Developing resources for te reo Māori Text To Speech synthesis system

Jesin James[1], Isabella Shields (Ngāti Porou)[1], Rebekah Berriman (Ngāi Tahu)[1], Peter J. Keegan (Waikato-Maniapoto, Ngāti Porou)[2], and Catherine I. Watson[1]

[1] Department of Electrical, Computer, and Software Engineering, University of Auckland jesin.james@auckland.ac.nz, ishi836@aucklanduni.ac.nz, rber798@aucklanduni.ac.nz, c.watson@auckland.ac.nz
[2] Te Puna Wānanga, University of Auckland p.keegan@auckland.ac.nz

Abstract. Te reo Māori (the Māori language of New Zealand) is an under-resourced language in terms of availability of speech corpora and resources needed to develop robust speech technology. Māori is an endangered indigenous language which has been subject to revitalisation efforts since the late 1970s, which are well known internationally. The Māori community recognises the need for developing speech technology tools for the language, which will improve its study and usage in wider and more digital contexts. This paper describes the development of speech resources in Māori to build one of the first Text To Speech synthesis system for the language. A speech corpus, extended dictionary and a parametric speech synthesiser are the main contributions of the study. To develop these resources, text processing, segmentation and alignment, letter to sound rules creation were also done with existing resources that were modified to be used for Māori. The acoustic similarity of synthesised speech vs natural speech was measured to evaluate the speech synthesis system statistically. Future work required is described.

Keywords: Under-resourced language · TTS system · te reo Māori

1 Introduction

Speech technology is a dominant research field and technology that talks is becoming the new norm. However, a majority of these studies are happening only in a few privileged languages (1.4% of world languages) [10]. Languages that do not have the technical support to develop speech processing technologies are termed as under-resourced languages. Te reo (means *language*) Māori[3] is the indigenous language of New Zealand and has official language status. However, New Zealand English is the dominant language in the country. The 2018 New Zealand census reports[4] that there are 185955 Māori speakers, which accounts

[3] We thank Te Hiku Media for their generous support in funding this project (https://tehiku.nz/). We thank the MAONZE research group [1] for their encouragement.
[4] https://www.stats.govt.nz/information-releases/2018-census-totals-by-topic-national-highlights-updated

for only 4% of the total population (4.58 million). There is a lack of speech and language resources in Māori, making it under-resourced. In 1998, Laws [12] developed a diphone-based synthetic Māori voice. A Festival-based TTS system was planned [13], but its development was stopped in 2003. The lexicon resources developed by Laws are sadly lost, although a copy of the recorded diphones remains. Since Laws' pioneering work, a lot more tools have become available to developing speech and language resources. But, no more work was done on Māori. In this paper, we present the development of speech resources for Māori (preliminary work reported in [18]), with the larger aim of developing a Text To Speech synthesis (TTS) system.

2 Motivation : te reo Māori revitalisation

Māori language spoken in New Zealand derives from the languages/dialects spoken by arrivals from Eastern Polynesian region of the South Pacific, 800 years ago. It was heavily influenced by English speakers who began arriving in the early 1800s. Since the early 1900s, most Māori were taught in English and were discouraged from speaking Māori in schools. Between 1950-1980 there was a sharp decline in the number of fluent Māori speakers [9][20]. From the 1980s there have been strong Māori revitalisation efforts (known internationally [2]) involving both Māori community initiatives, support from New Zealand Government in education, the media, official contexts and other organisations. Te Hiku Media is an example, who in addition to providing media services to the local Māori, are involved in developing digital tools (like speech tools) for the wider community. The Māori community is aware of sound change over time. It is regarded as a result of the break in intergenerational transmission; thus, this sound change has been regretted [20]. The MAONZE research group [1] has been doing extensive research into sound change in Māori [14] [21]. A Māori pronunciation aid (MPAi) is being developed [20], and research is focused on the ways to provide feedback to people from the aid. In this context, a module that can produce synthesised Māori speech can provide speech feedback to MPAi users. Also, a TTS system can be used by people not so proficient in Māori to listen to how words/sentences are spoken. This has potential applications like e-readers, human-computer interaction systems where the technology used in New Zealand currently is English-based. The new generation of Māori users are all exposed to the latest technology. If Māori-based interfaces are available to them, te reo Māori use can be boosted. Given all these applications, and the larger aim to revitalise the language, speech resource development in Māori is essential.

