Text-to-speech system for low-resource languages: A case study in Shipibo-Konibo (a Panoan language from Peru)

Daniel Menéndez
Postgraduate School
Pontificia Universidad Católica del Perú
dmenendez@pucp.edu.pe

Héctor Erasmo Gómez
Chana Research Group
Postgraduate School
Pontificia Universidad Católica del Perú
hector.gomez@pucp.edu.pe

Abstract

This paper presents the design and development of the first Text-to-Speech (TTS) model and speech dataset for Shipibo-Konibo, a low-resource indigenous language spoken mainly in the Peruvian Amazon. Despite the challenge posed by data scarcity, data was gathered and structured for the dataset, thus the TTS model was trained with over 4 hours of recordings and 3,025 written sentences. The test results demonstrated an intelligibility rate (IR) of 88.56% and a mean opinion score (MOS) of 4.01, confirming the quality of the generated audio using Tacotron 2 and HiFi-GAN. This study highlights the potential for extending this approach to other indigenous languages in Peru, contributing to their documentation and revitalization.

1 Introduction

With over 7,000 languages worldwide (SIL, Accessed March 14 2024), many of them face extinction, threatening linguistic diversity and indigenous knowledge (Evans and Levinson, 2009; Spiegelhalter et al., 2002; Campbell and Rehg, 2018). In the particular case of Peru, official statistics recognize 48 indigenous languages, while Glottolog lists 90 (Hammarström et al., 2021). Among these, Shipibo-Konibo, part of the Pano family, is spoken by approximately 40,000 people[1]; however, NLP efforts for the language face challenges such as at least two orthographic traditions and limited digital resources. For more details about the Shipibo-Konibo communities and its current writing and speech systems, see Appendix A.

Thus, to continue previous efforts made in projects like Huqariq (Zevallos et al., 2022), where a combined speech corpus of Quechua, Aymara, and Shipibo-Konibo was collected, this paper presents the development of the first Shipibo-Konibo TTS model. The study includes dataset creation, model selection, training, and evaluation. The goal is to facilitate future NLP applications for Shipibo-Konibo and other indigenous languages. Furthermore, we aim to support language revitalization by offering audio resources that facilitate pronunciation practice beyond the classroom. Additionally, integrating synthesized speech into educational tools such as dictionaries and verb conjugators enhances language accessibility as it was done in other countries (Pine et al., 2022).

2 Speech synthesis for low-resource languages

Developing a TTS model for low-resource languages presents challenges, primarily the lack of structured data. This limitation impacts training strategies, requiring transfer learning and specialized neural architectures as Transformer-TTS (Li et al., 2019) or Glow-TTS (Kim et al., 2020). Additional challenges include the estimation of computational resources, as renting the necessary capacity was the only viable option, leading to additional expenses within our limited budget.Another challenge arose in selecting the most suitable metrics for this low-resource scenario. While MOS is the most commonly used in such cases, the evaluations were also designed to extract computable data on intelligibility (Xu et al., 2020).

3 Data Collection

No existing Shipibo-Konibo speech dataset was available, so creating one from scratch was necessary. Texts published after 2015 alphabet normalization were prioritized for relevance, and corresponding audio recordings were collected.

[1] The 2017 Peruvian census estimates the total Shipibo-Konibo population at 34,000, but the actual figure is expected to be higher (INEI, 2018)

Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 1–7
May 4, 2025 ©2025 Association for Computational Linguistics

3.1 Text Compilation

Key sources included the Shipibo-Konibo translation of The Little Prince (*Jatibi Ibo Bake*) and some bilingual educational materials from the Peruvian Ministry of Education, resulting in a corpus of 3,025 sentences.

3.2 Audio Compilation

Audio recordings were done following the LJ Speech (Ito and Johnson, 2017) dataset requirements:

- Sampling rate of 22,050Hz or higher.
- Single speaker.
- The sentences must contain diverse phonemes.
- Audio duration must be between 1-10 seconds.
- Audio segments must not have long silence at the beginning or at the end.
- Audio segments must not contain long pauses.

The recording sessions featured a native Shipibo-Konibo speaker with prior experience in voice documentation, which were conducted over three months in intervals of up to two hours to ensure vocal consistency.

By the end of all the sessions, the final dataset exhibited the characteristics shown in Table 1.

