Development of a Māori Database for Speech Perception and Generation

Mark R. Laws
Knowledge Engineering Laboratory, Department of Information Science,
University of Otago, P.O.Box 56 Dunedin, New Zealand.
Email: maaka@kel.otago.ac.nz

Abstract - Māori speech data collection and analysis is an ongoing process, as new and existing data sets are continuously accessed for many different experimental speech perception and generation processing tasks. A data management system is an important tool to facilitate the systematic techniques applied to the speech and language data. Identification of the core components for Māori speech and language databases, translation systems, speech recognition and speech synthesis have been undertaken as research themes. The latter component will be the main area of discussion here. So to hasten the development of Māori speech synthesis, joint collaborative research with established international projects has begun. This will allow the Māori language to be presented to the wider scientific community well in advance of other similar languages, many times it's own size and distribution. Propagation of the Māori language via the information communication technology (ICT) medium is advantageous to it's long term survival.

1. INTRODUCTION

This paper reports on the small contribution that is being made to assist people to acquire proper pronunciation of the Māori language through the development of a speech synthesis system. Utilising ICT as the means to widely distribute the spoken language will reflect the multi-disciplinary nature of this research. The paper will also outline future directions for the Māori speech database management system and associated perception and generation applications.

2. THE MĀORI LANGUAGE

Māori belongs to the Proto-Eastern-Polynesian group of languages. Māori has been spoken in New Zealand for hundreds of years, it has undergone many changes since the arrival of the first English speaking populace [1]. More recently, the language was identified as being “in crisis of survival” [2] then and now it is in a continual state of decay. Much vocabulary known to native speakers has begun to deplete, so its usage has been replaced by either English words or Anglicism’s of those Māori terms [3]. Today, to stem this tide and revitalise the language, there has been an overall commitment shown by many to learn and use Māori in all aspects of social, educational and cultural exchanges.

The main phonological difference between English and Māori is the pronunciation of the vowel system. Māori vowels are fairly constant—whether being spoken on its own as either short or long, together as double vowels or allophones and diphthongs. Small variations in vowels do occur at the acoustic-phonetic level where co-articulation between the vowels and/or consonants can colour initial or preceding vowels. The second noted difference is in the slight pronunciation variations of four consonants [4]. The principles of primary stress assignment in Māori words are pitch fall and the length of the first vocalic element in diphthongs—there are six main stress features, which can alter the stops, fricatives, the liquid, nasals, and vowels. Stress may also include an increase in loudness [3]. The Māori orthographic representation has not changed much from its early developments established in the 1800’s.

Māori has a phonemic writing system of letter-to-sound rules. The Māori phonemic unit number is small by comparison to English. For example, there are only twenty-four Māori phonemes compared to forty-five NZ English phonemes [3, 4]. There are five distinct phonemic vowel sounds in Māori, vowel length can be short or long; there are five long vowels that correspond to twice the length of the five short vowels. Therefore, there are ten vowels in pairs. There are only four diphthongs (see Table 2). There are ten consonant phonemes, consisting of three stops (plosives), two fricatives, three nasals, one liquid and one semivowel glide (see Table 1). Two consonants have orthographic symbols with two Roman letters, *wh* for /f/ and *ng* for /ŋ/ [3].

The phonotactics of Māori are as follows; Consonants can only appear in word-initial (CV) and word-content (VCV) positions, they cannot appear as word-final or in clusters within words; Vowels can occur in all three positions (VC, CVC, CV), and in multiple clusters (e.g. CVV, CVVV, CVVVV). These statements can now both be collapsed into the Māori syllable (C)V, where (C) is optional [3].

Table 1: Māori consonant features, these are very similar to the English linguistic features.

	Type	Articulation	Voicing	Example
p	stop	labial	-	*papa*
t	stop	alveolar	-	*tiki*
k	stop	velar	-	*kaka*
f	fricative	labio	-	*whare*
h	fricative	glottal	-	*hoa*
m	nasal	labial	+	*mana*
n	nasal	alveolar	+	*noa*
ŋ	nasal	velar	+	*ngā*
w	approximate	labial	+	*waka*
r	liquid	alveolar	+	*ringa*
#	silence		-	

Table 2: These linguistic features are for the Maori vowels.