Māori phonology: (Details in [14], [11]) Māori, as with other Polynesian languages did not have an indigenous writing system. The Roman script was used to write Māori since the early 1800s. Māori uses macrons to differentiate between long and short vowels. The language has ten consonants <p, t, k, m, n, ŋ, f, w, ɾ, h> and five short vowels <i, e, a, o, u>. Vowel length is phonemic, i.e., there are five contrasting long vowels < ī, ē, ā, ō, ū >. Research points out that the timing unit in Māori as the mora, a unit consisting of a short vowel plus any preceding consonant. There are no consonant clusters.

3 Te reo Māori Resources development

As Māori is under-resourced, many resources needed for the TTS system had to be developed. Knowledge of Māori language is essential to build these resources.

Māori speech corpus: To build a parametric speech model for Māori, an appropriate speech corpus was developed. The transcript for the corpus was sourced from a collection of Māori myths and legends called Ngā Mahi a Ngā tūpuna [7]. These source files were processed, and text cues for recording were produced. Lines were split according to the presence of [. ? ! ; :]. The source files use old Māori alphabet, where long vowels were represented without macrons (e.g. old alphabet 'aa', new alphabet 'ā'). Conversion to the current Māori alphabet was made by replacing long vowels. Some occurrences of double vowels should not be replaced with macros as the sequence occurs across a morpheme boundary (e.g. whakaaro) and they were hand-corrected. A basic phonetic transcription was created via Python-based coding. This produced a new file which replaced macron symbols (e.g. 'ā') with a symbol and a colon to indicate length (e.g. 'a:'), and 'r' with 'ɾ', and digrams 'ng' and 'wh' with'ŋ' and 'f', respectively. Each line of the transcript was split into separate prompt files for recording.

Lexicon: An existing dictionary from the MAONZE project was used as the starting point. All words which appeared in the transcript but not in the original dictionary were added using Python scripts that performed comparison. The created lexicon contains over 10000 words and 18000 names, along with a phonetic transcription of words, syllable boundaries and stress mark up. The latter two are determined using Bigg's stress rules [3] and the division of words into morae. The automated rules were checked on 959 hand-transcribed rules, and the accuracy was 95%.

Corpus recording: Recordings took place in a WhisperRoom Sound Isolation chamber. The speaker was recorded using a Rode Lavalier Lapel microphone kept approximately 15 cm from their mouth. The microphone was connected to a Roland OCTA-CAPTURE. Audacity® [5] was used for audio capture at 44.1 kHz sampling frequency. A computer monitor displaying the prompts was set up on a desk inside the recording enclosure. The pre-amplifier gain was set to 37.5 dB. A middle-aged male Māori speaker was used to record the corpus. The speaker was given time to read over each phrase and familiarise with it before recording. Long sentences were read as phrases by speaker, using his own judgement to determine 'natural' phrase boundaries within sentences. The audio playthrough was enabled so that audio quality and correctness of the utterance could be assessed during recording. While the recording was trimmed and saved by the recorder, the speaker read the following sentence to be recorded. Each recording was saved in stereo WAV format. Recording sessions were limited to approximately 200 sentences to ensure speaker comfort. Each session took about 2-2.5 hours, accounting for a break in between the session. In total, 1030 sentences were recorded. The first 800 of these correspond to the first 800 lines of

[5] Audacity® software is copyright © 1999-2019 Audacity Team.

the transcript. The final 230 were selected from the remaining transcript after a basic analysis of phone and diphone coverage.

