Voice Activation Detection for Transcription of Indigenous Languages

Rolando Coto-Solano
Dartmouth College
rolando.a.coto.solano
@dartmouth.edu

Mikaela Browning
Dartmouth College
Mikaela.Browning.26
@dartmouth.edu

Thomas Corrado
Dartmouth College
thomas.r.corrado.25
@dartmouth.edu

Sally Akevai Tenamu Nicholas
Waipapa Taumata Rau
University of Auckland
ake.nicholas@auckland.ac.nz

Abstract

Voice Activity Detection (VAD) is the first step in a workflow intended for the automated transcription of Indigenous and low-resource languages. However, VAD's effectiveness when detecting voices in fieldwork settings remains untested. Fieldwork recordings have very different noise and interference conditions from the datasets that mainstream VAD models have been trained for, and so they might fail when confronted with this type of linguistic data. This paper tests different algorithms using data from two typologically distinct Indigenous languages: Bribri from Costa Rica and Cook Islands Māori from Polynesia. We compare energy-based methods (PyDub), GMM-based methods (WebRTC VAD), and two neural-network based methods (Silero and SpeechBrain) against human-annotated transcriptions. Our results indicate that hybrid architectures like that of SpeechBrain obtain the best results (89% accuracy for Bribri and 94% for Cook Islands Māori). However, no system performed well when tagging non-speech segments, which might indicate a bias towards marking the natural noise in a fieldwork setting as a false-positive for voice. With these findings we hope to inform the selection of VAD tools when implementing ASR workflows.

1 Introduction

Work in Indigenous language documentation faces important bottlenecks, one of which is the transcription of audio recordings. Language departments and researchers usually collect audio from fieldwork and documentation efforts, in the hopes of transcribing it in the future. The information in the recordings can be used for a range of purposes, from the creation of educational materials to its analysis as linguistic research. However, transcribing these recordings represents a major hurdle. Usually only a few experts can type out these transcriptions, and this work is very time consuming, with estimates of up to 50-human work hours to transcribe an hour of recording (Durantin et al., 2017; Shi et al., 2021).

There have been efforts to alleviate the transcription bottleneck by incorporating automated speech recognition (ASR) into Indigenous language documentation workflows. Fine-tuning custom speech models for a specific Indigenous language is becoming increasingly common because of the assumption that these models will accelerate transcription and thereby release the worker's time for more urgent work. However, there is little research about the models' actual use in documentation workflows. There is some evidence (Prud'hommeaux et al., 2021) that ASR does accelerate transcription, but there are reports (Teikitohe, Personal Communication) that a big part of this acceleration comes not from the transcription itself, but from one of its ancillary tasks: the separation of voice and non-voice sections in the recording.

Voice activation detection (VAD) is the task of identifying the presence of absence of human speech in a section of an audio recording. In a cascading language documentation workflow, VAD would be the first stage of the processing, where the computer identifies which parts of the signal are actual speech. These segments should ideally then by sent to language ID (in case of code-switching), diarization, and finally to the appropriate transcription model.[1] A language worker can perform this task manually on software like ELAN (Wittenburg et al., 2006), but this can often be time-consuming on its own. Automating this task would be highly desirable to increase the efficiency of transcription

[1] It is possible to have an end-to-end model that performs these tasks in a single pass, but these are not usually available for low-resource languages.

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 177–184

in Indigenous languages.

Despite these potential advantages, there is no research on VAD for Indigenous languages. Moreover, recordings in Indigenous languages usually involve soundscapes that are significantly different from those for recordings from majority languages like English. Recordings for Indigenous languages usually come from fieldwork environments, which might include sounds of animals or nature (e.g. chickens, rain) interspersed or interfering with the recorded speech. These recordings might also include song and other forms of oral arts that might be underrepresented in English VAD-training datasets. Because of these differences, it might not be straightforward to simply use existing VAD models for work in Indigenous languages.

