Evaluating Self-Supervised Speech Representations for Indigenous American Languages

Chih-Chen Chen[*1], **William Chen**[*2], **Rodolfo Zevallos**[3], **John E. Ortega**[4]
Taipei Medical University[1], Carnegie Mellon University[2]
Universitat Pompeu Fabra[3], Northeastern University[4]
williamchen@cmu.edu

Abstract

The application of self-supervision to speech representation learning has garnered significant interest in recent years, due to its scalability to large amounts of unlabeled data. However, much progress, both in terms of pre-training and downstream evaluation, has remained concentrated in monolingual models that only consider English. Few models consider other languages, and even fewer consider indigenous ones. In this work, benchmark the efficacy of large SSL models on 6 indigenous America languages: Quechua, Guarani , Bribri, Kotiria, Wa'ikhana, and Totonac on low-resource ASR. Our results show surprisingly strong performance by state-of-the-art SSL models, showing the potential generalizability of large-scale models to real-world data.

Keywords: Indigenous Languages, Low-resource, Self-Supervised Learning

1. Introduction

In recent years, the fields of Natural Language Processing (NLP) and speech processing has witnessed remarkable advancements, with applications ranging from machine translation and sentiment analysis to voice assistants and chatbots. These developments have predominantly focused on widely spoken languages such as English and Mandarin. However, the vast linguistic diversity represented by indigenous languages across the globe remains largely unexplored in the context of language processing.

Indigenous languages are the ancestral tongues of diverse communities with rich cultural heritage and profound connections to the environment in which they are spoken. These languages often exhibit distinct linguistic characteristics, deviating from the structures and conventions of widely studied languages. By expanding the scope of language processing to include indigenous languages, we can foster linguistic inclusivity and empower indigenous communities to participate in the digital era while preserving their linguistic and cultural identities.

One compelling reason to focus on indigenous languages in language processing is the potential for social impact. Many Indigenous communities face challenges related to limited access to information and technology, which further exacerbate social and economic disparities. By developing NLP models and applications tailored to indigenous languages, we can bridge the digital divide and enable these communities to leverage technology for communication, education, and cultural preservation. Such efforts have the potential to enhance language revitalization efforts, foster inter-generational transmission of knowledge, and promote cultural preservation within indigenous communities.

We evaluate the effectiveness of different multilingual self-supervised learning (SSL) models on Automatic Speech Recognition (ASR) for several indigenous American languages: Quechua, Bribri, Guarani, Kotiria, Wa'ikhana, and Totonac. We first introduce general characteristics of American indigenous languages and discuss the challenges in modeling them due to their unique linguistic natures. We then provide a brief overview of each language to this study, highlighting some key linguistic properties. As our other core contribution, we discuss research in American indigenous languages in both the fields of NLP and speech processing, hoping to create a bridge in the literature for communities.

2. American Indigenous Languages

American Indigenous Languages encompass a diverse range of language families and isolates, each with its own linguistic features. The languages of this region exhibit a remarkable variety of phonological, morphological, syntactic, and semantic structures, reflecting the rich linguistic diversity of the continent. A persistent challenge in modelling these languages, similar to many indigenous languages, is the frequency of code-switching. Coupled with the lack of both linguistic and electronic resources for these languages, creating language technologies for indigenous languages remains a significant challenge despite the exponential progress in NLP and speech processing.

Broadly speaking, the indigenous languages of the Americas are morphologically-rich, often exhibiting agglutinative or polysynthetic structures. These languages tend to have extensive systems of affixation, where morphemes are added to roots to convey meaning and grammatical information. For

LREC-COLING 2024, pages 6444–6450
20-25 May, 2024.

example, in Quechua or Guarani, complex words can be formed through the addition of numerous affixes to a single root. This makes them particularly challenging for NLP and language modelling tasks, due to the higher frequency of rare words.

2.1. Quechua

Quechua is a family of closely related languages spoken by around 10 million people across South America. While primarily spoken in the Andean regions. Quechua is considered one of the most widely spoken indigenous language families in the Americas. While there are regional variations, Quechua languages share many common linguistic characteristics. These variations are broadly separated into two distinct categories: Quechua I and Quechua II. The former refers to the varieties of Quechua spoken in the central parts of Peru, while the latter is spoken in Southern Peru, Bolivia, and Colombia.

2.2. Bribri

Bribri, also known as the Bribri-Poró language, is spoken by the Bribri people of Costa Rica. It belongs to the Chibchan language family, which is primarily found in Central America. The Bribri language specifically falls under the Guaymí subgroup of the Chibchan family. Geographically, the Bribri language is primarily spoken in the Talamanca region of Costa Rica, specifically in the southern parts of Limón and northern parts of Puntarenas provinces. It is a tonal language, meaning that pitch variations can distinguish between different words or meanings.