Diphone coverage: A basic analysis of diphone coverage was undertaken after the first 800 sentences were recorded. 30 occurrences were arbitrarily chosen as the goal for each insufficiently covered diphone. Sentences from the remaining unrecorded transcript that had instances of these were selected. Comparison of the frequency of occurrence of long and short vowels largely reflects rates reported in [16]. Fig.1 gives the diphone coverage of the corpus developed. Diphone pairs such as /wu/ and /wo/ only appear in loanwords. Diphones /fu/ and /fo/, although not well covered in the corpus, are very rare. There are few instances where the corpus insufficiently covers a diphone pair, such as 'ŋge' (/ŋe/).

Alignment: Montreal Forced Aligner (MFA Version 1.0.0) [15] was used to align text with the recordings. Recordings in WAV format, phonetic transcriptions in TextGrid format and dictionary are needed for the alignment. Manual checking of alignment and hand corrections were done. For converting TextGrid to .lab format (needed for MaryTTS voice creation), the phone tier is extracted, and each symbol and its corresponding end time are stored in the .lab format.

4 Te reo Māori TTS system development and analysis

This project used MaryTTS [17] (Java-based) to generate a Hidden Markov Model (HMM) - based (parametric speech synthesis) Māori voice. A deep-learning approach was not taken as a sufficiently large corpus is unavailable. An open-source speech synthesiser with good speech quality and support for new languages was needed for this project; which lead to the choice of MaryTTS.

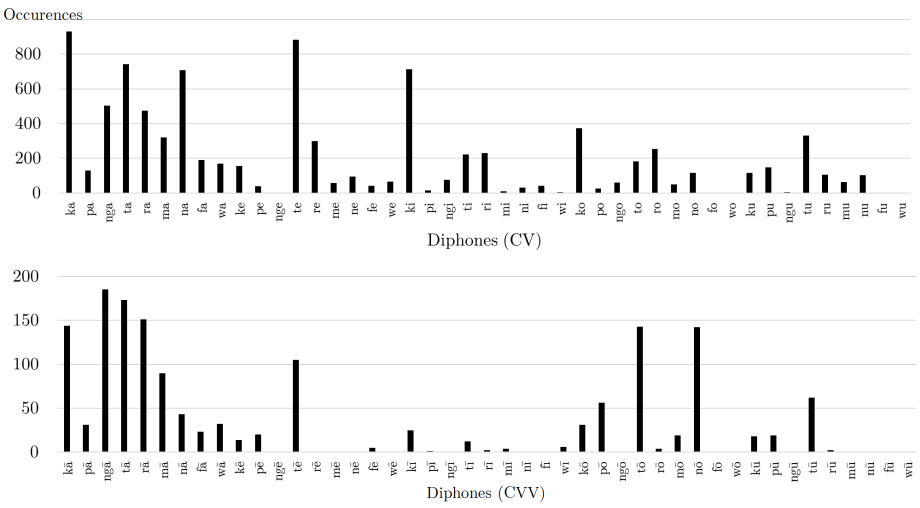

Fig. 1. Occurrences of diphones in the Māori corpus

Adding a new language to MaryTTS: MaryTTS New Language Support was used to develop the Māori TTS system. The locale name used for Māori was 'mi' following Windows locale codes. Requirements for new language addition are the allophone list, lexicon, letter to sound rules and language corpus. The lexicon and language corpus were built, as described previously. An allophone list was added specifying the features: vowel length, vowel height, vowel frontness, lip rounding, consonant type, place of articulation, consonant voicing. 39 allophones (28 for vowels, 10 for consonants, one for silence) are identified for Māori.

Creation of letter to sound rules: MaryTTS transcription tool was used to create letter to sound (LTS) rules using the dictionary and allophone set. First, a manual specification of all letters and corresponding phones that can be used to render them are created. The complete lexicon that is used for training is then aligned to the corresponding phones based on the mapping table. Then a classification and regression tree was built using the training data for LTS rules.

