Work in Progress: Text-to-speech on Edge Devices for te Reo Māori and ʻŌlelo Hawaiʻi

Tūreiti Keith, Gianna Leoni, Keoni Mahelona, Hina Puamohala Kneubuhl, Stephanie Huriana Fong, Peter-Lucas Jones

Te Reo Irirangi o Te Hiku o Te Ika (Te Hiku Media); Awaiaulu, Inc.; Pae Tū Ltd.
1 Melba St., Kaitāia, Aotearoa; 2667 ʼAnuʻu Pl., Honolulu, Hawaiʻi; 29 Noall St., Tāmaki, Aotearoa
{tureiti, gianna, keoni, peterlucas}@tehiku.co.nz, hina@awaiaulu.com, stephanie@paetultd.com

Abstract

Existing popular text-to-speech technologies focus on large models requiring a large corpus of recorded speech to train. The resulting models are typically run on high-resource servers where users synthesise speech from a client device requiring constant connectivity. For speakers of low-resource languages living in remote areas, this approach does not work. Corpora are typically small and synthesis needs to run on an unconnected, battery or solar-powered edge device. In this paper, we demonstrate how knowledge transfer and adversarial training can be used to create efficient models capable of running on edge devices using a corpus of only several hours. We apply these concepts to create a voice synthesiser for te reo Māori (the indigenous language of Aotearoa New Zealand) for a non-speaking user and ʻōlelo Hawaiʻi (the indigenous language of Hawaiʻi) for a legally blind user, thus creating the first high-quality text-to-speech tools for these endangered, central-eastern Polynesian languages capable of running on a low powered edge device.

Keywords: text-to-speech, voice synthesis, edge device, Māori, Hawaiʻi

1. Introduction

Although text-to-speech (TTS) technologies to support the non-speaking and low-vision communities have existed for many years, the languages typically supported are colonial or high-resource languages. For those who wish to use voice synthesis in languages like te reo Māori or ʻōlelo Hawaiʻi, two endangered (Moseley, 2012), central-eastern Polynesian languages, the typical option available is either 1) near unintelligible reproduction using another language's synthesiser or 2) being forced to use the language of the coloniser, who was ultimately responsible for the near extinction of the language.

The two people who inspired us to begin this work are a non-speaking woman who wishes to communicate with her friends, family and the wider community in te reo Māori and a now legally blind man whose work analysing and reviving ʻōlelo Hawaiʻi has been hindered by tools that cannot read his language to him.

To the best of our knowledge, this work represents the first and only neural TTS system for either te reo Māori or ʻōlelo Hawaiʻi targeting an edge device. The one and only TTS implementation we are aware of in the literature for either language is a MaryTTS implementation of te reo Māori (Schröder & Schröder, 2003; James et al., 2020). Our organisation, Te Hiku Media, published a FastPitch-based model for te reo Māori as part of our Papa Reo Natural Language Processing APIs in 2022 (Łańcucki, 2021; Te Hiku Media, 2022); however, this model is not capable of running on a lower powered edge device.

Recent lightweight neural acoustic models, including SpeedySpeech (4.3M parameters; Vainer & Dušek, 2020), BVAE-TTS (12M; Lee, Shin & Jung, 2020), Talknet 2 (13M; Beliaev & Ginsburg, 2021), PortaSpeech (6M; Ren, Liu & Zhao, 2021), and LightSpeech (1M; Luo et al., 2021), stand out for their compact sizes compared to established high-quality TTS systems like Tacotron 2 (28.2M parameters; Shen et al., 2018), Fastspeech 2 (27M; Ren et al., 2020) and VITS (29.09M; Kim, Kong & Son, 2021). However, these lightweight models specifically focus on converting text to mel-spectrograms. To synthesise waveforms, they require an additional neural vocoder, which can inflate the model size depending on the chosen vocoder model. On the other hand, complete end-to-end neural TTS models include LiteTTS (13M parameters; Nguyen et al., 2021), which relies on generative adversarial networks, as well as, Mini-VITS (5.2M; Kawamura et al., 2023), Nix-TTS (5.23M; Chevi et al., 2023), and Piper-TTS (7M; Hansen, 2023) which employ knowledge distillation to compress a VITS model. Among the end-to-end models, LiteTTS directly reports a Mean Opinion Score (MOS) of 3.84, while Nix-TTS reports a Comparative MOS (CMOS) of -0.27 when compared to VITS, which itself has a MOS score of 4.43. Notably, Piper TTS benefits from a well-supported and well-documented project. It has been successfully applied to over a dozen languages and features a training framework designed for transfer learning—a crucial advantage for under-resourced languages.