Characteristics	Value
Number of sentences	3025
Total duration	$4h37m14s$
Minimum sentence duration	$1.08s$
Maximum sentence duration	$12.1s$
Average sentence duration	$5.1s$

Table 1: Details of the audio clips collected from the Shipibo-Konibo language.

Finally, all the sentences were trimmed and resampled to 22.05KHz. Long silences were removed, volume, speed, and text were normalized as required by the Tacotron 2 model (Shen et al., 2018).

4 The proposed TTS model

A previous evaluation of many TTS models led to the selection of Tacotron 2 as a result of its success in low-resource environments, showing promising results using datasets of less than 3 hours (Debnath et al., 2020; Dasare et al., 2022; Gopalakrishnan et al., 2022) and the vibrant state of development also. Previous models like Voice Loop 2 (Taigman et al., 2017), FastSpeech2 (Ren et al., 2020) and Transformer-TTS (Li et al., 2019) were discarded due to factors such as their fall into disuse or their inferior performance.

4.1 Tacotron 2

Tacotron 2 (Shen et al., 2018) was chosen for its encoder-attention-decoder architecture, which improves speech quality. Our model was implemented in PyTorch and trained using transfer learning techniques.

4.2 The HiFi-GAN vocoder

Unlike the original Tacotron 2 approach using WaveGlow (Prenger et al., 2019), this study employed HiFi-GAN (Kong et al., 2020) due to its recent superior performance. HiFi-GAN, a GAN-based vocoder, generates waveforms from Tacotron 2 spectrograms.

A summary of the entire proposed model is shown in Figure 1.

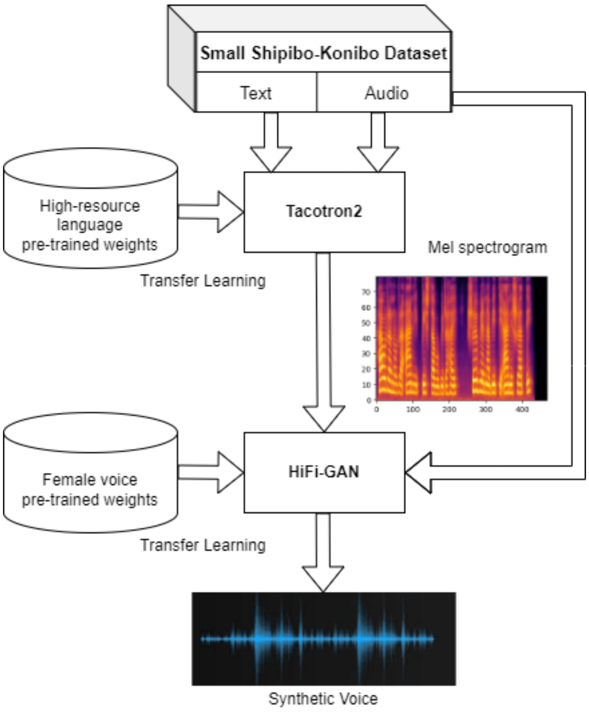

Small Shipibo-Konibo Dataset
Text
Audio
High-resource language pre-trained weights
Tacotron2
Transfer Learning
Mel spectrogram
Female voice pre-trained weights
HiFi-GAN
Transfer Learning
Synthetic Voice

Figure 1: Architecture of the proposed TTS model.

5 Experimentation and Results

Tacotron 2 and HiFi-GAN were trained separately, as the vocoder relied on the spectrogram predictions generated by the fully trained Tacotron 2 model. Hyperparameter tuning was optimized using insights from similar projects (Debnath et al., 2020; Dasare et al., 2022).

5.1 Training the Tacotron 2 Model

We used Google Colab with an Nvidia A100 GPU for the training process, and we fine-tuned a pre-trained Latin American Spanish model from Hugging Face (Cedillo, 2023), where the best hyperparameters achieved are displayed on tables 2 and 3, showing promising encoder-decoder alignment from the first epoch.

Hyperparameters	Value
Epochs	225
Batch size	16
Gate threshold	0.5
Decoder dropout	0.1
Attention	0.1

Table 2: Final Tacotron2 Hyperparameters.

Optimizer Parameters	Value
Learning rate	3.10^{-4}
β_1	0.9
β_2	0.999
Weight decay	1.10^{-5}

Table 3: Parameters used for the Adam optimizer.