Voice building: The voice building process in MaryTTS was followed with the language resources. The linguistic features extracted are ToBI accents; quinphones for each phone; part-of-speech of each word and features of each phone. This is *Text analysis and Linguistic analysis*. HMM-based voice building was done based on these features. The pitch range was set to 50-300 Hz (for male speaker). Mel-generalised cepstrum coefficients, log of fundamental frequency and strength were the speech features modelled. The final step builds the language-based speech model. This model will be used for *Waveform Generation*. An end-to-end Māori TTS system was set up as in Fig. 2, with Māori text input, and the output is synthesised speech. Native Māori speakers listened the synthesised speech produced, and they commented that the pronunciations were in alignment with those expected from Māori speakers. A client-server model-based speech synthesiser for Māori was set up in the University of Auckland robotspeech server and is available for the various research activities at the university. We are working on making the online TTS system accessible to the wider public. The lexicon and speech corpora will not be made available publicly, as we are guided by Te Hiku media's data sovereignty stance[6]

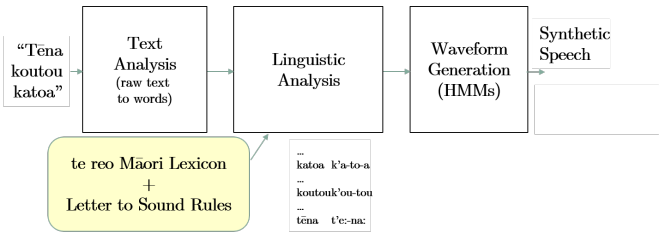

Fig. 2. The end-to-end Māori Text To Speech synthesis system

[6] https://tehiku.nz/te-hiku-radio/te-putahi/12707/dr-tahu-kukutai-keoni-mahelona-data-sovereignty

Signal-based quality diagnosis: To evaluate the synthesised speech, a simplified version of the signal-based quality diagnosis (mel-cepstral distortion - MCD) described in [5] was implemented. MCD measures the difference between two sequences of mel cepstra as shown in Fig. 3. Segmentation is done at the word level (using MFA), for the original and synthesised speech. MFCCs are extracted using [4]. Difference in timing of the two sequences is aligned by Dynamic Time Warping (DTW) (based on [6]). Consider the synthesised speech and reference original speech MFCCs as a time series $X = (x_1, x_2...x_N)$ and $Y = (y_1, y_2...y_M)$ respectively. A correspondence between their elements is established by a warping curve $\Phi = (\phi_t, \psi_t); t = 1, ...T$. ($T$ depends on the lengths M and N). The optimal warping curve is the one that minimises the distance between the two time series, represented by: $\hat{\Phi} = (\hat{\phi}_t, \hat{\psi}_t) = (\phi_t, \psi_t)\arg\max = \sum_{t=1}^{T} \frac{d(x_{\phi_t}, y_{\psi_t}) m_{t,\Phi}}{M_\Phi}$ [19]. Here $m_{t,\phi}$: *local weighting coefficient*, M_ϕ: *path-dependent normalisation* $= \sum_{t=1}^{T} m_{t,\Phi}$, d: *local distance*. The minimum cumulative distance is obtained as: $D(X,Y) = \sum_{t=1}^{T} \frac{d(x_{\hat{\phi}_t}, y_{\hat{\psi}_t}) m_{t,\hat{\Phi}}}{M_{\hat{\Phi}}}$. The distance measure obtained is the *average per-step distance* along the warping curve.

432 words from 30 sentences in the Māori corpus were tested. Example of time series alignment obtained after warping is shown in Fig. 4. The alignment is perfect when it is between the same words (a), and the alignment is poor