In this paper, we will study the performance of state-of-the-art VAD models for fieldwork recordings in two very different Indigenous languages: Cook Islands Māori and Bribri from Costa Rica. By studying their performance, we hope to inform the choice of computer scientists working to implement Indigenous language documentation workflows and further accelerate this work.

1.1 Bribri and Cook Islands Māori

Bribri is a Chibchan language spoken in Southern Costa Rica. It is spoken by approximately 7000 people (INEC, 2011), and it is a vulnerable language (Sánchez Avendaño, 2013), spoken by few children. The language has publicly available audio corpora (Flores-Solórzano, 2017). There has been work on Bribri NLP, including speech recognition (Coto-Solano, 2021; Ebrahimi et al., 2022b; Coto-Solano et al., 2024), machine translation (Feldman and Coto-Solano, 2020; Mager et al., 2021; Ebrahimi et al., 2023b,a; Jones et al., 2023; Chiruzzo et al., 2024; De Gibert et al., 2025), natural language inference (Ebrahimi et al., 2022a; Kann et al., 2022), forced alignment (Coto-Solano and Flores-Solórzano, 2016; Flores-Solórzano and Coto-Solano, 2017; Coto-Solano et al., 2022b), parsing (Coto-Solano et al., 2021; Karson and Coto-Solano, 2024), semantics (Solórzano, 2009; Coto-Solano, 2022), morphological segmentation and analysis (Flores-Solórzano, 2017, 2019; Anderson et al., 2025), diacritic restoration and spell-checking (Coto-Solano et al., 2025), digital keyboards (Flores-Solórzano, 2010) and digital dictionaries (Krohn, 2020, 2021).

Cook Islands Māori (CIM) is a Polynesian language, spoken by 12500 people in the Cook Islands and approximately 10000 people in the diaspora in New Zealand and Australia (Nicholas, 2018; Ministry of Finance and Economic Management, Government of the Cook Islands, 2021). It is also an endangered language, and in some islands like Rarotonga it is increasingly difficult to find children who speak the language. There are public corpora available (Nicholas, 2012). There is previous NLP work on Cook Islands Māori, including work on speech recognition (Foley et al., 2018; Coto-Solano et al., 2018, 2022a), forced alignment (Nicholas and Coto-Solano, 2019; Coto-Solano et al., 2022b), text-to-speech (James et al., 2024), parsing (Karnes et al., 2023) and diacritic restoration (Coto-Solano et al., 2025).

Bribri is a tonal language, with more phonemes than CIM. Additionally, CIM is related to languages like Hawaiian and Te Reo Māori from Aotearoa New Zealand, which are relatively well represented in speech foundation models such as Whisper (Radford et al., 2023).

2 Methodology

2.1 Data preparation

In order to perform these tests, we analyzed three audio files for each language. Each of the files contains a fieldwork recording for a different speaker of the language. In the case of Bribri, we selected recordings with a total duration of approximately 17 minutes, and for CIM, the recordings included a total of 73 minutes of audio. All the files had a corresponding ELAN transcription file which had been manually annotated and verified. From these files, we extracted the start and end time of each voice segment as determined by human annotators.

2.2 Evaluated models

The next step was to run the available audio files through the VAD models. We selected four models, using either signal processing energy-based approaches, or deep learning approaches. First, we used PyDub (Robert, 2011), which uses an amplitude threshold in decibels and a minimum silence duration as a way to separate silence from segments with potential speech. We used both the detect “silence" function, and its complement, the detect “non_silent" regions function.

Second, we used Silero VAD (Silero Team, 2021), a widely-used neural network-based system. It classified frames with a probability between zero and one for containing human speech, and it is

trained on a large multilingual corpus of more than 100 languages. Silero is used in many production pipelines (including in conjunction with Whisper (Radford et al., 2023)), and is trained to be relatively efficient. Third, we also tested WebRTC VAD (Wiseman, 2016). It uses a Gaussian Mixture Model to classify each frame based on spectral and energy features.

Finally, we used SpeechBrain VAD (Ravanelli et al., 2021). This model uses a convolutional, recurrent, dense neural network (CRDNN), which assigns Bayesian probabilities for speech presence with a neural network, and then adjusts these probabilities using energy-based thresholding.

