

Building Community-Centred NLP Resources for Puno Quechua

Elwin Huaman¹, Adrian Gamarra Lafuente², Johanna Cordova³, Anna Korhonen¹

¹University of Cambridge (UK), ²Stanford University (USA), ³ERTIM - Inalco (France)

Correspondence: elh97@cam.ac.uk, agamarra@stanford.edu, johanna.cordova@inalco.fr

Abstract

The preservation of under-resourced languages requires digital tools and resources shaped by and for their speakers. We present the first dedicated ASR resources for Puno Quechua (ISO 639-3: qxp): (1) the largest speech corpus for any single Quechua variety, consisting in 66 hours of recordings for scripted and spontaneous speech (including 36 hours of manually transcribed and validated data), collected via a participatory design campaign; (2) the first systematic ASR benchmark for Puno Quechua, evaluating state-of-the-art models and fine-tuning Whisper-base, wav2vec2-base, and XLS-R-300M, with and without continued pre-training (CPT); (3) an open release of all datasets and fine-tuned models.

1 Introduction

The revitalization of indigenous languages depends increasingly on digital tools that promote language use and confer economic and social value (Galla, 2016). Puno Quechua (qxp) is spoken by approximately 465,000 people in the Puno region of Peru¹, yet formal education is conducted almost exclusively in Spanish, leaving its speakers largely illiterate in their own language. This literacy gap prevents native speakers from interacting with text-input AI applications (ChatGPT or Gemini), excluding them from the growing digital ecosystem.

An Automatic Speech Recognition (ASR) system represents a more human-centred solution: rather than forcing community members to adapt to text-heavy interfaces, ASR can adapt technology to the oral-centric reality of communities. Despite this need, no dedicated ASR resources exist for Puno Quechua. Prior work on Quechua ASR tends to treat the language family as a homogeneous entity: the variant(s) in question are sometimes

¹Estimation based on 2017 Peruvian National Census, https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1563/

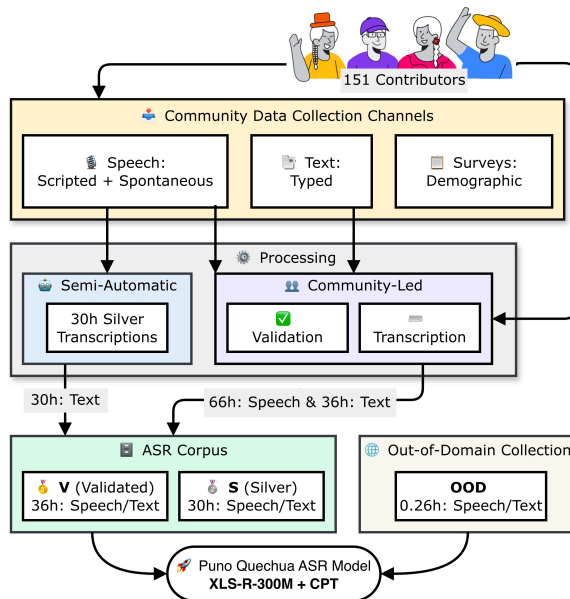


Figure 1: Puno Quechua ASR Pipeline.

not formally identified, or they are aggregated by linguistic group (e.g. Southern Quechua (Cardenas et al., 2018) or Collao Quechua (Paccotacya-Yanque et al., 2022)) without any proper examination of how their differences may affect practical applications. Furthermore, existing corpora suffer from data scarcity, restricted access (*Siminchik* by Cardenas et al., one of the largest corpora referenced in the scientific literature, has never been published in open access), and the absence of a variety-specific benchmark. Figure 1 shows an overview of the Puno Quechua ASR pipeline.

This paper makes three contributions: (1) the largest ASR corpus for a Quechua variety via participatory data collection; (2) the first systematic ASR benchmark for Puno Quechua; and (3) open-sourced code and splits,² datasets,³ and models.⁴

²<https://github.com/QuechuaBase/asr-puno-quechua>

³<https://mozilladatalcollective.com/datasets>

⁴<https://huggingface.co/QuechuaBase>

2 Background

Puno Quechua (qxp). It belongs to the Southern Quechua branch (QIIC) (Torero, 2002) and is characterised by a rich consonant inventory of 26 phonemes, including ejective, aspirated, and uvular stops, and three vowels (/a/, /i/, /u/). It exhibits Aymara influence in its phonology, suffix inventory, and vocabulary (Adelaar, 1987).