2. Method

This section describes our approach to creating three acoustic models for text-to-speech: two te reo Māori voices and a voice for ʻōlelo Hawaiʻi.

2.1 Language Codes

For consistency, we have adopted ISO 639-2 language codes for all languages in this article, as ʻōlelo Hawaiʻi is *not* defined in ISO 639-1 (Byrum, 1999). This means that readers used to seeing the ISO 639-1 code “es” for Spanish will see “spa” instead. Similarly, readers used to seeing “mi” for te reo Māori will see “mri” instead. The code for ʻōlelo Hawaiʻi is “haw”.

2.2 IPA Phonemisation

It is not unusual for an under-resourced language to lack some of the basic tools required for natural language processing. A basic IPA phonemiser for te reo Māori and ʻōlelo Hawaiʻi was one of the tools we built as part of this work. The popular eSpeak-ng package (Duddington, Dunn, 2015) claims to support the phonemisation of languages like te reo Māori, however, we were unable to find alignment between the literature (Harlow, 2007) and the outputs of the package, as such we developed our own phonemisers for this work.

The focus of the IPA phonemisers we developed is to first and foremost support encoding of the languages for speech synthesis, as opposed to accurately modelling the pronunciation of a particular regional variation of the language. This allows us to make some simplifications to the phonemisation in the literature, with little to no loss of information. Where qualitative analysis of the model output points to a loss of information at the phonemisation stage, we can modify the phonemiser to improve the model's performance.

Long vowels		Short vowels		Consonants		
IPA	desc.	IPA	desc.	IPA	desc.	lang.
aː	ā	a	a	ɾ	r	mri
eː	ē	e	e	n	n	haw, mri
iː	ī	i	i	f	wh	mri
oː	ō	o	o	ŋ	ng	mri
uː	ū	u	u	t	t	haw, mri
				m	m	haw, mri
				l	l	haw
				h	h	haw, mri
				v	w	haw
				ʔ	ʻokina	haw
				ɸ	wh	mri
				w	w	haw, mri
				p	p	haw, mri

Table 1: Combined IPA phonemes for te reo Māori (mri) and ʻōlelo Hawaiʻi (haw). Both languages use the same set of vowels.

Table 1 lists the combined IPA alphabet we considered when phonemising te reo Māori and ʻōlelo Hawaiʻi. This simplifies the IPA alphabets defined in the literature (Harlow, 2007; Parker Jones, Niebuhr, Ward, 2018) by 1) using the vowel set /a/, /e/, /i/, /o/, /u/ 2) not explicitly modelling diphthongs, 3) overloading variations in the pronunciation of the "t" in te reo Māori that depend on the following vowel and 4) overloading variations on the pronunciation of "w" in ʻōlelo Hawaiʻi that depend on its position by using only the /v/. Our overloading of /t/ and /v/ was based on the hypothesis that the model would learn any context-based variations from the data.

2.3 Knowledge Transfer

Due to the relatively small number of single-speaker recordings available for training a te reo Māori and ʻōlelo Hawaiʻi speech synthesiser, we chose to first train the model on an existing large and open dataset. The best choice for such a dataset is one where there is a large overlap of sounds between the languages. Anecdotal evidence of similarities between Spanish and te reo Māori was provided to us by Kāpō Māori Aotearoa New Zealand Ltd, who reported the use of Castilian Spanish screen-readers as a workaround for reading te reo Māori text. This suggests similarities between the linking of graphemes to phonemes in both languages. As such, we decided to investigate the phonological content of Castilian Spanish, te reo Māori and ʻōlelo Hawaiʻi.

Table 2 describes the results of this analysis. The first column (IPA) lists the union of IPA phonemes for both te reo Māori and ʻōlelo Hawaiʻi that we chose for this work, as discussed in Section 2.2. The remaining columns list the counts of these phonemes in each dataset. The phonemes for Spanish were generated by the eSpeak-ng phonemiser. See Section 2.4 for more information on the datasets used here.

