Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3872–3882
Marseille, 20-25 June 2022

Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori

Rolando Coto-Solano[1], Sally Akevai Nicholas[2], Samiha Datta[1], Victoria Quint[1], Piripi Wills[3], Emma Ngakuravaru Powell[4], Liam Koka'ua[3], Syed Tanveer[1], Isaac Feldman[1]

[1]Dartmouth College, New Hampshire, United States
[2]Massey University Te Kunenga ki Pūrehuroa, Auckland, Aotearoa New Zealand
[3]Kōrero Rororuia, Auckland, Aotearoa New Zealand
[4]University of Otago Te Whare Wānanga o Ōtākou, Dunedin, Aotearoa New Zealand
rolando.a.coto.solano@dartmouth.edu, s.nicholas@massey.ac.nz
emma.powell@otago.ac.nz, {willstutor,liamkokaua}@gmail.com
{samiha.datta.23,victoria.r.quint.22,syed.h.tanveer.21,isaac.c.feldman.23}@dartmouth.edu

Abstract

This paper describes the process of data processing and training of an automatic speech recognition (ASR) system for Cook Islands Māori (CIM), an Indigenous language spoken by approximately 22,000 people in the South Pacific. We transcribed four hours of speech from adults and elderly speakers of the language and prepared two experiments. First, we trained three ASR systems: one statistical, Kaldi; and two based on Deep Learning, DeepSpeech and XLSR-Wav2Vec2. Wav2Vec2 tied with Kaldi for lowest character error rate (CER=6$\pm$1) and was slightly behind in word error rate (WER=23$\pm$2 versus WER=18$\pm$2 for Kaldi). This provides evidence that Deep Learning ASR systems are reaching the performance of statistical methods on small datasets, and that they can work effectively with extremely low-resource Indigenous languages like CIM. In the second experiment we used Wav2Vec2 to train models with held-out speakers. While the performance decreased (CER=15$\pm$7, WER=46$\pm$16), the system still showed considerable learning. We intend to use ASR to accelerate the documentation of CIM, using newly transcribed texts to improve the ASR and also generate teaching and language revitalization materials. The trained model is available under a license based on the Kaitiakitanga License, which provides for non-commercial use while retaining control of the model by the Indigenous community.

Keywords: Automatic speech recognition, Cook Islands Māori, language documentation, low-resource languages

1. Introduction

Automatic Speech Recognition technology could dramatically accelerate language documentation of Indigenous languages and aid researchers and community members in the time-consuming task of manual transcription. In this paper we describe the creation of automatic speech recognition (henceforth ASR) for Cook Islands Māori, an Indigenous Polynesian language from the Realm of New Zealand[1] Section 1.1 describes the challenges of training ASR for extremely low-resource Indigenous languages, while section 1.2 reviews previous work in natural language processing for Cook Islands Māori. Section 2 presents the methodology for data processing and ASR training using statistical and Deep Learning techniques, and section 3 will describe the training results. Finally, section 4 will describe the future use of this ASR system, as well as licensing to ensure that the models remain under the control of Indigenous communities and contribute to the development of NLP for other Polynesian languages.

[1]The Realm of New Zealand includes the Cook Islands, Tokelau and Niue, as well as "New Zealand proper", or Aotearoa as it is known in Te Reo Māori, the Māori of New Zealand. We will refer to the territory of "New Zealand proper" as "Aotearoa New Zealand".

1.1. ASR for Language Documentation

The transcription of spoken audio recordings is a major bottleneck in language documentation. There is an urgent need to accelerate the process of transcription, so that communities who speak Indigenous and minoritized languages can tap into existing recordings of stories, traditions and genealogies to create language learning materials and thereby to contribute to the continued use of their languages. However, the process of transcribing these languages is substantially more time-consuming than it is with widely spoken languages like English. Transcribing an hour of a recording in an Indigenous language can take up to 50 hours of an expert's time (Shi et al., 2021).

Why is this process so slow? There are a series of challenges unique to the transcription of Indigenous languages. First, while languages like English have a large pool of transcribers who are experienced in writing the language, smaller languages are usually written only by a reduced number of specialists, typically limited to linguists and school teachers. This places severe limits on the availability of transcribers and also raises the cost of generating these transcriptions, given the higher level of expertise required.

Second, while English has an agreed-upon writing sys-

tem, many Indigenous languages might have multiple or no fixed orthographies. In practice this means that a language could be transcribed in numerous divergent ways depending on who is transcribing. These differences in orthographic representation can lead to profound schisms between groups and might even interact with pre-existing imbalances of power, where one group of speakers from a larger and/or wealthier background might wish to impose their writing style on others (Hinton, 2014). These orthographic conflicts add a layer of consideration that would not be necessary when working on NLP with a larger language.

Third, while English has millions of hours of audio from every imaginable genre, recorded in the voices of millions of different speakers, most languages only have a handful of recordings available, and making new ones is an expensive and complex endeavor. This means that there will always be a cap on how much data is available to be transcribed in a particular language.

Finally, while there is much English data that is readily usable by anyone, this should not be the case with data from Indigenous languages, where both the topics and the people in the recordings might need "restricted" access because of diverse restrictions that a community feels should be imposed on a recording or the knowledge it contains (Kukutai and Taylor, 2016).

