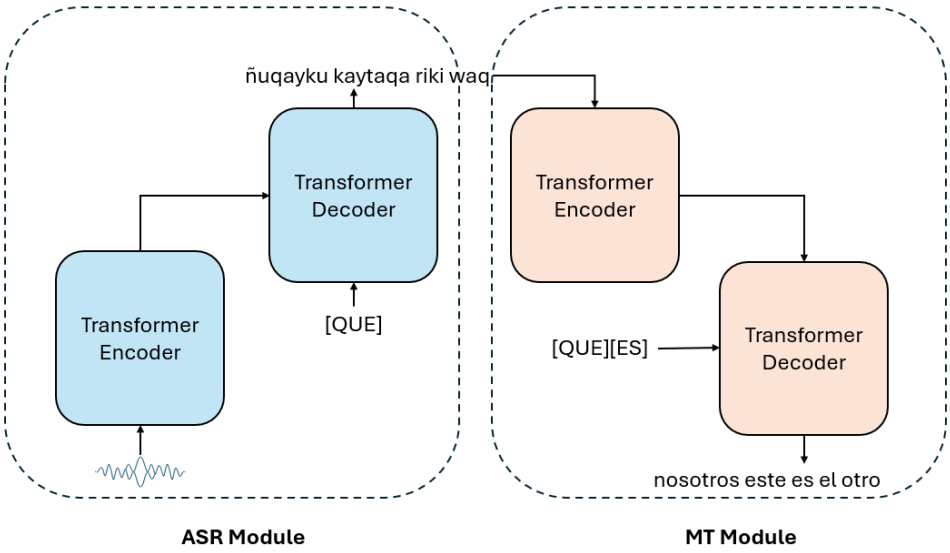

Figure 1: Overiew of the cascaded Contrastive 1 system. The input audio is passed into Whisper, which autoregressively generates a Quechua transcription. The transcription hypothesis is then passed to NLLB to be translated into Spanish.

On the decoder side, we incorporate CrossMamba, a unidirectional variant tailored for sequential processing without native cross-attention. CrossMamba simulates cross-attention by concatenating key and query sequences, retaining only the relevant portion of the output. This mechanism allows for effective integration of encoder context through a structured decoding pipeline: normalization, unidirectional Mamba (UniMamba), a second normalization step, CrossMamba integration, and a final feedforward refinement. We train both ConMamba and Conformer models using publicly available recipes[6], experimenting with small (S) and large (L) configurations (144/512 dimensions, 12+4/12+6 layers). Training is performed over 110 epochs using AdamW with a Noam scheduler (30k warm-up steps), and audio is tokenized with a BPE tokenizer trained for each language using SpeechBrain[7]. Once the speech is transcribed, we feed the resulting text into the machine translation model previously described, leveraging its capabilities to produce the final translated output in a cascaded speech translation setup.

3.4 Contrastive 1 System

The Contrastive 1 system is a simple ASR+MT cascade. We develop the ASR module by fine-tuning Whisper Large V3 (Radford et al., 2023) on the entire 48 hours of unconstrained Quechua ASR data in ESPnet (Watanabe et al., 2018). Whisper consists of a Transformer encoder and Transformer decoder (Vaswani et al., 2017). The bidirectional encoder receives mel audio features as input, whereas the decoder is conditioned on a language identity tag and the encoder output (Figure 1). The model is trained for 22K steps with the Adam optimizer (Kingma and Ba, 2015). We use a scheduler that linearly warms up the learning rate to a peak value of 1e-5 for 1500 steps, followed by exponential decay for the remainder of training (Vaswani et al., 2017). ASR inference is performed with greedy decoding, the results of which are then passed to the NLLB-based MT model described in Section 3.1.

