Development of Community-Oriented Text-to-Speech Models for Māori ‘Avaiki Nui (Cook Islands Māori)

Jesin James[1], Rolando Coto-Solano[2], Sally Akevai Nicholas[1], Joshua Zhu[1], Bovey Yu[1], Fuki Babasaki[1], Jenny Tyler Wang[1], Nicholas Derby[2], Jean Tekura Mason[3]

[1] The University of Auckland (Waipapa Taumata Rau), [2] Dartmouth College,
[3] Cook Islands Library & Museum Society
{jesin.james, ake.nicholas, jzhu709, byu116, fbab920, jwan553}@auckland.ac.nz
{rolando.a.coto.solano, nicholas.s.derby.24}@dartmouth.edu
library@cookislandsmuseum.com

Abstract

In this paper we describe the development of a text-to-speech system for Māori ‘Avaiki Nui (Cook Islands Māori). We provide details about the process of community-collaboration that was followed throughout the project, a continued engagement where we are trying to develop speech and language technology for the benefit of the community. During this process we gathered a group of recordings that we used to train a TTS system. When training we used two approaches, the HMM-system MaryTTS (Schröder et al., 2011) and the deep learning system FastSpeech2 (Ren et al., 2020). We performed two evaluation tasks on the models: First, we measured their quality by having the synthesized speech transcribed by ASR. The human produced ground truth had lower error rates (CER=4.3, WER=18), but the FastSpeech2 audio has lower error rates (CER=11.8 and WER=42.7) than the MaryTTS voice (CER=17.9 and WER=48.1). The second evaluation was a survey amongst speakers of the language so they could judge the voice's quality. The ground truth was rated with the highest quality (MOS=4.6), but the FastSpeech2 voice had an overall quality of MOS=3.2, which was significantly higher than that of the MaryTTS synthesized recordings (MOS=2.0). We intend to use the FastSpeech2 model to create language learning tools for community members both on the Cook Islands and in the diaspora.

Keywords: Speech synthesis, Text-to-speech, Cook Islands Māori

1. Introduction

Text-to-speech synthesis, or TTS, is used to transform an input string of text into a synthesized voice for a specific language. These systems are helpful as an assistive technology (Kewley-Port and M Nearey, 2020), as a tool for language learning (Cardoso et al., 2015), and as a way to add entertainment value to existing software such as chat agents and games (Cohn et al., 2019; Mayor et al., 2011). The quality of these synthetic voices has increased with the adoption of neural techniques in recent years (Vaswani et al., 2017), to the point where it is possible to make natural-sounding voices out of relatively small data sets. This allows for the expansion of TTS technology to under-resourced languages.

In this paper we describe the creation of a TTS system for Māori ’Avaiki Nui, also known as Cook Islands Māori (glottolog `raro1241`). This language, which we will refer to as CIM, is spoken by approximately 12500 people in the Cook Islands (Ministry of Finance and Economic Management, Government of the Cook Islands, 2021), and an additional 10,000 amongst the diaspora in Aotearoa New Zealand and Australia (Nicholas, 2018). CIM is an East Polynesian language, Indigenous to the Realm of New Zealand. The language is *endangered* in the main island of Rarotonga, which means most of the children do not speak the language and that the chain of intergenerational transmission is being broken. The language enjoys better health in the outer islands of the archipelago, where it is *vulnerable*. This means that most, but not necessarily all, children speak the language and that intergenerational transmission of the language still exists.

1.1. TTS in Indigenous Languages

Implementing natural language processing tools for Indigenous languages is a helpful step in aiding their revitalization and normalization. This is because they allow the language to extend to new domains of usage (Fishman, 2012). For example, a tool as simple as a virtual keyboard might help speakers write text messages in their language, thereby extending the language beyond traditional environments, and taking it to where younger people lead their lives. Other tools, such as speech recognition and parsing, can help create algorithms that understand the under-resourced language, thereby creating a “symbolic impact" (Galla, 2016) and impressing upon younger members of the community the idea that the language can be useful in every sphere of their lives. TTS systems in par-

LREC-COLING 2024, pages 4820–4831
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

ticular would be useful because of their potential as a teaching tool: They can provide a means for learners to listen to example words and sentences in the language, for example in a digital dictionary. It could also allow learners to have a conversation in their Indigenous language with a computer even when no speakers are available. In summary, NLP tools have the potential to bring these smaller languages into the digital domains where people lead their lives, and creating more opportunities for their everyday use.

The main challenge of course is that Indigenous languages often have few reliable datasets that could be used to train NLP tools, let alone TTS voices. Most languages have no available datasets (Joshi et al., 2020), and for many that do, lack of standardization and noise in the data is a major obstacle to their use in NLP tasks that are helpful to the community. Large existing datasets like the CMU-Wilderness dataset (Black, 2019) usually base their transcriptions on orthographies used for Bible making, with little consultation with the community and with unrepresentative corpora. This makes them unusable when trying to develop tools that will work for most community members. One solution would be for community members to create their own datasets, but this process is usually time-consuming and expensive, given the fact that there might be few people who can consistently transcribe the language and who have the time to do so. (Most people who have this knowledge are school teachers, who already are incredibly busy supporting their communities).