Because of the reasons mentioned above, generating the data needed for training ASR of Indigenous languages is substantially more complex that for languages with large resources. Despite these difficulties, there is an increasing body of research on how to adapt ASR to work effectively in Indigenous languages (Besacier et al., 2014; Jimerson and Prud'hommeaux, 2018; Adams et al., 2019; Foley et al., 2018; Gupta and Boulianne, 2020b; Gupta and Boulianne, 2020a; Zahrer et al., 2020; Thai et al., 2019; Partanen et al., 2020; Zevallos et al., 2019; Matsuura et al., 2020; Hjortnaes et al., 2020; Levow et al., 2021). Most systems try to transcribe the language into a vernacular orthography, but there are also efforts to generate transcriptions in the International Phonetic Alphabet (Michaud et al., 2019; Li et al., 2020), which would be beneficial for languages without writing systems. Researchers have found that ASR does accelerate the transcription pipeline (Prud'hommeaux et al., 2021). It can also be a significant benefit for the community by providing an opportunity for the transcribers to practice and ultimately connect with their language, as well as providing younger speakers with technological tools that might help encourage their participation in language work (Lillehaugen, 2016; Aguilar Gil, 2014).

1.2. Previous NLP work in Cook Islands Māori

Cook Islands Māori (Glottolog `raro1241`, henceforth CIM) is an East Polynesian language spoken in the South Pacific. It has 14,000 speakers in the Cook Islands and 7,000 in Aotearoa New Zealand (Nicholas, 2018)[2] as well as a small number of additional speakers elsewhere in the world. It is a vulnerable language (Moseley, 2010), but its vitality varies across the archipelago. In the central island of Rarotonga, which has the main airport and more contact with English-speaking countries, there are fewer children who speak CIM, and education is delivered mostly in English. On the other hand, on the smaller islands, also called the Pā 'Enua, CIM has considerably more vitality. It is still widely spoken by children, and it is commonly used in school and daily life (Nicholas, 2018).

CIM has had a writing system and a wide range of written materials since the mid 19th century (Nicholas, 2018), but a great deal of pre-processing is required before it can used for NLP, such as digitization, optical character recognition, and orthographic conformation. Furthermore, due to the shift towards English in the community over the last fifty years, little new material is produced in CIM, which is also not widely used in printed, audiovisual or social media. This means that there are limited existing CIM resources ready to train NLP tools. The largest resource is *Te Vairanga Tuatua* (Nicholas, Sally Akevai, 2012), a collection of audio recordings geared towards linguistic research, archived in the Paradisec repository (Barwick and Thieberger, 2012). It contains recordings from elders across the Cook Islands and Aotearoa New Zealand and is the basis of the most comprehensive grammar of the language (Nicholas, 2017). Approximately 25% of the recordings have been transcribed and this data was used to create the first NLP tool for CIM, a part-of-speech tagger (Coto-Solano et al., 2018). It has also been used for linguistic research using untrained forced alignment (Nicholas and Coto-Solano, 2019; Coto-Solano et al., 2022). A 309,301 word corpus that combines this data with other written sources (Simiona, 1979; Tanga et al., 1984; Taraare, 2000; YouVersion, 2014) was used to develop a predictive keyboard for mobile phones (Quint and Oh, 2021), a tool that could be immediately practical to community members. Figure 1 shows a screen capture of this application.

The transcribed sub corpus of *Te Vairanga Tuatua* was used to conduct the first ASR experiments with CIM. Approximately 60 minutes of transcribed audio served as training data for an ASR model using Kaldi (Foley et al., 2018); it obtained a word error rate (WER) of 64. After this initial attempt, a second experiment was conducted with the assistance of Caleb Moses and Dragonfly Data Science in Wellington, New Zealand. This experiment used transfer learning in DeepSpeech (Hannun et al., 2014). It took the same 60 minutes of existing CIM transcriptions and trained with the support of a large model for Te Reo Māori, which was itself trained on 300 hours of crowdsourced data from 1300 speakers (Te Reo Irirangi o Te Hiku o Te Ika, 2017). This

[2]There are 17,000 inhabitants in the Cook Islands (MFEM, 2016) and approximately 81,000 people of Cook Islands ethnicity in Aotearoa New Zealand (StatsNZ, 2018).

Figure 1: Predictive Keyboard for Cook Islands Māori (Quint and Oh, 2021)

model achieved a median character error rate (CER) of 31. Both of these results, encouraging but still far away from a practical documentation tool, indicated the need to continue the transcription of CIM data in order to improve ASR performance.

2. Methodology

In this section we will describe the preparation process of our data, which included manual transcription, input normalization and exclusion of data points with code-switching. We will then describe the technical aspects of our ASR experiments.

2.1. Data preparation

The data used for ASR training comprises glossed audio example sentences from the CIM grammar (Nicholas, 2017), transcribed recordings from *Te Vairanga Tuatua*, as well as transcribed audio collected from 2017 to 2020 (Nicholas, Sally Akevai, 2012). These recordings contain traditional stories, genealogies, and other narrations from adult and elderly speakers (aged 30 to 75) of the language. This is precisely the type of language we wish to transcribe because it is extremely valuable for language documentation and revitalization[3]. Because the rate of transcription was so slow, we initiated a second type of data collection: speakers reading traditional stories. These recordings were significantly faster to transcribe as the transcriber could use the source text as a guide and did not need to spend as much time deciphering the audio.[4]

Transcription of this type of data is a painstaking task. It was carried out by Sally Akevai Nicholas, Piripi Wills, Liam Koko'ua and Emma Powell over a period of 3 years. The transcribers are either native speakers or advanced learners of the language and belong to the Cook Islands community. The orthography chosen for the transcription was the one described in Nicholas (2013) and Nicholas (2017). Because the recordings were transcribed for use in linguistic description, the transcriptions were made using ELAN (Sloetjes and Wittenburg, 2008), and they included punctuation as well as a number of discursive and pragmatic tags (e.g. <VERSE> and <LAUGH>). These tags were removed to provide a transcription for training that only contained the phones of the language.

