QUESPA Submission for the IWSLT 2024 Dialectal and Low-resource Speech Translation Task

John E. Ortega[1], **Rodolfo Zevallos**[2], **William Chen**[3], **Ibrahim Said Ahmad**[1]
[1]Northeastern University, USA, [2]Universitat Pompeu Fabra, Spain
[3]Carnegie Mellon University, USA
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 2024: dialectal and low-resource speech translation. Two main submission types were supported in the campaign: *constrained* and *unconstrained*. This is our second year submitting our ST systems to the IWSLT shared task and we feel that we have achieved novel performance, surpassing last year's submissions. Again, we were able to submit six total systems of which our best (primary) *constrained* system consisted of an ST model based on the Fairseq S2T framework where the audio representations were created using log mel-scale filter banks as features and the translations were performed using a transformer. The system was similar to last year's submission with slight configuration changes, allowing us to achieve slightly higher performance (2 BLEU). Contrastingly, we were able to achieve much better performance than last year on the *unconstrained* task using a larger pre-trained language (PLM) model for ST (without cascading) and the inclusion of parallel QUE–SPA data found on the internet. The 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 19.7 BLEU. Additionally, we present the other four submissions (2 constrained and 2 unconstrained) which are part of additional efforts of hyper-parameter and configuration tuning on existent models and the inclusion of Whisper for speech recognition.

1 Introduction

Speech Translation (ST) has historically been a difficult task due to the lack of parallel data required to train neural end-to-end systems. As such, the traditional approach to this task has been to use a cascade of distinct modules, separating ST into the subtasks of Automatic Speech Recognition (ASR) and Machine Translation (MT). While this allows ST systems to benefit from the advances in Pretrained Language Models (PLMs) for ASR and MT, creating usable models for low-resource languages has remained a challenge due to the lack of support for these languages in PLMs. Findings from previous iterations of IWSLT (Antonios et al., 2022; Agarwal et al., 2023a) clearly show this phenomena: large-scale ensembling and multilingual supervised pre-training are required to even reach 15 BLEU (Papineni et al., 2002) in low-resource pairs such as Quechua–Spanish.

This year, the IWSLT 2024 (Agarwal et al., 2023b) evaluation campaign for low-resource and dialect speech translation has included several language pairs for which many teams have submitted to the *unconstrained* task. Some language pairs such as Bemba–English have recorded BLEU scores as low as 0.5. We feel that as second-time entries we are able to rely on previously built ST systems to leverage our work on the Quechua to Spanish (QUE–SPA) language pair.

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 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 high-resource language, 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

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 173–181
August 15-16, 2024 ©2024 Association for Computational Linguistics

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 three different institutions: Northeastern University, Carnegie Melon University, and Pompeu Fabra University. A new organizer has been introduced this year who has expertise in African languages. All of the IWSLT 2023 organizers have continued to work on the project; all of the previous organizers have had experience with the QUE–SPA language pair in the past. In this article, we report the QUESPA consortium submission for the IWSLT 2024 and once again focus on the low-resource task at hand by combining *all* the two dialects *Quechua I and II* into one.

The rest of this article is organized as follows. Section 2 presents the related work. The experiments for QUE–SPA low-resource track are presented in Section 3. Section 4 provides results from the six submitted systems and concludes this work.

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., 2023a) 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 low-resource languages. While this is generally done by pairing high-resource languages with low-resource ones, it can also be beneficial in settings where only low-resource languages 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. 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 low-resource languages.

3 Quechua-Spanish

In this section we present our experiments for the QUE–SPA dataset provided in the low-resource ST track at IWSLT 2024, identical to the dataset from

IWSLT 2023. As a reminder, the audio consists of contains 1 hour and 40 minutes of *constrained* speech along with its corresponding translations and nearly 48 hours of ASR data (with transcriptions) from the Siminichik (Cardenas et al., 2018) corpus. As an additional constrained setting, the dataset offers the QUE–SPA MT corpus 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).