ASR for Quechua (que). ASR development for Quechua faces several compounding challenges: (a) Data scarcity: labelled speech data are extremely scarce, and when corpora exist they often aggregate varieties under a macrolanguage label (que), conflating varieties that are not mutually intelligible; (b) Morphological complexity: Quechua is highly agglutinative, resulting in poor word-level WER metrics; (c) Low written literacy: most speakers are illiterate in Quechua, making community-driven transcription difficult, and (d) No variety-specific benchmark: to our knowledge, no prior work establishes a consistent evaluation benchmark for an individual Quechua variety. Recent work by Keren et al. (2025) covers 30+ Quechua varieties within omnilingual ASR, providing reference baselines including for qxp.

3 Participatory Design Data Collection

We collected the Puno Quechua speech corpus through a four-phase participatory design process (Huaman et al., 2025; Ulla et al., 2026; Spinuzzi, 2005; Wilson et al., 2025):

Planning. Identifying the ISO 639-3 code qxp, establishing partnerships with the National University of Altiplano Puno and the local community Illariy Ch’aska, and assessing community needs.

Preparation. Setting up data governance under CC0-1.0 licence, preparing seed sentences and questions covering agriculture, healthcare, and technology, and localising the Mozilla Common Voice platform to Puno Quechua.⁵

Collection. Voluntary, skill-based contributions such as reading, speaking, listening, or writing, with community-led validation and privacy-preserving processing data.

Deployment. Open release on Mozilla Data Collective,⁶ certificates of contribution, voucher incen-

⁵<https://commonvoice.mozilla.org/qxp>

⁶<https://mozilladatacollective.com/>

Dataset	Type	Validated (V)	Silver (S)
SCS-25	Scripted	30.5	-
SPS-3	Spontaneous	5.5	30.0
Total		36.0	30.0
out-of-domain (OOD) corpus			
Add_data	Radio	0.27	-

Table 1: Datasets collected for ASR, expressed in hours.

tives for participants, and impact assessment.

The campaign ran from January to February 2026. A total of 396 volunteers registered, 292 were confirmed, 151 contributed, and 31 completed the full campaign. The resulting speech data have been aggregated and released as Common Voice Scripted Speech v25 (including v23 and v24 by Huaman et al.) with a total of 34.81 hours (30.5 validated) and Common Voice Spontaneous Speech v3 (including v1 and v2 by Huaman et al.) with a total of 35.3 hours (5.18 validated).

Table 1 summarizes the processed data that has been collected and can be used for training and evaluating models.

4 ASR models for Puno Quechua (QXP)

Datasets. Two primary corpora were used: i) SCS-25 with 30.5 validated hours of scripted speech; and ii) SPS-3 with 5.5 validated hours of spontaneous speech (after excluding recordings longer than 30 seconds and adding 1 hour validated by a native speaker), supplemented by 30 hours of automatically generated silver transcriptions using omniASR_LLM_7B model. A small OOD corpus (Add_data, ~16 minutes) sourced from radio and social media, which was transcribed and validated manually by a native speaker, provides a third evaluation domain.

Foundation models. We fine-tuned three architectures: (a) Whisper-base (74M parameters), an encoder-decoder Transformer trained on 680k hours of supervised multilingual speech. We fine-tuned setting the transcript prefix to Spanish. The unbalanced setting (V) used a learning rate (LR) of 5×10^{-6} and the balanced setting (V+S) used a learning rate of 1×10^{-5} . Audio files longer than 30 seconds were excluded. (b) wav2vec2-base (95M parameters), a self-supervised convolutional-Transformer model pre-trained on 960 hours of Librispeech. Both configurations were trained with LR: 1×10^{-4} , with stronger attention dropout (0.1) for the unbalanced corpus (V) to mitigate overfitting.

Audio file longer than 20 seconds were excluded. and (c) XLS-R-300M (315M parameters), a multi-lingual wav2vec2 model pre-trained on 436k hours across 128 languages (Babu et al., 2022), making it the strongest starting point for low-resource languages with unusual phonological features such as ejectives and uvulars, and allophonic variations. A CTC projection head over a 46-character vocabulary (Puno Quechua Latin orthography) was added. Training runs for 20,000 updates with a tri-stage scheduler and LR: 5×10^{-5} . The encoder was frozen for the first 10,000 updates to prevent the randomly initialised CTC head from corrupting pre-trained representations before it stabilised. The best checkpoint was selected by validation WER.