There were a few issues to deal with regarding the orthography. Because the dataset was transcribed by a single team, there was little need to normalize the orthography in the texts. However, the glyph for the glottal stop was represented in a number of different ways when stored electronically, and it therefore needed to be normalized throughout the input. The glottal stop is ideally represented by the *saltillo* glyph ('), but it was also found in the input as an apostrophe, a typographic apostrophe, an HTML apostrophe (&quot) or a Hawaiian *'okina*. The solution adopted here was to convert all the glottal stop representations into a temporary 'q', so that the word *no'o* 'stay' was represented as `noqo` in the ASR input. A similar approach was adopted for the vowel length. CIM vowels can be long or short, and a long vowel is represented with a macron (e.g. *pā* [pa:] 'door'). In order to avoid Unicode issues when using older ASR training algorithms (see section 2.3 below), the vowel length was transcribed using the letter 'x' next to the long vowel. In this case, the word *pā* 'door' would be transcribed `pax` in the input given to the ASR. Both of these modifications (the glottal stop 'q' and the vowel length 'x') were later corrected in the ASR output.

For this experiment we excluded data with English words in it. This could constitute a single word in English (e.g. *Ko Chris ē tōna pupu* 'Chris and her group') or an entire phrase. There was relatively little English-CIM code-switching in the dataset (which was, after all, mostly made up of adult and elderly speakers retelling traditional stories), but leaving code-switching out is a major weakness of the system in the long run. This is because the Cook Islands are a multilingual territory where most of the population speaks both English and CIM. This multilingual tendency is even more pronounced in young speakers and in Aotearoa New Zealand so, while our system might have good performance with elders, it risks leaving out the speech of

[3]The demographics of the CIM speaking population makes a crowdsourcing campaign challenging. Most of our data would come from community members in Aotearoa New Zealand who have much better internet access than people in the Cook Islands. They are likely to be younger and learners rather than highly proficient speakers. This data would be valuable and it would increase our training set, but it would be different from the elders whose transcription we need to prioritize.

[4]These recordings also served a valuable language revitalization purpose in "re-oralizing" stories that had largely ceased to be passed down orally.

younger Cook Islanders. As the corpus grows, we intend to gather more data with code-switching into English for a more realistic representation of the linguistic circumstances of the archipelago.

The transcription effort yielded a total of 5033 sentences (36,390 words) recorded in 237 minutes of audio (3 hrs 57 mins). The recordings had a median duration of 2.3 seconds, with a minimum of 0.3 seconds and a maximum of 15.1 seconds. The transcriptions had a median size of 6 words (29 chars), with a minimum of one word (1 character) and a maximum of 42 words (193 chars). Figure 2 shows the distribution for word and character length of the transcriptions, as well as the duration in seconds of the recordings. There are a total of 10 speakers in the corpus who come from 4 different islands: Rarotonga, Penrhyn, Ma'uke and 'Atiu.

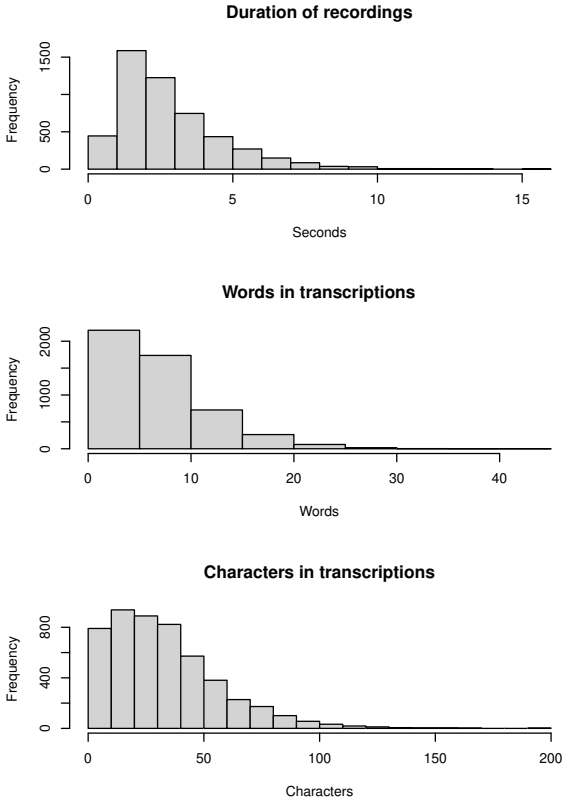

Figure 2: Length of CIM recordings and transcriptions

2.2. First Experiment: ASR Training

For the first experiment we include all of the speakers in all of the three sets (training, testing and validation). This helps us measure a system that could be used to transcribe more data from the same speakers, who have provided numerous recordings which are still untranscribed.

We chose three algorithms to train our data. The first was a statistical learning algorithm, based on Hidden Markov Models (HMM) with Gaussian Mixture Models (GMM), instantiated in Kaldi (Povey et al., 2011). While this approach might be older than Deep Learning methods, statistical learning can make better use of smaller datasets and can therefore provide an advantage with the relatively small amount of data available for CIM training. We used a trigram language model and the default hyperparameters for training (triphone transitions, 2000 HMM leaves, 10,000 Gaussian components and 35 training iterations)[5].

For the second approach we chose a Connectionist Temporal Classification (CTC) algorithm (Graves et al., 2006), implemented using two Deep Learning based systems. The first one was DeepSpeech (Hannun et al., 2014), which has already been used for Te Reo Māori from Aotearoa New Zealand (Moses et al., 2020; Mahelona, 2020). DeepSpeech uses bidirectional recurrent neural networks (BiRNN) to transform the audio signal into a sequence of glyphs, and then CTC to transform this sequence into a potential transcription in the chosen orthography. DeepSpeech was trained using a KenLM (Heafield, 2011) trigram language model, with training, validation and testing batch sizes of 200/60/60. The optimal model was chosen after 40 training epochs; the optimum was usually reached after 25-35 epochs.

