Multilingual Models for ASR in Chibchan Languages

Rolando Coto-Solano*, Tai Wan Kim*, Alexander Jones* and Sharid Loáiciga†

* Dept. of Linguistics, Dartmouth College

† Dept. of Philosophy, Linguistics, and Theory of Science, University of Gothenburg

rolando.a.coto.solano@dartmouth.edu

{tai.wan.kim.21,alexander.g.jones.23}@dartmouth.edu

sharid.loaiciga@gu.se

Abstract

We present experiments on Automatic Speech Recognition (ASR) for Bribri and Cabécar, two languages from the Chibchan family. We fine-tune four ASR algorithms (Wav2Vec2, Whisper, MMS & WavLM) to create monolingual models, with the Wav2Vec2 model demonstrating the best performance. We then proceed to use Wav2Vec2 for (1) experiments on training joint and transfer learning models for both languages, and (2) an analysis of the errors, with a focus on the transcription of tone. Results show effective transfer learning for both Bribri and Cabécar, but especially for Bribri. A post-processing spell checking step further reduced character and word error rates. As for the errors, tone is where the Bribri models make the most errors, whereas the simpler tonal system of Cabécar is better transcribed by the model. Our work contributes to developing better ASR technology, an important tool that could facilitate transcription, one of the major bottlenecks in language documentation efforts. Our work also assesses how existing pre-trained models and algorithms perform for genuine extremely low resource-languages.

Resumen

Modelos multilingües para reconocimiento de voz en lenguas chibchas. En este artículo presentamos experimentos sobre reconocimiento de voz en bribri y cabécar, dos lenguas de la familia chibchense. Se refinaron modelos usando cuatro algoritmos: Wav2Vec2, Whisper, MMS y WavLM, usando datos monolingües para cada lengua. El mejor rendimiento se obtuvo con Wav2Vec2. A continuación se completaron (1) experimentos de entrenamiento conjunto y de aprendizaje por transferencia para ambos idiomas, y (2) un análisis de los errores en la transcripción, enfocado en los errores tonales. Los resultados muestran que el aprendizaje por transferencia es efectivo para ambos idiomas, pero que es mejor para el Bribri. Se usó una revisión ortográfica post-procesamiento para reducir aún más los errores por palabra y por caracter. Para el bribri los tonos son el rasgo fonológico en el que el modelo comete la mayor cantidad de errores. Esto contrasta con el cabécar, que tiene un sistema tonal más simple, y que el modelo maneja mejor. Nuestro trabajo contribuye al desarrollo del reconocimiento de voz, una herramienta que facilita la transcripción, uno de los impedimentos durante la documentación lingüística. Nuestro trabajo también describe el desempeño de los modelos pre-entrenados al trabajar con lenguas con recursos extremadamente bajos.

1 Introduction

This paper presents experiments on Automatic Speech Recognition (ASR) for Bribri and Cabécar, two languages belonging to the Chibchan family and spoken in Costa Rica. The data available for modeling is extremely limited, with only 143 minutes for Bribri and 54 minutes for Cabécar. In such a context, conventional Natural Language Processing (NLP) tools relying on deep neural networks trained from scratch are not an option, but leveraging pre-trained models offers a solution.

Our first steps will be to compare the performance of different algorithms when tackling these extremely under-resourced languages. We used four ASR algorithms to finetune their base model and thereby train monolingual models for Bribri and Cabécar. We then used the best performing algorithm, Wav2vec2, to train two additional types of models: i) a joint model simultaneously trained on both languages, and ii) a transfer learning model initially trained on one language and subsequently on the other. Then, we evaluated all models on each language independently. Next, we applied a unigram spell checker to correct the final output. Our results indicate that transfer learning is effective for both Bribri and Cabécar, with better performance in transferring from Cabécar to Bribri. In contrast, the joint models did not surpass the mono-

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
June 16-21, 2024

lingual or transfer models. The post-processing spell-checking contributed to reducing the Word Error Rate (WER) for Cabécar. Finally, we performed an error analysis of the monolingual and transfer models to test our intuition that linguistic tone presents more issues during transcription than other aspects of Chibchan phonology.

ASR is potentially a valuable tool for field linguists working in language documentation and revitalization, given that their primary data source comprises recorded interviews and elicitation sessions with L1 speakers. ASR is also promising for developing tools directly accessible to the members of these indigenous communities. As speech serves as a more immediate form of communication than writing, ASR-based tools could be used for capturing diverse communication interactions between speakers, encouraging the use of the language.

The challenges inherent in working with these languages are substantial, distinguishing it from conventional NLP tasks. The audio files can present less-than-ideal recording conditions, with background noise, and the speakers may be elderly community members, potentially with communication difficulties. The impact of tokenization tools remains unclear, especially considering that the writing systems may not be stable.

On the other hand, engaging with these linguistically diverse languages not only stress-tests our models but also encourages inclusivity in research. The complexities involved in addressing their unique linguistic features contribute to a broader understanding of the capabilities and limitations of ASR in diverse linguistic contexts.

1.1 ASR for Low-Resource Languages

Automatic Speech Recognition (ASR) has gained recognition as a valuable tool for field linguists working in language documentation. While the technology itself is not new (cf. Besacier et al., 2014, for a survey on early approaches), achieving good results requires training on large corpora. Many of the languages studied by field linguists, however, are low-resourced, with some– such as Bribri and Cabécar–considered acutely low-resourced (Jimerson and Prud'hommeaux, 2018). This categorization implies that these languages are spoken by only a handful of individuals and often lack a standardized writing system, which complicates the creation of written materials.

In this context, pre-trained models have emerged as a game changer, as they provide a robust acoustic model from their pre-training data from several high-resourced languages. Thus, numerous studies have concentrated on bootstrapping the available data from the low-resourced target language to enhance the text language model. Common techniques include self-training with data generated by the model under training itself or text-to-speech systems (Bartelds et al., 2023), and data augmentation with external sources such as dictionaries (Hjortnaes et al., 2020; Arkhangelskiy, 2021).

Interestingly, some studies have found that a good match of domains is more important than model size for good performance (Liu et al., 2023; Arkhangelskiy, 2021). Additionally, factors such as transliteration and standardization to common orthographies (e.g., changing French *-eur* to Spanish *-or* in a FR→ES system) (Khare et al., 2021) enhances system performance.