Continued Pre-Training (CPT). For XLS-R-300M, we additionally performed CPT on the 65 hours of unlabelled Puno Quechua audio prior to fine-tuning. CPT adapts the model’s acoustic representations to the target language without requiring transcriptions, and has demonstrated consistent gains in low-resource settings (DeHaven and Billa, 2022; Mutisya and Mugane, 2026). Clips shorter than 1 second or longer than 15 seconds were excluded from training. We train for 10,000 updates (LR: 1×10^{-4} , polynomial decay, 1,000-step warmup), selecting the best checkpoint by validation loss. The best checkpoint occurs at update 9,000 with a validation loss of 2.249. Two models were fine-tuned from the CPT checkpoint (ft_xlsr_validated and ft_xlsr_silver); using the identical protocol described above for fine-tuned XLS-R-300M.

Reference baselines. We evaluated the omni-ASR model family (Keren et al., 2025), which combines a wav2vec2-style encoder with either CTC decoding (CTC_300M, CTC_7B; up to 6.5B parameters) or an LLM decoder (LLM_300M, LLM_7B; up to 7.8B parameters), and explicitly supports qxp. We also evaluated MMS-1b-a11 (1B parameters) by setting the language parameter to Cuzco Quechua (quz) for inference, the closest supported variety to Puno Quechua.⁷

5 Experiments and Results

5.1 Baseline with off-the-shelf models

As shown in Table 2, hybrid ASR-LLM models outperform CTC-only variants. The most balanced

Model	Scripted		Spontaneous		OOD	
	WER	CER	WER	CER	WER	CER
omniASR						
CTC_300M_v2	47.8	10.3	29.0	4.4	41.0	6.0
CTC_7B_v2	35.4	7.4	18.1	2.7	34.5	5.7
LLM_300M_v2	25.9	5.8	17.9	2.9	24.4	3.9
LLM_7B_v2	26.6	6.2	11.1	1.9	23.7	4.1
MMS						
mms-1b-a11	35.0	5.3	36.4	6.5	38.0	6.2

Table 2: Performance of SOTA off-the-shelf models on 1,000 files samples for qxp (Scripted, Spontaneous) and on OOD.

model across domains, omniASR LLM_7B_v2, achieves a 20.1% mean WER. Notably, all omni-ASR models handle spontaneous speech more accurately than scripted speech and OOD (the CER is significantly better), maybe because the sentences in this corpus are very short (often between 3 and 5 words), and don’t provide enough context.

The MMS model, despite being trained on a different Quechua variety (Cuzco Quechua, quz), remains competitive, reaching 5.3% CER on scripted speech.

5.2 Results with fine-tuned models

Training setup. Data were split 70/25/5 (train/dev/test). We compared two training configurations:

- **validated-only** (V, 36 hours) with 3× upsampling of spontaneous to compensate for class imbalance,
- **validated-plus-silver** (V+S, 66 hours) including silver spontaneous transcriptions.

Fine-tuning evaluation used three test sets: scripted (1.53 h), spontaneous (0.27 h), and OOD (0.27 h). All the experiments were conducted on a 48GB L40S single GPU.

Fine-tuned and CPT results. Table 3 reports WER and CER across all fine-tuned conditions. Three findings stand out. First, CPT yields consistent gains on scripted speech: XLS-R+CPT (V) achieves 1.19% WER versus 2.06% without CPT (a relative improvement of 42%). Second, silver data is the decisive factor for spontaneous speech: XLS-R+CPT trained on V+S reduces spontaneous WER (13.6% → 3.15%, a relative reduction of 77%); the same pattern holds without CPT (13.6% to 6.68%, a relative reduction of 51%). Third, a clear OOD generalisation gap persists for all fine-tuned models: silver models achieve ~35-54% WER versus

⁷Both belong to the Collao linguistic subgroup and share a similar phoneme inventory and writing system.