For the third approach we chose XLSR (Conneau et al., 2020), which is derived from the Wav2Vec2 architecture (Baevski et al., 2020). For simplification purposes, we will refer to the XLSR-Wav2Vec2 combination as *Wav2Vec2*. This algorithm uses Convolutional Neural Networks to encode the audio (Mohamed et al., 2019) and then transforms it into quantized audio embeddings using information from 54 languages. These embeddings are then used by a transformer (Vaswani et al., 2017) to decode the audio signal into a transcription. This family of algorithms has proven effective in training from reduced datasets; e.g. producing acceptable results after training on only 10 minutes of English input. Therefore, we predict it would be able to learn effectively from our small dataset. Wav2Vec2 was trained using its default settings: Using Transformers 4.4.0, with a learning rate of $3x10^{-4}$, batch size 16, attention and hidden dropout 0.1, layer drop 0.1, an intermediate layer size of 4096, a hidden layer size of 1024 (24 hidden layers), 16 attention heads, and a total of 2400 steps. Beyond this point, the system began to overfit.

For each of the three systems (Kaldi, DeepSpeech and Wav2Vec2), we took the 5033 sentences from all 10 speakers and randomly shuffled the sentences to create 20 train/validation/test sets for each of the systems. We split the sentences into training, validation and testing sets with 80%, 10% and 10% of the sentences (4027, 503 and 503 sentences respectively). For each of the twenty rounds we extracted the median word error rate (WER) and median character error rate (CER) of the

[5] A Kaldi monophone model was also trained, but its results, CER=17 and WER=33, were substantially inferior to those of the triphone model.

test set. We then calculated an average and a standard deviation from those twenty medians; those are the numbers reported below.

Training for Kaldi took approximately 40 minutes for each of the 20 runs, for a total of 13.3 hrs. It was carried out on a personal laptop computer using one Intel i7 CPU. Training for DeepSpeech took approximately 65 minutes per model (total of 21.7 hrs) using the HPC infrastructure at Dartmouth College, which provided 16 parallel CPUs per run. Finally, each round of Wav2Vec2 training took approximately four hours to complete (total of approximately 80 hrs) using Google Colab with one NVIDIA Tesla P100 GPU.

2.3. Second Experiment: Held-Out Speakers

The first experiment, described above, allows us to study how the system will perform when transcribing new audio from the same speakers. In the second experiment we study the system's generalization capabilities, specifically how it would behave when it encounters a completely new speaker.

In order to study this we made six partitions of the data where the training+validation and testing sets contained different speakers. This way, the ASR would train on one subset of speakers, and would perform its testing on a different set of speakers that it hadn't encountered before. For example, for the first two partitions, we collected speakers whose files would add up to approximately 10% of the files in the dataset. We then excluded them from the training+validation sets, constructing those by randomly shuffling sentences from all of the other speakers. Partition #1, for example, restricts the speakers {A,K,R,T2} to the test set (10% of the files); we then took the rest of the speakers {B,J,T1,T3,M1,M2} and randomly shuffled their files into a training set (80%) and a validation set (10%). This made these partitions the same in size as the train-valid-test model in the first experiment.

There were four speakers whose files encompassed more than 10% of the dataset. When this was the case, these speakers were placed in the test set, even if this made the test set larger and the training set smaller. For example, in partition 5, speaker J has 27% of the files. In this case the test set only contained the files from speaker J, and the train+val sets were random shuffles of the files from the other nine speakers. Because only 73% of the total files are now left for training and validation, the training set contains 90% of those files (65%), and the validation set contains 10% of those files (8%). Table 3 in section 3.2 contains detailed information on the speakers and sizes of each partition.

We trained five times, using five random shuffles per partition. We only performed the training using the XLSR-Wav2Vec2 system, using the same hyperparameters and stopping conditions as in the first experiment. Each round of training took approximately four hours (for total of approximately 30 hrs) using Google Colab with one NVIDIA Tesla P100 GPU.

Splitting the data this way ensures two things. First, that the system had no access to data from held-out speakers during the training phase. Second, that there are words in the test set that the language model hasn't seen before. This ensures that the model is facing new words and that it is not simply functioning as a forced aligner. For example, the first random shuffle of partition 6 contained a total of 522 unique words: 361 of them were present in both the testing and the train+val sets, but 161 were found only in the testing set. This ensures that the model accounts for words that were unseen in the training.

3. Results

3.1. Results of ASR Training

Figure 3 shows the results of the first experiment. It shows the word and character error rates for the three ASR systems trained on the entirety of the data, where all of the speakers were present in the train+val+test sets. The first noticeable pattern is that DeepSpeech had much lower performance than the other systems. The twenty DeepSpeech models had a mean median of CER=22, and a mean median of WER=41. In the case of WER, this error is approximately almost double than that for the other systems. This is not a surprise given that DeepSpeech belongs to a family of Deep Learning algorithms that need a large mass of data to train correctly (Goodfellow et al., 2016; Glasmachers, 2017). The group working on Te Reo Māori has reached approximately WER=10 using DeepSpeech (Mahelona, 2020), but this might be because of their large mass of data (between 300 and 400 hours).