2.3 Evaluation

After the recordings were tagged using the different algorithms, we compared them with the manual tagging using the following method. First, we split the recordings into 10ms windows. We annotated each of those 10ms windows with whether they were inside of an ELAN annotation in the manual transcription or not. Figure 1 shows an example of this. In this figure, the region between 20ms and 50ms contains the segment /a/, and the region between 0ms and 20ms is considered to not have any human voice at all.

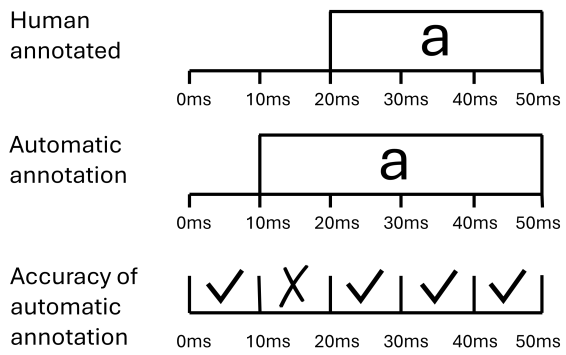

Figure 1: Example of accuracy evaluation. The human-annotated version is split in 10ms intervals and tagged for presence or absence of annotations. The automatic version is split in the same way, and then the two versions are compared. If both regions have the same value (presence or absence of voice), then the region is considered to be accurately tagged.

Next, we also extracted the annotation boundary information for the automatic transcriptions. In the example in figure 1, the region between 10ms and 20ms is mistakenly labeled as having speech, whereas this wasn't so in the gold-standard, human annotated version. With this information, we compared the automatic and human-annotated versions of the voice detection and calculated accuracy, precision, recall and F1. In the example in figure 1, the accuracy of the automatic annotation would be 80%. We calculated the previously mentioned metrics (accuracy, precision, recall, F1) for each recording, and then we report the average and standard deviation for each language.

3 Results

Table 1 shows the accuracy results for the studied algorithms. The PyDub algorithm, which relies only on amplitude and duration cues, had the lowest performance, with accuracies of approximately 13% for Bribri and 9% for CIM.

	Bribri	CIM
PyDub	13 ± 5	9 ± 6
PyDub (Silence)	12 ± 3	9 ± 6
Silero VAD	55 ± 2	74 ± 3
WebRTC VAD	84 ± 9	73 ± 3
SpeechBrain VAD	$\mathbf{89 \pm 4}$	$\mathbf{94 \pm 1}$

Table 1: Accuracy for VAD for two low-resource languages by algorithm

The Silero and WebRTC algorithms had similar accuracy for CIM (approx. 73%), but they were very different for Bribri: 55% for Silero, versus 84% for WebRTC. The best performing algorithm was SpeechBrain, which combined neural network and signal processing methods, and had an accuracy of 89% for Bribri and 94% for CIM.

Table 2 shows the results for precision, recall and F1 by language, algorithm, and type of segment (speech versus non-speech). The main pattern observable in the table is the fact that the performance for non-speech sections is much worse than that for speech sections. All systems might be too aggressive in trying to maximally tag every segment as a potential speech segment. For example, when tagging Bribri, the PyDub precision for speech is 99%, but the precision for non-speech is 5%; this indicates that the system is simply failing to separate speech from non-speech.

The best results are again obtained with SpeechBrain, which gets very high F1s for speech (94% for Bribri and 97% for CIM), and the more acceptable results for non-speech detection (29% for Bribri and 46% for CIM).