2.3. Guarani

The Guarani language is an indigenous language spoken by the Guarani people in South America. It is a member of the Tupi-Guarani language family, which encompasses several languages across Brazil, Paraguay, Argentina, and Bolivia. Guarani is one of the most widely spoken indigenous languages in the Americas, with 4-6 million speakers. Guarani is mainly distributed in Paraguay, where it has official status alongside Spanish. It is also spoken in parts of northeastern Argentina, southeastern Bolivia, and southern Brazil. Like many other American languages, Guarani is agglutinative.

2.4. Kotiria

Kotiria, also known as Wanano, is an indigenous language spoken by the Kotiria people who reside in the Vaupés region of Colombia and Brazil. Kotiria is part of the larger Eastern Tukanoan language family, which includes several other indigenous languages spoken in the northwest Amazon region. Kotiria language is primarily spoken in the upper and middle basins of the Vaupés River, which runs through the Amazon rainforest. The language is concentrated in remote areas of the Colombian Vaupés Department and the Brazilian state of Amazonas. Like Bribri, it is both agglutinative and tonal.

2.5. Wa'ikhana

Also known as Cubeo, Wa'ikhana is spoken by the Cubeo people in the northwest Amazon region, primarily in Colombia and Brazil. Wa'ikhana belongs to the Tucanoan language family, which encompasses several indigenous languages spoken in the northwest Amazon. Wa'ikhana is also both agglutinative and tonal.

2.6. Totonac

The Totonac language is an indigenous language spoken by the Totonac people in Mexico. It belongs to the Totonacan language family, which is primarily found in the states of Veracruz, Puebla, and parts of Hidalgo in eastern Mexico. Totonac is both agglutinative and tonal.

3. Indigenous Languages in Language Processing

3.1. Community Efforts

In the field of NLP, several initiatives have been started to encourage further research in indigenous languages. While the majority are workshops for general low-resource NLP (Ortega et al., 2021, 2022), newer efforts have also targeted indigenous languages (Mager et al., 2021b, 2023; Orife et al., 2020; Nekoto et al., 2020). For American indigenous languages specifically, the AmericasNLP (Mager et al., 2021b, 2023) community has helped driven research by improving the visibility of authors from indigenous communities. AmericasNLP also hosts an annual shared task, similar to those found in machine and speech translation workshops (Ebrahimi et al., 2023; Mager et al., 2021a), to further integrate state-of-the-art methods with indigenous languages.

In speech processing, research for indigenous languages is more *ad hoc*, with numerous decentralized efforts from a variety of research groups. Contrary to NLP, indigenous languages play a more common role in SOTA models (Chen et al., 2023; Babu et al., 2021; Radford et al., 2022; Pratap et al., 2020; Zhang et al., 2023) and benchmarks (Conneau et al., 2023; Shi et al., 2023b; Gales et al., 2014). Annual challenges, primarily for speech

translation, also help bring SOTA methods to these languages (Agarwal et al., 2023).

3.2. Research for Quechua

Quechua has received substantial attention in NLP, driven by its larger population and resources compared to other American languages. Early work focused on morphological analysis using finite state transducers (Rios Gonzales and Castro Mamani, 2014; Rios; Rios Gonzales and Göhring, 2013) and toolkits development (Rios, 2015; Rios et al., 2008; Rios, 2011). Neural methods were later applied, initially in machine translation (Ortega and Pillaipakkamnatt, 2018; Ortega et al., 2020; Chen and Fazio, 2021), and more recently in masked language models (Zevallos et al., 2022b).

However, Quechua speech processing has seen fewer studies due to limited available data. Siminchik (Cardenas et al., 2018) was the first Quechua speech corpus, although full data release didn't occur. Huqariq (Zevallos et al., 2022a), a multilingual collection of Peruvian languages, including Quechua, remains unreleased. Quechua was featured in speech processing challenges like AmericasNLP 2022 and IWSLT 2023 (Agarwal et al., 2023), with the latter marking the first evaluation using Transformer-based methods (Vaswani et al., 2017). Participants mostly relied on pre-trained SSL models (E. Ortega et al., 2023), but the potential of other SSL approaches for Quechua remains unexplored, as participants mainly used XLSR 53 (Conneau et al., 2020) or XLS-R 128 (Babu et al., 2021).