We present the six submissions for both the *constrained* and *unconstrained* as follows:

1. a primary constrained system that uses a direct ST approach with a extra small transformer (Vaswani et al., 2017; Wang et al., 2020);

2. a contrastive 1 constrained system that uses a direct ST approach with a medium (default) transformer (Vaswani et al., 2017; Wang et al., 2020) along with several data augmentation techniques;

3. a contrastive 2 constrained system that uses a direct ST approach with a medium (default) transformer (Vaswani et al., 2017; Wang et al., 2020) without data augmentation techniques;

4. a primary unconstrained system consisting of a SpeechT5 model fine-tuned for speech translation with one data augmentation technique;

5. a contrastive 1 unconstrained system consisting of a SpeechT5 model fine-tuned for speech translation with two data augmentation techniques;

6. a contrastive 2 unconstrained system consisting of a Whisper (Radford et al., 2023) ASR model fine-tuned for speech translation and cascaded with the NLLB MT system.

We present the experimental settings and results for all systems starting off with constrained systems in Section 3.1 and continuing with the unconstrained systems in Section 3.2. Finally, we offer results and discussion in Section 4.

3.1 Constrained Setting

Identical to last year, the IWSLT 2024 constrained setting for QUE–SPA consists of two main datasets.

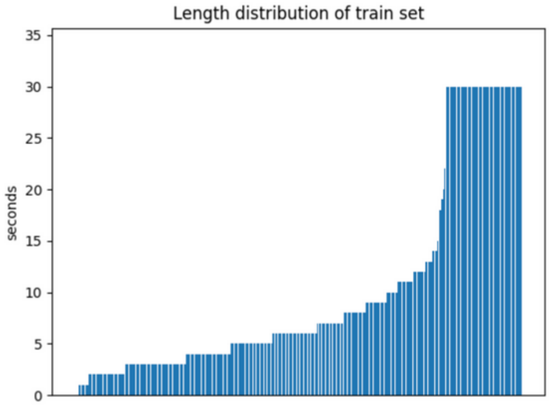

Figure 1: Training set audio lengths vary from 1 to 30 seconds while validation and test set are 30 seconds long.

First, the speech translation dataset consists of 1 hour and 40 minutes divided into 573 training files, 125 validation files, and 125 test files where each file is a .wav file with a corresponding transcription and human-validated translation from Siminchik (Cardenas et al., 2018). Secondly, there is a MT data set combined by previous work (Ortega et al., 2020) which consists of 100 daily magazine article sentences and 51140 sentences which are of religious context in nature.

This year, one of the findings we observed is that the dataset has uneven distributions between training and validation/test. The training set largely consisted of utterances shorter than 20 seconds (Figure 1), while the validation and test set was almost exclusively 30 seconds long inputs. This is something that the organizers plan to rearrange for next year's challenged, but this type of mismatch can be considered a hurdle due to the difference (smaller utterances result in unbalance). For this submission, we found it somewhat difficult to train direct ST systems under the constrained settings. However, we present the following systems that mitigate the concern considering that our results outperform the best performing *constrained* systems.

Development of the *Primary*, *Contrastive 1*, and *Contrastive 2* systems consisted of an extension of the original ST systems built in IWSLT 2023. During development, several experiments led us to the best performing systems (Primary and Contrastive 2). The developmental process is documented in Table 1 for a historical way of showing our path to the final systems.

Model Type	Optimization	Learning Rate	Checkpoint	BLEU
s2t_transformer	Adam	0.002	best of the last 10 on 500 epochs	0.9
s2t_transformer_xs	Adamax	0.0001	best of the last 10 on 500 epochs	0.6
2t_transformer_xs	Adamax	0.0001	best of the last 10 on 500 epochs	0.6
s2t_transformer_xs	Adam	0.0001	best of the last 10 on 500 epochs	0.7
s2t_transformer_large	Adam	0.001	best of the last 10 on 500 epochs	0.0
s2t_transformer	Adamax	0.001	best of the last 10 on 500 epochs	1.0
s2t_transformer	Adamax	0.001	best of the last 10 on 400 epochs	1.0
s2t_transformer	Adamax	0.001	best of the last 10 on 300 epochs	1.0
s2t_transformer	Adamax	0.001	best of the last 10 on 200 epochs	1.0
s2t_transformer	Adamax	0.001	best of the last 10 on 100 epochs	1.0
s2t_transformer	Adamax	0.001	avg of the last 10 on 400 epochs	1.4

Table 1: BLEU scores on developmental models for the *contrained* settings using beam size of five.