		Bribri			CIM		
		Precision	Recall	F1	Precision	Recall	F1
Speech	PyDub	99 ± 1	9 ± 6	16 ± 10	33 ± 47	0 ± 1	1 ± 1.0
	PyDub (Sil)	99 ± 1	7 ± 5	12 ± 8	33 ± 47	0 ± 1	1 ± 1
	Silero	99 ± 1	53 ± 1	69 ± 1	97 ± 2	73 ± 3	83 ± 2
	WebRTC	95 ± 1	87 ± 10	91 ± 6	98 ± 2	72 ± 3	83 ± 2
	SpeechBrain	98 ± 1	91 ± 5	$\mathbf{94 \pm 2}$	98 ± 1	96 ± 1	$\mathbf{97 \pm 1}$
Non-speech	PyDub	5 ± 1	97 ± 5	10 ± 2	9 ± 6	100 ± 0	16 ± 10
	PyDub (Sil)	5 ± 1	97 ± 4	10 ± 2	9 ± 6	100 ± 0	16 ± 10
	Silero	9 ± 3	84 ± 16	16 ± 5	21 ± 13	81 ± 14	32 ± 15
	WebRTC	8 ± 2	20 ± 17	10 ± 4	21 ± 13	80 ± 1	31 ± 15
	SpeechBrain	21 ± 6	52 ± 30	$\mathbf{29 \pm 13}$	38 ± 6	67 ± 20	$\mathbf{46 \pm 3}$

Table 2: VAD for two low-resource languages by algorithm

4 Discussion

The results reveal several important patterns when applying VAD to Indigenous language data. Perhaps the most interesting result, as mentioned above, is the relatively poor performance of all systems when detecting non-speech. This might indicate that VAD systems are biased towards over-detecting the noise in fieldwork environments as speech, which might result in numerous false positives. The energy-based thresholding methods were particularly ineffective in this experiment. Unlike studio-quality audio, field recordings contain sounds that can carry substantial energy but are not speech. The aforementioned chickens and rain are examples of this, but also wind and noise generated from household appliances ranging from air conditioning to light bulbs with poor electric insulation. The main recommendation from this would be to exercise caution when using general-purpose audio tools in language documentation work, as they might require extensive adaptation to get used to fieldwork audio conditions.

The divergences between Silero and WebRTC on Bribri, despite their similar performance on CIM, might be related to the differences in the recordings themselves. The Bribri recordings have much higher levels of interference and environmental noise (publicly available Bribri example versus CIM example). Silero's lower performance on Bribri (55%, compared to 84% for WebRTC) may reflect differences on what each system recognizes as speech. Silero is trained on a larger multilingual corpus, possible from audiobooks, movies, and audio recorded in urban settings. Therefore, its expectations about speech might be too different from those present in fieldwork conditions. On the other hand, WebRTC's GMM-based approach, which relies on lower-level spectral and energy features rather than learned acoustic representations might give it an advantage and make it more robust in unfamiliar sound environments, precisely because it makes fewer assumptions above what speech should shoud like. The fact that CIM recordings have roughly the same accuracy suggests that the CIM data might resemble the acoustic conditions that both tools were designed for.

SpeechBrain's high performance likely comes from its hybrid architecture, which combines neural network posterior probabilities and energy-based post-processing. However, its low F1 scores for non-speech reveal that even SOTA methods have weaknesses when analyzing fieldwork data. All systems struggled to distinguish environmental noise from speech, which is one of the core challenges in analyzing fieldwork audio. Future work might need to focus on fine-tuning existing VAD models on appropriate environmental sounds.

5 Conclusions

This paper presents an evaluation of voice activation detection (VAD) systems on fieldwork recordings for Indigenous languages. Our experiment shows that the choice of algorithm matters: energy-based methods like PyDub might be unsuitable for fieldwork data processing, while hybrid neural-network and energy-based architectures provide the best results. The massive differences in results indicate that the choice of VAD algorithm is not trivial, and it should be treated as an important step in the preprocessing for speech recognition. Our experiment also indicates that VAD systems are systematically weak when confronted with the en-

vironmental noise present in fieldwork recordings, with F1 results being much lower than those for actual human speech. In summary, the soundscapes involved in fieldwork might not be well represented in current VAD training data, and don't appear to be well understood by SOTA VAD algorithms.