Base model	Dataset	Scripted		Spontaneous		OOD		Mean	
		WER	CER	WER	CER	WER	CER	WER	CER
whisper-base	V	8.57	1.38	26.2	4.13	54.7	10.8	29.8	5.43
whisper-base	V+S	3.81	0.60	17.1	2.74	42.0	7.77	21.0	3.70
wav2vec2-base	V	5.84	0.77	21.6	3.06	54.2	10.3	27.2	4.71
wav2vec2-base	V+S	7.37	0.96	13.9	1.70	50.2	9.45	23.8	4.03
xls-r-300m	V	2.06	0.30	13.6	1.71	35.5	6.03	17.1	2.68
xls-r-300m	V+S	4.36	0.57	6.68	0.81	28.9	4.35	13.3	1.91
xls-r + CPT	V	1.19	0.19	13.6	1.73	35.0	6.09	16.6	2.67
xls-r + CPT	V+S	2.11	0.30	3.15	0.41	27.4	4.55	10.9	1.75

Table 3: Performance of foundation models fine-tuned on validated corpus (V) and on complete corpus (V+S). WER and CER are expressed in %.

~27-50% for validated-only models, but the omniASR LLM_7B_v2 still outperforms all fine-tuned systems on OOD (WER: 23.7%), indicating that fine-tuning on in-domain data comes at some cost to out-of-domain robustness.

6 Discussion

Silver data as the decisive factor for spontaneous speech. Across all model architectures, including V+S silver transcriptions drastically reduces spontaneous WER. The effect is most noted for XLS-R+CPT (a relative reduction of 77%) and is consistent even for Whisper-base (26.2% → 17.1%, a relative reduction of 34.7%) and wav2vec2-base (21.6% → 13.9%, a relative reduction of 35.65%). This confirms that the low validation rate of spontaneous speech (14.7%) is a bottleneck for improving ASR system’s performance, and that automatically generated silver transcriptions, despite their lower quality, provide crucial coverage of speech variation.

Effect of CPT on scripted speech. CPT on unlabelled Puno Quechua audio provides consistent gains on scripted speech regardless of data configuration. The relative improvement for XLS-R (V) (2.06% → 1.19%, a relative improvement of 42%) demonstrates that adapting the pre-trained acoustic model to the target-language’s phonology characteristics is valuable even when the same data are later used for fine-tuning (Getman et al., 2024). The CPT benefit is somewhat diminished when silver data is added (4.36% vs. 2.11% without CPT), suggesting that silver data partially compensates for the lack of language-specific pre-training.

OOD generalisation gap. A clear generalisation gap exists for all fine-tuned models on out-of-domain data. The omniASR LLM_7B_v2 achieves

the best OOD WER (23.7%), outperforming the best fine-tuned system (XLS-R + CPT, V+S: 27.4%). This suggests that task-specific fine-tuning on a narrow domain comes at the cost of robustness to unseen acoustic conditions and speaking styles. However, there is a substantial resource disparity involved. The omniASR LLM_7B_v2 operates at ~7.8B parameters and requires ~30 GB of VRAM at inference⁸, while our XLS-R+CPT competitive model operates at just 317M parameters and ~2GB of VRAM, making it deployable on commodity hardware. Closing the OOD gap for our model will therefore require not simply more training data, but exploration of lightweight strategies that preserve cross-lingual generalisation while remaining deployable in low-resource settings.

7 Conclusions and Future Work

This paper makes three concrete contributions.

Largest Puno Quechua ASR corpus. We have constructed, to the best of our knowledge, the largest corpus ever prepared for ASR in a single Quechua variety. The corpus comprises 66 hours of recordings for scripted and spontaneous speech (including 36 hours of manually transcribed and validated data), supplemented by 30 hours of automatically transcribed silver spontaneous speech, and 0.27 hours of out-of-domain annotated data. The data were collected through a four-phase participatory design process involving 151 native speakers and released openly via Mozilla Data Collective under a CC0-1.0 licence. The participatory methodology ensured that the corpus reflects domains directly relevant to the community.

⁸<https://huggingface.co/facebook/omniASR-LLM-300M>

Systematic ASR benchmark for Puno Quechua.

We establish the first variety-specific ASR benchmark for Puno Quechua, evaluating SOTA models (omniASR CTC and LLM variants up to 7B parameters; MMS-1b-all) and fine-tuned foundation models (Whisper-base, wav2vec2-base, XLS-R-300M, with and without CPT) across scripted, spontaneous, and out-of-domain test sets. Key findings are: (a) silver transcriptions are the decisive factor for spontaneous speech performance, reducing WER by up to 77% relative; (b) continued pre-training on unlabelled Puno Quechua audio yields consistent gains on scripted speech; (c) omniASR models outperform all fine-tuned systems on out-of-domain data, revealing a generalisation gap that remains an open challenge.