1.2 Transfer Learning

Transfer learning is a widely adopted technique to leverage knowledge from one low-resource language to another. In the context of Machine Translation (MT), it is common practice to pre-train on a pair of languages with high resources and subsequently to fine-tune on the pair with limited resources (Zoph et al., 2016; Kocmi and Bojar, 2018). The effectiveness of this approach depends on the linguistic similarities between the source and target languages. Notably, a greater overlap in vocabulary between the source languages tends to result in more significant gains, as evidenced by studies like Nguyen and Chiang (2017) or Dabre et al. (2017). However, such overlap is not a necessary condition to see improvements. Interestingly, when pre-training on ASR data and then fine-tuning on speech-to-text translation, even a non-related language helps, as the acoustic model plays a main role in the overall gain (Bansal et al., 2019).

1.3 Joint Learning

Another common method of transfer learning across languages is multilingual neural machine translation (NMT), i.e., translation systems that incorporate more than two languages. In contrast to the approach mentioned in §1.2 where a parent model undergoes training on a high-resource language pair and a child model is subsequently fine-tuned on a low-resourced pair, in multilingual NMT, multiple languages are fed simultaneously into the model (Dabre et al., 2020). A commonly used configuration involves complete parameter sharing among languages, where corpora are concatenated, and sequences are distinguished by a special language tag token. This approach works well in data-rich settings, even with a massive amount of languages (Bapna et al., 2022; Fernandes et al., 2023). In low-resource settings, additional considerations such as language similarity are crucial for success (Huang et al., 2023). A recent technique involves using adapters—additional feed-forward layers added to the transformer while keeping the transformer's parameters unchanged (Rebuffi et al., 2017; Houlsby et al., 2019). This technique has proven beneficial for low-resourced languages, as reported by Chronopoulou et al. (2023), who reports gains of up to 2 BLEU points for several languages. Ranathunga et al. (2023) provides an overview of NMT work for low-resourced languages.

Similar to NMT, research on multilingual ASR also finds that adding new languages presents challenges, especially when working with limited data, given that the base models are predominantly pretrained on English data (Srivastava et al., 2023). Strategies such as data augmentation with self-training (Srivastava et al., 2023) and transliterating the input into a common writing system have been found effective (Verma et al., 2023). Research on massively multilingual models by Tjandra et al. (2023) propose architectural modifications in the input and output embedding space. They find that the tokenization has an essential role in unifying the scripts across diverse languages and maintaining a relatively small vocabulary size, thereby directly influencing performance. They also report that grouping languages based on their similarity benefits performance.

1.4 Chibchan Languages

Here we will study two Indigenous languages from Costa Rica: Bribri and Cabécar. They belong to the Chibchan family, which has languages in Honduras, Costa Rica, Panama and Colombia. Bribri is spoken by approximately 7000 people, and Cabécar is spoken by approximately 11000 in Southern Costa Rica (INEC, 2011). Both languages are classified as vulnerable (Moseley, 2010; Sánchez Avendaño, 2013), which means that there are children in the community who are no longer learning the language. There is some NLP work for these languages. Notably, the AmericasNLI corpus (Mager et al., 2021) includes data from Bribri. There is also work on machine translation (Feldman and Coto-Solano, 2020; Ebrahimi et al., 2023a; Jones et al., 2023; Ebrahimi et al., 2023b, 2022b), forced alignment (Coto-Solano and Solórzano, 2016; Solórzano and Coto-Solano, 2017; Coto-Solano et al., 2022), speech recognition (Coto-Solano, 2021), dependency parsing (Coto-Solano et al., 2021), natural language understanding (Ebrahimi et al., 2022a; Kann et al., 2022) and analysis of word embeddings (Coto-Solano, 2022).

2 Experiments

In this section we will describe the data sources, the algorithms used for ASR training, and the methods for evaluating the experiments.

2.1 Data

The Bribri data includes 143 minutes of transcribed audio from 28 different speakers. This comes from an online oral corpus (Flores-Solórzano, 2017) and from the recordings included in the book *Sẽ́bliwak Francisco García ttṑ Las palabras de Francisco García* (Jara Murillo, 2022). These two sources use different orthographies. For example, in the first one, the word *mother* is written 'a̱mì̱', whereas in the second one it can be written 'ãmĩ̀' or even 'mì'. We standardized the transcriptions into a uniform orthographic representation, which can then be converted to either of the human orthographies in use. The Bribri transcriptions contain 20674 words in 2653 utterances (7.8 words per utterance), with a total of 3422 unique words. The utteraces are 3.2 ±2.1 seconds long on average. The recordings include data from three different dialects of Bribri: Amubri, Buenos Aires and Salitre.

The Cabécar data totals 54 transcribed minutes from 13 different speakers. The data comes from interviews included in a Cabécar dictionary (González Campos and Obando Martínez, 2020). They contain 8602 words in 594 utterances (14.4 words per utterance). The transcriptions contain 1206 unique words, and each of the utterances is 5.5 ±2.4 seconds long. The data contains only one dialect of Cabécar, the Chirripó dialect.

2.2 Monolingual Training

Our first task was to establish a baseline for the Chibchan multilingual models. We tested four algorithms: XLSR-53 Wav2Vec2 (Conneau et al., 2020), Whisper (Radford et al., 2022), MMS (Pratap et al., 2023), and WavLM (Chen et al., 2022). We fine-tuned to train separate models for each of the two Chibchan languages. (In the case of MMS, we also tested the inference from the pre-trained model for Cabécar). These models are not strictly monolingual, as the base multilingual models come pretrained with acoustic data from numerous other high-resource languages. But, in order to facilitate the comparison with the Chibchan-Transfer and the Chibchan-Joint models, we will refer to these as the *monolingual* models. For each of these we used the maximum of available data: 143 minutes for Bribri and 54 for Cabécar. We randomly shuffled the data several[1] times to make different train/dev/test sets, with ratios of 80%, 10% and 10%. With this we will obtain an average word and character error rate for each algorithm. We will compare the performance of these algorithms and use the best-performing one for the rest of the experiments. (The hyperparameters for each algorithm can be found in Appendix B).

In order to study the progression of training with different masses of data, we randomly selected audio files up to certain time durations, so that we could train for smaller datasets and observe the reduction in error rate. For Bribri, we made random sets of [5, 10, 15, 30, 45, 60, 90, 120, 154] minutes. For Cabécar, we made random sets of [5, 10, 15, 30, 45, 54] minutes. For each of these, we made ten randomly shuffled train/valid/test sets with the data available, and split that data into 80%, 10% and 10%. For each of these ten training runs, we selected the earliest model that had the lowest validation word error rate (WER) before overfitting. Then, we used that model to get the median of the character error rate (CER) and the WER of the test files. Finally, we averaged the value of those ten medians; these are the values presented in table 2 and figures 1 and 2.

