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 2. Massey University Te Kunenga Ki Pūrehuroa
3. Kōrero Rororuia 4. University of Otago Te Whare Wānanga o Ōtakou
Language Resources and Evaluation Conference LREC 2022

Summary

(1) ASR for Language Documentation

(2) Cook Islands Māori

(3) Data collection and normalization

(4) Two experiments
    (i) Training with all speakers
    (ii) Training with held-out speakers

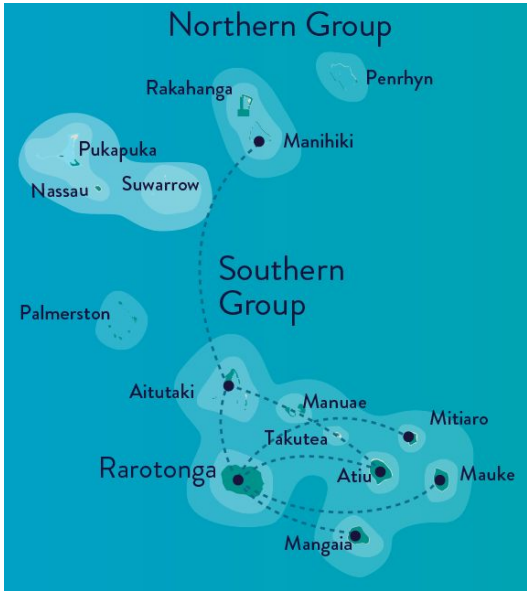
(5) Licenses and documentation workflow

ASR for Language Documentation

Transcribing text in an Indigenous language is an extremely specialized and expensive process: It can take up to 50hrs to transcribe one hour (Shi et al. 2021).

Using ASR would accelerate transcription (Prud'hommeaux et al. 2021) and help in the creation of linguistic and educational materials.

Cook Islands Māori



East Polynesian

Whangaunga tāta ki
te Reo Māori o Aotearoa

Endangered

Indigenous to the
Realm of New Zealand

14K speakers (+8K in NZ)

Data Collection and Normalization

Data source:
Te Vairanga Tuatua

Large
Linguistically rich
Under annotated
Transcription bottleneck

Previous work on CIM NLP

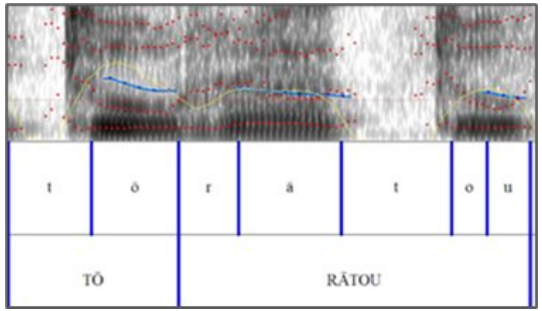
POS Tagging

92% accuracy (Random forest)
http://cimpos.appspot.com/pos.jsp

ka tanu a Tere i te taro
tam v det nprop acc det n

Untrained Forced Alignment

8% error for center of words



First ASR Experiment

Transfer model from Dragonfly and Te Hiku Media's DeepSpeech, with 1 hour CIM audio. 31% CER

And then the pig got tangled
ē e tāpeka'ia te puaka
e tāpaka te pu'ka

Data Collection and Normalization

Recordings transcribed by Ake, Piripi Wills, Emma Powell and Liam Koka'ua

Two main challenges:

-Transforming transcriptions for ASR algorithms

-Accounting for variation (e.g. code-switching)

Speaker 1 Māori Transcription	Ana Andrew	29.126	31.067	1.941	I runga i te 'enua ko Tupuaki,
Speaker 1 Māori Transcription	Ana Andrew	31.635	32.731	1.096	i te tuātau ta'ito,
Speaker 1 Māori Transcription	Ana Andrew	33.202	37.468	4.266	tē no'o ra tēta'i māpū māro'iro'i, ko Rū tōna ingoa.
Speaker 1 Māori Transcription	Ana Andrew	38.356	39.477	1.121	Kāre ia i te ariki,
Speaker 1 Māori Transcription	Ana Andrew	39.932	42.371	2.439	ē kāre katoa aia i te tamaiti nā te ariki,
Speaker 1 Māori Transcription	Ana Andrew	42.617	43.383	0.766	inārā,

Data Collection and Normalization

5033 files: 237 minutes (~4 hrs)
Median: 6 words (29 chars), 2.3 seconds

10 speakers, 4 islands (Rarotonga, Tongareva, Ma'uke, 'Atiu)

```
Speaker 1 Māori Transcription   Ana Andrew     29.126  31.067  1.941   I runga i te 'enua ko Tupuaki,
Speaker 1 Māori Transcription   Ana Andrew     31.635  32.731  1.096   i te tuātau ta'ito,
Speaker 1 Māori Transcription   Ana Andrew     33.202  37.468  4.266   tē no'o ra tēta'i māpū māro'iro'i, ko Rū tōna ingoa.
Speaker 1 Māori Transcription   Ana Andrew     38.356  39.477  1.121   Kāre ia i te ariki,
Speaker 1 Māori Transcription   Ana Andrew     39.932  42.371  2.439   ē kāre katoa aia i te tamaiti nā te ariki,
Speaker 1 Māori Transcription   Ana Andrew     42.617  43.383  0.766   inārā,
```

