Giving Voice to the Constitution: Low-Resource Text-to-Speech for Quechua and Spanish Using a Bilingual Legal Corpus

John E. Ortega[1] **Rodolfo Zevallos**[2] **Fabrício Carraro**[3]
[1]Northeastern University, USA
[2]Universitat Pompeu Fabre, Spain
[3]Barcelona Supercomputing Center, Spain
j.ortega@northeastern.edu, rodolfojoel.zevallos@upf.edu, fabricio.carraro@bsc.es

Abstract

We present a unified pipeline for synthesizing high-quality Quechua and Spanish speech for the Peruvian Constitution using three state-of-the-art text-to-speech (TTS) architectures: *XTTS v2*, *F5-TTS*, and *DiFlow-TTS*. Our models are trained on independent Spanish and Quechua speech datasets with heterogeneous sizes and recording conditions, and leverage bilingual and multilingual TTS capabilities to improve synthesis quality in both languages. By exploiting cross-lingual transfer, our framework mitigates data scarcity in Quechua while preserving naturalness in Spanish. We release trained checkpoints, inference code, and synthesized audio for each constitutional article, providing a reusable resource for speech technologies in indigenous and multilingual contexts. This work contributes to the development of inclusive TTS systems for political and legal content in low-resource settings.

1. Introduction

Indigenous Andean communities in South America often face barriers where crucial information, such as laws and other political issues, is only communicated in the official high-resource language of the government (Spanish). One indigenous community found in Peru is a prime example of this notion in which the government intended to address barriers by translating its constitution from Spanish to Quechua, the largest indigenous language in Peru. For most Peruvians, the Peruvian Constitution is central to civic life and is made officially available online[1][2][3][4]; public portals[5] and legal repositories provide the authoritative Spanish text and cataloged Quechua versions and translations. (Bewes, 1920)

In order to facilitate further work in the political arena for Quechua, the authors of this paper target two principal goals: (1) **accessibility**—produce clear audio renditions of the Peruvian Constitution in Quechua and Spanish for radio, legal aid, screen readers, and civic outreach; and (2) **reusability**—publish aligned text and synthesized speech that others can adopt for ASR/ST training, evaluation, and augmentation in Quechua-focused research.

This short paper is submitted to the **PoliticalNLP**[6] workshop at LREC 2026 and follows the conference formatting and anonymization guidelines. The proposed resource is lightweight enough to be developed by a student team, yet sufficiently broad to support practical experimentation in Quechua and Spanish speech technologies.

Section 2 reviews relevant prior work. Section **??** introduces the role of TTS in low-resource political and legal domains. Our modeling framework and training strategy are presented in Section 3, followed by experimental results in Section 4. Finally, we discuss ethical considerations and data-related aspects in the concluding sections.

2. Related Work

IWSLT QUE–SPA (2023–2025). Since 2023, the International Conference on Spoken Language Translation (IWSLT)[7] (Agostinelli et al., 2025; Ahmad et al., 2024; Agarwal et al., 2023) has included *Quechua→Spanish* within its Low-Resource/Dialect track. The organizers released a curated ST set of ∼1h40m of parallel Quechua speech with Spanish translations and, for unconstrained participation, tens of hours (∼48–60h) of fully transcribed Quechua drawn from Siminchik along with several machine translated (MT) texts. (Zevallos et al., 2022b; Ortega et al., 2020; Cardenas et al., 2018)

In the past three years, the *Quechua→Spanish* submissions to IWSLT have shown consistent BLEU (Papineni et al., 2002) score improvements via multilingual pretrained models (e.g., SpeechT5

[1] https://www.congreso.gob.pe/biblioteca/constituciones_peru
[2] https://www.wipo.int/wipolex/en/legislation/details/21225
[3] https://biblioteka.sejm.gov.pl/konstytucje-swiata-peru/?lang=en
[4] https://www.constituteproject.org/constitution/Peru_2021
[5] https://biblioteka.sejm.gov.pl/konstytucje-swiata-peru/?lang=en
[6] https://sites.google.com/view/politicalnlp2026
[7] https://iwslt.org/

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 64–69
16 May 2026, Palma, Mallorca (Spain).