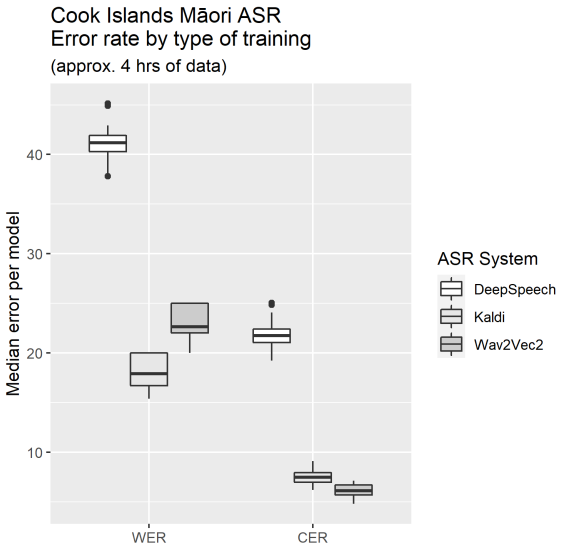

Figure 3: Median WER and CER for Cook Islands ASR by type of training

Table 1 shows the mean and standard deviation for the WER and CER in the three systems. These numbers are

the average of the medians from each of the 20 models run for each system. Kaldi still achieves a slightly higher performance in transcribing complete words, as measured by the WER (mean: 18±2, median: 18), but Wav2Vec2 follows relatively closely, (mean: 23±2, median: 23). Moreover, Wav2Vec2 is virtually tied with Kaldi when the error is measured by the number of correctly transcribed characters: Kaldi has a CER of 7.5±0.8 (median: 7.5), and Wav2Vec2 has the slightly lower CER of 6.1±0.7 (median: 6.1). This means that, out of every 100 characters in the ASR output, approximately 93 are transcribed correctly by both systems.

	WER	CER
Kaldi	**17.9 ± 1.7**	7.5 ± 0.8
DeepSpeech	41.1 ± 2.0	21.9 ± 1.6
Wav2Vec2	22.9 ± 2.0	**6.1 ± 0.6**

Table 1: Average medians for word and character error rate for CIM trained with three ASR systems

In order to examine the performance of these systems more closely, table 2 shows example transcriptions from each of them. Wav2Vec2 is generally more accurate and could provide better results within a language documentation pipeline. The analysis of accuracy for specific phones remains part of our future work, but an initial examination of the data shows that most of the errors in Wav2Vec involve the modification of vowels (e.g. ***mo**to**k**a*/*m**ou**t**a**k**ā* 'car', *k**e**t**u*/*k**i**t* 'dig', *'**o**ki*/*'**a**ki* 'stay'), the insertion or deletion of a glottal stop (e.g. *ka iroiro* / *kā'iro'i roa* 'will be mixed up'), or the lengthening of a short vowel (e.g. *motok**a***/*moutok**ā*** 'car', k**a** / k**ā**'iro'i 'will be mixed up'). Kaldi has many of these errors, but it also tends to replace entire words, probably influenced by its heavy reliance on its language model (e.g. *ki* 'to' instead of *i* 'in'; *'oki* 'stay' instead of *au* 'I'; *ki* 'to' instead of *ketu* 'dig').

Despite these issues, the transcriptions produced by Kaldi and Wav2Vec2 could be useful as a "first pass" that would then be corrected by a human expert. This promises to be faster than the current completely manual transcriptions, thereby providing a virtuous circle that improves the ASR and increases the amount of transcribed materials available for the generation of linguistic and language learning materials.

3.2. Second Experiment: Held-Out Speakers

Table 3 shows the results of the second experiment, where speakers were systematically held out of the training and validation sets with the purpose of testing how the model would transcribe new speakers. These held-out models had higher error rates than those which saw all the speakers during training. When averaged across the six partitions, the median CER for held-out speakers was 14.9±7.2, and the median WER was 46.4±15.6.

The results for each partition show a large amount of variation depending on which speaker is held out. For example, partition 4 (speaker B held-out) had a performance of CER=6 and WER=25, similar to the best performing models in the first experiment. On the other hand, partition 6 (speaker T1 held out) had a much higher error rate, with CER=23 and WER=66. This was also the case with partition 3 (speaker M1), with median CER=25 and WER=65. These two speakers (M1 and T1) are the only representatives from their islands, which might explain the lower performance of the models. Moreover, speaker T1 has a faster speech rate (2.5 words per second) than the average for the rest of the speakers (2.1±0.4 words per second), which could also account for the difficulties of the model.

The results of the second experiment show that, despite reduced performance, the system is still able to perform transcriptions that could help accelerate language documentation. For three of the partitions the CER rates remained below 12, indicating that, roughly, only one in every nine letters of the transcription would be incorrect. Even averaging across partitions, the median CER was 15, which roughly corresponds to one in every seven letters being mistranscribed. This rate is still helpful and could represent a major contribution to the difficult task of transcribing Cook Island Māori even when the system faces a speaker it has never trained for before.

4. Discussion

The results show evidence that Deep Learning solutions can perform similarly to statistical learning when it comes to the very small datasets involved in Indigenous language speech recognition. Wav2Vec2, which uses Transformers and significant pretraining, obtained results similar to those of Kaldi. This is relevant because the initial tests of Wav2Vec2 (Baevski et al., 2020) used simulated low-resource conditions, using truncated corpora of English instead of actually low-resource languages. Our results show that systems like XLSR-Wav2Vec2 could also deliver promising results for extremely low-resource languages like Cook Islands Māori, even when the system is attempting to transcribe new speakers. The fact that a Deep Learning solution worked is also relevant because, although the training process of Wav2Vec2 takes substantially longer time and could potentially emit more CO_2 (Strubell et al., 2019), these Deep Learning solutions are also much easier to train and maintain, which could allow for their wider use by more communities.

