# Yachachiq–Yachakuq: Teacher–Student Transfer and Candidate Reranking for Puno Quechua ASR

Arturo Oncevay⋆, Aaron Arana, Flavio Corvetto, Jafeth Cuyo,
Daniel Malca, Pool Pocco, Eliezer Villarreal, and Juan Zavala

Chana Research Group, Pontificia Universidad Católica del Perú (PUCP), Lima, Perú
{first}.{last}@pucp.edu.pe

**Abstract.** We describe the Chana (PUCP) submissions to the Puno Quechua ASR shared task at SIMBig 2026. A LoRA-adapted 7B LLM-ASR teacher (*yachachiq*) trained on gold transcripts labels extra audio and distils token-level knowledge into a 300M-encoder LLM-ASR student (*yachakuq*). We submit three systems defined by inference memory and speed, from a Chana-S that fits a 4 GB consumer GPU to a Chana-L that needs one 24 GB GPU, and in every class we pool candidate transcripts from several models and pick one with a pairwise reranker. Because training prompts are read by many speakers, we select all systems on held-out sentences and speakers rather than on seen prompts. On held-out data, distillation improves the student by up to 2.9 WER points on unseen scripted speech over a same-data cross-entropy model, and reranking gains a further 0.9–1.6 points over majority voting in every class. We also collected new Puno radio audio and generated morphological text variants, but neither improved the submitted systems. Each class still leaves 4–5 WER points of oracle headroom, which makes candidate selection the most promising direction for future work.

**Keywords:** Speech recognition · Quechua · Knowledge distillation · N-best reranking · Low-resource ASR.

## 1 Introduction

Puno (Collao) Quechua is spoken by about 474 000 people, mainly in the Puno region of southern Peru [14], yet has few public speech resources [5, 26, 14]. The shared task provides scripted and spontaneous Common Voice recordings with a curated official spelling, and scores scripted and spontaneous test speech [24].

Two tensions shape our work. First, the strongest open models (7B LLM-ASR) are too large for the hardware available in the region. Second, the training prompts are read many times by many speakers: models memorise them, and scores on seen sentences overstate quality. The official test confirms this concern: it has new speakers and almost only new sentences (Section 3).

Our pipeline is a teacher–student one (Figure 1). A gold-only 7B teacher labels and filters extra audio; a small LLM-ASR student learns from the teacher's token distributions; and in each memory class we rerank a pool of candidates from several models.

⋆ After the first author, authors are listed in alphabetical order.

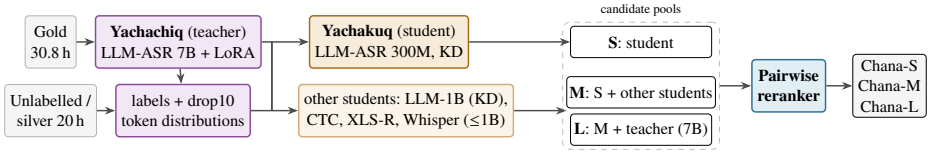



**Fig. 1.** Our pipeline. A gold-only 7B teacher labels and filters extra audio and provides token distributions for distilling a 300M-encoder student (Chana-S); other students trained on the same labels join the medium pool, and the teacher itself joins the large pool. Per memory class, a pairwise reranker picks one transcript from the pool.

Our contributions are: (1) system classes defined by inference memory and speed and a held-out-sentence selection protocol, supported by an analysis of the official test set; (2) token-level distillation that transfers most of the teacher's spontaneous-speech skill into a 3.7 GB model; (3) per-class candidate pooling with a pairwise reranker, with an oracle analysis of the headroom left; (4) new Puno radio and YouTube audio with a filtering pipeline, reported with negative results outside the submissions; and (5) a reproducible cost table (memory and real-time factor).

## 2 Background and Related Work

**Quechua speech.** Earlier Quechua ASR work used the Siminchik radio corpus [5] and Huqariq [26]. Shared tasks have covered Quechua text in AmericasNLP machine translation [17] and Southern Quechua in IWSLT speech translation [1], but not Puno speech. Common Voice [2] added several Quechua varieties, and Puno Quechua now has a dedicated corpus and benchmark [14, 13], with fine-tuned Whisper, wav2vec2 and XLS-R baselines.

