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

Rolando Coto-Solano¹, Sally Akevai Nicholas², Samiha Datta¹, Victoria Quint¹, Piripi Wills³, Emma Ngakuravaru Powell⁴, Liam Koka’ua³, Syed Tanveer¹, Isaac Feldman¹


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
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

**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'i a 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)
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)
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

### Cook Islands Māori ASR
Error rate by type of training (approx. 4 hrs of data)

<table>
<thead>
<tr>
<th>ASR System</th>
<th>WER</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaldi</td>
<td>17.9 ± 1.7</td>
<td>7.5 ± 0.8</td>
</tr>
<tr>
<td>DeepSpeech</td>
<td>41.1 ± 2.0</td>
<td>21.9 ± 1.6</td>
</tr>
<tr>
<td>Wav2Vec2</td>
<td>22.9 ± 2.0</td>
<td>6.1 ± 0.6</td>
</tr>
</tbody>
</table>
## Experiment 1: ASR with all speakers

<table>
<thead>
<tr>
<th>English</th>
<th>Target</th>
<th>WER</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>One day I was just sitting in my car</td>
<td>i tētā'i rā tē no'o 'ua ara au i roto i tōku motoka</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Kaldi</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DeepSpeech</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wav2Vec2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was sure that it was the pig who had rooted (it up)</td>
<td>kua kite ra 'oki au ē nā te puaka i ketu</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kaldi</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>DeepSpeech</td>
<td>55</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Wav2Vec2</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td>Absolutely, it will get mixed up</td>
<td>āe 'oki ka iroiro atu</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kaldi</td>
<td>80</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>DeepSpeech</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Wav2Vec2</td>
<td>40</td>
<td>23</td>
</tr>
</tbody>
</table>
Experiment 2: Held-Out Speakers

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

Partition 1
Training and validation sets: T1, T3, M1, M2, B, J (90% of files)
Test set: A, K, T2, R (10% of files)

Partition 5
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

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

**Partition 5**
Meaning: *From morning till night.*
Target: mei te pōpongi mai e pō
Inference: mei te pūpongi 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)
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.
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.
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.