(Ao et al., 2022), SeamlessM4T (Barrault et al., 2023)) and synthetic data, with 2024–2025 findings documenting steady gains and practical recipes for both cascaded and end-to-end ST (Ortega et al., 2025, 2024, 2023; Gow-Smith et al., 2023). These results validate the impact of *domain-aligned* and *synthetic* resources for low-resource speech tasks.

AmericasNLP (2021–2025). Alternatively, another workshop dedicated to the preservation of American indigenous languages called AmericasNLP[8] has consistently included Quechua as a language for MT and ASR tasks since its inception in 2021.

The AmericasNLP shared tasks and findings report measurable gains in Spanish↔Quechua MT using pretrained multilingual models and carefully curated corpora, including *legal-domain* content such as constitutions/laws (Ebrahimi et al., 2023; Ahmed et al., 2023). These insights support our choice of the Peruvian Constitution as a high-value, stable source for speech synthesis and evaluation to be used for future tasks in the political arena.

Quechua ASR and augmentation. For Quechua ASR, Zevallos et al. (2022a) proposed TTS-driven augmentation and reported ∼8.7% absolute word error rate (WER) reduction; Zevallos et al. (2022a) corroborated improvements using wav2letter++ (Pratap et al., 2019) with synthetic data. In broader Indigenous ASR, self-supervised models (e.g., XLS-R (Babu et al., 2021), mHuBERT (Boito et al., 2024)) transfer surprisingly well, indicating that additional labeled pairs from TTS can be highly effective (Chen et al., 2024; Romero et al., 2024a,b). Their techniques were used in the latest IWSLT team QUESPA submission. (Ortega et al., 2025)

TTS frameworks for low-resource. Recent open-source TTS ecosystems have lowered the barriers to developing speech synthesis systems for low-resource languages. Coqui[9] TTS enables cross-lingual synthesis through *XTTS* (Casanova et al., 2024), supporting multilingual speaker adaptation and facilitating knowledge transfer in unbalanced Quechua–Spanish scenarios. In parallel, diffusion- and flow-matching approaches, such as *F5-TTS* (Chen et al., 2025) and *DiFlow-TTS* (Nguyen et al., 2025), improve robustness and prosodic modeling, making them well suited for low-resource and noisy data conditions.

3. Method and Settings

3.1. Corpus and Normalization

We distinguish between speech–text corpora used for model training and text-only resources employed exclusively for evaluation. For Quechua, we utilize the Siminchik (Cardenas et al., 2018) and Lurin (Zevallos et al., 2022a) corpora, which provide approximately 97.5 and 83.3 hours of fully transcribed Southern Quechua speech, respectively. The Lurin corpus comprises approximately 8,000 read-speech sentences collected from lexicographic and literary sources, including the Collao–Spanish and Chanka–Spanish dictionaries. Together, these datasets constitute approximately 180 hours of aligned speech–text data.

Both corpora contain a mixture of long utterances (≈ 30s) and very short segments (≈ 1s). Since extremely short clips provide limited prosodic information and may negatively affect duration and alignment modeling in neural TTS systems, we apply duration-based filtering and discard all segments shorter than one second and some of low quality (e.g., music, noise, silence). This quality-control step reduces the total amount of training data by approximately 140 hours, resulting in a final curated corpus of approximately 40 hours.

All Quechua transcriptions are normalized using a morphological parser and normalizer to convert surface forms into standardized Southern Quechua representations, thereby reducing orthographic variability and data sparsity in this highly agglutinative language. Additional cleaning procedures include segmentation refinement, removal of misaligned samples, and consistency checks across morphological boundaries, following prior work on low-resource Quechua data preparation.

[8] https://turing.iimas.unam.mx/americasnlp/

[9] https://github.com/coqui-ai/TTS

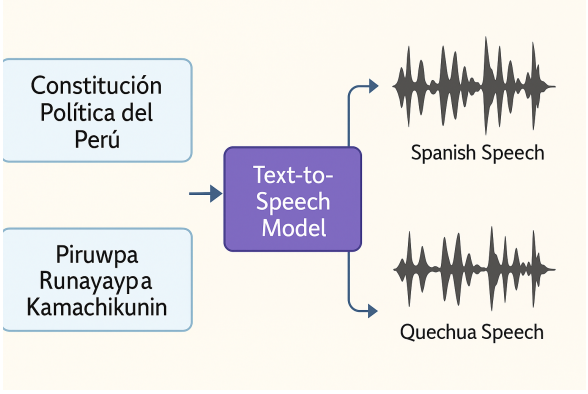

Figure 1: Spanish and Quechua Text to Speech Model.