2.3 Transfer and Joint Training

In order to carry out the transfer learning experiments, we first selected the best performing Bribri and Cabécar models as our foundation. We then used these models to continue fine-tuning on the other language. We used a similar evaluation procedure as the one described for the monolingual training: We trained five randomly shuffled sets for each of the time durations (e.g. 5 sets made up of 5 minutes of Bribri each, another 5 sets totaling 10 minutes of Bribri each, etc). We trained the model and extracted the CER and WER for the test set, and then averaged the values across the five runs.

The next experiment we conducted was to train a (Chibchan) multilingual joint model. All of the transcriptions for both languages were poooled together. From these, we randomly selected files until the set reached a total duration of [5, 10, 15, 30, 45, 60, 90, 120, 150, 197] minutes. We carried this process out three times for each time point. Each of those was split into 80%, 10% and 10% for the train/valid/test sets. The evaluation was then performed separately for each language, measuring the CER and WER for the Bribri and Cabécar test transcriptions

2.4 Post-training corrections

End-to-end models do not depend on statistical n-gram models to correct the orthographic output of the algorithm. This advantage is very desirable in high resource languages, as it allows the program to spell unknown words, while still learning the language's orthography. However, in low resource scenarios, the data might not suffice for the system to learn the orthography well (and in previous, statistical ASR, the system would be restricted to the words occurring in its n-gram language models).

As a simple way to overcome these issues, we applied a unigram spell checker to correct the output. We started with Norvig's (2021) statistical unigram spell check algorithm, and made one modification: If (i) the source sentence and the ASR hypothesis transcription have the same number of words, and (ii) the word in source$_i$ is not the same as the word in hyp$_i$, then we will assume that the word hyp$_i$ is a spelling mistake and it will be changed to a different, existing word in the corpus. This allows the system to preserve the words that are already correctly transcribed, while giving it a chance to improve its results by changing potential nonwords.

We applied the spell checking to the output of each of our experimental conditions: (i) the monolingual models for Bribri and Cabécar, (ii) the Bribri-to-Cabécar and Cabécar-to-Bribri transfer models, and (i) the joint models, evaluated on the Bribri and Cabécar test transcriptions. We then calculated the CER and WER between the source and the corrected sentences.

3 Results

In this section we will present four results: (i) the quantitative measurements of the character and

[1]For each language there were randomly shuffled 10 sets for Wav2Vec2, and 5 for the other algorithms.

word error rates for models trained on Bribri and Cabécar data, using different algorithms, (ii) the differences in performance between the monolingual, joint and transfer models, (iii) a qualitative analysis of the differences in the transcriptions from the different models, and (iv) an error analysis of different components of Bribri phonology and how those are transcribed by the best performing models.

3.1 Algorithm Comparisons

Table 1 shows the average results for the monolingual models, trained on all the available data (143 minutes for Bribri and 54 minutes for Cabécar). Each of the numbers is the average and standard deviation from the training runs for each language.

	Bribri		Cabécar	
	CER	WER	CER	WER
W2V2	**34** ±2	**76** ±2	**21** ±2	**48** ±4
Whisper	36 ±3	79 ±4	29 ±2	63 ±3
MMS-FT	NA	NA	59 ±2	83 ±3
MMS-Inf	NA	NA	49 ±1	95 ±1
WavLM	95±4	100±0	164±9	100±0

Table 1: Average and standard deviation of the median error for ASR algorithms: XLSR-53 Wav2Vec2 (W2V2), Whisper, MMS using Fine-Tuning on Cabécar (MMS-FT), MMS using only inference for Cabécar (MMS-Inf), and WavLM. These are trained on 143 minutes of Bribri and 54 minutes of Cabécar.

Wav2Vec2 was the best performing model for both languages. This difference in performance might be due to the fact that Wav2Vec2 has a more balanced dataset in the pretraining, which might allow it to observe a wider range of linguistic behavior in low resource languages. Whisper, for example, has a larger proportion of English compared to other languages, which might bias it against understanding linguistic characteristics distinct from those found in English. The other base models are probably pretrained on poor quality data (e.g. bibles), and therefore they might be a poor match to the way the language is written in the community.

Because Wav2Vec2 had the better performance, we used this algorithms to perform the training of the joint and transfer models.

3.2 Joint and Transfer Training Results

Figure 1 shows the error rates for Bribri models trained using different amounts of data, starting with only 5 minutes of data, and ending with all of the data available (143 minutes). As expected, results improve as the training data increases. For example, in the case of the monolingual data, 5 minutes of Bribri result in an average of CER=61 and WER=97, whereas the maximum amount of data reduces the average error to CER=34 and WER=76. The learning shows steady progress, but the rate of error reduction starts to slow down after 60 minutes of data, particularly for the CER. Figure 2 shows that this trend is also visible in the Cabécar monolingual models, but here the performance is overall better, despite having less data to train from. When using only 5 minutes, Cabécar has an average CER=39 and WER=74, which is later reduced to CER=21 and WER=48 when using all of the data (54 minutes). The Cabécar data shows a faster rate of error lowering. When both of them are measured at 45 minutes of training data, both Bribri and Cabécar have similar rates of character error reduction: ΔCER=18 for Cabécar ($\text{CER}_{Cab:45}$=21) and ΔCER=20 for Bribri ($\text{CER}_{Br:45}$=41). However, the Cabécar has a greater word error reduction: ΔWER=25 ($\text{WER}_{Cab:45}$=49) compared to only ΔWER=14 for Bribri ($\text{WER}_{Br:45}$=83).

The joint training models produced gains for Bribri and loses for Cabécar. When measured on the highest amount of data (54 minutes for Cabécar and 143 for Bribri), the transcriptions from the Bribri joint model had CER=33 and WER=73, which are improvements of ΔCER=1 and ΔWER=3 compared to their equivalent monolingual models. On the other hand, Cabécar had worse performance: ΔCER=-4 and ΔWER=-12, which means that the monolingual models were better. Table 2 summarizes these results.

Transfer learning showed improvements for Bribri and neutral results for Cabécar. This might be because of the differences in the datasets: Bribri has a more complex and less "clean" dataset, as it includes songs, interviews with more noise, and wider intralinguistic variation. Therefore, it might not contribute as much to the Cabécar data, whereas the smaller but cleaner Cabécar set might contribute more to the Bribri. The transfer from Cabécar to train on Bribri data showed improvements of ΔCER=2 (from $\text{CER}_{Br:Monolingual}$=34 to $\text{CER}_{Br:Transfer}$=32) and ΔWER=6 (from $\text{WER}_{Br:Monolingual}$=76 to $\text{WER}_{Br:Transfer}$=70). The best overall model for Bribri was this transfer model. The transfer into Cabécar, on the other hand, showed no improvement at all.