**Methods we build on.** Massively multilingual models (MMS [22], XLS-R [3], Whisper [23], and omniASR [19] with CTC and LLM-ASR families) give strong starting points that LoRA [12] can adapt. Pseudo-labelling with teacher filtering is standard for low-resource ASR [20, 25], as is distillation, token-level with soft targets [11] or sequence-level [15]. System combination ranges from ROVER [7] and minimum-Bayes-risk selection [8] to pairwise reranking [4]; quality estimation predicts a hypothesis' error without a reference [18].

## 3 Task and Data

**Data.** The task data are the Common Voice Puno Quechua releases (Scripted Speech 27.0 and Spontaneous Speech 5.0), collected in a community campaign [14, 13, 24]. The training set has 30.8 h of gold speech (25.9 h scripted from 35 speakers, 4.9 h spontaneous) and 20.0 h of spontaneous silver (zero-shot omniASR-LLM-7B transcripts). We score WER after the organisers' spelling mapping ($\text{WER}^k$) with paired bootstrap tests [16].

**Our splits.** We do not use the Common Voice dev/test splits or those of the organisers' repository: in Common Voice, dev and test scripted clips read the same prompts

**Table 1.** Data splits (v1 manifests). Silver = spontaneous clips with model transcripts. Unseen = sentences not in train.

| Split | Clips | Hours | Speakers | Use |
|---|---|---|---|---|
| Train gold scripted | 19 427 | 25.9 | 35 | teacher, students |
| Train gold spontaneous | 1 065 | 4.9 | 46 | teacher, students |
| Train silver spontaneous | 4 573 | 20.0 | 54 | students |
| Held-out scripted | 1 443 | 1.9 | 35 | teacher (final) |
| Dev scripted (unseen) | 948 (124) | 1.4 | 21 | selection |
| Dev newsent (ours) | 731 | 1.0 | 34 | selection |
| Dev spontaneous | 110 | 0.6 | 12 | selection |
| Test scripted (unseen) | 909 (96) | 1.4 | 21 | one look |
| Test spontaneous | 197 | 0.9 | 10 | one look |
| Official test | 659 | 1.0 | – | submission |

as train, most dev/test spontaneous clips are silver, and the released baseline checkpoints [13] had seen many of them. We re-split the pooled data (Table 1) with three rules. (i) Dev and test speakers are disjoint from train and from each other. (ii) 150 scripted sentences are held out *entirely*: every recording of them leaves train. Recordings by dev/test speakers become the *unseen* part of dev and test scripted; recordings by train speakers form a *held-out* set. (iii) Dev and test spontaneous use gold clips only. From the held-out set we build `dev_newsent`: 731 recordings of 74 sentences that occur in neither train nor test. Its speakers are seen in training, so it isolates sentence memorisation; unseen dev scripted (new speakers, new sentences) and dev spontaneous complement it. Systems are selected on the mean of these three, with one look at our test per system. We also correct a few gold transcripts where a build bug split words or dropped hyphens (6 clips, taken from Spontaneous 5.0).

**The official test.** The official test (659 clips, 0.99 h; 438 scripted, 221 spontaneous) was released without references, so we profiled it against our data. No test clip matches any of our 187 speakers by speaker embedding: the median best-match cosine is 0.54 (maximum 0.71), below the 0.66 typical of the nearest *different* speaker among our own clips. Of the scripted sentences, 95% are new; only 21 of 438 are in or close to our training text. The clips are shorter (scripted 3.2 vs 4.8 s; spontaneous 8.2 vs 18.9 s) and have less edge silence (0.26 vs 0.44 of the clip). Trimming our dev clips to this profile costs the 7B 0.34 $\text{WER}^k$ on scripted speech ($p = 0.002$). The official test thus measures new speakers on new sentences, which is the regime our selection protocol targets.