For Spanish, we leverage a large collection of open-source speech corpora curated for high-quality text-to-speech training (Guevara-Rukoz et al., 2020). The combined dataset includes professionally recorded and crowdsourced read speech spanning multiple Spanish dialects, such as Peninsular, Argentinian, Chilean, Colombian, Peruvian, Puerto Rican, and Venezuelan varieties, as well as TEDx Spanish recordings. In total, the training set comprises approximately 218 hours of annotated Spanish audio. All audio samples undergo resampling, amplitude normalization, and sentence-level segmentation prior to training to ensure consistency and precise alignment between speech and transcriptions across sources.

In addition to training resources, we compile authoritative Spanish text of the Peruvian Constitution and catalogue Quechua versions/segments referenced by public repositories (Ministerio de Justicia y Derechos Humanos, 2020). We retain Title/Article headings and segment into sentences to produce *article-aligned* JSON (IDs, language tags). We normalize punctuation, numbers (cardinals/ordinals, dates), legal abbreviations (e.g., *Art.*, §), and expand references and Roman numerals. For Quechua, we adopt a rule-based grapheme-to-phoneme front end that preserves orthographic conventions in the source and marks morphological hyphenation minimally.

3.2. Models and Training

We synthesize audio using three SOTA text-to-speech systems, XTTS-v2[10], F5-TTS[11], and DiFlow-TTS[12], trained or fine-tuned following their respective official recipes and optimized on NVIDIA H100 GPUs for efficient convergence and reproducibility. XTTS-v2 uses a GPT-based encoder-decoder stack with standard Coqui[13] hyperparameters, AdamW optimization, cosine learning rate scheduling, and FP16 mixed precision; pre-processing follows Coqui dataset tools, and pretrained checkpoints accelerate convergence while enabling multilingual and voice-conditioned synthesis. F5-TTS adopts the flow-matching paradigm, with text processed through ConvNeXt (Woo et al., 2023) encoders and denoised via the Diffusion Transformer backbone; it preserves default flow-matching objectives, step schedules, and sampling strategies such as Sway Sampling for inference efficiency. DiFlow-TTS leverages discrete flow matching over factorized codec tokens for zero-shot synthesis, maintaining discrete flow losses and factorized prediction heads for prosody and acoustics, following open-source training procedures. Across all models, learning rates (1e-4 and 5e-5), weight decay, gradient clipping, mixed precision, and batch sizing are tuned for stability and generalization; training spans tens to hundreds of thousands of steps with early stopping based on validation loss and audio quality. Generated audio is saved as 22.05–24 kHz WAV files with consistent amplitude normalization, sentence alignment, and metadata to support systematic evaluation and reproducibility.

[10] https://github.com/coqui-ai/TTS

[11] https://github.com/SWivid/F5-TTS

[12] https://github.com/Tobertz-max/DiFlow-TTS

[13] https://github.com/coqui-ai/TTS

3.3. Evaluation

We adopt an objective and proxy-perceptual evaluation protocol tailored to low-resource TTS settings and fully aligned with the reported results. All metrics are computed on Quechua audio synthesized from constitutional text.

- **Perceptual Quality (Proxy)**: We report **UTMOS**, an objective estimator of mean opinion scores, to approximate perceived naturalness of synthesized speech without requiring large-scale human evaluations.
- **Speaker Consistency**: Cross-utterance speaker stability is measured using **SIM-O**, computed as the cosine similarity between speaker embeddings extracted from synthesized speech. Higher values indicate more consistent voice characteristics across generated samples.
- **ASR-based Intelligibility**: We report Word Error Rate (**WER**) on synthesized speech using language-specific ASR systems. For Quechua, we employ the **QUESPA** self-supervised ASR model (Ortega et al., 2025), which is specifically designed for low-resource Andean languages. Both systems are used off-the-shelf without additional fine-tuning.
- **Prosodic Accuracy**: To evaluate prosodic stability, we report root mean squared error for fundamental frequency (**RMSE**$_{F0}$) and energy (**RMSE**$_{E}$), computed between reference and synthesized signals. Lower values indicate more accurate modeling of pitch contours and loudness dynamics.