Spell checking also had complex and sometimes opposite effects: When it comes to the word error, it

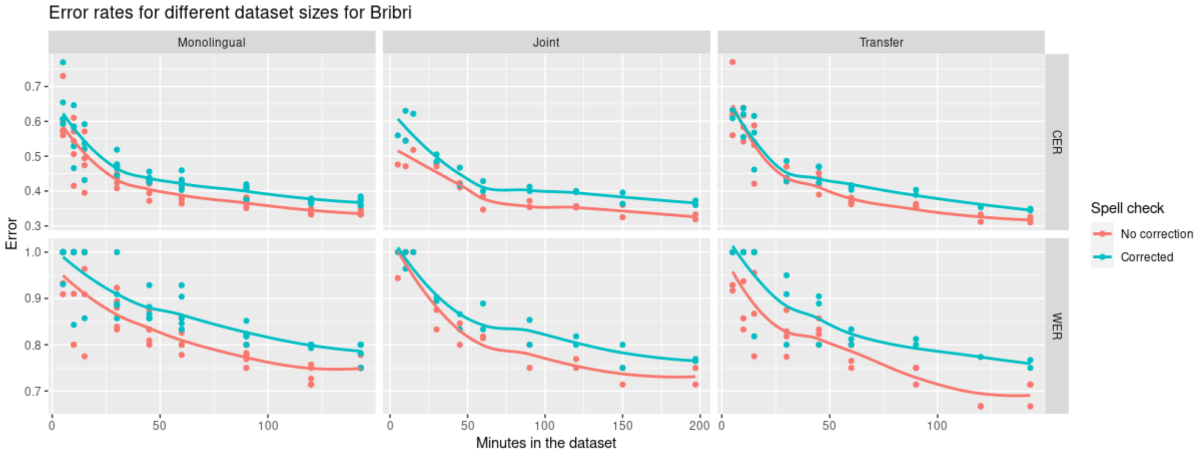

Figure 1: Average error rates for Bribri experiments

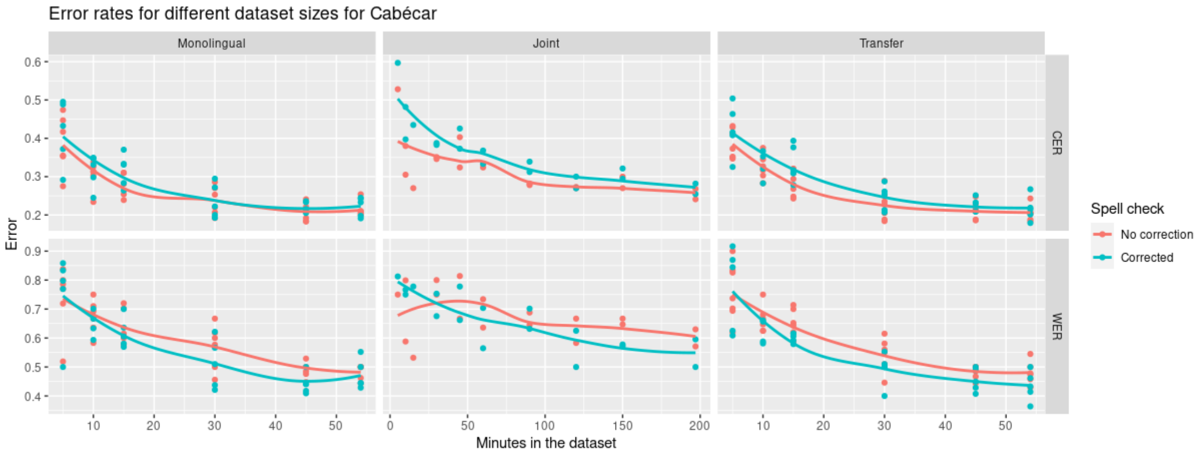

Figure 2: Average error rates for Cabécar experiments

aided the transcription of Cabécar, and it produced worse results in the transcription of Bribri. But when it comes to the character error rate, spell checking increased the error for both languages. As for Bribri, the CER became worse by ΔCER=-3∼4, and the WER became worse by ΔCER=-3∼-6. On the other hand, in the case of Cabécar, the CER also became worse (ΔCER=-1∼2), but the WER became better by ΔWER=1∼5. In fact, the best overall model for Cabécar was the model that combined transfer with the spell checking.

3.3 Qualitative results

Table 3 shows examples of transcriptions for Bribri, going from the better to the worst performing conditions. The spoken sentence “e’ ta̲ bua’ i wȍ ta̲’ mía̲ shalê̂lê̂ ta̲ sulû i mía̲" *[if you do it like that] you’ll get a good harvest; if it’s far apart it’ll go bad* is best transcribed by the transfer model, which has several words completely correct (e.g. “bua’" *good* and “mía̲" *to go*), and several words with transcriptions that are close enough to the actual word to be understandable (e.g. *tsulû instead of “sulû" *bad*, and *shaalè̀lế instead of shalê̂lê̂ *far apart*). The transfer model only produced one major mistake, fusing the words “wȍ ta̲’" *to have a harvest* into *ȍa̲. The transcription from the joint model is not as good but it is still better than the monolingual transcription. The joint model fuses some words together (*wȍkta̲ instead of “wȍ ta̲’" *to have a harvest*, but, if read out loud, it is still understandable. When examined at the character level, both the transfer and joint transcriptions show issues with the tone, confusing the low-rising tones in shalê̂lê̂ with high or falling tones.