The training process took about 225 epochs until it stalled at a validation loss of 0.1434 (Figure 2) and encoder-decoder alignment as shown in Figure 3.

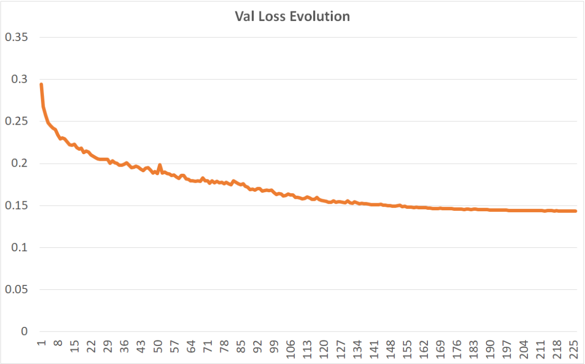


Figure 2: Validation loss function evolution.

5.2 Training the HiFi-GAN vocoder

The vocoder was trained after Tacotron 2 using a transfer learning strategy with a pre-trained universal female voice model to achieve faster convergence with the native speaker's voice. After 34 epochs, loss stabilization indicated convergence.

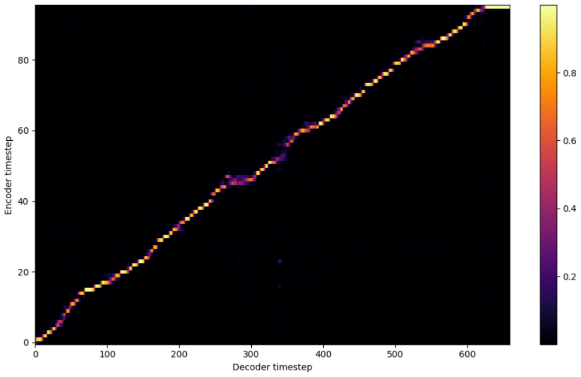


Figure 3: Epoch 225 final encoder-decoder alignment.

5.3 Evaluation and Results

For model evaluation, we extracted 400 additional sentences from the book Koshi Shinanya Ainbo (The Testimony of a Shipibo Woman) (Valenzuela and Rojas, 2005), a pre-2015 Shipibo-Konibo book with an alphabet easily adaptable to the modern one. It narrates the life and traditions of Mrs. Ranin Ama and her community.

As an initial inference test, Figure 4 presents a five-word phrase from the book, lasting three seconds: Nokon titan ea axeani jawékibo ("The things my mother taught me").

Objective metrics like PESQ (Rix, 2003) and POLQA (Beerends et al., 2013) require extensive high-quality reference data, which is unavailable for Shipibo-Konibo. Instead, we followed subjective evaluations by several native speakers, despite being more time-consuming. The Mean Opinion Score (MOS) is the most commonly used metric in low-resource scenarios, while the Intelligibility Rate (IR) can also provide valuable insights.

The Intelligibility Rate (IR) measures how accurately listeners transcribe synthesized speech, calculated as the percentage of correctly identified words. The Mean Opinion Score (MOS) evaluates speech quality based on clarity, naturalness, and fluency, rated on a scale from 1 to 5 (see Table 4). The average score given by the evaluators is used to determine the final mean opinion score.

Value	Descripción
1	Unacceptable or very poor quality
2	Poor quality
3	Acceptable or adequate quality
4	Good quality
5	Excellent quality

Table 4: MOS metric scale.