## 4 System Classes

We define classes by what a system needs at inference: peak GPU memory and decoding time. Chana-S is a single small model (peak 3.7 GB, real-time factor 0.05), suited to a consumer GPU. Chana-M pools several small and 1B models; run one after another, its largest member needs 7.8 GB (LLM-1B with beam 8), so it runs on the same hardware as S but decodes several times slower. Chana-L adds the 7B and needs one 24 GB GPU. Memory, not parameter count, decides where a system runs: the “300M” LLM-ASR

student has 1.63B parameters. Peak memory is `max_memory_allocated` at batch 1 on the longest test clip; Figure 3c gives decoding times.

## 5 Method

### 5.1 Teacher (Chana-L): LoRA on a 7B LLM-ASR

We adapt omniASR-LLM-7B [19] with LoRA (rank 16) in two stages: a first stage on gold plus teacher-labelled audio, then a gold-only stage on gold, gold segments and held-out recordings, with scripted prompts down-weighted to 0.3, a learning rate of $2.5\times10^{-5}$ and a fixed 500 steps, without checkpoint selection. Adding silver data did not help the teacher on unseen or spontaneous dev, nor did any post-training we tried (continued training, augmentation, minimum-WER training): every continuation of the final recipe was about 0.3 WER worse. Retraining from the same start reproduces the model (test unseen 10.47→10.47, spontaneous 6.85→6.89). The final stage also trains on our held-out recordings, which read the unseen dev and test prompts (by other speakers), so for clean new-sentence results we use the checkpoint before that stage (7B-start).

### 5.2 Teacher labelling and filtering

The teacher transcribes extra audio. We drop the 10% of clips on which it disagrees most with an independent CTC model (drop10). Filters across different model families ranked bad labels better, but did not yield better students.

### 5.3 Student (Chana-S): token-level KD into LLM-300M

The student is omniASR-LLM-300M (a 300M encoder with an LLM decoder; 1.63B parameters in total). Its loss is $\alpha\,\mathrm{CE}(y) + (1-\alpha)\,\mathrm{KL}(p_T^{(\tau)} \| p_S^{(\tau)})$ over the teacher's top-8 tokens per step, cached offline, with $\alpha = 0.5$ and $\tau = 1$ on gold plus drop10 silver, for 4 000 steps (100 s of audio per batch, learning rate $10^{-5}$), with checkpoints selected every 500 steps on the dev mean. The seed is chosen on dev (seed 3: 9.87 vs 9.99 and 10.17), before any test look. Adding radio data or more steps never improved unseen scripted speech.

### 5.4 Other students (Chana-M members)

Chana-M adds students from other families, trained on the same gold data and teacher labels: two more distilled LLM-300M students (other seeds and temperatures), an LLM-1B student distilled the same way, omniASR CTC-300M and CTC-1B, XLS-R-300M and XLS-R-1B with CTC heads, and Whisper-small and large-v3-turbo, each 1B or smaller. None of them alone beats the distilled LLM-300M, but their errors differ, which is what pooling exploits.

**Table 2.** Candidate pools per class: candidates per clip and dev oracle $\text{WER}^k$ (newsent / unseen / spont).

| Class | Members | Cand. | Vote → oracle |
|---|---|---|---|
| S | KD LLM-300M student | ∼8 | 9.78→4.73 / 11.23→5.41 / 8.98→4.98 |
| M | S + other students | 41–53 | 8.79→2.21 / 10.40→3.95 / 8.24→4.25 |
| L | M + 7B teacher | 44–56 | 8.76→2.00 / 10.40→3.74 / 8.24→4.04 |

### 5.5 Candidate pools and pairwise reranking (all classes)

For each class we pool the candidates of every model that fits its memory budget: beam-8 lists plus 8 temperature-0.8 samples from the LLM-ASR models, and 8-best lists from the CTC and Whisper models (Table 2). A pairwise reranker (gradient-boosted trees) compares the 16 candidates with the most votes, two at a time. Its features are vote fraction and rank per member, minimum-Bayes-risk agreement, length and duration ratios, teacher-forced log-probabilities from three distilled students, and edit-type features of the pair. It is trained on dev with 2-fold speaker-disjoint cross-validation and frozen before the official test. In the large class, the 7B is one member among about 50 candidates per clip; plain voting there is 2 points worse than the 7B alone, because twelve small models outvote it, so the reranker must know who proposed each candidate.