	Dataset			
IPA	spa_male	mri_male	mri_female	haw_female
a	77710	3979	1228	594
aː	0	3036	982	432
e	79247	3596	1179	555
eː	0	1149	211	114
f	23973	0	0	0
h	0	3209	990	495
i	70571	3638	1175	563
iː	1	333	100	59
k	64029	3586	1184	564
l	66650	0	0	518
m	62837	2722	857	475
n	72138	2783	984	514
ŋ	7271	1699	744	0
o	77715	3451	1121	557
oː	0	1791	583	247
p	55550	1755	531	369
ɸ	0	1291	500	0
ɾ	71621	3142	1026	0
t	67294	3576	1176	0
u	50871	3297	1102	525
uː	0	533	158	152
v	0	0	0	257
w	28561	1015	397	0
ʔ	0	0	0	479

Table 2: Phonemes counted in single speaker datasets. Low phoneme counts, between 0 and 100, are highlighted on a linear scale.

The data in Table 2 demonstrates a significant overlap between the phonemic sounds of the three languages. The short vowels (listed in Table 1) are represented in all three languages. However, this cannot be said for the long vowels (also listed in Table 1). We have hypothesised that the /ː/ would be sufficiently modelled by the Polynesian data as a lengthening of the short vowel. Similarly, the consonants /k/, /m/, /n/ and /p/ are found in all three datasets. The /w/ sound is present in all datasets; however, due to our decision to represent this sound with a /v/ (Section 2.2), this is not listed in the table and won't therefore be subject to knowledge transfer from the Spanish or te reo Māori models.

Of the phonemes listed in Table 2, there are a total of 2 phonemes from ʻōlelo Hawaiʻi that aren't represented at all in the other datasets: /ʔ/, the ʻōkina or glottal stop and the /v/ sound. For te reo Māori, only the /ɸ/ sound is not found in the other datasets.

Despite significant overlap of /f/ and /w/ across the datasets, we chose to phonemise 'wh' in te reo Māori as /ɸ/ rather than /f/, and 'w' in ʻōlelo Hawaiʻi as /v/. Our goal was to train these specific sounds from the Polynesian data only; however, fine-tuning /f/ and /w/ may produce improved results, which will be the subject of future experiments.

2.4 Data Curation

Four datasets were used in the work. Public domain single-speaker data in Spanish and data recorded specifically for this project in te reo Māori and ʻōlelo Hawaiʻi. Table 3 summarises the length of each dataset in minutes and the source of the data. We used approximately 99.4 hours of a male Spanish voice, 5.5 hours of a male Māori voice, 2.4 hours of a female Māori voice and 58 minutes of a female voice speaking ōlelo Hawaiʻi.

We obtained single-speaker Spanish data from the public domain via LibriVox (LibriVox, 2005).

The female te reo Māori data was sourced from recordings made by Pae Tū Ltd, specifically for this work. These recordings were performed in a recording studio by the co-author and broadcaster Stephanie Huriana Fong and sound engineer Ed Waaka.

The male te reo Māori data was sourced from recordings made by Te Hiku Media from recorded interviews of, and readings by, broadcaster and co-author Peter-Lucas Jones in our radio studios.

The ʻōlelo Hawaiʻi data was carefully curated, prepared, read and recorded by co-author, Hina Puamohala Kneubuhl of Awaiaulu, Inc.

Our Data Team at Te Hiku Media curated and prepared the data for readings in te reo Māori. This team also performed quality checks of transcripts in both te reo Māori and ʻōlelo Hawaiʻi to ensure that they match the audio recorded. Each utterance was reviewed by two independent reviewers.

2.5 The Acoustic Model

After evaluating several models for this task (see Section 1), we followed the example set by coqui.ai (Coqui, 2020) and Nabu Casa choosing the end-to-end VITS-based model, specifically the Piper TTS (Hansen, 2023) training framework which was designed to target the Raspberry Pi 4, and supported by Nabu Casa (Nabu Casa, 2019). Given that a low-powered edge device is the target, we chose the x-low model which uses knowledge distillation to compress a VITS model to 7.07M 32-bit floating-point parameters and uses a 256-character alphabet.

Dataset	Minutes	Source
spa_male	5,966.22	LibriVox
mri_female	146.08	Pae Tū Ltd.
mri_male	333.17	Te Hiku Media
haw_female	58.36	Awaiaulu, Inc.

Table 3: The number of minutes in and the source of each dataset

2.6 The Training Process

The models were trained in a Kubeflow pipeline developed for our NVIDIA A100 servers. We chose to train on a single GPU with 80GB of GPU memory. Due to the end-to-end nature of the VITS model, the pipeline is of relatively simple linear design with fetch, data preparation, training and publishing components.