	Length	Height	Position	Round	Example
i	short	high	front	-	*pipi*
i:	long	high	front	-	*pīpī*
e	short	mid	front	-	*keke*
e:	long	mid	front	-	*kēkē*
a	short	low	back	-	*kaka*
a:	long	low	back	-	*kākā*
o	short	low	front	+	*koko*
o:	long	low	front	+	*kōkō*
u	short	mid	back	+	*ruru*
u:	long	mid	back	+	*rūrū*
ei	dipth	mid	front	-	*hei*
ai	dipth	low	mid	-	*kai*
ou	dipth	mid	mid	-	*kau*
ɔ	dipth	low	back	+	*poi*

3. SPEECH ANALYSIS

This analysis investigates existing computational linguistic and information processing techniques applied to the Māori language. Overall, the speech data and higher-order linguistic knowledge is analysed in terms of basic phonetic and acoustic characteristics, features, labels and prosodic models. It describes the methodology designed for the construction of a Māori diphone database used with the MBROLA project. A speech synthesiser based on the concatenation of diphones that reads text via a phoneme transcribed list appended with parameters for pitch and duration, to reproduce the best possible synthesised output [5, 6].

3.1. Diphones

It has been noted that the term 'diphone' seems to be used only amongst the speech synthesis research fraternity, as it is not a common term used in linguistics by phoneticians or phonologists. But based on its derived structure, the diphone best describes two parts of two adjoining phonemes. It is all the phonemic transcriptions taken between all the possible phonemes used in a language [6, 7, 8]. There are over 1400 diphones for English, and less than 300 for Māori. Silence or pauses are also classed as phonemes, they are also important units within the diphone inventory.

The unique approach to deriving all the diphones is by segmenting the transitions starting from one phonemes stable position (usually in the middle) across the articulation channel to the middle of the next phonemes stable position. Because diphones contain the all important co-articulation information between phonemes, the concatenative speech synthesis system provides a smoother transition between all phoneme units to give a more natural reconstruction of word formations. The phoneme and diphone symbolic representation is based on the SAMPA notation [6, 7, 8], which also uses '_' for silence and '–' for the phoneme separator.

3.2. Māori Diphones

Identifying all the diphones in Māori included mapping all the possible combinations of vowel and consonant clusters (see Table 3) [5].

Table 3a: The Māori diphone inventory showing all the possible CV and VV combinations, it also identifies the impossible ones.

	e	i	A	o	u	eI	aI	OI	@U	_
p	p-e	p-i	p-A	p-o	p-u	p-eI	p-aI	p-OI	p-@U	
t	t-e	t-i	t-A	t-o	t-u	t-eI	t-aI	t-OI	t-@U	
k	k-e	k-i	k-A	k-o	k-u	k-eI	k-aI	k-OI	k-@U	
f	f-e	f-i	f-A	f-o	f-u		f-aI		f-@U	
h	h-e	h-i	h-A	h-o	h-u	h-eI	h-aI	h-OI	h-@U	
m	m-e	m-i	m-A	m-o	m-u	m-eI	m-aI		m-@U	
n	n-e	n-i	n-A	n-o	n-u	n-eI	n-aI	n-OI	n-@U	
N	N-e	N-i	N-A	N-o	N-u		N-aI	N-OI	N-@U	
r	r-e	r-i	r-A	r-o	r-u	r-eI	r-aI	r-OI	r-@U	
w	w-e	w-i	w-A				w-aI		w-@U	
e	e-e		eA	e-o	eu					e-_
i	i-e	i-i	i-A	i-o	i-u		i-aI			i-_
A	A-e		A-A	A-o	A-u		A-aI		A-@U	A-_
o	o-e		o-A	o-o	o-u					o-_
u	u-e	u-i	uA	u-o	u-u					u-_
eI										eI-_
aI	aI-e	aI-i	aI-A	aI-o						aI-_
OI			OI-A	OI-o						OI-_
@U	@U-e	@U-i	@U-A	@U-o					@U-@U	@U-_
_	_-e	_-i	_-A	_-o	_-u	_-eI	_-aI		_-@U	_-_

Table 3b: Possible VC combinations

	p	t	k	f	h	m	n	N	r	w
e	e-p	e-t	e-k	e-f	e-h	e-m	e-n	e-N	e-r	e-w
i	i-p	i-t	i-k	i-f	i-h	i-m	i-n	i-N	i-r	i-w
A	A-p	A-t	A-k	A-f	A-h	A-m	A-n	A-N	A-r	A-w
o	o-p	o-t	o-k	o-f	o-h	o-m	o-n	o-N	o-r	o-w
u	u-p	u-t	u-k	u-f	u-h	u-m	u-n	u-N	u-r	u-w
eI		eI-t	eI-k		eI-h	eI-m	eI-n	eI-N	eI-r	
aI	aI-p	aI-t	aI-k	aI-f	aI-h	aI-m	aI-n	aI-N	aI-r	aI-w
OI	OI-p	OI-t	OI-k			OI-m	OI-n	OI-N	OI-r	
@U	@U-p	@U-t	@U-k	@U-f	@U-h	@U-m	@U-n	@U-N	@U-r	@U-w
_	_-p	_-t	_-k	_-f	_-h	_-m	_-n	_-N	_-r	_-w

The diphone inventory tables show the required combinations that will match the Māori diphone sound samples used by MBROLA. A set of Māori words each containing a predetermined diphone was also generated to create a 229 list of words (see Table 5 for an example).