The quality of transcription degrades for the other methods. The transcriptions are progressively more difficult to read, and the monolingual transcription has a number of nonwords which hinder comprehension. The corrected transcriptions are even worse, despite being made up of words that

		CER		WER	
	Type of model	Not corrected	Corrected	Not corrected	Corrected
Bribri	Monolingual	34	37	76	79
	Joint	33	37	73	77
	Transfer	**32**	35	**70**	76
Cabécar	Monolingual	**21**	22	48	47
	Joint	25	27	60	55
	Transfer	**21**	22	48	**43**

Table 2: Average error rates for Bribri and Cabécar models at the maximum number of train/valid/test items (Joint = 197 minutes, other Bribri models = 143 minutes; other Cabécar models = 54 minutes)

Condition	Sentence	CER	WER
Source	e’ ta̱ bua’ i wö̀ ta̱’ mía̱ shalë̂lë̂ ta̱ sulû i mía̱		
Translation	*you’ll get a good harvest; if it’s far apart it’ll go bad*		
Transfer	e’ ta̱ bua’ i ö̀a̱ mía̱ shaalë̀lë́ ta̱ tsulû mía̱	20	42
Joint	e’ ta̱ bua’ i wö̀kta̱ mía̱ shé klë́lë ta̱ tsulù i mí̱	20	50
Monolingual	e’ ta̱ bua’ i ö̀ a̱ mía̱ sh lë̂lè ta̱ tsulûû mìa̱	20	58
Transfer (corrected)	e ta̱ bua i ta̱ mía̱ chaléla ta̱ sulû mí	30	58
Joint (corrected)	e ta̱ bua i dö̀ka̱ mía̱ i shalë̀blë ta̱ sulû mi̱	34	67
Monolingual (corrected)	e ta̱ bua rè ta̱ mía̱ shka ë̀alè’ ta̱ sulû mí̱	34	75

Table 3: Example transcriptions for Bribri

Condition	Sentence	CER	WER
Source	bö́tkä i rä i ká cha ijé cha		
Translation	*These two are not his*		
Transfer (corrected)	bó̱tkë́i rä i ká cha ijé cha	11	25
Monolingual (corrected)	bótäwkë́ i rä i kicha ijë cha	22	37
Monolingual	buótkë́ i rä i ká chá̱ ijé chá̱	26	38
Joint (corrected)	bá kë́ jí rä ká cha ijé cha	30	50
Transfer	bóäkë́b kä́ i kácsha ijé cha	33	62
Joint	bá kë́ ñí rää́ iká cha ijé cha̱	33	75

Table 4: Example transcriptions for Cabécar

exist. This might be because the corrections were not done with a probabilistic language model, but with a unigram spell checker. An example of this problem is the monolingual corrected transcription, which can be roughly translated as *and then good fixed, go walk only maybe, bad from*. At this point, comprehensibility is severely compromised.

Table 4 shows examples of Cabécar transcriptions, where the conditions are again ordered from better to worse. All models showed some confusion at the beginning of the phrase, and often fused the words “bö́tkä i” *these two*. There is also some phone confusion, particularly with the central vowel ä. Here, as mentioned in the previous section, the corrected models achieve much better performance, and it is the non-corrected models that produce unintelligible transcriptions. For example, the joint model’s transcription can be roughly translated as *you no flat, no-that of, his no*. Interestingly, neither the Bribri nor the Cabécar joint models produced words in the wrong language. The Bribri outputs from the joint model only contain Bribri or Bribri-like words, and the Cabécar output only contains Cabécar words. This is a positive signal that the model could differentiate between the two languages during the learning process.

3.4 Error Analysis

Given that many languages in the Americas are tonal, an important and common question arises among linguists who might use ASR to accelerate documentation: Does ASR systems transcribe tones worse than other types of sounds? The Wav2Vec2 base model is pretrained on tonal languages such as Mandarin, Cantonese, Kinyarwanda, Lao, Vietnamese and Zulu, so a priori it should learn to distinguish tones adequately. However, does it show lower accuracy in transcribing tones relative to consonants or vowels? This section seeks to answer this question.

The first part of the experiment was to transform the Bribri and Cabécar transcriptions into strings that separate the different phonemes they contains. Table 5 has examples of such strings. In the "all" condition, we included the same transcriptions we used in the previous sections. In the "tone" condition, we inserted a character for each of the orthographic tonal diacritics in the languages. In Cabécar and Bribri those where H (high) and L (non-diacritic low). Bribri also has the orthographic tones F (falling), G (glottal low rise) and R (low rise)[2]. Therefore, the phrase *Ìs be' shkèna̠* "How are you?" would be transformed into `H G HL`.

We repeated this process with other aspects of Bribri phonology. We used a "consonant" representation where *Ìs be' shkèna̠* becomes `s b shkn`. The "vowel" representation includes the vocalic qualities of each of the vowels (`i e ea`), and the "nasal" representation contains information for whether the vowels in the text are oral or nasal (`O O ON`).

	Transcription
Translation	How are you?
All	`ìs be' shkèna̠`
Tone	`H G HL`
Consonant	`s b shkn`
Vowel	`i e ea`
Nasal	`O O ON`

Table 5: Different transcriptions of the Bribri phrase *How are you?*. These are used to analyze the different types of errors in the model.

We performed these transformations on both the gold standard transcriptions and the hypotheses produced by the transfer models. We used the models trained on the maximum data, 143 minutes for Bribri and 54 for Cabécar. (The results for the monolingual model are in Appendix A). We then calculated the word and character error rates for each of these subrepresentations. Here, the "word" error rate means that the total number of words where at least one tone was wrong, for example. The "character" error rate means a mistake in a single tonal marker, for example. The medians for these results are in table 6.

	Bribri		Cabécar	
Error	WER	CER	WER	CER
All	71	35	50	22
Tone	57	30	40	17
Consonant	54	29	38	16
Vowel	53	26	42	20
Nasal	48	19	37	13

Table 6: Medians for the error rates of the transfer models for different subrepresentations of the data.

First we will analyze the data for Bribri. A Kruskall-Wallis[3] test shows that there are significant differences between the different subrepresentations ($\chi^2(4)$=226, p<0.00001). A post-hoc analysis with a Benjamini Hochberg (BH) correction (Benjamini and Hochberg, 1995) revealed that Bribri tones have significantly more errors (WER=57, CER=30) than the consonants (WER=54, CER=29, p<0.005), the quality of vowels (WER=53, CER=26, p<0.005) or the oral/nasal contrasts (WER=48, CER=19, p<0.00001).

Let's now analyze the Cabécar data. A Kruskall-Wallis test shows that there are significant differences between the different phoneme types ($\chi^2(4)$=89, p<0.00001). A post-hoc analysis with a BH correction shows that Cabécar tones (WER=40, CER=17) have fewer errors than the vowel qualities (WER=42, CER=20, p<0.05). On the other hand, tones have more errors than the nasality contrasts (WER=37, CER=13, p<0.005), and there was no significant difference between the tones and the consonants (WER=38, CER=16, p=0.07).

This evidence suggests that tones are more difficult for the model to transcribe if the tonal system is complex enough. The Bribri tonal system is the more complex of the two. It has five orthographic indications for tone (high, falling, apostrophe low rising, glottal low rising, no-diacritic low) and the model makes more errors when transcribing these tonal marks than when transcribing consonants or vowels. On the other hand, the Cabécar tonal sys-

[2]Bribri has five phonological tones (Jara, 2018; Coto-Solano, 2015): high, falling, low rise, low and neutral. The low rise is expressed orthographically with a apostrophe (a glottal stop), or a circumflex diacritic. The low and neutral tones are both unmarked in the orthography, and they are both expressed by using just the vowel without any diacritics.