Despite these difficulties, there are a few Indigenous languages for which TTS has been developed. These include languages in Canada like Ojibwe (Hammerly et al., 2023; Pratap et al., 2023), Plains Cree (Harrigan et al., 2019), Nakyen'kéha/Iroquoian, Gitksan and SENĆOŦEN/Saanich (Pine et al., 2022). TTS has also been developed for Cherokee (Conrad, 2020), Navajo (Sproat and Shih, 1997), Quechua (Zevallos, 2022; Zevallos et al., 2022), Rarámuri (Urrea et al., 2009), Sámi languages (Hiovain-Asikainen and Moshagen, 2022; Makashova, 2021) and Kalaallisut (Oqaasileriffik, 2020). As for the languages of Polynesia, there is work on TTS systems for te reo Māori (James et al., 2020; Laws, 2003; Shields et al., 2019).

1.2. NLP for Cook Islands Māori

There has been previous work on natural language processing for CIM. There are both statistical and deep-learning based models for CIM speech recognition (Foley et al., 2018; Coto-Solano et al., 2022a), as well as work on untrained forced alignment at the phoneme level (Nicholas and Coto-Solano, 2019; Coto-Solano et al., 2022b). There is also work on parsing using Universal Dependencies (Karnes et al., 2023) and part-of-speech tagging (Coto-Solano et al., 2018).

2. Methodology

In this paper we will test two types of TTS models. The first one will be the HMM-based MaryTTS (Schröder et al., 2011). Given the data limitations for CIM, it is possible that the older probability-based methods might have better performance with this smaller dataset. The second method we will use is the deep learning-based FastSpeech2 (Ren et al., 2020). This system uses an encoder-decoder architecture to transform strings into synthesized spectrograms. There is a growing body of evidence that low-resource conditions can be successfully modeled by neural methods, and therefore we will use this for the CIM data.

In order to evaluate the performance of the models, we conducted two experiments. First, we evaluated the intelligibility of the voice using automatic speech recognition. This provides us with a quantitative measure of the difference between a human's voice and the synthesized samples. In the second experiment, we carried out a survey where we asked speakers of CIM to evaluate both human-produced and synthetic samples and to report on the quality of the voices.

2.1. Data preparation

In order to train the TTS models we used 1.5 hours of transcribed CIM speech from a single speaker of CIM, a resident of Rarotonga with roots in Ngā Pū Toru (Ma'uke island of the Cook Islands archipelago) (There is more information about the speaker in Section 4.2 below). These recordings were transcribed according to the orthography in (Nicholas, 2018), with a first pass by students of CIM, and a second pass by an expert speaker and writer of CIM. The transcriptions were time-aligned at the phrase level.

After this, we created a dictionary with the words of the corpus. This dictionary contained the individual words, and a mapping of the words into a phoneme representation. Table 1 shows examples of words in CIM and their corresponding mappings.

Word	Phonemes	English
tēta'i	t eː t a ʔ i	one
runga	ɾ u ŋ a	*top*
mānga	m aː ŋ a	*a little bit*

Table 1: Example of CIM words and their mapping to phonemes (only the first two columns are included in the dictionary)

Using this dictionary and the phrase-level alignments, we created transcriptions that were aligned at the phoneme level. These were formatted as Praat TextGrids (Boersma and Weenink, 2023), and made using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017).

2.2. Text-to-Speech Models

The phoneme-level transcriptions were used as input for the training of both TTS algorithms. The first one used was MaryTTS (Schröder et al., 2011). This is based on Hidden Markov Models (HMMs): It extracts the probabilities of transition between phone n-grams, and it uses the words in the corpus to calculate the transition probabilities between word n-grams. Using these two sets of probabilities, the model attempts to assign the transcription that has the highest probability of matching the speech signal. This method is popular with low-resource languages because these probabilities can be calculated for very small quantities of data, without having to calculate the large number of connections in a neural network. The data was randomly split into 97.1% training and 3.9% validation sets. This training took approximately one hour of processing time. The hyperparameters for MaryTTS are in Appendix 1.

The second system used was FastSpeech2 (Ren et al., 2020). This algorithm uses deep learning to transform a string of phonemes into a waveform. It uses an encoder to encode phoneme embeddings, and has an additional module to predict pitch, energy and phoneme duration, potentially making recordings sound more natural. This data was also randomly split into 97.1% testing and 3.9% validation sets. It took approximately 10 hours to train a full model[1]. The hyperparameters for the model are in Appendix 2.

2.3. Evaluation

We used both the models from both the MaryTTS and the FastSpeech2 algorithms to create a synthetic version of 28 sentences in CIM. These were accompanied by their ground truth audio (a version of the phrase spoken by the same person we used for the training data), and a target transcription of the ground truth, made by an expert writer of CIM.

We ran all the utterances through an ASR model for CIM (Coto-Solano et al., 2022a). The ASR model was trained using four hours of transcribed audio, using the Wav2Vec2 algorithm (Baevski et al., 2020). From this model we got an automated transcription for all the utterances, and then we compared those automated transcriptions with the human expert's target transcription. From this we calculated the character error rate (CER) and the word error rate (WER) of the automated transcriptions for the three types of recordings (human ground truth, sentence synthesized with MaryTTS, and the same sentence synthesized with FastSpeech2).