As part of our future work, we need to test more algorithms, for example pyannotate.audio (Bredin and Laurent, 2021) and Cobra VAD (Picovoice, 2024), as well as ASR systems that have the VAD incorporated in them such as WhisperX (Bain et al., 2023). We also need to fine-tune these algorithms using fieldwork audio to measure any potential improvements. We also intend to test both on a wider range of languages, and on languages that are more similar phonologically to one another. Bribri and CIM are very different, and therefore the performance gap observed between the two languages might be due to Bribri's larger phonological inventory. This should be tested in future experiments.

We hope that this work encourages language documentation workers who might be experimenting with ASR to more closely consider the different steps involved in incorporating speech recognition into their workflows, with the hopes that it actually helps save time and enables the workers to focus on their goals of revitalization and reclamation.

Limitations

The work presented here was done on relatively small amounts of data, particularly for Bribri (only 17 minutes), so these results need to be further tests with larger masses of data. Moreover, the specific noise conditions need to be controlled, with a more refined experiment also measuring data for these languages recorded in a laboratory setting. This will help further tease out the specific language effects from the fieldwork setting effects. Finally, the sample also needs Indigenous languages from non-tropical settings, where other types of environmental disruptions might also be present.

References

Carter Anderson, Mien Nguyen, and Rolando Coto-Solano. 2025. Unsupervised, semi-supervised and llm-based morphological segmentation for bribri. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 63–76.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Proc. Interspeech 2023*.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Rolando Coto-Solano. 2021. Explicit tone transcription improves ASR performance in extremely low-resource languages: A case study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online. Association for Computational Linguistics.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467.

Rolando Coto-Solano and Sofía Flores-Solórzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de costa rica. *Kánina*, 40(4):175–199.

Rolando Coto-Solano, Tai Wan Kim, Alexander Jones, and Sharid Loáiciga. 2024. Multilingual Models for ASR in Chibchan Languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8521–8535, Mexico City, Mexico. Association for Computational Linguistics.

Rolando Coto-Solano, Daisy Li, Manoela Teleginski Ferraz, Olivia Sasse, Cha Krupka, Sharid Loáiciga, and Sally Akevai Tenamu Nicholas. 2025. Diacritic restoration for low-resource indigenous languages: Case study with bribri and cook islands māori. *arXiv preprint arXiv:2512.19630*.

Rolando Coto-Solano, Sharid Loáiciga, and Sofía Flores-Solórzano. 2021. Towards universal dependencies for bribri. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 16–29.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022a. Development of automatic

speech recognition for the documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882.

Rolando Coto-Solano, Sally Akevai Nicholas, Brittany Hoback, and Gregorio Tiburcio Cano. 2022b. Managing data workflows for untrained forced alignment: examples from Costa Rica, Mexico, the Cook Islands, and Vanuatu. *The Open Handbook of Linguistic Data Management*, 35.

Rolando Coto-Solano, Sally Akevai Nicholas, and Samantha Wray. 2018. Development of natural language processing tools for Cook Islands Māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 26–33.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

G. Durantin, B. Foley, N. Evans, and J. Wiles. 2017. Transcription survey. *Paper presented at the Australian Linguistic Society Annual Conference*.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022a. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023a. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, and 15 others. 2022b. Findings of the Second AmericasNLP Competition on Speech-to-Text Translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Abteen Ebrahimi, Arya D. McCarthy, Arturo Oncevay, John E. Ortega, Luis Chiruzzo, Gustavo Giménez-Lugo, Rolando Coto-Solano, and Katharina Kann. 2023b. Meeting the needs of low-resource languages: The value of automatic alignments via pretrained models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3912–3926, Dubrovnik, Croatia. Association for Computational Linguistics.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sofía Flores-Solórzano. 2010. Teclado Chibcha: Un software lingüístico para los sistemas de escritura de las lenguas bribri y cabécar. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, 36(2):155–161.

Sofía Flores-Solórzano. 2017. Corpus oral pandialectal de la lengua bribri. `http://bribri.net`.

Sofía Flores-Solórzano. 2017. *Un primer corpus pandialectal oral de la lengua bribri y su anotacion morfologica con base en el modelo de estados finitos*. Ph.D. thesis, Universidad Autónoma de Madrid.