[3]The distributions do not meet the assumptions of normality, and therefore a non-parametric test was used.

tem is simpler. It only has two orthographic tones (high and no-diacritic low), and the model makes fewer errors when marking it, relative to the Cabécar vowels. It might also be the case that the Cabécar data is more internally consistent and the Bribri data has more variation (see section 4), and that this makes the Cabécar tones even easier to learn.

From these results we can infer that tones can be more difficult to transcribe than other aspects of a language's orthography, particularly if the recordings are in a more conversational style, and if the tonal system is relatively complex.

4 Discussion

Regarding the type of learning, transfer learning proved effective for both Bribri and Cabécar, with superior performance for Bribri. The joint models did not surpass the effectiveness of the monolingual ones, but at least for Bribri, they closely followed the performance of the transfer-based models. Regarding the post-processing, spell-checking only contributed to reducing the WER for Cabécar.

Contrary to findings in the multilingual NMT literature, where transfer approaches involving sequential training of languages tend to underperform compared to joint approaches where both languages are input simultaneously, our results present the opposite trend–the transfer approach yielded better results. The result that both transfer and joint models outperformed in Bribri compared to Cabécar strongly suggests that the "cleanliness" of the data is a very important and determinant factor. The Cabécar data, characterized by fewer genres and speakers from a single geographic region, provides a more consistent signal that may result in fewer structural errors. Despite having more Bribri data, the model seemingly grasps Cabécar better, enabling better generalization to the Bribri data.

In contrast to multilingual NMT techniques, our joint model does not use language tags when inputting both languages. Introducing language tags would require modifications to the audio files. We posit that such modifications could enhance the model's ability to distinguish between the languages more accurately, especially given the subtle "bumps" observed in the Cabécar plots which might be hinting a potential confusion. Additionally, this change might improve regularization by effectively increasing the number of training examples, aligning with the patterns observed in the multilingual NMT literature (Huang et al., 2023). These tags will be added in future work.

The correction post-processing was more effective for Cabécar, suggesting once again that variation in the Bribri data is high, and that the Cabécar data has greater consistency. The Cabécar data also has lower lexical diversity, a factor which might directly impact the spell-checker's performance.

Critically, these improvements concerned the WER, but the CER did not show much gain. One plausible explanation is that the acoustic model may already be saturated, meaning that it does not benefit as much from additional data. Another possibility is that the pre-trained model lacks sufficient typologically relevant knowledge for the Chibchan languages under examination. For instance, linguistic features like tones may be insufficiently represented. This potential deficiency serves as an example of how applying our methods in the context of languages divergent from the typical scope of languages in NLP papers allows us to assess the effectiveness of our tools.

There is potential future work in expanding this dataset to perform zero-shot transcriptions of other, even lower-resourced Chibchan languages such as Malecu and Ngäbere, and hopefully including them into a large pan-Chibchan multilingual model in the future. Such a model would enhance the transcription accuracy of the existing languages.

5 Conclusions

We presented experiments on multilingual ASR learning for Bribri and Cabécar, two extremely low-resource languages spoken in Costa Rica. We have shown that cross-lingual learning does take place between languages, specifically from Cabécar to Bribri. We have also shown that the quality of the data is highly significant in resource-constrained settings where each training data point holds considerable weight.

Our results are promising but we have not explored the full potential of this technique. For example, further data processing could involve the inclusion of language tags and their corresponding audio equivalents. Another possibility is is selecting a subset of the Bribri data that aligns more closely with the Cabécar data in terms of domain. Additionally, future research could consider zero-shot evaluation on languages within the same linguistic family.

Limitations

There are a number of limitations in this paper. First of all, the models attempt to be representative of the languages, but they include relatively few speakers, and in the case of Cabécar, only one of its spoken dialects. Care needs to be taken to include a larger diversity of speakers so that the model doesn't place implicit preference on one variety of Cabécar over others.

Another important limitation is the amount of computing power needed to train these models. The experiments presented here were exhaustive, but they were also time consuming: The XLSR-53 Wav2Vec2 experiments took 455 GPU hours, using an Nvidia Tesla K80 GPU in a HPC infrastructure. (This training was performed in parallel on 5∼7 GPUs, and it took approximately one week). The inference is relatively quick and can be performed on free cloud-based platforms, but it still requires a GPU to run. This could be prohibitive for communities who might want to implement this system in an offline environment.

Finally, the Bribri data is open, but the Cabécar data is only partially available, and will be released in the future. This might make it difficult for other teams to work on Cabécar ASR.

Ethics Statement

This work seeks to develop and test language technology for Indigenous languages of the Americas, and in doing so becomes part of the history of those languages and the people that speak them. The authors recognize that this history is marred by the violence and dispossession of colonialism, including the deliberate destruction of indigenous cultures and languages. We thus take it as our responsibility that our work does good, not harm, to the linguistic communities it concerns.

Acknowledgements

We want to thank Prof. Guillermo González Campos for his assistance with the Cabécar data. This research was funded in part by the Swedish Research Council (VR) grant (2014-39) for the Centre for Linguistic Theory and Studies in Probability (CLASP).

References

Timofey Arkhangelskiy. 2021. Low-resource ASR with an augmented language model. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 40–46, Syktyvkar, Russia (Online). Association for Computational Linguistics.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic Speech Recognition for Under-Resourced Languages: A Survey. *Speech Commun.*, 56:85–100.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020.

Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Rolando Coto-Solano. 2015. The Phonetics, Phonology and Phonotactics of the Bribri Language. In *2nd International Conference on Mesoamerican Linguistics. California State University, Los Angeles*.

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.

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, 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, Brittany Hoback, and Gregorio Tiburcio Cano. 2022. 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 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.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

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, et al. 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.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, et al. 2022b. Findings of the Second AmericasNLP Competition on Speech-to-Text Translation. In *NeurIPS 2022 Competition Track*, 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.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling Laws for Multilingual Neural Machine Translation. In *Proceedings of the 40 th International Conference on Machine Learning*, PMLR 202, Honolulu, Hawaii.

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

Guillermo González Campos and Freddy Obando Martínez. 2020. *Diccionario Escolar del Cabécar de Chirripó - Ditsá duchíwák kté chulíí i yuäklä*. Universidad de Costa Rica, Vicerrectoría de Acción Social, Sede del Atlántico.