3.1.1 Primary System

The **Primary** System is similar to previous work (Ortega et al., 2023). The dataset has not changed since their work and our system consists of the use of a direct ST approach.

Again, we use the Fairseq (Ott et al., 2019) toolkit to perform direct ST using the 573 training files, a total of 1.6 hours of audio. The use of feature extraction through log mel-filter bank (MFB) features and is still based on the S2T approach by (Wang et al., 2020). Identically, we generate a 1k unigram vocabulary for the Spanish text using SentencePiece (Kudo and Richardson, 2018), with no pre-tokenization. This year's model consists of a convolutional feature extractor and transformer encoder-decoder (Vaswani et al., 2017), also known as the "extra-small transformer", (s2t_transformer_xs) with 6 encoder layers and 3 decoder layers. Error is measured using cross entropy and optimization is done using Adam. Our model was run for 500 epochs with a learning rate of .0002. For this submission, the main difference is that we use a device that allows us to **average** the 10 last checkpoints through PyTorch[1]. We compared the average to the best of the last 10 checkpoints and found that the average performed better.

3.1.2 Contrastive 1 System

The **Contrastive 1** system is based on a transformer much like the Primary system. However, Contrastive 1 uses two novel techniques introduced that were not present in the IWSLT 2023 QUESPA submission (Ortega et al., 2023): (1) a new model size which contains more layers and (2) five new data augmentation techniques based on the data at hand.

As was done in the Primary system, the Fairseq (Ott et al., 2019) toolkit is used to perform direct ST on the training data of 1.6 hours of audio. Identical feature extraction techniques are used via the log mel-filter bank (MFB) features from the S2T approach in previous work (Wang et al., 2020). Also, we generate a 1k unigram vocabulary for the Spanish text using SentencePiece (Kudo and Richardson, 2018), with no pre-tokenization.

The first main difference is the model. The Contrastive 1 model consists of a convolutional feature extractor and transformer encoder-decoder (Vaswani et al., 2017); but it uses the medium-sized transformer, also known as the "transformer", (s2t_transformer) with 12 encoder layers and 6 decoder layers. Additionally, Contrastive 1 has 8 decoder attention heads as opposed to 4 in the Primary system.

The second difference we consider a *major* difference – the use of data augmentation to increase the input size. Augmentation techniques were used from previous work using LibRosa[2]. More specifically, code can be found online[3] to reproduce our experiments. The increase in the input training dataset increased four fold using the following four techniques for augmentation: *Noise*, *Roll*, *Time*, and *Pitch*. The noise addition (augmentation) is done using an aggregation of 0.009. The Roll adjustment is of $sr/10$. Time is through a stretch factor of 0.4 and Pitch is of -5. With the increase of input size, experiments ran slower yet were not of significant impact. We save further iterations of data augmentation as future work as we believe that is has had an impact here.

Error is measured using cross entropy and optimization is done using Adam. Other hyperparameter choices that were not the same as the Primary submission include the exclusion of **SpecAugment** (Park et al., 2019) as an audio aug-