**Language-model and dialect features (all classes).** Each candidate also gets Quechua language-model scores [10] (a Puno-weighted word 4-gram, a character 6-gram and a morph-segment 4-gram, trained only on training text and cross-fitted on the reranker's speaker folds) and per-candidate rates of Puno dialect markers (spellings and suffixes that distinguish Puno from neighbouring Collao and Cusco usage). They lower the dev mean of every class (S $-0.54$, $p = 0.007$; M $-0.60$, $p < 0.001$; L $-0.81$) and help on our test for S ($-0.37$, $p = 0.03$) and L ($-0.24$); for M our test is flat. Every reranker is averaged over 5 bagged runs, because its result moves by about 0.2 with its training seed alone. Word-level scores on the span where two candidates differ and an audio head trained on minimal pairs mined from real confusions (86% pair accuracy) did not pass our selection rule.

## 6 Results

**Reranking.** Table 3 and Figure 2 show each class before and after reranking, on clean models. Against majority voting over the same pool, the pairwise reranker lowers the dev mean by 0.4 (S and M) and 0.8 (L), and the LM and dialect features add a further 0.5–0.8, for totals of 0.9 (S, 10.00 → 9.08), 1.0 (M, 9.14 → 8.16) and 1.6 (L, 9.13 → 7.57). On our test, the final systems improve spontaneous speech over voting in every class (S $-0.9$, M $-0.5$, L $-1.5$) and unseen scripted speech for S and L ($-1.1$ and $-1.8$), while M stays level. The gains are largest on new sentences and spontaneous speech, the two conditions that dominate the official test.

**Headroom and cost.** Figure 3 puts accuracy next to its price. Panel (a) shows that the reranker recovers 14% (S), 23% (M) and 41% (L) of the gap between voting and

**Table 3.** Systems by class, WER$^k$. Dev mean = newsent, unseen and spontaneous dev. Clean = no held-out prompts in training. *Chana-L itself trained on the unseen prompts (unseen speakers only). †Spontaneous references were post-edited from 7B output, which likely favours the 7B family.

| Class | System | Peak GB | Dev mean | Test unseen | Test spont |
|---|---|---|---|---|---|
| S | KD LLM-300M, greedy | 3.7 | 10.41 | 13.09 | 9.66 |
| | + pool, reranked | 6.5 | 9.62 | 12.57 | 9.63 |
| | + LM and dialect features (Chana-S) | 6.5 | 9.08 | 12.30 | 9.24 |
| M | pool, vote | 7.8 | 9.14 | 12.57 | 8.55 |
| | reranked | 7.8 | 8.76 | 12.30 | 8.16 |
| | + LM and dialect features (Chana-M) | 7.8 | 8.16 | 12.57 | 8.09 |
| L | 7B-start alone (clean) | 18.2 | – | 10.99 | 7.86 |
| | + pool, reranked (clean) | ≤24 | 8.38 | 10.99 | 7.12 |
| | + LM and dialect features (clean) | ≤24 | 7.57 | 10.73 | 6.85 |
| | Chana-L (final 7B)* | 18.2 | – | 10.47 | 6.85 |
| – | best CTC-300M (no KD) | 1.0 | | 18.59 | 10.78 |
| – | omniASR-LLM-300M zero-shot | 3.7 | | 19.90 | 11.74 |
| – | omniASR-LLM-7B zero-shot | 18.2 | | 20.16 | 4.00† |

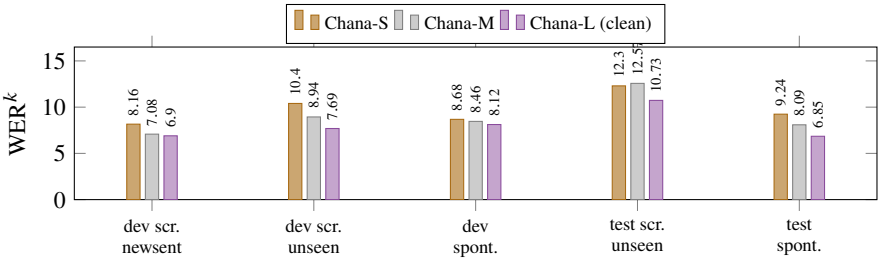



**Fig. 2.** Submitted systems by split: scripted (scr.) new sentences and unseen sentences, and spontaneous speech. Dev values are out-of-fold for the reranker; Chana-L uses the clean 7B-start pool.