Nils Hjortnaes, Timofey Arkhangelskiy, Niko Partanen, Michael Rießler, and Francis Tyers. 2020. Improving the language model for low-resource ASR with online text corpora. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 336–341, Marseille, France. European Language Resources association.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP . In *Proceedings of the International Conference on Ma-*

chine Learning, Proceedings of Machine Learning Research, page 2790–2799.

Kaiyu Huang, Peng Li, Jin Ma, Ting Yao, and Yang Liu. 2023. Knowledge transfer in incremental learning for multilingual neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15286–15304, Toronto, Canada. Association for Computational Linguistics.

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

Carla Victoria Jara. 2018. *Gramática de la lengua bribri*. E-Digital ED.

García Segura Alí Jara Murillo, Carla. 2022. Sëbliwak Francisco García ttö Las palabras de Francisco García. `https://www.lenguabribri.com/las-palabras-de-francisco`.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Katharina Kann, Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, John E Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo A Giménez-Lugo, et al. 2022. AmericasNLI: Machine translation and natural language inference systems for Indigenous languages of the Americas. *Frontiers in Artificial Intelligence*, 5:995667.

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Proceedings of Interspeech 2021*, pages 1529–1533.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Zoey Liu, Justin Spence, and Emily Prud'Hommeaux. 2023. Studying the impact of language model size for low-resource ASR. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 77–83, Remote. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez Lugo, Ricardo Ramos, et al. 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.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Peter Norvig. 2021. How to Write a Spelling Corrector. `https://norvig.com/spell-correct.html`.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural Machine Translation for Low-Resource Languages: A Survey. *ACM Comput. Surv.*, 55(11).

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *dvances in Neural Information Processing Systems*.

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

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.

Tejes Srivastava, Jiatong Shi, William Chen, and Shinji Watanabe. 2023. EFFUSE: Efficient Self-Supervised Feature Fusion for E2E ASR in Multilingual and Low Resource Scenarios.

Andros Tjandra, Nayan Singhal, David Zhang, Ozlem Kalinli, Abdelrahman Mohamed, Duc Le, and Michael L. Seltzer. 2023. Massively Multilingual ASR on 70 Languages: Tokenization, Architecture, and Generalization Capabilities. In *ICASSP 2023 -*

2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5.

Tushar Verma, Atul Shree, and Ashutosh Modi. 2023. ASR for Low Resource and Multilingual Noisy Code-Mixed Speech. In *Proc. INTERSPEECH 2023*, pages 3242–3246.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Error analysis for the monolingual models

We analyzed the errors of the monolingual models trained with Wav2Vec2. Table 7 shows a summary of the results. The most interesting result is that the transfer model did reduce the tonal error rate. For the monolingual Bribri data the median tonal error rate was WER=62 and CER=38. Compare this to the results for the Cab→Br transfer learning: WER=57 and CER=30. The transfer did not help the Br→Cab model. The tonal error for both the monolingual and transfer models was WER=40 and CER=17. This adds to the results that indicate that Bribri benefitted the most from the transfer of information from another Chibchan language.

	Bribri		Cabécar	
Error	WER	CER	WER	CER
All	76	38	50	22
Tone	62	32	40	17
Consonant	58	30	40	16
Vowel	58	29	43	20
Nasal	53	20	37	14

Table 7: Medians for the error rates of the monolingual models trained XLSR-53 Wav2Vec2.

The statistical patterns found for the transfer learning also appear here. A Kruskall-Wallis test shows that there are significant differences in the monolingual Bribri dataset (χ^2(4)=587, p<0.00001). A post-hoc analysis with a Benjamini Hochberg correction revealed that Bribri tones show have significantly more errors (WER=62, CER=32) than consonants (WER=58, CER=30, p<0.00001) the quality of vowels (WER=58, CER=29, p<0.00001) or the nasalization marks (WER=53, CER=20, p<0.00001).

As for the Cabécar data, a Kruskall-Wallis test shows that there are significant differences in the data (χ^2(4)=89, p<0.00001). A post-hoc analysis with a BH correction shows that Cabécar tones (WER=40, CER=17) have fewer errors than the vowel qualities (WER=43, CER=20, p<0.05). The tone has more errors than the vowel nasality markings (WER=37, CER=14, p<0.005), and there were no significant difference between the tones and the consonants (WER=40, CER=16, p=0.15).

These results have the same pattern as those for the transfer learning errors: The more complex Bribri tones appear to cause the most difficulties for the ASR models.

Figure 3 show the error rates for the tones, consonants, and vowel features for both Bribri and Cabécar, across different masses of training data, when trained using Wav2Vec2. As for Bribri tones, they are the most difficult linguistic feature to transcribe in every type of training, and with all but the smallest masses of data. (When the model only has 10 minutes of data, vowels have worse rates than tones for the monolingual and transfer learning models). As for the Cabécar models, vowels have consistently higher error rates for all conditions.

B Model hyperparameters

We use the following hyperparameters with the Wav2Vec2 v2 models trained in our experiments:

```
from transformers import Wav2Vec2ForCTC,
     Wav2Vec2Processor,
    Wav2Vec2FeatureExtractor,
    Wav2Vec2CTCTokenizer,
    TrainingArguments

tokenizer = Wav2Vec2CTCTokenizer(
    "./vocab.json",
    unk_token="[UNK]",
    pad_token="[PAD]",
    word_delimiter_token="|" )

feature_extractor =
    Wav2Vec2FeatureExtractor(
    feature_size=1,
    sampling_rate=16000,
    padding_value=0.0,
    do_normalize=True,
    return_attention_mask=True )

processor = Wav2Vec2Processor(
    feature_extractor=feature_extractor,
     tokenizer=tokenizer)

model = Wav2Vec2ForCTC.from_pretrained(
    "facebook/wav2vec2-large-xlsr-53",
    attention_dropout=0.1,
    hidden_dropout=0.1,
    feat_proj_dropout=0.0,
    mask_time_prob=0.05,
    layerdrop=0.1,
    ctc_loss_reduction="mean",
```