The speech recordings of the Māori words were done in a recording studio with good acoustics, using high quality industry standard sound equipment (e.g. DAT). The native speaker first practised the intelligibly and pronunciation of all the words. To maintain constant pitch throughout the session, we first recorded a single vowel sound utterance (e.g. 'A') on a separate recorder as an audible reference. This was played at the beginning

of every fifth word utterance to assist the speaker to keep at a steady pitch. The speech was saved directly to DAT at 48khz 16bit mono.

Once the recording was completed the speech was transferred to computer, the entire file was processed for normalisation and then down-sampled to 22050hz, 16bit, mono. The file was then divided into the many word labelled units and saved in separate manageable files, this made the diphone segmentation process an easier task to perform. Figure 1 shows how the diphones are processed using an application called Diphone Studio, which was specifically designed to construct diphone databases for MBROLA [9].

Each file was hand segmented with three sampling points, indicating the left, the middle and right boundaries. These three boundaries represent the start of the first phoneme sample, the crossover point between the two phonemes and the end of the second phoneme sample.

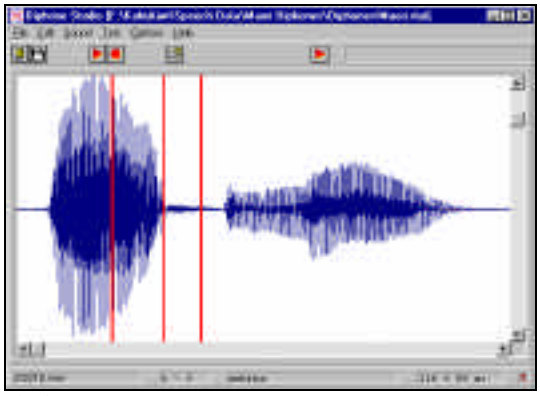

Fig 1. An example of the diphone unit (A-f) being segmented from the word "awhina", the sampling points are shown in milliseconds.

To monitor the quality of the segmentation, the pitch and the energy, each diphone was checked by testing it with a number of words/phrases (e.g. A hAkA mAnA pArA tAwAa NA FA). This first evaluation step only gave a very crude reproduction of the speech, but in terms of checking all adjoining diphones, it was very successful–the diphones performed well within the required specifications.

Table 4: The Māori diphone data file generated by the Diphone Studio tool.

Diphone File	Diphone Label	Left Boundary	Right Boundary	Middle Boundary
d0.raw	_-_	2205	8082	5145
d1.raw	_-@U	2205	9417	5801
d2.raw	_-A	2205	7701	4962
...	...	...	...	...
d100.raw	h-eI	1835	5791	2939
d101.raw	h-i	2205	5489	3375
d102.raw	h-o	2205	7259	4298
...	...	...	...	...
d227.raw	w-aI	2205	8959	3646
d228.raw	w-e	1620	5331	2547
d229.raw	w-i	1593	4179	2747

Post-processing of all the diphones took place with the entire inventory being compiled into i) a data file containing all the diphone details (see Table 4), and ii) each diphone was extracted from their original word examples and saved in their '.RAW' format (e.g. headerless wave files). The files contain no other linguistic information.

The diphone database was then sent to the MBROLA project team for further processing, where the data analysis and re-synthesis procedures were performed, then a compiled database was returned for evaluation.

4. MĀORI SYNTHESISER

The Māori diphone database named 'mb1' was tested using the MBROLIN and MBROLI speech synthesis tools. MBROLIGN is a prosody transplantation tool that generates '.pho' files. MBROLI is a 'phoneme-to-speech' tool that reads '.pho' files and produces speech [6, 10].

The initial results are very acceptable given the duration, pitch and pattern point adjustments to the '.pho' test files were made to compensate for any slight variations. The wave files produced had a very close resemblance to the original Māori speaker's voice. The following table and figure outline the phrase *kia ora* being processed and tested.

Table 5: Māori word examples, wave files, diphone files and labels that make up the phrase 'kia ora'.