This first method of evaluation provides us with a quantitative measurement of the clarity of the synthetic recordings. In addition to this, we also wanted to measure the human perception of the synthesized speech. We created an online questionnaire so that speakers of CIM could listen to the three types of recordings and report their opinion across four dimensions: (i) overall quality, (ii) naturalness, (iii) speaking rate and (iv) intelligibility. We asked the participants to listen to 10 recordings for each of the three conditions (ground truth, FastSpeech2, MaryTTS); the recordings were presented in a random order. The participants' answers were measured along a Likert scale from 1 to 5, where 5 was considered “best". A total of 15 speakers of CIM answered the questionnaire. Their ages ranged between 30 and 66 years old, and their fluency levels in CIM ranged from learner to completely fluent.

3. Results

In this section we present the results of the ASR and opinion evaluations of the synthetically generated speech, compared to the human-produced ground truth. In general, the deep-learning based FastSpeech2 system performed better than the HMM-based MaryTTS in both experiments.

3.1. ASR evaluation results

Figure 1 shows a summary of the results; Table 2 shows the averages and standard deviations for the error rates of the two ASR systems, compared to the ground truth.

A statistical experiment confirms that the FastSpeech2 system performs significantly better than MaryTTS. An ANOVA test showed that there were significant differences in the character error rates (F(2,81)=15.5, $p<0.00001$). A Bonferroni-corrected post-hoc test revealed that the ground truth had significantly lower CER when compared to FastSpeech2 ($\mathrm{CER}_{\mathrm{GT}}$ = 4.3, $\mathrm{CER}_{\mathrm{FS2}}$ = 11.8, $p<0.01$) and MaryTTS ($\mathrm{CER}_{\mathrm{MT}}$ = 42.7, $p<0.00001$). However, FastSpeech2 does have a significantly lower error rate than MaryTTS ($\mathrm{CER}_{\mathrm{FS2}}$ = 11.8, $\mathrm{CER}_{\mathrm{MT}}$ = 42.7, $p<0.05$).

These improvements of the FastSpeech2 over MaryTTS were not found for the word error rate. An ANOVA test showed that there were significant differences (F(2,81)=11.7, $p<0.00001$), but

[1] This was trained using an Apollo Workstation with a dual 3090 GPU, 256GB of RAM, and an AMD Threadripper 3970X CPU with 32 cores and a 128MB cache.

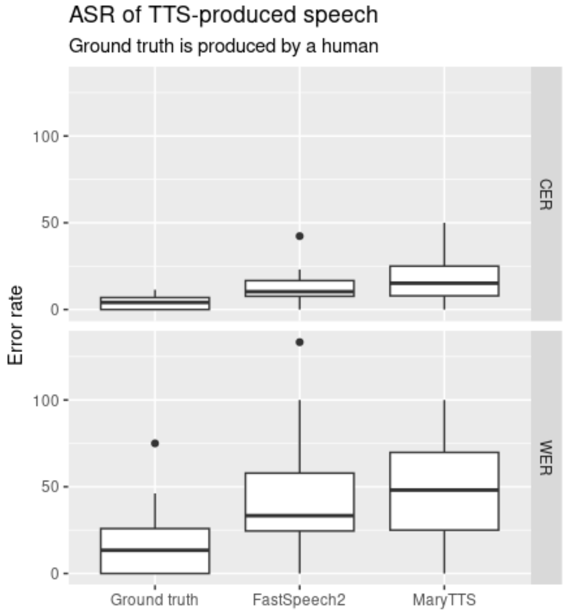

Figure 1: Error rates for speech samples when transcribed by an ASR system

	CER	WER
Ground truth	4.3 ±3.8	17.6 ±18.3
FastSpeech2	11.8 ±8.3	42.7 ±29.9
MaryTTS	17.9 ±12.9	48.1 ±26.6

Table 2: Average and standard deviation for the error rates of the transcription of speech samples

these were present only between the ground truth and each of the training conditions. A Bonferroni-corrected post-hoc test indicated that the ground truth has significantly lower word errors than the speech generated by both the FastSpeech2 and the MaryTTS models (WER$_{\mathrm{GT}}$ = 17.6, WER$_{\mathrm{FS2}}$ = 42.7, WER$_{\mathrm{MT}}$ = 48.1; p$_{\mathrm{GT/FS2}}$=0.005, p$_{\mathrm{GT/MT}}$=0.00001). However, there was no significant difference in the WER when transcribing from the synthetic systems (p$_{\mathrm{FS2/MT}}$=1.0).

Table 3 has examples of transcriptions with high, low, and average error rates for the three conditions.

3.2. Text-to-Speech evaluation results

A total of 15 participants listened to ten samples from each of the experimental conditions (ground truth recordings, speech synthesized from FastSpeech2, and speech synthesized from MaryTTS). They had to provide their opinion about the sample's overall quality, naturalness, speaking rate and intelligibility. This opinion was recorded in a scale from 1 to 5, where 5 is best. Figure 2 summarizes the distribution of answers to the opinion surveys for each of the questions, and table 4 provides the average and standard deviation for the opinion scores.