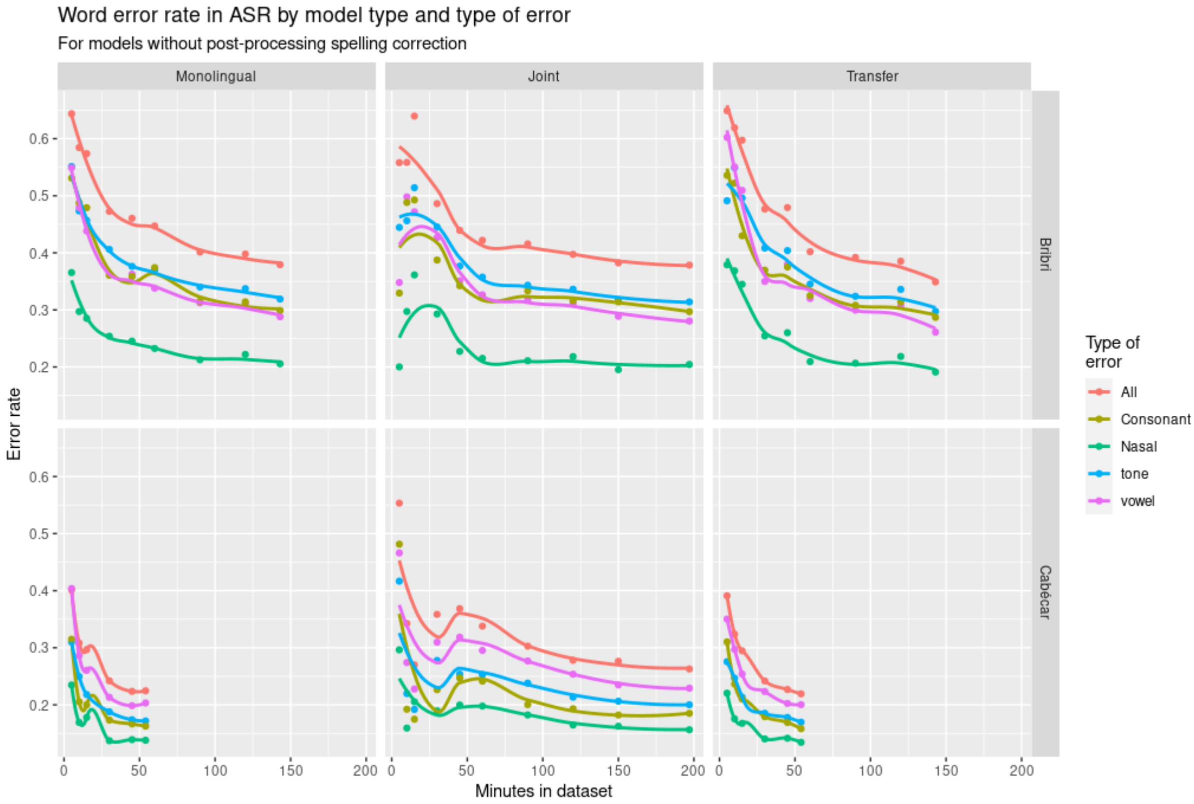

Figure 3: Average error rates for specific types of phonemes and linguistic features.

```
    pad_token_id=processor.tokenizer.
   pad_token_id,
    vocab_size=len(processor.tokenizer),
    ignore_mismatched_sizes=True )

training_args = TrainingArguments(
  output_dir=folderModelFiles,
  group_by_length=True,
  per_device_train_batch_size=2,
  gradient_accumulation_steps=1,
  evaluation_strategy="steps",
  num_train_epochs=30,
  fp16=True,
  save_steps=400,
  eval_steps=100,
  logging_steps=50,
  learning_rate=3e-4,
  warmup_steps=500,
  save_total_limit=38 )
```

For any parameter not explicitly listed, we use the default value.

These are the hyperparameters for Whisper:

```
modelType = "openai/whisper-large-v2"

training_args = Seq2SeqTrainingArguments
   (
    output_dir=folderModelFiles,
    per_device_train_batch_size=16,
    gradient_accumulation_steps=1,
    learning_rate=1e-5,
    warmup_steps=500,
    max_steps=4000,
    gradient_checkpointing=True,
    evaluation_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=True,
    generation_max_length=225,
    fp16=True,
    save_steps=400,
    eval_steps=100,
    logging_steps=25,
    report_to=["tensorboard"],
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False )
```

These are the hyperparameters for MMS:

```
from transformers import
   TrainingArguments, Trainer

    training_args = TrainingArguments(
            output_dir=output_path,
      group_by_length=True,
      per_device_train_batch_size=16,
      evaluation_strategy="steps",
      num_train_epochs=4,
      gradient_checkpointing=True,
      fp16=False,
      save_steps=200,
      eval_steps=100,
      logging_steps=100,
      learning_rate=1e-3,
      warmup_steps=100,
      save_total_limit=2 )

    trainer = Trainer(
```

```
        model=model,
        data_collator=data_collator,
        args=training_args,
        compute_metrics=compute_metrics,
        train_dataset=data_train,
        eval_dataset=data_test,
        tokenizer=processor.feature_extractor )

    from safetensors.torch import save_file as safe_save_file
    from transformers.models.wav2vec2.modeling_wav2vec2 import WAV2VEC2_ADAPTER_SAFE_FILE

    adapter_file = WAV2VEC2_ADAPTER_SAFE_FILE.format(target_lang)
    adapter_file = os.path.join(training_args.output_dir, adapter_file)

    safe_save_file(model._get_adapters(), adapter_file, metadata={"format": "pt"})
```

These are the hyperparameters for WavLM:

```
processor = AutoProcessor.from_pretrained("patrickvonplaten/wavlm-libri-clean-100h-base-plus")
model = WavLMModel.from_pretrained("patrickvonplaten/wavlm-libri-clean-100h-base-plus")

model = AutoModelForCTC.from_pretrained(
    "patrickvonplaten/wavlm-libri-clean-100h-base-plus",
    ctc_loss_reduction="mean",
    pad_token_id=processor.tokenizer.pad_token_id )

training_args=TrainingArguments(
    per_device_train_batch_size=4,
    gradient_accumulation_steps=2,
    learning_rate=1e-5,
    warmup_steps=500,
    max_steps=1800,
    gradient_checkpointing=True,
    fp16=True,
    group_by_length=True,
    evaluation_strategy="steps",
    per_device_eval_batch_size=4,
    save_steps=600,
    eval_steps=100,
    logging_steps=25,
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False )
```

The technical specifications for the computer running the Wav2Vec2 models are available in the Limitations section. The Whisper and WavLM models were run on the same computers. Each of the WavLM models took approximately 40 minutes to train, for a total of seven hours. Each of the Whisper models took approximately 5 hours to train, for a total of 50 hours. The MMS models were run on a AMD Ryzen Threadripper PRO 5975WX 32-Cores, using 2 NVIDIA GeForce RTX 3090 Ti GPUs and 256GB of ECC DDR4 memory. Each run with the adapter took approximately 5 minutes, for a total of 50 minutes.