Sofía Flores-Solórzano. 2019. La modelización de la morfología verbal bribri. *Procesamiento del Lenguaje Natural*, 62:85–92.

Sofía Flores-Solórzano and Rolando Coto-Solano. 2017. Comparison of two forced alignments systems for aligning bribri speech. *CLEI Electronic Journal*, 20(1):2–1.

Ben Foley, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan Van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, and 1 others. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *SLTU*, pages 205–209.

INEC. 2011. X Censo Nacional de Población y VI de Vivienda 2011 - Territorios Indígenas - Principales Indicadores Demográficos y Socioeconómicos.

Jesin James, Rolando Coto-Solano, Sally Akevai Nicholas, Joshua Zhu, Bovey Yu, Fuki Babasaki,

Jenny Tyler Wang, and Nicholas Derby. 2024. Development of community-oriented text-to-speech models for māori ʻavaiki nui (cook islands māori). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4820–4831.

Alex Jones, Rolando Coto-Solano, and Guillermo González Campos. 2023. TalaMT: Multilingual machine translation for Cabécar-Bribri-Spanish. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 106–117, Singapore. Association for Computational Linguistics.

Katharina Kann, Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, John E. Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo A. Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Elisabeth Mager, Vishrav Chaudhary, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, and Ngoc Thang Vu. 2022. AmericasNLI: Machine translation and natural language inference systems for Indigenous languages of the Americas. *Frontiers in Artificial Intelligence*, Volume 5 - 2022.

Sarah Karnes, Rolando Coto-Solano, and Sally Akevai Nicholas. 2023. Towards universal dependencies in cook islands māori. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 124–129.

Jessica Karson and Rolando Coto-Solano. 2024. Morphological Tagging in Bribri Using Universal Dependency Features. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 56–66.

Haakon Krohn. 2020. Elaboración de una base de datos en XML para un diccionario bribri–español español–bribri en la web. *Porto das Letras*, 6(3):38–58.

Haakon S. Krohn. 2021. Diccionario digital bilingüe bribri. http://www.haakonkrohn.com/bribri.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. 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*, pages 202–217, Online. Association for Computational Linguistics.

Ministry of Finance and Economic Management, Government of the Cook Islands. 2021. Census 2021: Key findings. https://www.mfem.gov.ck/statistics/census-and-surveys/census/267-census-2021.

Sally Akevai Nicholas. 2012. Te Vairanga Tuatua o te Te Reo Māori o te Pae Tonga: Cook Islands Māori (Southern dialects).

Sally Akevai Nicholas and Rolando Coto-Solano. 2019. Glottal variation, teacher training and language revitalization in the cook islands. In *Proceedings of the 19th International Congress of Phonetic Sciences, University of Melbourne, Australia*, pages 3602–3606.

Sally Akevai Te Namu Nicholas. 2018. Language contexts: Te Reo Māori o te Pae Tonga o te Kuki Airani also known as Southern Cook Islands Māori. *Language Documentation and Description*, 15:64.

Picovoice. 2024. Cobra: On-device voice activity detection engine. https://github.com/Picovoice/cobra.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation & Conservation*, 15.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 202:28492–28518.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624.

James Robert. 2011. Pydub. https://github.com/jiaaro/pydub.

Carlos Sánchez Avendaño. 2013. Lenguas en peligro en Costa Rica: vitalidad, documentación y descripción. *Revista Káñina*, 37(1):219–250.

J. Shi, J. D. Amith, R. Castillo García, E. G. Sierra, K. Duh, and S. Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yolóxochitl Mixtec. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, page 1134–1145.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Sofía Flores Solórzano. 2009. Los mamiferos en la clasificacion etnobiologica de la comunidad de amubre. *Estudios de Lingüística Chibcha*, 28:7–47.

John Wiseman. 2016. py-webrtcvad: Python interface to the webrtc voice activity detector. https://github.com/wiseman/py-webrtcvad.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).