the oracle on unseen dev; the remaining 4–5 points are in the lists but not found. Panel (b) shows diminishing returns in memory: going from 3.7 to 7.8 GB takes the dev mean from 10.4 to 8.2, while tripling memory again for the 7B buys a further 0.6. Panel (c) shows that time is the main cost of pooling: Chana-M fits the same 8 GB GPU as Chana-S but takes about 3.7 times as long on the official test, and Chana-L, with the 7B, only 13% longer than Chana-M.

**Distillation.** At equal data, token-level KD beats cross-entropy by 1.0–2.9 on unseen scripted test speech (15.97 vs 13.09–14.92 over three seeds) and by up to 0.8 on spontaneous test (9.78 vs 9.01–9.66). The student scores 10.19 on seen and 13.09 on unseen test sentences; we encourage reporting this memorisation gap. Seeds vary by up to 1.8 on unseen test, so we select on dev.

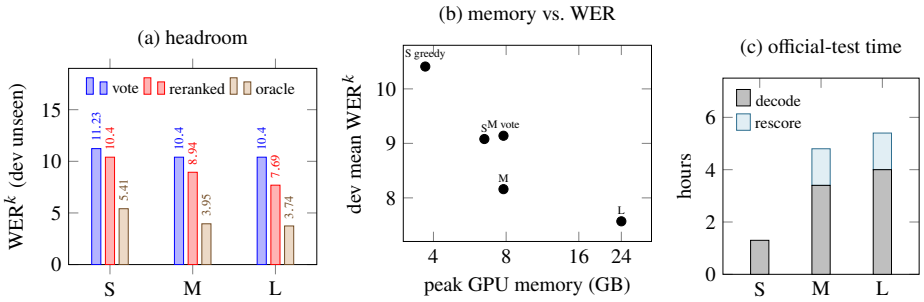



**Fig. 3.** (a) Unseen dev scripted WER$^k$ per class: majority vote, pairwise reranker and oracle choice from the same pool, with LM and dialect features (L: clean 7B-start pool). (b) Dev mean against peak inference memory; S, M and L are the reranked pools. (c) Time to process the 0.99 h official test on one A10G: decoding and the reranker's teacher-forced rescoring (included in S's decode time).

## 7 Additional Experiments (Not in the Submissions)

**Radio and YouTube collection.** We recorded Puno broadcasts (Onda Azul, Pachamama Radio, TV Perú Ñuqanchik) and collected Quechua tales from YouTube. We segment whole recordings ourselves and filter them (69 h raw, 12.9 h Quechua by language ID, 8.6 h kept): file hashes, language ID [22], teacher labels with drop10, audio fingerprints against repeats such as advertisements [9], text cues, and speaker embeddings [6] against every dev, test and official-test voice. With earlier batches this gives 20.5 h of radio-domain silver. As student training data it improves two radio test sets but not our scripted or spontaneous held-out sets, so it is not in the submissions.

**Other corpora.** We also tried the Onda Azul (Puno radio) part of Siminchik [5]. Validating it by ear first, we found an encoding bug that had blanked ñ and accents in 640 of 1 505 transcripts, excluded 54 Spanish-dominant clips and recovered 216 unlabelled segments (1.5 h). Its human transcripts hurt an independent radio set (+4.0 WER for a CTC-300M) because of their spelling, while teacher labels on the same clips help (−2.1). Neither improves our scripted or spontaneous sets (scripted is even worse, +0.6 to +1.4), so it is not used.

**Quality-estimated silver.** The audio head of Section 5.5 ranks pseudo-labels better than teacher–CTC disagreement (correlation with label WER 0.40 vs 0.24). With the radio silver, keeping the 75% it scores best keeps the full radio gain and removes the 0.7 cost of unfiltered radio on dev scripted speech (−0.63, $p = 0.009$). Spontaneous and unseen scripted speech do not move significantly, so the filter is not submitted.