[1] https://pytorch.org/

[2] https://librosa.org/

[3] https://colab.research.google.com/gist/keyurparalkar/5a

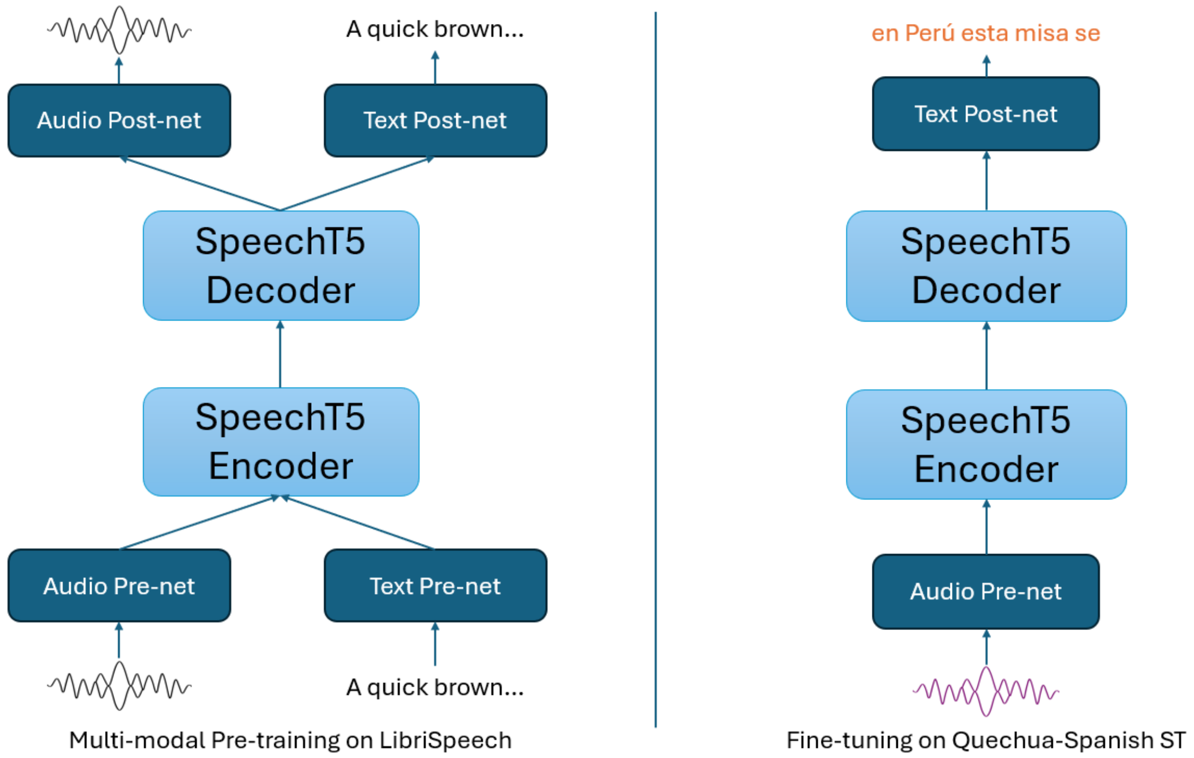

Figure 2: The best-performing *unconstrained* speech translation pipeline. We use a pre-trained SpeechT5 (Ao et al., 2022) on English, and fine-tune it on direct Quechua-to-Spanish ST.

mentation technique and the choice of 200 epochs as opposed to 500 used in the Primary submission. An average checkpoint method was used identical to the one in the Primary system (average of the last 10 checkpoints using Pytorch).

3.1.3 Contrastive 2 System

The **Contrastive 2** System is identical to the constrained Primary system in Section 3.1.1 with one main difference – model size. The model size of this Contrastive 2 system uses a medium-sized transformer known as the "transformer", (s2t_transformer) with 12 encoder layers and 6 decoder layers identical to the Contrastive 1 system defined in Section 3.1.2. All other hyperparameters were identical to the Primary system with the exception of the number of epochs which was 400 as opposed to 500.

3.2 Unconstrained Setting

Just like in IWSLT 2023, 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[4] (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. Finally, all the datasets described in this section allowed for further fine-tuning of the previously trained end-to-end speech translation model.

3.2.1 Primary System

The Primary 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 encoder/decoder's pre/post-

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

Team **QUESPA** BLEU and CHRF Scores			
Constrained			
System	**Description**	**BLEU**	**CHRF**
primary	mfb + s2t-extrasmall + avg	2.0	30.0
contrastive 1	mfb + s2t-med + aug + avg	1.3	30.9
contrastive 2	mfb + s2t-med + avg	1.4	30.3
Unconstrained			
System	**Description**	**BLEU**	**CHRF**
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 2024.

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[5] for Speech Translation using the SpeechT5 fine-tuning recipe[6] for Speech-Translation with the same hyperparameter settings. We used the 48 hours of audio provided by the organizers. We applied nlpaug a data augmentation technique (noise, distortion, duplication)[7] (Ma, 2019), resulting in a total of 96h: 48h original + 48h synthetic data.