This evaluation setup enables consistent comparison across models while remaining feasible in a low-resource bilingual scenario, where large-scale subjective evaluations are often impractical.

4. Results

Table 1 reports the quantitative evaluation of the three TTS systems considered in this work. In line

Model	#Params	UTMOS ↑	SIM-O ↑	WER ↓	RMSE_{F0} ↓	RMSE_{E} ↓
XTTS-V2	470M	3.22	0.53	0.19	21.03	0.021
F5-TTS	336M	3.23	0.60	0.19	15.17	0.017
DiFLOW-TTS	164M	3.31	0.49	0.16	10.24	0.011

Table 1: Objective and perceptual evaluation of Quechua synthesized speech using constitutional text as input. All metrics are computed on audio generated from Quechua text of the Constitution, leveraging cross-lingual training with Spanish as a high-resource language. Higher is better for UTMOS and SIM-O, while lower is better for WER, RMSE_{F0}, and RMSE_{E}.

with our low-resource setting, all models are trained or adapted using a bilingual setup combining Spanish (high-resource) and Quechua (low-resource) data, with the goal of assessing how cross-lingual transfer impacts perceptual quality, intelligibility, and prosodic stability in Quechua synthesis.

Across all metrics, DiFLOW-TTS achieves the most favorable trade-off between model size and synthesis quality. Despite being the smallest model (164M parameters), it obtains the highest UTMOS score (3.31) and the lowest WER (0.16), indicating improved naturalness and intelligibility under limited Quechua supervision. Crucially, it also yields the lowest RMSE_{F0} and RMSE_{E}, suggesting that cross-lingual acoustic patterns learned from Spanish are transferred more effectively to stabilize prosody in Quechua.

F5-TTS, with 336M parameters, shows competitive perceptual quality (UTMOS 3.23) and achieves the highest SIM-O score (0.60), reflecting stronger speaker consistency across languages. While its WER matches that of XTTS-V2, its substantially lower prosodic errors point to a more robust exploitation of shared phonetic and rhythmic structure between Spanish and Quechua, even when trained with limited target-language data.

XTTS-V2, the largest model evaluated (470M parameters), provides stable baseline performance in the cross-lingual scenario but does not fully capitalize on parameter scale. Its comparatively higher F0 and energy reconstruction errors indicate less precise prosodic transfer, highlighting that larger models do not necessarily yield better outcomes in low-resource bilingual settings.

Overall, these results underline the importance of cross-lingual learning over model scaling for low-resource TTS. Leveraging Spanish data enables all systems to synthesize intelligible and natural Quechua speech; however, architectures explicitly designed to control temporal and prosodic dynamics exhibit superior transfer efficiency. This makes such models particularly suitable for low-resource languages, where data scarcity and deployment constraints are central considerations.

5. Ethics & Limitations

We avoid any personally identifiable or celebrity-like voices; voices are synthetic or consented. We document dialect scope (e.g., Cusco Collao influence), orthographic choices, and intended use. We encourage Indigenous data governance practices and feedback from Quechua media/community groups[14]. The TTS is not a substitute for professional legal interpretation and may require style tuning for ceremonial or court settings.

6. Conclusion

In this work, we set out to make the Peruvian Constitution accessible in Quechua through high-quality synthesized speech, addressing a concrete gap at the intersection of language rights and speech technology. Given the severe data scarcity of Quechua, we adopt a bilingual training strategy that leverages Spanish as a high-resource language to enable effective cross-lingual transfer for TTS.

Our results show that cross-lingual TTS models trained on Spanish and Quechua data can generate intelligible and natural Quechua speech from legal-domain text, with competitive perceptual quality and stable prosody. Notably, model architectures that explicitly control temporal and prosodic dynamics achieve better performance despite substantially fewer parameters, underscoring that architectural design and data alignment are more critical than scale in low-resource settings.