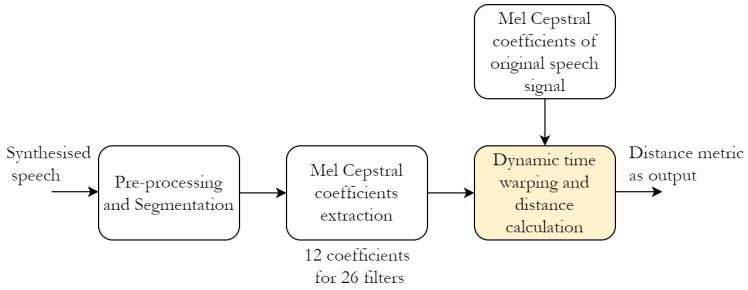

Fig. 3. Acoustic similarity measure for synthesised speech

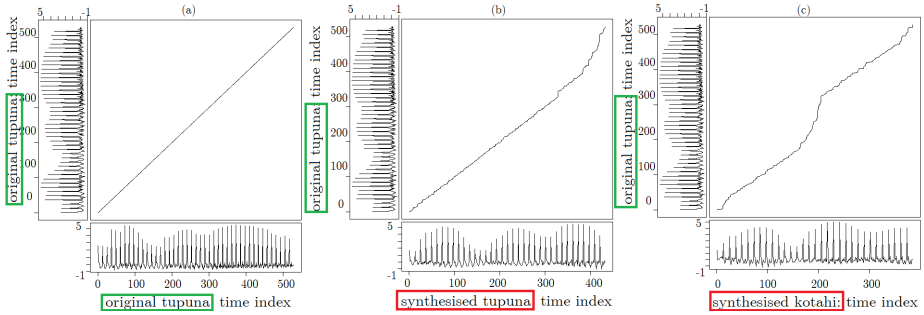

Fig. 4. Time series alignment using DTW of MFCCs. (a) Same words (b) natural vs synthesised version of same word, (c) natural vs synthesised version of two words [6].

(a)

Synthesised word	wehewehe	ao	ēnei	rāua	raro	mawehe
wehewehe	34.84	83.38	53.84	70.61	74.49	55.64
ao	76.27	26.57	76.51	50.31	57.94	70.14
ēnei	54.58	83.32	29.31	78.29	75.43	60.84
rāua	72.41	45.54	75.69	46.49	50.27	64.2
raro	69.83	52.28	74.01	50.12	31.59	67.36
mawehe	46.12	63.39	59.87	59.67	69.73	32.5

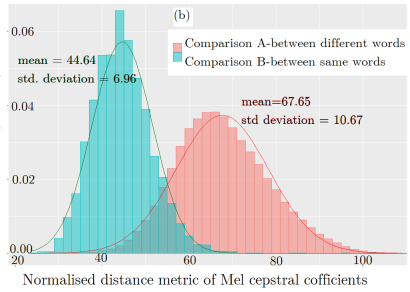

Fig. 5. (a) Confusion matrix of distance between original and synthetised MFCCs. (b) The probability distribution of distance between original and synthesised MFCCs.

when the original and synthesised versions are from different words (c). The alignment is good for the original and synthesised version of the same word (b), which is an indication of the performance of the TTS system. Normalised distance measure between the original and synthesised MFCC time series is then taken as described in Fig. 3 [8]. Fig. 5 (a) shows the confusion matrix of distance measures for 6 example words. It can be seen that the distance is comparatively lower when the original and synthesised MFCCs are of the *same word* (see Fig. 4). Fig. 5 (b) shows the probability distribution of the distance measures for: *Comparison A:* Between MFCCs of original and synthesised versions of *different words* and *Comparison B:* Between MFCCs of original and synthesised versions of *different words* for all words tested. It is clear that *Comparison B* results in larger distance compared to *Comparison A*. This statistically shows the acoustic similarity between the words in the original and synthesised speech signals.

5 Conclusion and future work

This paper describes the development of te reo Māori TTS system. Māori is under-resourced, and development of speech technology resources is critical for its revitalisation. A speech corpus was developed containing 1030 sentences. The Māori lexicon was built with 10000 words and 18000 names. Existing tools (like Montreal Forced Aligner) were customised to segment and align the text and speech signals in the corpus. These resources were then used to develop an end-to-end Māori TTS system, where Māori text is entered, and the output is synthesised speech. Acoustic similarity analysis of the synthesised speech was done. Perceptual testing of the synthesised speech will be conducted. Future work will focus on expanding the Māori lexicon and checking entries, especially where orthography does not align with the current Māori phonetics. There is also a need to incorporate modern Māori pronunciations (like the occurrence of affrication and semi-vowels not present in traditional Māori) into the lexicon. Also, the effect of phraseology on the assignment of stress/syllable structure needs to be added to the linguistic analysis. In the real world, code-switching is common in spoken Māori; therefore, any useful Māori TTS system will need to accommodate New Zealand English. We hope to implement this by 2021. [18]