3.5 Contrastive 2 System

The Contrastive 2 System for the unconstrained setting consists of a pre-trained model called SpeechT5 (Ao et al., 2022) , which was trained on 960 hours of audio from LibriSpeech. SpeechT5 consists of 12 Transformer encoder blocks and 6 Transformer decoder blocks, with a model dimension of 768, an internal dimension (FFN) of 3,072, and 12 attention heads. Additionally, the voice encoder's pre-net includes 7 blocks of temporal convolutions. Both the pre-net and post-net of the voice decoder used the same configuration as in Shen et al. (2018), except that the number of channels in the post-net is 256. For the text en-

[6] https://github.com/xi-j/Mamba-ASR

[7] https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriSpeech/Tokenizer

Team **QUESPA** BLEU and CHRF Scores			
Unconstrained 2025			
System	**Description**	**BLEU**	**CHRF**
primary	mamba asr + nllb mt	14.8	51.8
contrastive 1	whisper-v3 asr + nllb mt	15.0	52.4
contrastive 2	speechT5 + anlp + da-tts + nlpaug* + quz	26.7	48.6
Unconstrained 2024			
primary	speechT5 + aug	16.0	52.2
contrastive 1	speechT5 + anlp + da-tts + nlpaug*	19.7	43.1
contrastive 2	whisper asr + nllb mt	11.1	44.6

Table 2: Team QUESPA results for the Quechua to Spanish low-resource task at IWSLT 2025.

coder/decoder's pre/post-net, a shared embedding layer with a dimension of 768 is utilized. For vector quantization, two codebooks with 100 entries each are used for the shared codebook module. The model was trained using the normalized training text from the LibriSpeech language model as unlabeled data, which contains 400 million sentences. Training was optimized using Adam (Kingma and Ba, 2015), with a learning rate that linearly increases during the first 8% of updates up to a maximum of 0.0002.

We fine-tuned SpeechT5[8] for Speech Translation using the SpeechT5 fine-tuning recipe[9] for Speech-Translation with the same hyperparameter settings. We used the 48 hours of audio provided by the organizers (anlp). We applied a data augmentation technique called *nlpaug* (noise, distortion, duplication)[10] (Ma, 2019), resulting in a total of 96h: 48h original + 48h synthetic data + 15 hours of Quechua Collao (Paccotacya-Yanque et al., 2022) (quz) .

4 Results and Discussion

Results are presented in Table 2. When compared to IWSLT 2024 (Ahmad et al., 2024; Ortega et al., 2024), it is clear that Speech Translation as a task is best performed using a multi-lingual transformer such as the Speech T5 model. Addtionally, by fine-tuning the Speech T5 model, we were able to increase the score by a dramatic 7 BLEU points by the addition of data found online. Additionally, the introduction of the latest Whisper model (version 3) seems to show promising increases when compared to last year's result by this team.

[8] https://github.com/microsoft/SpeechT5

[9] https://github.com/microsoft/SpeechT5/tree/main/SpeechT5

[10] https://github.com/makcedward/nlpaug

5 Conclusion and Future Work

Our submission to the IWSLT 2025 (Abdulmumin et al., 2025) evaluation campaign for low-resource and dialect speech translation has included novelties based on the most state-of-the-art techniques for ASR and ST. The addition of three new characteristics: 1) a new Quechua Collao corpus (referred to as quz) and 2) the introduction of a stateless ASR model (Mamba) along with 3) a machine translation case study. These three new inclusions have brought to light what MT systems, corpus, and ASR models work best with the language pair when compared to last year's work.

Next year, we plan to include more human annotation and experimentation with the model presented here since the BLEU score achieved (26.7) warrant further investigation and annotation. We also believe that we have localized a Speech Translation recipe that we allow further iterations of data in the future to achieve even better performance.

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, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, 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, Xutail 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 Ojha Kr., 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)*. Association for Computational Linguistics.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr. Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemánek, and Rodolfo Zevallos. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Yannick Estéve, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel M. Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alex Waibel, Changhan Wang, and Shinji Watanabe. 2022. FINDINGS OF THE IWSLT 2022 EVALUATION CAMPAIGN. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland. Association for Computational Linguistics.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP 2*, page 21.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, and Solo Farabado Cissé. 2025. Smol: Professionally translated parallel data for 115 under-represented languages. *Preprint*, arXiv:2502.12301.