The results also show that ASR could be successfully applied to a language documentation pipeline for CIM. Getting to 4 hours of transcription took a substantial amount of effort, but the ASR will become part of a virtuous cycle, where new recordings are transcribed automatically, giving them a "first pass" that could reduce the work of the transcribers. This could make the transcription process quicker, and thereby provide more data to continue training and improving the model. Prud'hommeaux et al. (2021) have con-

English	*One day I was just sitting in my car*		
Target	i tēta'i rā tē no'o 'ua ara au i roto i tōku motoka	WER	CER
Kaldi	ki tēta'i rā tē no'o 'ua ara 'oki i roto i tōku motoka	15	9
DeepSpeech	i tēta'i a te no'o ara i roto i tōku motoka	31	18
Wav2Vec2	i tēta'i rā tē no'o 'ua ara au i roto i tōku moutakā	8	5

English	*I was sure that it was the pig who had rooted (it up)*		
Target	kua kite ra 'oki au ē nā te puaka i ketu	WER	CER
Kaldi	kua kite rā 'oki au e nā te puaka i ketu	18	5
DeepSpeech	kite rāi koe i nā te puaka i ki	55	38
Wav2Vec2	kua kite rā 'aki au ē nā te puaka i kit	27	10

English	*Absolutely, it will get mixed up*		
Target	āe 'oki ka iroiro atu	WER	CER
Kaldi	'aere ka'iro i roa atu	80	50
DeepSpeech	āe ki ka'iro 'oki roa te	100	50
Wav2Vec2	āe 'oki kā'iro'i roa atu	40	23

Table 2: CIM ASR Output Examples for three ASR systems

Partition	Train-Validation-Test Splits (#files and %)	WER	CER	Test speaker(s)	% total files	% total time
1	4036 - 504 - 493	32.9 ± 0.9	8.4 ± 0.2	A	3.7	3.4
	80% - 10% - 10%			K	3.6	4.5
				T2	2	4.5
				R	0.5	1.0
2	4007 - 500 - 526	40.1 ± 1.9	11.0 ± 0.5	T3	6.9	7.6
	80% - 10% - 10%			M2	3.4	7.2
3	3849 - 481 - 703	64.5 ± 3.1	24.5 ± 1.0	M1	14.0	8.0
	76% - 10% - 14%					
4	3769 - 419 - 845	25.0 ± 0.0	5.9 ± 0.3	B	17.0	18.5
	75% - 8% - 17%					
5	3268 - 408 - 1357	50.0 ± 0.0	16.4 ± 0.5	J	30	27
	65% - 8% - 27%					
6	3532 - 392 - 1109	65.9 ± 1.9	23.0 ± 0.2	T1	22	15
	70% - 8% - 22%					
Average		46.4 ± 15.6	14.9 ± 7.2			

Table 3: Average medians for CER and WER with speakers held out of the training and validation sets

ducted experiments designing precisely such a pipeline for the Seneca Onödowá'ga: language and found that ASR significantly accelerates transcription[6]. Figure 4 shows a potential workflow for this virtuous cycle. Our next step is to implement an easy interface so that the transcription can be carried out by linguists and community members without having to interact with the Kaldi or Wav2Vec2 code. Such an interface could, for example, provide an option to upload an audio file, or for the user to record themselves, and then conduct the ASR processing in the background.

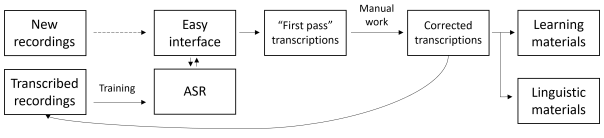

Figure 4: Example of cycle for ASR in language documentation

There is much future work that is still needed to improve these models. The first change will be to incorporate a language model to constrain the output and decrease the word error rate of Wav2Vec2, bringing it closer to the Kaldi results. A second necessary change is to increase the length of transcribed recordings so we

[6]Prud'hommeaux et al. (2021) also found an interesting result: while the transcription itself was faster, some of the transcribers preferred to do the work completely by hand because it let them get closer to the data, which made them feel a stronger connection to the language. This is highly relevant because we seek to attract youth to language through technology, and keeping this connection is an important part of the continuity of our NLP work.

can measure the performance of the data when handling longer stretches of speech. As shown in figure 2 above, half of the recordings used for this paper contained only between one and six words, so the system needs more training on longer stretches of speech. A third necessary step is to include data from the islands that are not in the present corpus (e.g. Aitutaki, Mitiaro, Mangaia, Manihiki, Rakahanga), as well as testing the performance of the system in the closely related but distinct language of Pukapuka (Glottolog `puka1242`).

Another future experiment will examine if the transcription of vowel length has an effect on error rates. As described in section 2.1, the input of the system was manipulated so that the vowel length was represented with a glyph different from the vowel glyph. For example, the orthographic form *pā* was written `pax` in the input, with the 'x' representing the lengthening of the vowel. There isn't consensus on best practices for representing suprasegmental features such as vowel length in ASR input for extremely low-resource languages. Research in languages with nasal vowels such as Portuguese and Hindi has indicated that nasality is likely best represented in the same glyph as the vowel because nasality alters vocalic quality (Meinedo et al., 2003; Jyothi and Hasegawa-Johnson, 2015). As for tone, there are contradictory results, but results from Bribri (Coto-Solano, 2021) indicate that representing the tone separate from its vowel leads to lower ASR error. This might be because tonal trajectories don't have as strong of an effect on the quality of the vowel. Given this, what would be the best representation for vowel length in CIM ASR input? Is the input `pax` optimal, where the vowel and the length are represented separately, or would it be better to represent this word with only two glyphs, where the second one has the information for both the vowel and the length (e.g. `pā`)? Coto-Solano (2021) provides evidence that differences in transcription can lead to differences in WER, even using Deep Learning, so, given the paucity of data for these languages, every bit of advantage could be useful.