References

1. homepages.engineering.auckland.ac.nz/~cwat057/MAONZE/index.html
2. http://www.maoriMay12020language.info/mao_lang_abib.html
3. Biggs, B.: Let's learn Māori: a guide to the study of the Māori language. Wellington: A.H. A.W. Reed (1969)
4. Ellis, D.P.W.: PLP and RASTA (and MFCC), and inversion) in Matlab. Online web resource. (2005), www.ee.columbia.edu/~dpwe/resources/matlab/rastamat
5. Falk, T., Moller, S.: Towards signal-based instrumental quality diagnosis for text-to-speech systems. vol. 15, pp. 781 – 784. IEEE (2008)
6. Giorgino, T.: Computing & visualizing dynamic time warping alignments in R:the dtw package. In: J. Stat. Softw. vol. 31(7), pp. 1–24. Open Access Statistics (2009)
7. Grey, S.G.: Ngā Mahi a Ngā tūpuna. New Plymouth, 3^{rd} edition (1928)
8. Hall, K.C., Allen, B., Fry, M., Johnson, K., Lo, R., Mackie, S., McAuliffe., M.: Phonological corpustools, version 1.3. [computer program].
9. Harlow, R.: Māori: A Linguistic Introduction. Cambridge University Press (2007)
10. James, J., Watson, C.I., Gopinath, D.P.: Exploring text to speech synthesis in non-standard languages. In: Int. Conf. Speech Science Tech., Australia. pp. 213–16 (2016)
11. Keegan, P., Watson, C., Maclagan, M., King, J., Harlow, R.: The role of technology in measuring changes in the pronunciation of Māori over generations. In: Language Endangerment in the 21st Century: Globalisation, Technology & New Media, NZ. pp. 65–71 (2012)
12. Laws, M.R.: A bilingual speech interface for New Zealand English to Māori, school = Doctoral dissertation, University of Otago, New Zealand. Ph.D. thesis (1998)
13. Laws, M.R.: Speech data analysis for diphone construction of a Māori online Text-to-speech Synthesizer. In: IASTED Int. Conference, USA. pp. 103–108 (2003)
14. Maclagan, M., Harlow, R., King, J., Keegan, P., Watson, C.: Acoustic Analysis of Māori:Historical Data. In: Conf.: Australian Linguistic Soc. p. 104 (2005)
15. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In: Interspeech, USA. pp. 498–502 (2017)
16. Rácz, P., Hay, J., Needle, J., King, J., Pierrehumbert, J.B.: Gradient Māori phonotactics. In: Te Reo. vol. 59, pp. 3–21 (2016)
17. Schroder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. In: Int. J. of Speech Tech. pp. 365–77. Springer (2003)
18. Shields, I., Watson, C., Keegan, P., Berriman, R., James, J.: Creating a synthetic te reo māori voice. In: Int. Conf. on Language Tech. for All, Paris (2019), https://lt4all.org/media/papers/P1/136.pdf
19. Tormene, P., Giorgino, T., Quaglini, S., Stefanelli, M.: Matching incomplete time series with dynamic time warping: An algorithm and an application to post-stroke rehabilitation. In: Art. intelligence in medicine. vol. 45, pp. 11–34. Elsivier (2008)
20. Watson, C., Keegan, P., Maclagan, M., Harlow, R., King, J.: The motivation and development of MPAi, a Māori pronunciation aid. In: Annual Conference of the International Speech Communication Association, Stockholm. pp. 2063–2067 (2017)
21. Watson, C., Maclagan, M., J, K., Harlow, R., , Keegan, P.: Sound change in Māori and the influence of New Zealand English. In: Journal of the International Phonetic Association. vol. 46(2), pp. 185–218. Cambridge University Press (2016)