**Morphological augmentation.** A third of the remaining scripted errors (34% for the 7B, 37% for Chana-L) are one-suffix slips on the right stem; *-n* alone is a quarter of them (*mana*/*manan*). Following [21], we generated about 50k one-change morphological variants of training sentences with LLMs (about US$6.5), keeping forms on which two models agree. They did not help as reranker LM features, as a minimal-pair classifier, inside beam search, or as text-only decoder training: they contain under 2% of the missing reference forms; where the form is known, the speaker often said the reduced

form while the reference is the written prompt. No use changed WER significantly, so none is submitted.

**Selection beyond the reranker.** None of the following beat the submitted rerankers: training them on 712 more held-out clips (pooled +0.04 to +0.15); a wider 7B beam (16), which raises the oracle but not the reranked result; and word- or morph-unit ROVER and word- or character-level MBR, plain or weighted by reranker scores or pairwise win rates. To distil the pool's choice into the student, we decoded the 20 h of silver audio with the 7B (the other members trained on it); retraining the student on 7B-only reranked targets (27.6% of silver labels changed) helps spontaneous speech a little (test −0.42) but hurts unseen scripted sentences (test +1.83). None of these passed our selection rule, so the submissions keep the rerankers of Section 5.5.

## 8 Discussion

**What transfers.** Distillation transfers the teacher's spontaneous-speech skill, and reranking adds a consistent gain in every class. New scripted sentences stay hardest. Most remaining errors are 1–3 characters apart (ejectives and aspirates, the suffix *-n*, loans and names); 1–2% are spellings the references themselves vary.

**Where each class runs.** By peak GPU memory (Fig. 3b), Chana-S fits a 4 GB consumer GPU (8 GB with reranking), Chana-M an 8 GB GPU, and Chana-L a single 24 GB GPU such as an NVIDIA A10G. Chana-M stays within 8 GB only because its members run one after another, so it decodes much more slowly than S on the same GPU (3.4 vs 1.3 h for the official test; Figure 3c).

**Limitations.** Spontaneous references were post-edited from 7B output (zero-shot 7B: 4.00 on test spontaneous), which may favour the 7B family. Chana-L's unseen-sentence scores are not clean (Section 5.1). Our test sets are small: differences below about 0.5 are rarely significant per split.

## 9 Conclusion

We built Puno Quechua ASR around a simple idea: one strong model teaches, small models learn and vote. The recipe has four parts: (1) a 7B LLM-ASR teacher adapted with LoRA on gold transcripts only; (2) teacher labels for extra audio, filtered and cached as token distributions; (3) a same-family 300M student distilled from them, running in 3.7 GB; and (4) in each memory class, a pairwise reranker with Quechua and Puno-dialect features that picks among the candidates of every model that fits. One recipe thus serves GPUs from 4 to 24 GB, with gains at every size.

In future work, we will target the 4–5 WER points of oracle headroom that every class leaves. Many of the misses are one-suffix slips on the right stem, where the speaker reduced what the written prompt keeps: scoring that slot in context, or asking native speakers to judge such pairs, would address them directly. We also expect recordings from new speakers to help the reranker more than further candidates, and plan to extend the recipe to other Quechua varieties.