3.2.2 Contrastive 1 System

The Contrastive 1 system is nearly identical to the Primary System for the unconstrained setting. However, we used the 48 hours described in 3.2, totally translate to Spanish. Moreover, we added 19 minutes of Guarani and 29 minutes of Bribi, along with their translations as described 3.2. Additionally, we applied two data augmentation techniques: (1) nlpaug (Ma, 2019) and (2) DA-TTS (Zevallos et al., 2022), which involves generating synthetic text and audio using a delexicalization algorithm and a TTS system for the source language (Quechua). These two data augmentation techniques generated 48 hours and 48 hours respectively. We used in total 151h and 48 min: 55h (new dataset) + 48 min (ANLP dataset) + 48h nlpaug + 48h DA-TTS.

3.2.3 Contrastive 2 System

The Constrastive 2 system is a new introduction this year for our team. We felt that the Whisper (Radford et al., 2023) ASR model would outperform QUESPA's 2023 cascaded system (Section 4 Table 1, *called fleurs+lm+floresmt*) (Ortega et al., 2023). However, despite the use of the same machine translation system (floresmt) (NLLB Team et al., 2022), we were unable to achieve better performance.

We use a Whisper ASR model that has been pre-trained on multiple languages (multi-lingual). In total, the Whisper model is trained on 680,000 hours of which 117,000 is multilingual, including nearly 96 languages. We use the medium variant of Whisper, which has 770M parameters. In our experiments, we fine-tune the Whisper model on the ASR training data, as the first part of an ASR+MT cascade. The output from Whisper (Quecha text) is then used as input to the same MT system from last year (called floresmt) that translates that Quechua text to Spanish.

4 Results and Discussion

Results are presented in Table 2. The constrained systems continue to be a difficult problem to solve with our best-performing system scoring a maximum of 2 BLEU (1.96 when measured with two decimal points). It is clear that a constrained system of this nature could not be deployed in the wild at this point. Nonetheless, it is a promising increase of nearly 1 BLEU point when compared to IWSLT 2023 results. Additionaly, the novel addition of data augmentation has proven to be a good first step to solving the constrained problem. In past

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

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

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

IWSLT tasks and in the current one, constrained systems are not realistically able to achieve much more than 5 BLEU points when the audio data is less than 5 hours in length.

For the unconstrained setting, our findings have shown that in the past year several novel PLMs have been created that surpass previous models. It is clear that Speech Translation as a task is becoming more solvable with pre-trained techniques that perform transfer learning. The combination of the Microsoft Speech T5 model with data augmentation as shown in Figure 2 is a new approach that has not been applied to the QUE–SPA language pair in the past and can be considered the best performing system as of this date to our knowledge. Previous systems based on w2vletter (Pratap et al., 2019) performed well but did not surpass the Microsoft Speech T5 Model in our experiments.

5 Conclusion and Future Work

Our submission to the IWSLT 2024 (Agarwal et al., 2023b) 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. More specifically, we have been successful by changing the sizes of the models in the *constrained* setting and changing the type of models in the *unconstrained* setting. Additionally, we have shown that different data augmentation techniques can be used for increased performance on both tasks.

We save for future work the experimentation of data augmentation techniques which seem to be the most advantageous novelty in this year's submission. In our opinion, data augmentation can be used for benefits in both the unconstrained and constrained tasks. Our plan for future IWSLT tasks is to experiment with and without features like SpecAugment, roll addition, and more.

References

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. 2023a. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

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. 2023b. Findings of the IWSLT 2024 Evaluation Campaign. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*. Association for Computational Linguistics.

Anastasopoulos Antonios, Barrault Loc, Luisa Bentivogli, Marcely Zanon Boito, Bojar Ondřej, Roldano Cattoni, Currey Anna, Dinu Georgiana, Duh Kevin, Elbayad Maha, et al. 2022. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157. 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.

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.

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.

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

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

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 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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

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.

Vineel Pratap, Awni Y. Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. 2019. Wav2letter++: A fast open-source speech recognition system. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6460–6464.

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.

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.

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.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

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