A final future experiment should explore applying these ASR techniques to different Indigenous languages. The encouraging results from experiment 1 (WER=23, CER=6) might be influenced by the fact that CIM has a relatively small phonemic inventory (9-12 consonants, 5 short vowels and 5 long vowels) and relatively simple (C)V phonotactics. Moreover, CIM has few morphological inflections, so the system has relatively less phonemic, phonotactic and morphological variation to learn. This methodology needs to be tested on a typologically wider range of languages to confirm its general applicability.

One important aspect of this work has to do with data sovereignty. The audio is available freely (Nicholas, Sally Akevai, 2012), and the trained model is available for use by third parties (`https://github.com/Akevai/CIM-ASR-Models`). However, we need to safeguard the data sovereignty of the Indigenous community that the data belongs to. The members of the Cook Islands community should manage the data, so the model was deposited in a GitHub account belonging to one of the Indigenous researchers in the team (Nicholas). Similar efforts have been made to ensure that members of the Cook Islands community retain control of the audio files, the transcriptions and the trained models. In the same vein, we seek ways to bring the fruits of this work to language teachers and other members of the CIM speaking community so that they might find new and creative ways to use it. Likewise, we seek collaborations with other Pacific communities, so that these models can be used to accelerate the development of speech recognition tools for additional Polynesian languages, in particular by using these models for transfer learning. Because of this, we are using a localization of the Kaitiakitanga License (`https://github.com/TeHikuMedia/Kaitiakitanga-License`), which allows for non-commercial use of these models while retaining permission for their use within the Indigenous community. If you wish to use these models, please contact the authors for information on how to do so.

5. Conclusions

In this paper we have described the process of transcription, data preparation, and training of automatic speech recognition for an Indigenous language of Polynesia, Cook Islands Māori. The best performing systems, trained using XLSR-Wav2Vec2 and Kaldi, can transcribe short utterances of CIM with a character error rate of CER=6, and a word error rate of between WER=18 and WER=23. Even when speakers are held out, the character error rate can oscillate between CER=6 and CER=25, which is potentially adequate to accelerate the transcription of new recordings. The paper also provides evidence that Deep Learning can work in truly low-resource environments and with minority/Indigenous languages. We will continue the work to expand this dataset to include a wider geographical and age coverage of the Cook Islands population, but this work will be made faster and easier by incorporating ASR into the documentation pipeline.

6. Acknowledgments

This work would not be possible without the ongoing support of the many CIM speakers who have given so much of their time and knowledge to this project. Of particular note amongst this group are Jean Te Kura Mason, our most prolific contributor and literary expert, along with the students of the first cohort of the Diploma in Cook Islands Māori at the University of the South Pacific. We also wish to thank Prof. Tyler Peterson and Prof. Samantha Wray for their support at multiple stages of this project. Finally, this project is supported by the Marsden Fund Council from the Aotearoa New Zealand Government (21-MAU-018), managed by The Royal Society Te Apārangi.

7. Bibliographical References

Adams, O., Wiesner, M., Watanabe, S., and Yarowsky, D. (2019). Massively Multilingual Adversarial Speech Recognition. *arXiv preprint arXiv:1904.02210*.

Aguilar Gil, Y. E. (2014). ¿Para qué publicar libros en lenguas indígenas si nadie los lee? E'px.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Barwick, L. and Thieberger, N. (2012). Keeping records of language diversity in Melanesia: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). University of Hawai'i Press.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic Speech Recognition for Under-resourced Languages: A Survey. *Speech communication*, 56:85–100.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Coto-Solano, R., Nicholas, S. A., and Wray, S. (2018). Development of Natural Language Processing Tools for Cook Islands Māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 26–33.

Coto-Solano, R., Nicholas, S. A., Hoback, B., and Tiburcio Cano, G. (2022). Managing Data Workflows for Untrained Forced Alignment: Examples from Costa Rica, Mexico, the Cook Islands, and Vanuatu. In Andrea Berez-Kroeker, et al., editors, *The Open Handbook of Linguistic Data Management*, vol. 35. MIT Press, Cambridge.

Coto-Solano, R. (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.

Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., Ellison, T. M., van Esch, D., Heath, S., Kratochvil, F., Maxwell-Smith, Z., Nash, D., Olsson, O., Richards, M., San, N., Stoakes, H., Thieberger, N., and Wiles, J. (2018). Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 205–209.

Glasmachers, T. (2017). Limits of End-to-end Learning. In *Asian Conference on Machine Learning*, pages 17–32. PMLR.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Gupta, V. and Boulianne, G. (2020a). Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2521–2527.

Gupta, V. and Boulianne, G. (2020b). Speech Transcription Challenges for Resource Constrained Indigenous Language Cree. 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 362–367.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep Speech: Scaling up End-to-End Speech Recognition. *arXiv preprint arXiv:1412.5567*.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Hinton, L. (2014). Orthography wars. In Michael Cahill et al., editors, *Developing Orthographies for Unwritten Languages*, page 139–168. SIL International, Dallas, TX.

Hjortnaes, N., Partanen, N., Rießler, M., and Tyers, F. M. (2020). Towards a speech recognizer for komi, an endangered and low-resource uralic language. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 31–37.

Jimerson, R. and Prud'hommeaux, E. (2018). ASR for Documenting Acutely Under-resourced Indigenous Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jyothi, P. and Hasegawa-Johnson, M. (2015). Improved Hindi broadcast ASR by adapting the language model and pronunciation model using a priori syntactic and morphophonemic knowledge. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kukutai, T. and Taylor, J. (2016). *Indigenous Data Sovereignty: Toward an Agenda*, volume 38. Anu Press.