4. Experimental Setup

4.1. Data

We experimented with six indigenous American languages: Quechua , Bribri, Guarani, Kotiria, Wa'ikhana, and Totonac . The Quechua is based on the Siminchik corpus (Cardenas et al., 2018), which contains recordings of two Quechua II dialects: Chanca Quechua (spoken mainly in Ayacucho and surroundings) and Collao Quechua (spoken in Cusco and Puno). Siminchik (Cardenas et al., 2018) consists of crowd-sourced transcriptions of radio recordings from these regions, totaling 97 hours of audio. The audio clips were segmented to a maximum of 30 seconds, and the transcripts underwent punctuation removal, casing normalization, and interjection standardization due to dialectal differences. Additionally, the ASR transcripts were normalized using a finite state transducer-based toolkit (Rios Gonzales and Castro Mamani, 2014) adhering to the Chanca dialect's spelling. The data for Bribri, Guarani, Kotiria and Wa'ikhana were obtained from the 2022 edition of AmericasNLP challenge (Mager et al., 2021b, 2023). Since the official test splits remain hidden, we created our own by dividing the provided validation sets. The Totonac data was obtained from a prior study on efficiently fusing SSL models with spectral featrures(Berrebbi et al., 2022) All datasets adhere to the ML-SUPERB format (Shi et al., 2023a), with 1-hour and 10-minute training sets, a 10-minute validation set, and a 10-minute testing set, obtained by randomly sampling from the appropriate split in the original dataset. This assesses semi-supervised model performance for indigenous languages in real-world scenarios due to limited labeled and unlabeled data.

4.2. Self-Supervised Models

We evaluate three SSL models on each language, along with log-Mel filterbank features (FBANK). The models are described as follows:

4.2.1. XLSR 53

XLSR 53 (Conneau et al., 2020) is trained on 56k hours of multilingual data for 53 languages, which are pre-dominantly European. It uses the 317M parameter wav2vec architecture (Schneider et al., 2019), which consists of a convolutional feature extractor and Transformer encoder (Vaswani et al., 2017) trained with contrastive loss.

4.2.2. XLS-R 128

XLS-R 128 (Babu et al., 2021) is the large-scale extension of XLSR 53, trained on 436k hours of multilingual data across 128 languages. It instead uses the wav2vec 2.0 (Baevski et al., 2020) architecture, which also includes a convolutional feature extractor and Transformer encoder, but is trained with both contrastive and codebook prediction losss.

4.2.3. mHuBERT

mHuBERT (Lee et al., 2022) builds off of the HuBERT (Hsu et al., 2021) architecture, which uses an iterative approach to SSL. HuBERT models are trained to predict discrete representations of masked speech. After each iteration of pretraining, hidden representations are extracted from the model and clustered using k-means, creating the discrete targets for the next round of pretraining. mHuBERT was trained multilingually on 3 languages: Spanish, French, and Italian, each 4.5k hours of data. It uses the 95M parameter HuBERT Base architecture, which modifies the wav2vec 2.0 design for pure codebook prediction.

Table 1: Evaluation of SSL models on each indigenous language on the 10-minute set, measured in character error rate (CER ↓).

Model	Hours	Quechua	Bribri	Guarani	Kotiria	Wa'ikhana	Totonac	Average
XLSR 53	56k	47.8	54.6	37.6	64.2	83.3	29.6	52.9
XLS-R 128	436k	**42.5**	**49.5**	**27.5**	**51.2**	**62.2**	**27.7**	43.4
mHuBERT	13.5k	47.7	54.3	35.2	64.8	84.8	30.1	52.8

Table 2: Evaluation of SSL models on each indigenous language on the 1 hour set, measured in character error rate (CER ↓).

Model	Hours	Quechua	Bribri	Guarani	Kotiria	Wa'ikhana	Totonac	Average
XLSR 53	56k	37.5	49.5	31.5	49.9	62.4	26.0	42.8
XLS-R 128	436k	**34.0**	**44.1**	**24.0**	**43.4**	**55.1**	**20.6**	**36.8**
mHuBERT	13.5k	37.1	49.2	32.0	50.6	62.3	26.1	42.8

4.3. Training Settings

We conduct all experiments using the ESPnet (Watanabe et al., 2018) toolkit with the official settings of the ML-SUPERB competition (Shi et al., 2023a). The SSL model is used as a frozen feature extractor, such that the hidden representation of each layer is obtained. The layer-wise outputs of combined via a weighted sum, where the weight is learned during training. These outputs are then down projected to a hidden size of 80 and then augmented with SpecAug (Park et al., 2019), before being used as the model inputs of a Transformer encoder (Vaswani et al., 2017). The Transformer consists of 2 layers, each with a hidden size of 256, 8 attention heads, and a feed-forward size of 1024. Models are trained with CTC loss (Graves et al., 2006) and the Adam optimizer (Kingma and Ba, 2015), with a constant learning rate of 0.0001. Models are trained for a maximum of 15,000 steps and the 5 best checkpoints are averaged for inference, which is performed with CTC greedy decoding.

5. Results

Our experimental results are presented in Tables 1 and 2 for the 10-minute and 1-hour settings respectively. Models are evaluated in character error rate (CER).