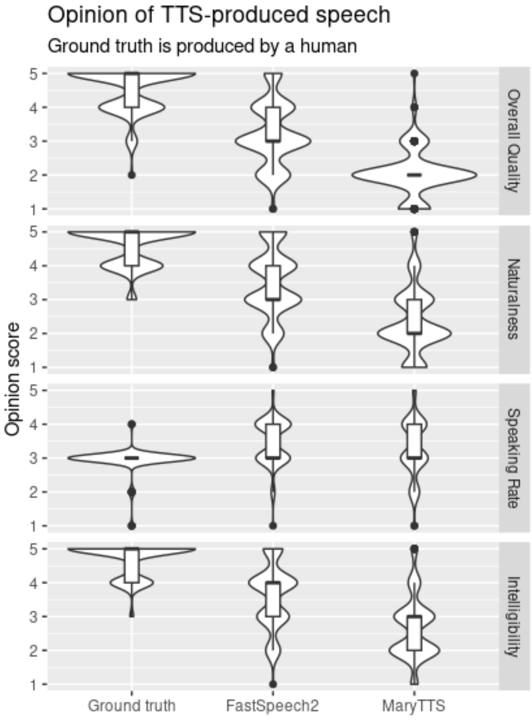

Figure 2: Opinion of recordings in the survey

As the figure shows, the samples synthesized using FastSpeech2 had higher scores than those from MaryTTS in three of the questions (overall quality, naturalness and intelligibility). Let's look first at overall quality. A Kruskal-Wallis test reveals that there are significant differences between the conditions (χ^2(2)=293, p<0.00001). A Bonferroni-corrected post-hoc Dunn test confirmed that the ground truth has significantly higher quality than the synthesized samples (MOS$_{\mathrm{GT}}$=4.6, MOS$_{\mathrm{FS2}}$=3.2, MOS$_{\mathrm{MT}}$=2.0, p$_{\mathrm{GT/FS2}}$<0.00001, p$_{\mathrm{GT/MT}}$<0.00001), and that the FastSpeech2 samples have significantly higher quality than the MaryTTS samples (p$_{\mathrm{FS2/MT}}$<0.00001).

FastSpeech2 also performed better in naturalness of speech. A Kruskall-Wallis test was used to measure the difference between the conditions. There were significant difference between them (χ^2(2)=241, p<0.00001). A Bonferroni-corrected post-hoc Dunn test found that, while the ground truth has significantly higher naturalness values (MOS$_{\mathrm{GT}}$=4.6, MOS$_{\mathrm{FS2}}$=3.5, MOS$_{\mathrm{MT}}$=2.4, p$_{\mathrm{GT/FS2}}$<0.00001, p$_{\mathrm{GT/MT}}$<0.00001), the FastSpeech2 samples have significantly higher naturalness scores than those produced by MaryTTS (p$_{\mathrm{FS2/MT}}$<0.00001).

The CIM utterances synthesized by FastSpeech2 also scored higher in the intelligibility scores. A

	High (worst) WER		
English	*He is coming to Motutapu and Tinirau.*		
Target	tē 'aere mai ra a kae ki motutapu ki a tinirau	WER	CER
Ground truth	tē 'aere mai ra akae ki mutu tapu ki a tinirau	36	7
FastSpeech2	tē 'aere mairaka ē ki motu tapu ki atini rau	73	17
MaryTTS	ē 'are maika a kae ki motū taku ki ei atini rava	82	28

	Average WER		
English	*The change to English in Aitutaki is happening.*		
Target	tē 'akamata nei te neke'anga ki kō i te reo papa'ā ki aitutaki	WER	CER
Ground truth	tē 'akamatanei te neke'anga ki kō i te reo papa ā ki a 'itutaki	46	7
FastSpeech2	tē 'akamata nei te neke'anga ki kō i te reo papa akia i tūtaki	31	11
MaryTTS	tēi 'akamata nei tenekē'anga kia ō i te reo papa aki a tūptaki'i	69	21

	Low (best) WER		
English	*Mana is on the island.*		
Target	tei runga a mana i tēta'i motu	WER	CER
Ground truth	tei runga a mana i tēta'i motu	0	0
FastSpeech2	tei runga mana i tēta'i motu	14	7
MaryTTS	tei runga a mana i tēta'i motu	0	0

Table 3: ASR examples for utterances from two TTS models and their corresponding human-uttered ground truth

	GT	FS2	MT
Overall quality	4.6 ±0.6	3.2 ±0.9	2.0 ±0.7
Naturalness	4.6 ±0.6	3.5 ±1.0	2.4 ±1.0
Speaking rate	2.9 ±0.4	3.5 ±0.6	3.3 ±0.8
Intelligibility	4.7 ±0.5	3.6 ±0.9	2.6 ±0.9

Table 4: Average and standard deviation of opinion scores for the three experimental conditions (GT: ground truth, FS2: FastSpeech2, MT: MaryTTS) and the four questions in the survey.

Kruskall-Wallis test determined that there were differences between the conditions (χ^2(2)=241, p<0.00001). Again, the ground truth was rated as having significantly higher intelligibility according to the Bonferroni-corrected Dunn test (MOS$_{\mathrm{GT}}$=4.7, MOS$_{\mathrm{FS2}}$=3.6, MOS$_{\mathrm{MT}}$=2.6, p$_{\mathrm{GT/FS2}}$<0.00001, p$_{\mathrm{GT/MT}}$<0.00001), and FastSpeech2 was rated as more intelligible than MaryTTS (p$_{\mathrm{FS2/MT}}$<0.00001).