Table 4 summarises the four training phases performed to produce the two te reo Māori and the ʻōlelo Hawaiʻi models and the number of epochs trained at each stage. The ordering of the training runs determines the direction of knowledge transfer. For example, the te reo Māori models reused knowledge of Spanish phonemes, while the ʻōlelo Hawaiʻi model in turn reused knowledge of te reo Māori.

Training Phase	Dataset	Epochs
1. Initial train	spa_male	157
2. Fine-tune	mri_male	9304
3. Fine-tune	mri_female	10539
4. Fine-tune	haw_femle	10000

Table 4: The training phases

3. Trials on an Edge Device

To trial the male te reo Māori voice we worked with TalkLink Trust a provider of technology solutions to the non-speaking community. They provided us with an Accent 1000 device from PRB-Satillo running Windows 11 and the NuVoice software for non-speaking users.

The VITS model, a PyTorch implementation, was converted to an optimised onnx model of approximately 20 MB. The model was wrapped in a C interface to the onnx runtime version 1.16 and C++ interfaces to the Windows SAPI version 5.4 interface. An installer was also developed to register the resulting library and the te reo Māori voice with the operating system and the SAPI engine. We installed this to the Accent 1000 and provided the voice to TalkLink for testing with the NuVoice software.

We measured the real-time factor of synthesis (synthesis_duration / audio_duration) on the Accent 1000 as being approximately 0.5. The library synthesises per sentence, which allows it to maintain the prosodic elements of speech; however, this impacts response time when synthesising longer sentences.

4. Initial Findings

The te reo Māori models went through an initial qualitative assessment with attention to special cases in pronunciation that were not captured in the phonemisation. A detailed analysis of the ʻōlelo Hawaiʻi model has yet to be performed.

4.1 General Comments

The quality of the female te reo Māori and ʻōlelo Hawaiʻi models demonstrate clear pronunciation with some glitching where sentences are not correctly terminated with punctuation. Adjacent punctuation generates noise which may be attributed to some recordings of the male Māori speaker made outside of the studio and indicates that these recordings should be removed from the dataset. We observed some cases where the male māori speaker does not pronounce the ‘r’.

As we have observed good performance with the female Māori speaker, whose voice is fine-tuned on the Māori male voice using high-quality studio recordings, we believe more and better (studio) quality data of the Māori male voice will resolve these issues. Alternatively, fine-tuning the female Māori voice with the male Māori voice data only recorded in the studio may also resolve some of these issues.

An initial review of the ʻōlelo Hawaiʻi model returned positive results, however, a reviewer noted that the \l\ seemed overly elongated and the emphasis on some three and four-syllable words was not in the correct place, reflecting a more te reo Māori pronunciation than a ʻōlelo Hawaiʻi pronunciation.

4.2 “Whakairo”

The word “whakairo” (“to carve”) is composed of the prefix “whaka” and the noun “iro”, the joining of which builds the diphthong “ai” with emphasis on (in bold) “whakairo” and a corresponding shortening of the diphthong (Harlow, 2015). A contra example is captured by the word “whakairi” (“to hang”), here the emphasis is (in bold) “whakairi”.

Despite our phonemiser not explicitly accounting for the variation in pronunciation observed in “whakairo” and “whakairi” as spoken by the voice artists, both female and male te reo Māori models have successfully learnt this difference from the data.

4.3 “[k]i a ia”

The Māori grammar requires that the particle “a” is placed before proper nouns and pronouns in many situations. The pronunciation of this particle lengthens and is emphasised when placed before the pronoun “ia” (“she / he / it”) or “koe” (“you” - singular) (Biggs, 1998). The female Māori model has learnt this contextual difference in the pronunciation from the data. The male model also demonstrates this pronunciation; however, the male model did not lengthen or emphasise the “a” in “I a ia” when placed at the beginning of the synthesised text.

4.4 “Ta”, “te”, “to” vs “ti”, “tu”

In general, the pronunciation of the consonant ‘t’ in te reo Māori changes depending on the vowel that follows, this is a consequence of a slightly different tongue position in the case of “ta”, “te” and “to” vs the tongue position when pronouncing “ti” and “tu” (Harlow, 2015). Additionally, there are slight variations on this depending on the region from which the speaker comes.