Levow, G.-A., Ahn, E. P., and Bender, E. M. (2021). Developing a Shared Task for Speech Processing on Endangered Languages. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 96–106.

Li, X., Dalmia, S., Li, J., Lee, M., Littell, P., Yao, J., Anastasopoulos, A., Mortensen, D. R., Neubig,

G., Black, A. W., et al. (2020). Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Lillehaugen, B. D. (2016). Why write in a language that (almost) no one can read? Twitter and the development of written literature. *Langauge Documentation and Conservation*.

Mahelona, K. (2020). Te Reo Māori Speech Recognition: A story of community, trust, and soverignty. Paper presented at Natives in Tech 2020. https://papareo.nz/#matauranga.

Matsuura, K., Ueno, S., Mimura, M., Sakai, S., and Kawahara, T. (2020). Speech corpus of Ainu Folklore and End-to-End Speech Recognition for Ainu language. *arXiv preprint arXiv:2002.06675*.

Meinedo, H., Caseiro, D., Neto, J., and Trancoso, I. (2003). AUDIMUS. media: a Broadcast News speech recognition system for the European Portuguese language. In *International Workshop on Computational Processing of the Portuguese Language*, pages 9–17. Springer.

MFEM. (2016). Census of Population Dwellings 2016 Results. Cook Islands Ministry of Finance and Economic Management. http://www.mfem.gov.ck/statistics/census-and-surveys/census/142-census-2016.

Michaud, A., Adams, O., Cox, C., and Guillaume, S. (2019). Phonetic lessons from automatic phonemic transcription: preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data. In *ICPhS XIX (19th International Congress of Phonetic Sciences)*.

Mohamed, A., Okhonko, D., and Zettlemoyer, L. (2019). Transformers with Convolutional Context for ASR. *arXiv preprint arXiv:1904.11660*.

Moseley, C. (2010). Atlas of the world's languages in danger (3rd edn. ed.). *Online: http://www. unesco. org/languages-atlas*.

Moses, C., Thompson, M., Mahelona, K., and Jones, P. L. (2020). Scoring Pronunciation Accuracy via Close Introspection of a Speech Recognition Recurrent Neural Network. Paper presented at NeurIPS 2020.

Nicholas, S. A. and Coto-Solano, R. (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.

Nicholas, S. A. (2013). Orthographic reform in Cook Islands Māori: Human considerations and language revitalisation implications. Paper presented at the 3rd International Conference on Language Documentation and Conservation (ICLDC), Honolulu, Hawai'i.

Nicholas, S. A. (2017). Ko te Karāma o te Reo Māori o te Pae Tonga o Te Kuki Airani: A Grammar of Southern Cook Islands Māori.

Nicholas, S. A. (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*, 15:36–64.

Partanen, N., Hämäläinen, M., and Klooster, T. (2020). Speech Recognition for Endangered and Extinct Samoyedic languages. *arXiv preprint arXiv:2012.05331*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Prud'hommeaux, E., Jimerson, R., Hatcher, R., and Michelson, K. (2021). Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation & Conservation*, 15:491–513.

Quint, V. and Oh, Y. (2021). Cook Islands Māori Keyboard. https://keyman.com/keyboards/cim.

Shi, J., Amith, J., García, R. C., Sierra, E. G., Duh, K., and Watanabe, S. (2021). Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yoloxóchitl Mixtec. *arXiv preprint arXiv:2101.10877*.

Simiona, T. (1979). *'E au tua ta'ito nō te Kuki Airani*. University of the South Pacific, Suva, Fiji.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.

StatsNZ. (2018). Cook Islands Maori ethnic group. Stats NZ Tatauranga Aotearoa. https://www.stats.govt.nz/tools/2018-census-ethnic-group-summaries/cook-islands-maori.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for Deep Learning in NLP. *arXiv preprint arXiv:1906.02243*.

Tanga, T., Kura, P., Manu, N., Kakepare, M., Kapao, K., Tanga, T., Teiuto, U., Koronui, V. M., Ariki, R. M. K., Mana, N. T., Cameron, T., Koronui, N., Tanga, M., George, T., Rau, T., Mariri, T., and Bob, N. K. (1984). *Atiu nui maruarua: 'E au tua ta'ito*. Institute of Pacific Studies of the University of the South Pacific, Suva, Fiji.

Taraare, T. (2000). *History and Traditions of Rarotonga*. The Polynesian Society, Auckland N.Z. Richard Walter and Rangi Moekaa (eds).

Te Reo Irirangi o Te Hiku o Te Ika. (2017). kōreromāori.io. https://koreromaori.io/.

Thai, B., Jimerson, R., Arcoraci, D., Prud'hommeaux, E., and Ptucha, R. (2019). Synthetic data augmentation for improving low-resource ASR. In *2019*

IEEE Western New York Image and Signal Processing Workshop (WNYISPW), pages 1–9. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

YouVersion. (2014). Cook Islands Māori Revised New Testament: Digital Publication. Bible Society of South Pacific. [Accessed 2021-01-10].

Zahrer, A., Zgank, A., and Schuppler, B. (2020). Towards Building an Automatic Transcription System for Language Documentation: Experiences from Muyu. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2893–2900.

Zevallos, R., Cordova, J., and Camacho, L. (2019). Automatic Speech Recognition of Quechua Language Using HMM Toolkit. In *Annual International Symposium on Information Management and Big Data*, pages 61–68. Springer.

8. Language Resource References

Nicholas, Sally Akevai. (2012). *Te Vairanga Tuatua o te Te Reo Māori o te Pae Tonga: Cook Islands Māori (Southern dialects)*. Paradisec, Collection SN1 at catalog.paradisec.org.au. DOI: 10.4225/72/56E9793466307.