The one question where this behavior was different was in speech rate. The Kruskall-Wallis test detected significant differences between the conditions (χ^2(2)=59, p<0.00001), but the Bonferroni-corrected Dunn test only detected significant differences between the ground truth and the two synthetic conditions (MOS$_{\mathrm{GT}}$=2.9, MOS$_{\mathrm{FS2}}$=3.5, MOS$_{\mathrm{MT}}$=3.3, p$_{\mathrm{GT/FS2}}$<0.00001, p$_{\mathrm{GT/MT}}$<0.00001). The samples from FastSpeech2 and MaryTTS had very similar scores (3.5 versus 3.3), and these were not significantly different (p=0.07).

Finally, we were interested in measuring the reliability of the survey's answers. In order to measure this we used the Kendall's Coefficient of Concordance W for ordinal categorical data. The results show that the raters show substantial and significant agreement in their answers (15 raters, W=0.66, χ^2(119)=1184, p<0.00001).

4. Discussion

4.1. ASR Training Results

The results above confirm that a deep-learning TTS model can be trained for CIM, and that its performance is better than the equivalent HMM-based model. Despite the small amount of data available, the FastSpeech2 CIM model performs similarly to other low-resource neural TTS models (Lam et al., 2022; Xu et al., 2020; Guo et al., 2022; Nguyen et al., 2022; Tu et al., 2019), with MOS rates of around 3.5.

Why did the audio synthesized with FastSpeech2 have better CER than the MaryTTS utterances? Figure 3 shows spectrograms for the ground truth and the synthesized recordings of the phrase “mei Aotearoa ē ’Autirēria" *from New Zealand and Australia*. The FastSpeech2 synthesised speech has a higher degree of resemblance to the human voice when it comes to phonetic features. For example, in the FastSpeech2, the transition between the vowels of “mei" and the start of “ao" is fluid and without breaks. In the MaryTTS recording, this transition has a abrupt break between the two words. Likewise, in the ground truth, the transition between the end of “Aotearoa" and the word “ē" *and* is rela-

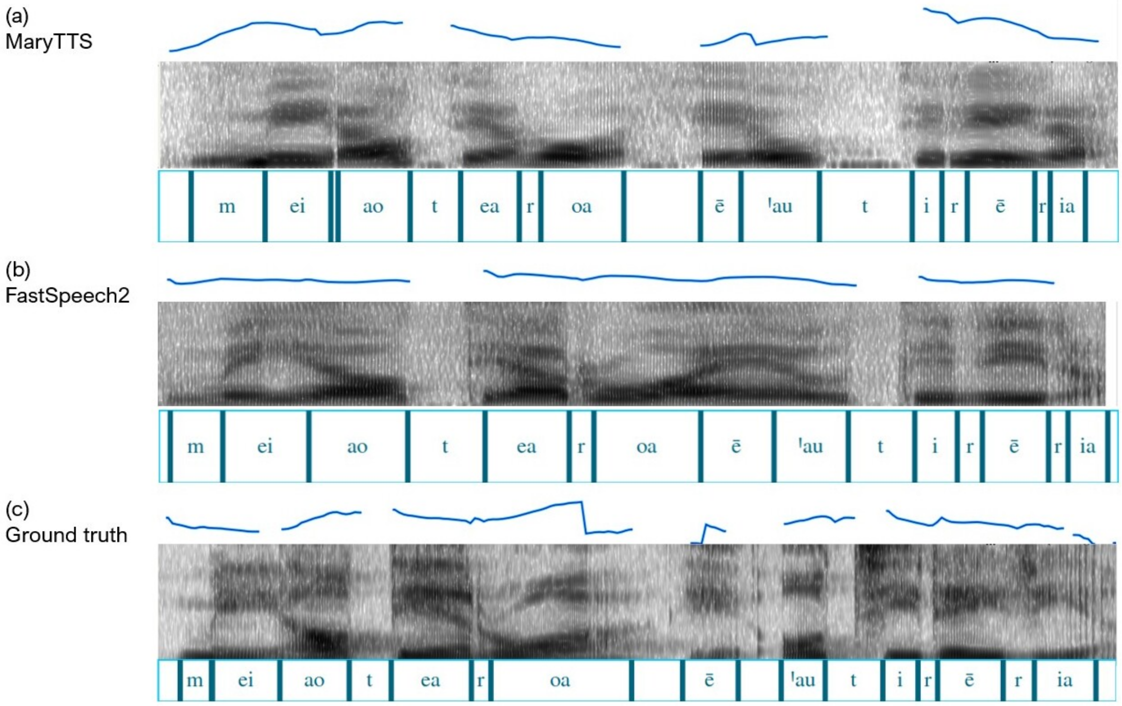

Figure 3: Spectrograms and pitch contours for the phrase “mei Aotearoa ē ’Autirēria" *from New Zealand and Australia*.