Both the female and male Māori models have learnt this difference from the data, further to this, there is a slight variation in tongue position used in the region from where the male speaker comes, this is also audible in the synthesised recordings.

5. Discussion

Through this work we have demonstrated that it is possible to train a 7M parameter TTS model to generate te reo Māori and ʻōlelo Hawaiʻi that runs on a Windows-based edge device for assistive technologies, the Accent 1000. This allows non-speaking and low-vision users from these language communities the opportunity to hear, for the first time, their own language expressed on these devices.

The initial qualitative findings demonstrate that the female te reo Māori model has good pronunciation of te reo Māori and is able to simulate key features of pronunciation that differentiate native from non-native speakers. This is despite having only 146 minutes of recordings for this voice. This demonstrates the benefits of transfer learning to fine-tune a TTS for an under-resourced language, in this case, transfer learning from over 99 hours of a Spanish voice and 5.6 hours of a male te reo Māori voice, languages with a significant overlap in phonemic content. The male te reo Māori voice demonstrated some anomalies which may be alleviated by better cleaning of the data.

Similarly, although a detailed analysis of ʻōlelo Hawaiʻi is to be performed, the model was positively received with specific comments around an elongated \l\ and incorrect emphasis on some three and four-syllable words, both of which may be due to the influence of transfer learning from Spanish and te reo Māori models. As less than 60 minutes of ʻōlelo Hawaiʻi were used to fine-tune the voice, we believe that, with additional data, these issues can be resolved.

Despite still being a work in progress, we believe that these models for te reo Māori and ʻōlelo Hawaiʻi could be of use to the wider Pacific community. The models produced here demonstrate how transfer learning from one central-eastern Polynesian language can be used to create a voice with a minimal amount of data from another language within the same family. Given that all Polynesian languages are under-resourced, models such as those produced in this work could form a basis for using transfer learning to fine-tune other central-eastern, eastern, and perhaps even wider Polynesian languages.

6. Conclusion

Inspired by two people from the non-speaking and low-vision communities who wish to have text-to-speech technology for te reo Māori and ʻōlelo Hawaiʻi, we created three synthetic voices using the VITS model and the Piper TTS training pipeline. We used public domain Spanish recordings to create a base model which we then fine-tuned for te reo Māori and ʻōlelo Hawaiʻi based on the high intersection of common IPA phonemes between the three languages. We developed tools to deploy these voices to edge devices running the Windows operating system and demonstrated usable real-time performance on an Accent 1000, assistive technology device. We analysed the performance of the synthetic voices and found that the female Māori voice fulfils our qualitative criteria, whereas the male Māori voice demonstrates some anomalies that may be alleviated through improvements to data quality.

7. Future Work

Based on the findings from the work we have performed thus far, we see the potential for improvement of the male te reo Māori voice. We believe we can obtain this through either additional training data or through fine-tuning the female Māori voice with only the high-quality portions of the male voice. Further recordings for both the female and male Māori voice are planned which we expect will improve the quality of both voices once added to the training dataset.

The noise produced by adjacent punctuation may be due to low-quality recordings of the male te reo Māori voice being included in the pipeline. Removal of these recordings and subsequent retraining of the model (from the 157th epoch) may resolve these issues.

For the ʻōlelo Hawaiʻi voice, we will work with native speakers to evaluate the model and make improvements if necessary. As less than an hour of data was available at the time of writing, we may need to increase the amount of training data to see improvements.

Further and more thorough testing of all voices is planned including a deeper qualitative analysis of the te reo Māori and ʻōlelo Hawaiʻi voices. We also plan to gather opinion scores from native speakers to assess the overall quality and acceptance of the voices.

While deployment to the edge device demonstrated a reasonable response time, due to the synthesis of speech at the sentence level, longer sentences can result in an unreasonable delay. As such we plan to investigate implementing synthesis at the sub-sentence level.

Although we have developed tools for Windows devices, many users in the non-speaking and low-vision communities rely on Apple's MacOS or iOS software. Unfortunately, neither of these operating systems allows for easy extension of their voice libraries, which means those wishing to introduce a voice to the Apple ecosystem must either engage directly with each of the existing producers of assistive software or build their own assistive technology

One important consideration is that virtually all speakers of te reo Māori and ʻōlelo Hawaiʻi are at least bilingual, speaking English as well. Given the need to communicate in both languages in a day-to-day context, it would be advantageous for users to be able to express themselves in both languages without having to switch voices. As such, we are designing a bilingual speech package that can be deployed to an edge device as a single voice. This will involve implementing reliable language detection for te reo Māori, ʻōlelo Hawaiʻi and English that is capable of distinguishing the language of words that appear in two or all languages e.g. "one" which is the number 1 [ˈwʌn] in English, but means "sand" in both ʻōlelo Hawaiʻi and te reo Māori.