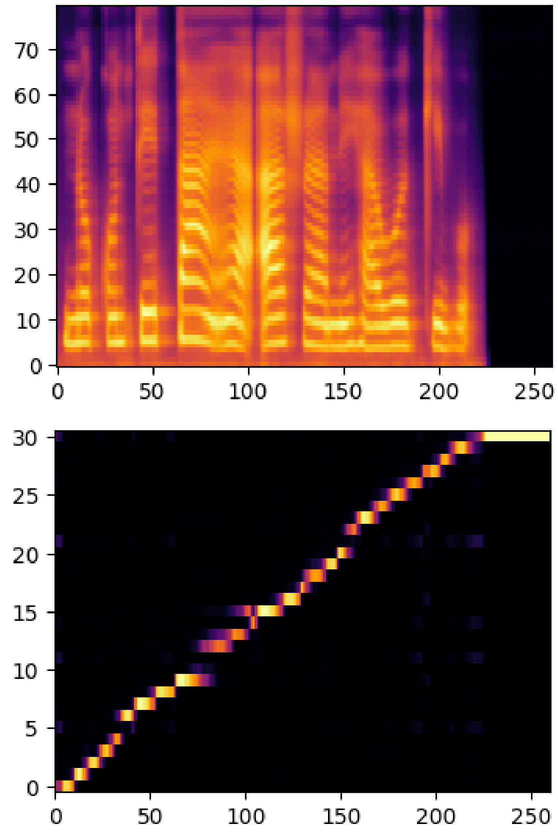


Figure 4: Spectrogram and encoder-decoder alignment graph of a 3-second synthetic phrase.

A total of 26 native Shipibo-Konibo speakers (11 men, 15 women), all under 30 and university-educated, participated as evaluators. All of them familiar with the use of PC and office software tools to make evaluations easier. To compare natural and synthetic speech, 25% of the evaluated phrases were natural samples, randomly included in a set of 20 audios per evaluator. Table 5 presents the intelligibility rate (IR) results for both speech types.

Type of voice	IR
Natural	83.45%
Synthetic	88.56%

Table 5: Intelligibility rate results.

Meanwhile, the results obtained from the same evaluators for the mean opinion score (MOS) are shown in Table 6.

Type of voice	MOS
Natural	3.75 ± 1.2
Synthetic	4.01 ± 1.09

Table 6: Mean opinion score results.

Furthermore, Figure 5 illustrates the comparison between the MOS distributions for natural and synthetic voices.

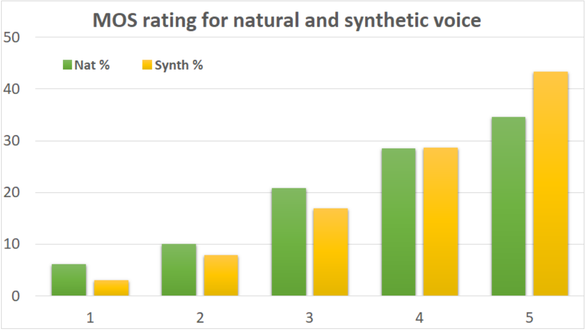


Figure 5: MOS percentage distribution for natural and synthetic voice.

5.4 Discussion

The results were highly positive, with an IR of 88.56% and MOS of 4.01, meeting expectations. Surprisingly, the synthetic voice outperformed the natural voice in both metrics.

However, a parallel qualitative analysis revealed issues such as ambiguous intonation, improper punctuation handling, and pronunciation inconsistencies. The encoder-decoder alignment analysis during sentence synthesis identified some recurring synthesis faults, particularly in the suffixes *titai*, *tiai*, *tian* and *wai*, due to insufficient training data. Additionally, significant pronunciation variations were observed across recordings for the sounds *iki*, *nai*, *non*, *ani*, *ja*, *ea*, *baon*, *kon*, *xe*, and *noa*.

6 Conclusions and future work

We successfully designed, developed, and evaluated the first TTS model for the Shipibo-Konibo language. For this task, we compiled speech corpus of over 4 hours and 3,025 labeled sentences.

The Tacotron 2 spectrogram predictor and HiFi-GAN vocoder were effectively trained, achieving an IR of 88.56% and an MOS of 4.01, which indicates that the synthetic speech samples surpassed the natural ones in some tests. These results show the potential of the model for other Panoan and Amazonian languages with similarly limited speech data.

Future work will focus on improving corpus quality, refining audio recording conditions, and incorporating more diverse sentence structures. This model serves as a foundation for adapting TTS systems to other indigenous languages within the Pano family and beyond.

Limitations

Despite promising results, this study has some limitations:

- The corpus consists of only 4 hours and 3,025 sentences, which may not fully capture the phonetic and prosodic variability of Shipibo-Konibo. A larger dataset could improve generalization.
- Our was trained on a young female single speaker, limiting voice diversity. Multi-speaker training for broader applicability should be included.
- Although our collaborator is a native Shipibo-Konibo speaker, she has been living in the capital, Lima, for several years. The distance from her community and the reduced daily use of her language have led to variations in her pronunciation, which are reflected during evaluations in the intelligibility rate (IR) and mean opinion score (MOS) metrics for natural voice.
- Due to the absence of high-quality reference data, PESQ and POLQA were not used. Subjective evaluations (IR and MOS) were employed, but they are more time-consuming and dependent on evaluator bias.
- While the model provides a framework for low-resource TTS development, its adaptation to other Panoan or Amazonian languages requires additional data and fine-tuning.
- Shipibo-Konibo has at least two writing conventions. The dataset prioritizes the 2015 standard, but variations may affect the model's usability in different communities.