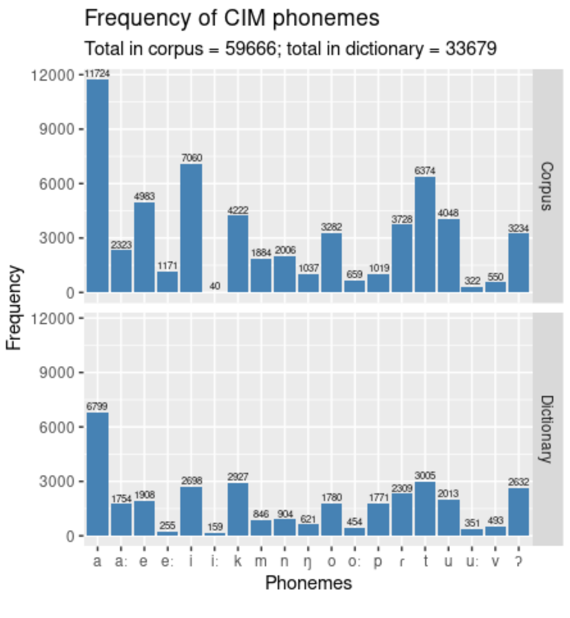

Figure 4: Frequency of phonemes in the corpus and the dictionary.

tively fluid, carried on by creaky voice between the two words. This transition is clearly present in the FastSpeech2 (albeit without the creaky voice), but, in the MaryTTS version, there is a clear and long cut from one word to the next. Another phonetic example is the air release of the aspirated /t/, which is clearer and stronger in the FastSpeech2. The /t/ in the word *Australia*, in both the ground truth and FastSpeech2, has a distinct explosion which is only faintly visible in the MaryTTS version. This is also the case with the flap /ɾ/. In the word *Australia*, the first flap is clearly voiced and has a relatively long occlusion (45 ms for FastSpeech2 and 44 ms for the ground truth), whereas the occlusion is relatively short for the MaryTTS voice (10 ms).

Given that FastSpeech2 performed better with CER, why did it have no gains in the WER performance? When transcribed using ASR, FastSpeech2 had WER=43 and MaryTTS had WER=48, a difference that was not statistically significant. It is possible that the more severe breaks between words that are visible in the MaryTTS did not degrade the ASR’s capacity to identify individual words, despite the phonetic issues in the HMM-based utterances.

There is an important pattern than should be kept in mind when generating CIM phrases using these models: Sounds with few tokens in the training set might have worse performance. Figure 4 shows the frequency of the CIM phonemes in the corpus used to train the TTS, as well as in a dictionary composed of all the unique words in the larger, Te Vairanga Tuatua corpus (Nicholas, 2018). It shows that there are almost no long /i:/ phonemes in the dataset. It appears only 40 times in the corpus, less than 0.01% of the total, and it appears only 159 times in the dictionary, about 0.05% of the sounds in the dictionary. This leaves the system vulnerable to misproducing this sound, particularly in the HMM-based systems, which would not be able to use the attention mechanism (Vaswani et al., 2017) to understand these sounds in context. This problem might not severely affect the other long sounds when synthesized using the deep learning system. The examples in figure 3 have a long /e:/ for the word *and*. This word has a duration of 128 ms in the ground truth and 131 ms in the FastSpeech, both of them very similar, but its duration is only 80ms in the MaryTTS.

The FastSpeech2 model was also perceived as being significantly better than MaryTTS. As explained above, there are a number of phonetic features that are more natural in the FastSpeech2 recording. In addition to this, figure 3 also shows the pitch contours for the words. Even though the FastSpeech2 has a flatter intonation, it also has fewer breaks and less variation between the highest and lowest pitches, making it sound more human-like. For example, in the word *Australia*, the ground truth starts at 158 Hz, goes up to 192 Hz, and ends at 131 Hz, with a difference of 34 Hz between the start and the peak. The FastSpeech2 version of this trajectory is 162-181-159 Hz, with a difference of 19 Hz between start and peak. But for the MaryTTS version, this trajectory is 150-298-162 Hz, with a much larger difference of 148 Hz between the start and the peak of the word.

The FastSpeech2 voice was indeed perceived as being better in three dimensions (quality, naturalness and intelligibility), but not on speaking rate. This might be an artifact of how the recordings were made. The human speaker that made the ground truth recordings regularly works in linguistic documentation, and therefore the person's rhythm might be slower than usual. This is one possible reason for why ground truth receives worse MOS in this rubric. Both of the synthetic voices get similar MOS for speaking rate (3.5 for FastSpeech2, 3.3 for MaryTTS), and these are not statistically different. This might be because the speaking rate is set to default in both of this, and they might both be generating a voice with a neutral speed.

4.2. Community-Oriented Work

The main components of this community-oriented work have been: (i) NLP work and tools that are requested by the community itself, not by outsider researchers, (ii) work that is meant to benefit the community, (iii) long-standing relations and reciprocity between the out- and in-community researchers, and (iv) the obligation that stewardship of data and the resulting models be kept within the community. We will discuss each of these in this section.

