Scoring pronunciation accuracy via close introspection of a speech recognition recurrent neural network

Calculating letter by letter confidence scores by applying speech recognition to a recording of a known sentence.

papa reo

Authors: Caleb Moses[2] Miles Thompson[1] Keoni Mahelona[1] Peter-Lucas Jones[1] (1 Te Hiku Media, 2 Dragonfly Data Science)

Our motivation

Te reo Māori is the language of the indigenous people of New Zealand. While it has been suppressed over a period of generations, there is a strong movement to revitalise the language. The New Zealand Government has pledged to ensure one million people are able to speak basic te reo Māori by 2040. This tool aims to contribute to the digital revitalisation of te reo Māori.

The Māori language

The Māori language is a member of the East-Polynesian branch of the Austronesian language family. It uses a phonemic alphabet, with 5 short vowels, 5 long vowels and 10 consonants:

a e i o u ā ē ī ō ū h k m n p r t w ng wh

The data

The training data for the speech recognition model consists of a database of crowd sourced labelled speech recordings, and separately a text corpus from a range of sources.

The audio data consists of:

198,000	speech recordings
400	hours of audio
5,000	unique sentences
2,200	speakers

The text data consists of:

4.8M	total words
20 MB	of text

About Papa Reo

Led by Te Hiku Media in partnership with Dragonfly Data Science, our research will lead the revitalisation of minority and pacific languages and the indigenisation of digital devices worldwide. Our core focus is in producing robust speech and language models for under resourced languages, especially indigenous languages, starting with te reo Māori.

NEURAL INFORMATION PROCESSING SYSTEMS te hiku media DRAGONFLY Data Science

What we did

We used speech recognition to calculate character-level confidence scores to provide instant feedback to second language learners of te reo Māori. The principle is illustrated below:

1. The user hits ‘record’ and reads a provided sentence to an app on their device.
2. A speech recognition model can then provide a score of their pronunciation accuracy.

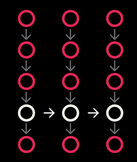
*means ‘the people of the land’

How it works

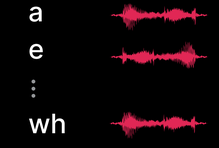
“kia ora” *

1. The user provides a voice recording along with its transcript.

*means ‘hello’

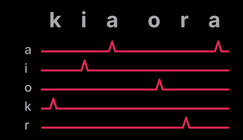


2. We apply a Mozilla DeepSpeech speech-to-text model previously trained on Māori language data.



3. Extract a stream of character probabilities from the acoustic model.

4. Align the character probability stream to the target sentence.



5. Reduce the stream to a simple character confidence score for each letter in the target sentence.

Results

We observed the model working with confident te reo speakers as expected. No detailed pronunciation data is used during training, but we are accumulating a devset to estimate the error of the pronunciation tool. We would like to improve the model further before putting it into production, and for this reason we are still working on this tool.