Experiment 1: ASR with all speakers

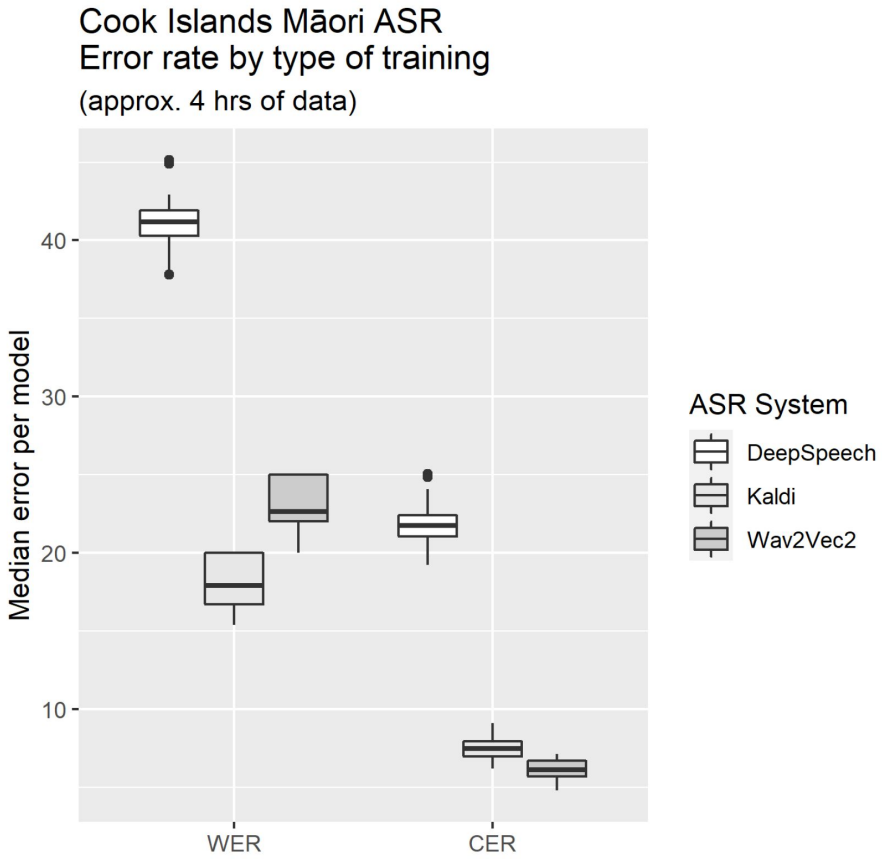
Training with three ASR systems:

- Kaldi (Povey et al. 2011)
- DeepSpeech (Hannun et al. 2014)
- Wav2Vec2 / XLSR (Conneau et al. 2020, Baevski et al. 2020)

Training/validation/test splits:
80% (4027 files), 10% (503 files), 10% (503 files)

Random permutations of all the data for all the speakers
x20 runs

Experiment 1: ASR with all speakers



	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**

Experiment 1: ASR with all speakers

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

Experiment 2: Held-Out Speakers

We split the sets to ensure that there were speakers who were never seen during training/validation.

<u>Partition 1</u>

Training and validation sets:	T1, T3, M1, M2, B, J	(90% of files)
Test set:	A, K, T2, R	(10% of files)

<u>Partition 5</u>

Training and validation sets:	A, B, J, K, T2, R, T3, M1, M2, T1	(70% of files)
Test set:	J	(30% of files)

Random permutations within the train/val sets.
x5 times

Experiment 2: Held-Out Speakers

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

Experiment 2: Held-Out Speakers

Partition 5

Meaning:	*From morning till night.*
Target:	mei te pōpongi mai e pō
Inference:	mei te pupongi mai ēpo
	(CER=16, WER=50)

Partition 1

Meaning:	*When we die we die, when we live we live.*
Target:	mē mate tātou kua mate mē ora kua ora
Inference:	mē mati tātou kua mate me ora kua ra
	(CER=8, WER=33)

Analysis

Deep Learning systems can indeed work with extremely low-resource languages.

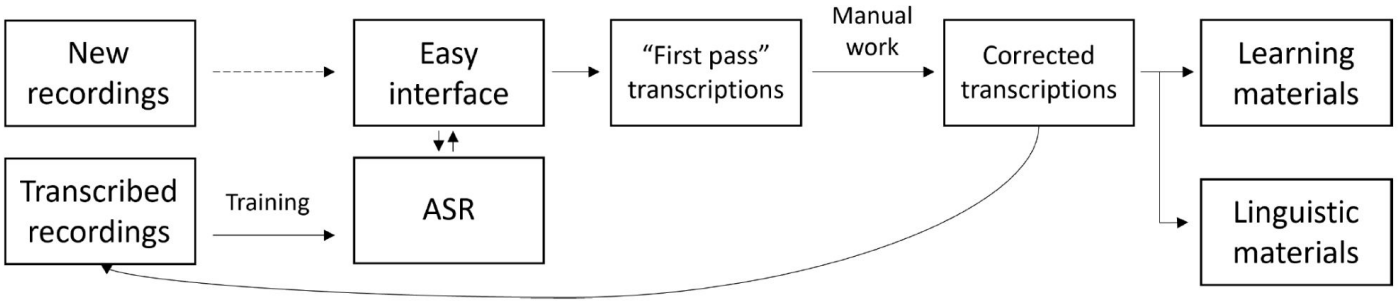
We need to test if different transcriptions will have an effect (e.g. should long vowels be *ā* or *ax*?)

CIM has relatively few phonemes (9~10 cons, 10 vowels) and few affixes. This might be enhancing the results.

Analysis

We need recordings from more islands.

We need to generate a virtuous cycle where transcriptions create revitalization materials and create further training materials for the ASR.



Analysis

Model and data are freely available.

Kaitiakitanga License:

- Non-commercial use allowed
- In consultation with Indigenous community

Objective:
Maintain data sovereignty.

Conclusions

(1) We built an ASR with CER~6 for Cook Islands Māori

(2) Deep Learning ASR methods are approaching a point of usability for low-resource languages.