One important part of this work is that it was born from the needs of the community, and it has been done in close collaboration with the community. The project started as a part of the Te Vairanga Tuatua (Nicholas, 2012), a project to compile an annotated corpus of CIM. This project started in 2012, and it is led by the third author (Nicholas), a member of the Cook Islands community, Whāngaingia e Taranaki in Aotearoa and Ngā Pū Toru (Ma'uke) in the Cook Islands. The objective of the project has been to collect narration and speech from elders and study the language's grammatical structure from those recordings. During the time of the project, Nicholas has worked with the ninth author (Mason) to document the language. Mason is an L1 speaker of CIM, with roots in Ngā Pū Toru (Ma'uke), and is the director of the Cook Islands Library & Museum Society. Mason is an expert in the history and genealogy of the Cook Islands archipelago, and her contribution to the project has been instrumental.

One byproduct of the project has been the training of an ASR system (Coto-Solano et al., 2022a) to assist in the transcription of the narrations. This was trained by the second author (Coto-Solano), a research of Costa Rican origin who has been working with Nicholas and Mason since 2017. Coto-Solano has collaborated with Nicholas in the creation of different NLP tools (see Section 1.2 above), and has collaborated with the community in general several ways, including teaching linguistics and NLP workshops at the University of the South Pacific in Rarotonga. These workshops are meant to transfer the knowledge of NLP to Cook Islands' programmers, and to train community members so that they conduct the NLP work in the future.

As the work progressed, there were enough recordings of Mason to train a TTS system using her voice samples. Author 1 (James) is an expert in TTS, of Indian origin, and has collaborated with Nicholas since 2019. James has also collaborated with Nicholas on NLP research on te reo Māori, the Indigenous language of Aotearea New Zealand, and James coordinated with Nicholas and Coto-Solano, and led a team of non-community researchers (authors four through eight) to train and evaluate TTS models. The human evaluation of the results (see Section 3.2) was done in collaboration with the University of South Pacific in Rarotonga, a local, Cook Islands academic institution. The final results of the project (i.e. the synthetic voice) have been reviewed by Mason, the community-member that the voice was modelled on.

During the process of training and testing our model, we have been acutely aware of concerns regarding data sovereignty. This is particularly sensitive for Indigenous communities (Kukutai and Taylor, 2016), who regularly see their data used in NLP and get no control over the tools made with their data, and no benefit from the use of said tools. The data and models resulting from this project are under the stewardship of Nicholas, a Cook Islander and a community leader. One part of Nicholas' commitment to this stewardship comes from the ethics permissions approved by the University of Auckland (permit UAHPEC26272) and by the Cook Islands Research Ethics Committee (permit #25/22a). However, the most important part of the stewardship commitment comes from community custom, in particular 'Ākono'anga (guardianship, equivalent to te reo Māori kaitiakitanga). This principle binds Nicholas into a network of long-term responsibilities towards her community. This system of community relations and reciprocity, which are often poorly understood by Institutional Review Boards, are the main set of regulations that the work is bound too.

As for public release of the data and models, we will release the TTS model publicly in the future, bound by the Kaitiakitanga license (Te Hiku Media, 2023). This license allows for non-commercial use in consultation with the Cook Islands community, in particular for applications that benefit the community. As for the training data, it will be deposited in the repository Paradisec (Nicholas, 2012), along with the rest of the Te Vairanga Tuatua materials, where it can be used by community members, or by other who request permission to use it from the community members. This system has been pioneered by entities in New Zealand (Te Reo Irirangi o Te Hiku o Te Ika, 2017), and we hope that it is a step forward in aligning NLP work with the concerns and needs of Indigenous communities.

Our main goal with this tool is to use the TTS voice to create language learning tools oriented towards the community of the Cook Islands. For example, the voice could be used for learning apps, where the example sentences could be read by a synthetic voice. It could also be installed into a chat agent, which could help learners and members of the diaspora learn Cook Islands Māori. Finally, the voice could be used to enrich the existing NLP tools for the language. It could be used for synthetic augmentation of ASR training corpora, and for creating devices that can both listen to CIM and then respond to the user in the language.

Another one of our long-term goals for future work is to collaborate with other Polynesian communities in creating TTS for their languages. Now that the CIM voice is trained, this could be used with transfer learning to train voices for languages with even fewer resources. Other major languages in Polynesia, like Tahitian, Tongan and Sāmoan are also chronically under-resourced and under-served by NLP, and we think that the existing work for CIM could be useful to help kickstart NLP throughout the region.

4.3. Limitations of the work

There are some important limitations in this work: The voice that is being synthesized (Mason's) corresponds to only one dialect of CIM. The language has a different dialect on each island, and there are known phonetic and lexical differences between them (Nicholas, 2018), for example in the realization of the glottal stops (Nicholas and Coto-Solano, 2019). The fact that the only TTS voice available comes from the dialect of Ma'uke might inadvertently help reify Ma'ukean as a de facto standard for NLP, changing the ecological balance of the different varieties of CIM. This is particularly sensitive for vulnerable languages, where the creation of NLP resources might have an oversized effect in the equilibrium between the variants of the language. In this research we are limited by the amount of data available; there is no speaker from the other islands for which have data to train an additional voice at this point. However, this has to be kept in mind for future work.

5. Conclusions and future work

In this paper we described the process of creating a TTS system for Cook Islands Māori. It showed that the training of a synthetic voice using deep learning algorithms in a low-resource environment is not only possible, but that it yields better results that using the older, statistical and HMM-based algorithms. This was verified by speakers of CIM, who listened to both synthetic voices and preferred the one generated with FastSpeech2, the deep learning algorithm used here.