Māori Word Examples	Speech Wave Files	Diphone Files	Diphone Labels
kete	#k2216.wav	d8.raw	_–k
kite	ki2216.wav	d127.raw	k-i
hia	iA2216.wav	d106.raw	i-A
hūia	A#2216.wav	d34.raw	A-_
o	#o2216.wav	d12.raw	_-o
ora	or2216.wav	d178.raw	o-r
rā	rA2216.wav	d192.raw	r-A

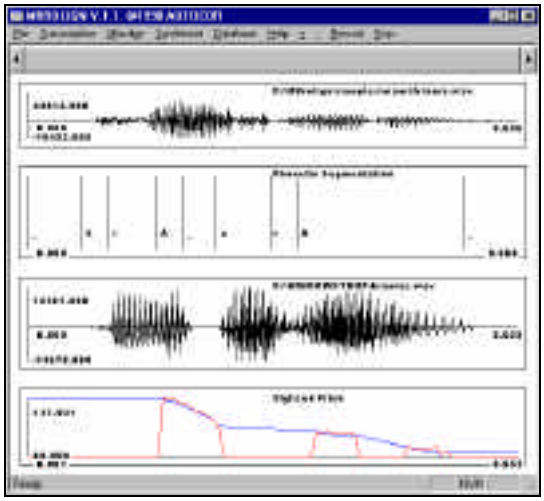

Fig 2. The test toolbox used for *kia ora* showing the original sound file, the transcription, the synthesised output file and pitch analysis.

The resulting synthesised speech file is a very good reproduction that far exceeds all other attempts made with the current text-to-speech synthesis tools to pronounce Māori correctly.

Preliminary testing is now being undertaken using the Festival Text-To-Speech (TTS) system [8]. Many MBROLA users prefer to use this system because it offers full TTS compatibility with current diphone databases. Specific files that are based on Māori speech and language parameters are being developed and refined to improve on the previous test results. Main areas of work to be done are on the prosodic modelling for duration, pitch, stress, voicing and intonation, also developing definitions for the phoneme inventory and lexicon are underway.

5. SUMMARY

This paper has outlined the first step to introducing the Māori language to the ICT arena, it has followed the path of many other languages development cycles by many dedicated researchers within the fields of computational linguistics and languages. Speech database management has allowed the data to be accessed and processed for a number of different applications–where speech synthesis of Māori is just the beginning.

Parallel to this research is another project that uses the phoneme as the basis to training evolving fuzzy neural networks (EfuNN) for speech-based phoneme recognition [11]. It has been noted that the diphone analysis may provide a more comprehensive unit of speech for processing with neural networks, given the boundary between adjoining phonemes is just as important in isolated and continuous speech perception as it is now in speech generation. The likelihood that the same diphone units could be used for a speech recognition system are very promising.

ACKNOWLEDGEMENTS

Support from the UOO-808 grant funded by the New Zealand Foundation for Research Science and Technology (NZFRST) and the NZFRST Tūāpapa Pūtaiao Māori Fellowship scheme.

Mr Ratu Tibble, for providing the speech data used in the Māori diphone database. Mr Barney Taiapa, Mr Dennis Mariu, Mrs Noi Hudson and Mrs Alva Kapa - Ph.D Māori Advisory Committee for supporting this research.

REFERENCES

[1] Benton, R. A. (1990) *The History and Development of the Māori Language,* Dirty Silence Aspects of Language and Literature in New Zealand, Essays arising from the University of Waikato Winter Lecture Series of 1990, Oxford University Press, Auckland.

[2] Harlow, R. (1990) *A Name and Word Index to Ngā Mahi a Ngā Tūpuna*, University of Otago Press, Dunedin.

[3] Bauer, W., Parker, W., Evans, T. (1993) *Maori Descriptive Grammars*, Routledge, London and New York.

[4] Ryan, P. M. (1997) *The Reed Dictionary of Modern Māori,* Reed Publishing (NZ) Ltd.

[5] Laws, M. (1998) *A Bilingual Speech Interface for New Zealand English to Māori*. Unpublished M.Sc. thesis, University of Otago.

[6] Dutoit, T., V, Pagel., N, Pierret., F, Bataille., O, Van Der Vrecken. (1996) "*The MBROLA Project: Towards a Set of High-Quality Speech Synthesisers Free of Use for Non-Commercial Purposes*". Proc ICSLP'96, Philadelphia, Vol. 3, pp. 1393-1396.

[7] Campbell, N. (1998) *Multi-Lingual Concatenative Speech Synthesis*. The 5th International Conference on Spoken Language Processing, Sydney Convention and Exhibition Centre, Darling Harbour.

[8] Black, A. and Taylor, P. (1997). *Festival Speech Synthesis System: system documentation*. Technical Report HCRC/TR-83, Human Communication Research Centre. University of Edinburgh, Scotland.

[9] Dirksen, A. and Menert, L. (1998*). Diphone Studio Manual*. Fluency Speech Technology, Berkelstraat 137, Utrecht, Netherlands.

[10] Malfrére, F. and Dutoit, T. (1997) *Speech Synthesis for Text-to-Speech Alignment and Prosodic Feature Extraction*. Proceedings of the International Symposium on Circuits and Systems, pp. 2637-2640.

[11] Kasabov, N. (1998) *Evolving fuzzy neural networks: Theory and applications for on-line adaptive prediction, decision making and control*. Australian Journal of Intelligent Information Processing Systems, vol. 5 (3), pp 154-160.