Similar to the results on the complete ML-SUPERB benchmark, XLS-R 128 (Babu et al., 2021) obtains the highest overall scores in both data settings. The results presented here are even more distinct: XLS-R 128 outperforms all other models every single task. This suggests the powerful generalizability of large-scale multilingual SSL: all evaluated languages (aside from Guarani) were unseen during pre-training. The distance between the other two models, XLSR 53 and mHuBERT, is much smaller, with only a difference of 0.1 average CER on the 10-minute track and no significant difference on the 1-hour track. A strong future research question would be to isolate the cause for the lack of difference, as one would expect the model trained on more languages to generalize better.

Overall, we find the results of our evaluation surprisingly strong. The average CER of XLS-R 128 on the 10-minute / 1 hour set is 43.4 / 36.8, only 3.7 / 6.2 CER higher than its average monolingual score on ML-SUPERB (39.7 / 30.6 CER) (Shi et al., 2023a). Our results suggest that multilingual pre-training generalizes well to unseen languages during fine-tuning, allowing them to receive the benefits of pre-training large-scale unlabeled data from high-resource languages

6. Conclusion

While the recent progress of deep learning in NLP and speech processing has significantly accelerated the development of language technologies, the progress has been unequally distributed. We are the first benchmark the effectiveness of large-scale speech SSL models on ASR for indigenous American languages such as Quechua, which are known to be among the most difficult for NLP. We find surprisingly positive results, showing the impressive generalization ability of large-scale multilingual SSL models on new languages.

Acknowledgements

The third author has been supported by a FI grant of the Catalan Funding Agency for Research and Universities (AGAUR).

7. Bibliographical References

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

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*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Dan Berrebbi, Jiatong Shi, Brian Yan, Osbel López-Francisco, Jonathan Amith, and Shinji Watanabe. 2022. Combining Spectral and Self-Supervised Features for Low Resource Speech Recognition and Translation. In *Proc. Interspeech 2022*, pages 3533–3537.

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

William Chen, Xuankai Chang, Yifan Peng, Zhaoheng Ni, Soumi Maiti, and Shinji Watanabe. 2023. Reducing barriers to self-supervised learning: Hubert pre-training with academic compute. In *Proc. Interspeech*.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In *Proc. LoResMT*, pages 20–31.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

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 *Proc. IWSLT*, pages 261–268.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proc. AmericasNLP*, pages 206–219.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA).

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*, pages 369–376.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. ICLR*.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan

Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022. Textless speech-to-speech translation on real data. In *Proc. NAACL*, pages 860–872.

Manuel Mager, Abteen Ebrahimi, Arturo Oncevay, Enora Rice, Shruti Rijhwani, Alexis Palmer, and Katharina Kann, editors. 2023. *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021a. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proc. AmericasNLP*, pages 202–217.

Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors. 2021b. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2144–2160.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

John Ortega, Atul Kr. Ojha, Katharina Kann, and Chao-Hong Liu, editors. 2021. *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*. Association for Machine Translation in the Americas.

John Ortega and Krishnan Pillaipakkamnatt. 2018. Using morphemes from agglutinative languages like Quechua and Finnish to aid in low-resource translation. In *Proc. LoResMT*, pages 1–11.

John E. Ortega, Marine Carpuat, William Chen, Katharina Kann, Constantine Lignos, Maja Popovic, and Shabnam Tafreshi, editors. 2022. *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Workshop 2: Corpus Generation and Corpus Augmentation for Machine Translation)*. Association for Machine Translation in the Americas.

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.

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. *Proc. Interspeech*, pages 2613–2617.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Annette Rios. Applying finite-state techniques to a native american language: Quechua.

Annette Rios. 2011. Spell checking an agglutinative language: Quechua.

Annette Rios. 2015. *A basic language technology toolkit for quechua*. Ph.D. thesis, University of Zurich.

Annette Rios, Anne Göhring, and Martin Volk. 2008. A quechua-spanish parallel treebank. *Lot occasional series*, 12:53–64.

Annette Rios Gonzales and Richard Alexander Castro Mamani. 2014. Morphological disambiguation and text normalization for Southern Quechua varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 39–47, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Annette Rios Gonzales and Anne Göhring. 2013. Machine learning disambiguation of Quechua verb morphology. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 13–18, Sofia, Bulgaria. Association for Computational Linguistics.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech*, pages 3465–3469.

Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, et al. 2023a. ML-SUPERB: Multilingual Speech Universal PERformance Benchmark. In *Proc. Interspeech*.

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. 2023b. Findings of the 2023 ml-superb challenge: Pre-training and evaluation over more languages and beyond. In *Proc. 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. *Proc. NeurIPS*, 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, Luis Camacho, and Nelsi Melgarejo. 2022a. 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, Marseille, France. European Language Resources Association.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022b. Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13, Hybrid. Association for Computational Linguistics.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.