Ethics statement

This study was conducted with ethical considerations in mind, ensuring respect for the Shipibo-Konibo community and its linguistic heritage. The native speaker who contributed to the dataset participated voluntarily, providing informed consent before any recording sessions. Additionally, she was fairly compensated for her time and contributions.

We acknowledge the importance of responsible data collection and ensure that all linguistic resources were gathered and processed with the utmost respect for cultural sensitivity. The project aligns with ethical guidelines for language documentation and preservation, aiming to empower indigenous communities by providing AI tools that support language revitalization.

Any future use of the dataset and TTS model will be governed by principles of transparency and community engagement, ensuring that the benefits of this research extend directly to the Shipibo-Konibo people.

Acknowledgments

We would like to express our gratitude to the *Chana Project* of the *Pontificia Universidad Católica del Perú (PUCP)* and prof. Roberto Zariquiey for their support and for providing us with literature in the Shipibo-Konibo language, which was essential for conducting the evaluations.

Additionally, we extend our appreciation to the *Universidad Intercultural de la Amazonía (UNIA)* for their support in providing facilities and student participation during the evaluation process.

Finally, we would like to sincerely thank Ms. Seleni Rojas, an activist dedicated to the preservation of Shipibo-Konibo culture, whose invaluable contribution made the creation of both the dataset and the TTS model possible.

References

John G Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl. 2013. Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment. *journal of the audio engineering society*, 61(6):366–384.

Lyle Campbell and Kenneth Rehg. 2018. Introduction. In Lyle Campbell and Kenneth Rehg, editors, *The Oxford Handbook of Endangered Languages*, pages 1–18. Oxford: Oxford University Press.

Rene Cedillo. 2023. taco2-checkpoints. https://huggingface.co/datasets/rmcpantoja/taco2-checkpoints/tree/main/es,[Accesed: (2024-02-09)].

Ashwini Dasare, KT Deepak, Mahadeva Prasanna, and K Samudra Vijaya. 2022. Text to speech system for lambani-a zero resource, tribal language of india. In *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Ankur Debnath, Shridevi S Patil, Gangotri Nadiger, and Ramakrishnan Angarai Ganesan. 2020. Low-resource end-to-end sanskrit tts using tacotron2, waveglow and transfer learning. In *2020 IEEE 17th*

India Council International Conference (INDICON)*, pages 1–5. IEEE.

Nicholas Evans and Stephen Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–492.

Thirumoorthy Gopalakrishnan, Syed Ayaz Imam, and Archit Aggarwal. 2022. Fine tuning and comparing tacotron 2, deep voice 3, and fastspeech 2 tts models in a low resource environment. In *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, pages 1–6. IEEE.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. Glottolog 4.4. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at http://glottolog.org. Accessed on 2021-05-20.

INEI. 2018. *Perú: Resultados definitivos de los Censos Nacionales 2017*. Instituto Nacional de Estadística e Informática, Lima, Perú.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7346–7359.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Antony W Rix. 2003. Comparison between subjective listening quality and p. 862 pesq score. *Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN'03), Prague, Czech Republic*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

SIL. Accessed March 14 2024. *Ethnologue: Languages of the World.* SIL International, https://www.ethnologue.com/.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde. 2002. Bayesian measures of model complexity and fit. *J. R. Statist. Soc. B*, 64(4):583–639.

Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani. 2017. Voiceloop: Voice fitting and synthesis via a phonological loop. *arXiv preprint arXiv:1707.06588*.

Pilar Valenzuela and Agustina Valera Rojas. 2005. *Koshi shinanya ainbo*. Fondo Editorial de la Facultad de Ciencias Sociales UNMSM.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2802–2812.

Rodolfo Zevallos, Luis Camacho, and Nelsi Melgarejo. 2022. Huqariq: A multilingual speech corpus of native languages of peru for speech recognition. *arXiv preprint arXiv:2207.05498*.