Open release of fine-tuned models. We release all fine-tuned model variants for Puno Quechua, including Whisper-base, wav2vec2-base, and XLS-R-300M under both V and V+S configurations, as well as the CPT checkpoint and the CPT-based fine-tuned models. The best-performing system, XLS-R-300M with CPT, fine-tuned on V+S, achieves 2.11% WER and 0.30% CER on scripted speech, and 3.15% WER and 0.41% CER on spontaneous.

Future research follows from these results. The enrichment of the corpus, in accordance with the quality requirements outlined above, must be continued. Another objective is to incorporate data representative of a wider range of domains, as we have begun to do through the annotation of data crawled from media sources. With the ongoing goal of developing tools that genuinely meet users' needs, we also aim to design more resource-efficient models (for example through quantization) that can be integrated into everyday applications, particularly mobile voice input systems.

Limitations

Despite the contributions presented in this paper, some limitations should be acknowledged.

Corpus size. Although the corpus is the largest for any single Quechua variety, only 14.7% of the 35.3 recorded hours of spontaneous speech have been validated and transcribed, reflecting the difficulty of manual annotation in a community with low written literacy rates in Quechua.

Out-Of-Domain. All fine-tuned models exhibit a clear generalisation gap relative to the off-the-shelf models. Expanding the diversity of training

and evaluation domains, e.g., radio, television, and social media, will be necessary to close this gap without sacrificing the parameter efficiency that makes our models deployable on commodity hardware.

Acknowledgements

This research was supported by UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1 under the UK government's funding guarantee for ERC Advanced Grants for the project entitled "Towards Globally Equitable Language Technologies (EQUATE)"; by netidee Förderung (www.netidee.at); by SILICON Stanford (silicon.stanford.edu); and by the French National Research Agency and Ministry of Higher Education, Research and Innovation (MESR).

References

- Willem F. H. Adelaar. 1987. Aymarismos en el quechua de puno. *Indiana*, 11:223–231. Published by Gebr. Mann Verlag.
- Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhota, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. [XLS-R: self-supervised cross-lingual speech representation learning at scale](#). In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2278–2282. ISCA.
- Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *IS-LP*, 2:21.
- Mitchell DeHaven and Jayadev Billa. 2022. Improving low-resource speech recognition with pre-trained speech models: Continued pretraining vs. semi-supervised training. *arXiv preprint arXiv:2207.00659*.
- Candace Galla. 2016. [Indigenous language revitalization, promotion, and education: function of digital technology](#). *Computer Assisted Language Learning*, 29:1137 – 1151.
- Yaroslav Getman, Tamas Grosz, Katri Hiovain-Asikainen, and Mikko Kurimo. 2024. Exploring adaptation techniques of large speech foundation models for low-resource asr: a case study on northern sami. In *Interspeech*.
- Elwin Huaman, Wendi Huaman, and Jorge Luis Huaman. 2025. Making an under-resourced language available on the wikidata knowledge graph: Quechua language. In *Information Management and Big Data*, pages 212–224, Cham. Springer Nature Switzerland.

- Elwin Huaman, Wendi Huaman, Jorge Luis Huaman, and Ninfa Quispe. 2026. Quechua speech datasets in common voice: The case of puno quechua. In *Information Management and Big Data*, pages 184–193, Cham. Springer Nature Switzerland.
- Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Dupenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, and 13 others. 2025. [Omnilingual ASR: open-source multilingual speech recognition for 1600+ languages](#). *CoRR*, abs/2511.09690.
- Hillary Mutisya and John Mugane. 2026. Continued pretraining for low-resource swahili asr: Achieving state-of-the-art performance with minimal labeled data. *arXiv preprint arXiv:2603.11378*.
- Rosa Y. G. 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.
- Clay Spinuzzi. 2005. The methodology of participatory design. *Technical communication*, 52(2):163–174.
- Alfredo Torero. 2002. *Idiomas de los Andes. Lingüística e historia*. Editorial horizonte.
- Petti Ulla, M. Claus Hannah, Barford Anna, Sadek Malak, Reichart Roi, and Korhonen Anna. 2026. [COACT – a community-centered, participatory and actionable roadmap for equitable language ai](#). In *PrePrint*.
- Marianne Wilson, David M. Howcroft, Ioannis Konstas, Dimitra Gkatzia, and Gavin Abercrombie. 2025. [Participatory design for positive impact: Behind the scenes of three NLP projects](#). In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 252–263, Vienna, Austria. Association for Computational Linguistics.