Beyond the quantitative evaluation, the released resources, including bilingual aligned text, reproducible inference pipelines, and synthesized audio of the entire Constitution; constitute a practical and reusable asset for the community. This bilingual legal TTS resource provides immediate value for Quechua accessibility while also serving as a domain-matched seed for downstream ASR and speech-to-text research, where synthetic data from trusted sources has been shown to yield measurable benefits.

Overall, this work demonstrates that high-impact speech resources for low-resource languages can be built using open tools and cross-lingual learn-

[14] https://latamjournalismreview.org

ing, even under limited supervision. We hope that this effort encourages further research on legally grounded, socially relevant speech technologies for indigenous and underrepresented languages.

7. Data and Prompt Availability

Due to the appendix constraint and for anonymity, we omit the data and prompts used. We will deliver them upon positive acceptance.

Acknowledgements

We thank the speaker communities and language workers associated with the Quechua work performed. We also thank the maintainers and staff of documentation archives and repositories consulted in this study for curating and providing access to materials and metadata. Finally, we thank the Political NLP 2026 reviewers for constructive feedback that improved the camera-ready version.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, et al. 2023. Findings of the iwslt 2023 evaluation campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61.

Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, et al. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481.

Ibrahim Sa'id Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, et al. 2024. Findings of the iwslt 2024 evaluation campaign. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11.

Nouman Ahmed, Natalia Flechas Manrique, and Antonije Petrović. 2023. Enhancing spanish–quechua machine translation with pretrained models and diverse data sources. In *AmericasNLP 2023*, pages 156–162.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 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.

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*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Wyndham A Bewes. 1920. The new constitution of peru (january 18, 1920). *Journal of Comparative Legislation and International Law*, 2(3):266–269.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mhubert-147: A compact multilingual hubert model. *arXiv preprint arXiv:2406.06371*.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pages 4978–4982.

Chih-Chen Chen, William Chen, Rodolfo Zevallos, and John E. Ortega. 2024. Evaluating self-supervised speech representations for indigenous american languages. In *Proc. LREC-COLING 2024*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271, Vienna, Austria. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Garcia Baltazar, María Elena Méndez 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 *AmericasNLP 2023*, pages 206–219.

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.

Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson. 2020. Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6504–6513, Marseille, France. European Language Resources Association (ELRA).

Ministerio de Justicia y Derechos Humanos. 2020. *Constitución Política del Perú en Castellano y Quechua*, cuarta edición oficial edition. Ministerio de Justicia y Derechos Humanos, Lima, Perú. Sistema Peruano de Información Jurídica (SPIJ).

Ngoc-Son Nguyen, Thanh VT Tran, Hieu-Nghia Huynh-Nguyen, Truong-Son Hy, and Van Nguyen. 2025. Diflow-tts: Compact and low-latency zero-shot text-to-speech with factorized discrete flow matching. *arXiv preprint arXiv:2509.09631*.

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, Rodolfo Joel Zevallos, William Chen, and Idris Abdulmumin. 2025. QUESPA submission for the IWSLT 2025 dialectal and low-resource speech translation task. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 260–268, Vienna, Austria (in-person and online). Association for Computational Linguistics.

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 2023*, pages 261–268.

John E. Ortega, Rodolfo Zevallos, William Chen, and Ibrahim Said Ahmad. 2024. Quespa submission for the iwslt 2024 dialectal and low-resource speech translation task. In *Proc. IWSLT 2024*, pages 173–181.

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.

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

Mónica Romero, Sandra Gómez, and Iván G. Torre. 2024a. Asr advancements for indigenous languages: Quechua, guarani, bribri, kotiria, and wa'ikhana. *arXiv*.

Mónica Romero, Sandra Gómez-Canaval, and Iván G. Torre. 2024b. Automatic speech recognition advancements for indigenous languages of the americas. *Applied Sciences*, 14(15):6497.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142.

Rodolfo Zevallos, Núria Bel, Guillermo Cámbara, Mireia Farrús, and Jordi Luque. 2022a. Data Augmentation for Low-Resource Quechua ASR Improvement. In *Interspeech 2022*, pages 3518–3522.

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