8. Bibliographical References

Beliaev, S., & Ginsburg, B. (2021). Talknet 2: Non-autoregressive depth-wise separable convolutional model for speech synthesis with explicit pitch and duration prediction. *arXiv preprint arXiv:2104.08189*.

Biggs, B. (1998). Let's learn Māori: A guide to the study of the Māori language. Auckland University Press.

Byrum, J. D. (1999). ISO 639-1 and ISO 639-2: International Standards for Language Codes. ISO 15924: International Standard for Names of Scripts.

Chevi, R., Prasojo, R. E., Aji, A. F., Tjandra, A., & Sakti, S. (2023, January). Nix-TTS: Lightweight and end-to-end text-to-speech via module-wise distillation. In *2022 IEEE Spoken Language Technology Workshop (SLT)* (pp. 970-976). IEEE.

Coqui. (2020). VITS - TTS 0.22.0 documentation. Docs.coqui.ai. Retrieved November 2, 2023, from https://docs.coqui.ai/en/latest/models/vits.html

Duddington, J., Dunn, R. H. (2015) *GitHub - espeak-ng/espeak-ng: eSpeak NG is an open source speech synthesizer that supports more than hundred languages and accents.* GitHub. https://github.com/espeak-ng/espeak-ng/

Hansen, M. (2023, January 11). rhasspy/piper. GitHub. https://github.com/rhasspy/piper

Harlow, R. (2007). *Maori: A linguistic introduction*. Cambridge University Press.

Harlow, R. (2015). A Māori reference grammar. Huia Publishers.

James, J., Shields, I., Berriman, R., Keegan, P. J., & Watson, C. I. (2020). Developing resources for te reo Māori text to speech synthesis system. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23* (pp. 294-302). Springer International Publishing.

Kawamura, M., Shirahata, Y., Yamamoto, R., & Tachibana, K. (2023, June). Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time fourier transform. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

Kim, J., Kong, J., & Son, J. (2021, July). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning* (pp. 5530-5540). PMLR.

Łańcucki, A. (2021, June). Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6588-6592). IEEE.

Lee, Y., Shin, J., & Jung, K. (2020, October). Bidirectional variational inference for non-autoregressive text-to-speech. In *International conference on learning representations*.

Luo, R., Tan, X., Wang, R., Qin, T., Li, J., Zhao, S., ... & Liu, T. Y. (2021, June). Lightspeech: Lightweight and fast text to speech with neural architecture search. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5699-5703). IEEE.

Nabu Casa Inc. (2019). Nabu Casa. Nabu Casa. https://www.nabucasa.com/

Nguyen, H. K., Jeong, K., Um, S. Y., Hwang, M. J., Song, E., & Kang, H. G. (2021, August). LiteTTS: A Lightweight Mel-Spectrogram-Free Text-to-Wave Synthesizer Based on Generative Adversarial Networks. In *Interspeech* (pp. 3595-3599).

Parker Jones, ʻŌ., Niebuhr, O., & Ward, N. G. (2018). Hawaiian. *Journal of the International Phonetic Association*, *48*(1).

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Ren, Y., Liu, J., & Zhao, Z. (2021). Portaspeech: Portable and high-quality generative text-to-speech. *Advances in Neural Information Processing Systems*, *34*, 13963-13974.

Schröder, M., & Schröder, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, *6*, 365-377.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4779-4783). IEEE.

Te Hiku Media. (2022, September). *Natural Language Processing Tools for te Reo Māori* [Review of *Natural Language Processing Tools for te Reo Māori*]. Papa Reo; Te Hiku Media. https://papareo.io/

Vainer, J., & Dušek, O. (2020). Speedyspeech: Efficient neural speech synthesis. *arXiv preprint arXiv:2008.03802*.

9. Language Resource References

LibriVox. (2005). Free public domain audiobooks read by volunteers from around the world. LibriVox. Retrieved November 2, 2023, from https://librivox.org/

Moseley, C. (2012). *The UNESCO atlas of the world's languages in danger: Context and process*. World Oral Literature Project.