Chih-Chen Chen, William Chen, Rodolfo Zevallos, and John Ortega. 2023a. Evaluating self-supervised speech representations for indigenous american languages. *arXiv preprint arXiv:2310.03639*.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. 2023b. Improving massively multilingual asr with auxiliary CTC objectives. *arXiv preprint arXiv:2302.12829*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

John E. Ortega, Rodolfo Zevallos, and William Chen. 2023. QUESPA submission for the IWSLT 2023 dialect and low-resource speech translation tasks. In

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 261–268, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E Ortega, Rolando Coto-Solano, et al. 2023. Findings of the americasnlp 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, Bashar Talafha, Alcides Alcoba Inciarte, Muhammad Abdul-Mageed, Luis Chiruzzo, Rolando Coto-Solano, Hilaria Cruz, Sofía Flores-Solórzano, Aldo Andrés Alvarez López, Ivan Meza-Ruiz, John E. Ortega, Alexis Palmer, Rodolfo Joel Zevallos Salazar, Kristine Stenzel, Thang Vu, and Katharina Kann. 2022. Findings of the second americasnlp competition on speech-to-text translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Edward Gow-Smith, Alexandre Berard, Marcely Zanon Boito, and Ioan Calapodescu. 2023. NAVER LABS Europe's multilingual speech translation systems for the IWSLT 2023 low-resource track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 144–158, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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

Xilin Jiang, Yinghao Aaron Li, Adrian Nicolas Florea, Cong Han, and Nima Mesgarani. 2024. Speech slytherin: Examining the performance and efficiency of mamba for speech separation, recognition, and synthesis. *arXiv preprint arXiv:2407.09732*.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. Bilex rx: Lexical data augmentation for massively multilingual machine translation. *Preprint*, arXiv:2303.15265.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015, Conference Track Proceedings*.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*.

Jonathan Mbuya and Antonios Anastasopoulos. 2023. GMU systems for the IWSLT 2023 dialect and low-resource speech translation tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 269–276, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

John E Ortega and Krishnan Pillaipakkamnatt. 2018. Using morphemes from agglutinative languages like quechua and finnish to aid in low-resource translation. *Technologies for MT of Low Resource Languages (LoResMT 2018)*, page 1.

John E Ortega, Rodolfo Joel Zevallos, Ibrahim Sa'id Ahmad, and William Chen. 2024. Quespa submission for the iwslt 2024 dialectal and low-resource speech translation task. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 125–133.

Rosa YG Paccotacya-Yanque, Candy A Huanca-Anquise, Judith Escalante-Calcina, Wilber R Ramos-Lovón, and Álvaro E Cuno-Parari. 2022. A speech corpus of quechua collao for automatic dimensional emotion recognition. *Scientific Data*, 9(1):778.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou

Zhang, Yui Sudo, Muhammad Shakeel, Jee-Weon Jung, Soumi Maiti, and Shinji Watanabe. 2023. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Jiatong Shi, William Chen, Dan Berrebbi, Hsiu-Hsuan Wang, Wei-Ping Huang, En-Pei Hu, Ho-Lam Chuang, Xuankai Chang, Yuxun Tang, Shang-Wen Li, Abdelrahman Mohamed, Hung-Yi Lee, and Shinji Watanabe. 2023. Findings of the 2023 ml-superb challenge: Pre-training and evaluation over more languages and beyond. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Shengkun Tang, Liqun Ma, Haonan Li, Mingjie Sun, and Zhiqiang Shen. 2024. Bi-mamba: Towards accurate 1-bit state space models. *Preprint*, arXiv:2411.11843.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

Rodolfo Zevallos, Nuria Bel, Guillermo Cámbara, Mireia Farrús, and Jordi Luque. 2022a. Data augmentation for low-resource quechua asr improvement. *arXiv preprint arXiv:2207.06872*.

Rodolfo Zevallos, Luis Camacho, and Nelsi Melgarejo. 2022b. Huqariq: A multilingual speech corpus of native languages of peru forspeech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5029–5034.

Xiangyu Zhang, Qiquan Zhang, Hexin Liu, Tianyi Xiao, Xinyuan Qian, Beena Ahmed, Eliathamby Ambikairajah, Haizhou Li, and Julien Epps. 2024. Mamba in speech: Towards an alternative to self-attention. *arXiv preprint arXiv:2405.12609*.