A About de Shipibo-Konibo community

The Shipibo-Konibo community is an indigenous Amazonian group located primarily in the Ucayali region of Peru, with additional populations in Loreto, Huánuco, Madre de Dios and urban areas such as Lima and Pucallpa. They are known for their rich cultural heritage, textile art, and deep spiritual connection to nature.

In 2015, the Peruvian Ministry of Education formalized an official alphabet as shown in Figure 7, though older orthographies still exist as can be seen on the Figure 8.

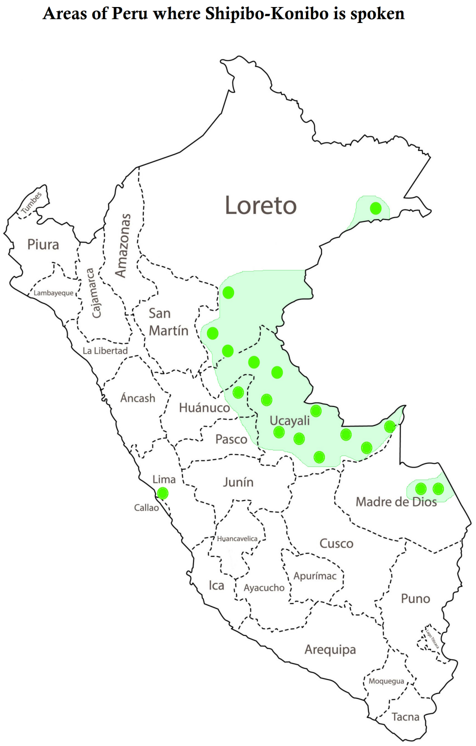
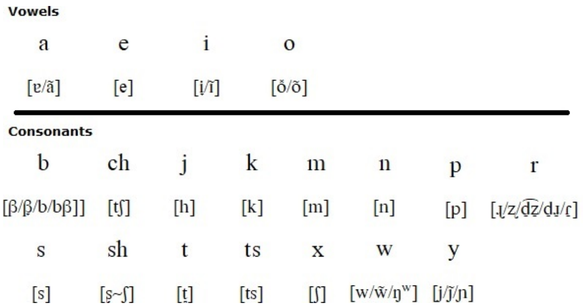


Figure 6: Map of Peru showing the areas where various Shipibo-Konibo communities are located.



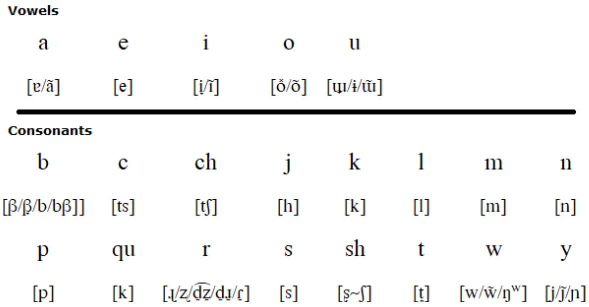
Vowels

a	e	i	o
[ɐ/ã]	[e]	[i̞/ĩ]	[o̊/õ]

Consonants

b	ch	j	k	m	n	p	r
[β/β̞/b/bβ]]	[tʃ]	[h]	[k]	[m]	[n]	[p]	[ɻ/ʐ/d͡ʐ/ɖɻ/ɾ]
s	sh	t	ts	x	w	y	
[s]	[s̺~ʃ]	[t̪]	[ts]	[ʃ]	[w/w̃/ŋʷ]	[j/j̃/ɲ]	

Figure 7: Current Shipibo-Konibo normalized graphemes and phonemes



Vowels

a	e	i	o	u
[ɐ/ã]	[e]	[i̞/ĩ]	[o̊/õ]	[ɯ̞/ɨ/ɯ̃]

Consonants

b	c	ch	j	k	l	m	n
[β/β̞/b/bβ]]	[ts]	[tʃ]	[h]	[k]	[l]	[m]	[n]
p	qu	r	s	sh	t	w	y
[p]	[k]	[ɻ/ʐ/d͡ʐ/ɖɻ/ɾ]	[s]	[s̺~ʃ]	[t̪]	[w/w̃/ŋʷ]	[j/j̃/ɲ]

Figure 8: One of the old Shipibo-Konibo graphemes and phonemes