## References


1. Agarwal, M., Agrawal, S., Anastasopoulos, A., et al.: Findings of the IWSLT 2023 evaluation campaign. In: Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023). pp. 1–61 (2023). https://doi.org/10.18653/v1/2023.iwslt-1.1
2. Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., Weber, G.: Common voice: A massively-multilingual speech corpus. In: Proceedings of the 12th Language Resources and Evaluation Conference (LREC). pp. 4218–4222 (2020), https://aclanthology.org/2020.lrec-1.520/
3. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., Auli, M.: XLS-R: Self-supervised cross-lingual speech representation learning at scale. In: Proceedings of Interspeech. pp. 2278–2282 (2022). https://doi.org/10.21437/Interspeech.2022-143
4. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine Learning (ICML). pp. 89–96 (2005). https://doi.org/10.1145/1102351.1102363
5. Cardenas, R., Zevallos, R., Baquerizo, R., Camacho, L.: Siminchik: A speech corpus for preservation of southern Quechua. In: Proceedings of the LREC 2018 Workshop on Collaboration and Computing for Under-Resourced Languages (CCURL) (2018), https://zenodo.org/records/3595354
6. Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: Proceedings of Interspeech. pp. 3830–3834 (2020). https://doi.org/10.21437/Interspeech.2020-2650
7. Fiscus, J.G.: A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 347–354 (1997). https://doi.org/10.1109/ASRU.1997.659110
8. Goel, V., Byrne, W.J.: Minimum Bayes-risk automatic speech recognition. Computer Speech & Language **14**(2), 115–135 (2000). https://doi.org/10.1006/csla.2000.0138
9. Haitsma, J., Kalker, T.: A highly robust audio fingerprinting system. In: Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR). pp. 107–115 (2002), https://dblp.org/rec/conf/ismir/HaitsmaK02.html
10. Heafield, K.: KenLM: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 187–197 (2011), https://aclanthology.org/W11-2123/
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. NIPS Deep Learning and Representation Learning Workshop, arXiv:1503.02531 (2015), https://arxiv.org/abs/1503.02531
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022), https://arxiv.org/abs/2106.09685
13. Huaman, E., Gamarra Lafuente, A., Cordova, J., Korhonen, A.: Building community-centred NLP resources for Puno Quechua. In: Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP). pp. 147–152. Association for Computational Linguistics, San Diego, California, USA (2026). https://doi.org/10.18653/v1/2026.americasnlp-6.13
14. Huaman, E., Huaman, W., Huaman, J.L., Quispe, N.: Quechua speech datasets in Common Voice: The case of Puno Quechua. In: Lossio-Ventura, J.A., Medina Rodriguez, R., Zilberman, J., Paz Espinoza, F.A., Alatrista-Salas, H. (eds.) Information Management and Big Data (SIMBig 2025). pp. 184–193. Communications in Computer and Information Science, Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-20322-9_13

15. Kim, Y., Rush, A.M.: Sequence-level knowledge distillation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1317–1327 (2016). https://doi.org/10.18653/v1/D16-1139
16. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 388–395 (2004), https://aclanthology.org/W04-3250/
17. Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I.V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N.T., Kann, K.: Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas. pp. 202–217 (2021). https://doi.org/10.18653/v1/2021.americasnlp-1.23
18. Negri, M., Turchi, M., de Souza, J.G.C., Falavigna, D.: Quality estimation for automatic speech recognition. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. pp. 1813–1823 (2014), https://aclanthology.org/C14-1171/
19. Omnilingual ASR team: Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. arXiv:2511.09690 (2025), https://arxiv.org/abs/2511.09690
20. Park, D.S., Zhang, Y., Jia, Y., Han, W., Chiu, C.C., Li, B., Wu, Y., Le, Q.V.: Improved noisy student training for automatic speech recognition. In: Proceedings of Interspeech. pp. 2817–2821 (2020). https://doi.org/10.21437/Interspeech.2020-1470
21. Pocco, P., Oncevay, A.: On the robustness of morphosyntactic transformation with large language models: The case of Quechua Collao. In: Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP). pp. 128–146. Association for Computational Linguistics, San Diego, California, USA (2026). https://doi.org/10.18653/v1/2026.americasnlp-6.12
22. Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.N., Conneau, A., Auli, M.: Scaling speech technology to 1,000+ languages. Journal of Machine Learning Research **25**(97), 1–52 (2024), https://jmlr.org/papers/v25/23-1318.html
23. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR 202 pp. 28492–28518 (2023), https://proceedings.mlr.press/v202/radford23a.html
24. Shared task organisers: Overview of the Puno Quechua ASR shared task at SIMBig 2026 (2026), to appear
25. Xu, Q., Likhomanenko, T., Kahn, J., Hannun, A., Synnaeve, G., Collobert, R.: Iterative pseudo-labeling for speech recognition. In: Proceedings of Interspeech. pp. 1006–1010 (2020). https://doi.org/10.21437/Interspeech.2020-1800
26. Zevallos, R., Camacho, L., Melgarejo, N.: Huqariq: A multilingual speech corpus of native languages of Peru for speech recognition. In: Proceedings of the 13th Language Resources and Evaluation Conference (LREC). pp. 5029–5034 (2022), https://aclanthology.org/2022.lrec-1.537/