As mentioned in the discussion, our priorities for future work include collaborating with other Polynesian communities to create synthetic voices of their languages, using the CIM voice to create learning materials that the community needs to support its language revitalization process, and integrating it into existing NLP systems for the language. We hope that this paper also serves as proof that, even with relatively little resources, complex NLP applications like text-to-speech are possible for Indigenous languages.

6. Ethical Statement

The main concerns in the project is to ensure that the system was (i) created by request of the Cook Islands community, and in consultation with this community, that (ii) it was a tool that worked for the community's needs, and that (iii) there is proper stewardship of the models and the data. The first concern has been addressed with the community during the ethics consultation and data collection period. The second concern was addressed using the MOS survey, and the third concern is being addressed by using the Kaitiakitanga license. We have also discussed our concerns about reinforcing the position of one dialect over others by training this tool. Finally, we have provided information about the energy usage during the training of the models (11 hours with a dual GPU configuration).

We encourage researchers in NLP to work in collaboration with communities when creating tools, as this helps researchers create tools that will have an actual use in the community and that could potentially have an impact on the vitality of the language.

7. Acknowledgments

This project is supported by the Marsden Fund Council from the Aotearoa New Zealand Government (21-MAU-018), managed by The Royal Society Te Apārangi. The authors would like to thank Debi Futter-Puati and the personnel at the University of the South Pacific Rarotonga for their collaboration during this project. We also thank the anonymous reviewers of the paper for their useful suggestions, as well as the Cook Islands Māori speakers who provided their time and effort for the perceptual evaluation, as well as the other speakers whose voices are recorded in the Te Vairanga Tuatua.

8. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33:12449–12460.

Alan W Black. 2019. CMU Wilderness Multilingual Speech Dataset. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5971–5975. IEEE.

Paul Boersma and David Weenink. 2023. Praat: doing phonetics by computer [Computer program]. Version 6.3.19.

Walcir Cardoso, George Smith, and C Garcia Fuentes. 2015. Evaluating text-to-speech synthesizers. In Critical CALL–Proceedings of the 2015 EUROCALL Conference, Padova, Italy, pages 108–113. Research-publishing. net.

Michelle Cohn, Chun-Yen Chen, and Zhou Yu. 2019. A large-scale user study of an alexa prize chatbot: Effect of tts dynamism on perceived quality of social dialog. In Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, pages 293–306.

Michael Conrad. 2020. Tacotron2 and Cherokee TTS.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022a. Development of automatic speech recognition for the documentation of Cook Islands Māori. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), page 3872–3882. European Language Resources Association.

Rolando Coto-Solano, Sally Akevai Nicholas, Brittany Hoback, and Gregorio Tiburcio Cano. 2022b. Managing data workflows for untrained forced alignment: examples from Costa Rica, Mexico, the Cook Islands, and Vanuatu. The Open Handbook of Linguistic Data Management, 35.

Rolando Coto-Solano, Sally Akevai Nicholas, and Samantha Wray. 2018. Development of natural language processing tools for Cook Islands Māori. In Proceedings of the Australasian Language Technology Association Workshop 2018, pages 26–33.

Joshua A Fishman. 2012. Language maintenance, language shift, and reversing language shift. The handbook of bilingualism and multilingualism, pages 466–494.

Ben Foley, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, et al. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In SLTU, pages 205–209.

Candace Kaleimamoowahinekapu Galla. 2016. Indigenous language revitalization, promotion, and education: Function of digital technology. Computer Assisted Language Learning, 29(7):1137–1151.

Haohan Guo, Fenglong Xie, Xixin Wu, Hui Lu, and Helen Meng. 2022. Towards high-quality neural TTS for low-resource languages by learning compact speech representations. arXiv preprint arXiv:2210.15131.

Christopher Hammerly, Sonja Fougère, Giancarlo Sierra, Scott Parkhill, Harrison Porteous, and Chad Quinn. 2023. A text-to-speech synthesis system for Border Lakes Ojibwe. In Proceedings of the Sixth Workshop on the

Use of Computational Methods in the Study of Endangered Languages, pages 60–65.

Atticus Harrigan, Timothy Mills, and Antti Arppe. 2019. A Preliminary Plains Cree Speech Synthesizer. In Proceedings of the Workshop on Computational Methods for Endangered Languages, volume 1.

Katri Hiovain-Asikainen and Sjur Moshagen. 2022. Building open-source speech technology for low-resource minority languages with Sámi as an example–tools, methods and experiments. In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, pages 169–175.

Jesin James, Isabella Shields, Rebekah Berriman, Peter J Keegan, and Catherine I Watson. 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, pages 294–302. Springer.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. arXiv preprint arXiv:2004.09095.

Sarah Karnes, Rolando Coto, and Sally Akevai Nicholas. 2023. Towards universal dependencies in Cook Islands Māori. In Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 124–129.

Diane Kewley-Port and Terrance M Nearey. 2020. Speech synthesizer produced voices for disabled, including stephen hawking. The Journal of the Acoustical Society of America, 148(1):R1–R2.

Tahu Kukutai and John Taylor. 2016. Indigenous Data Sovereignty: Toward an Agenda, volume 38. Anu Press.

Tuong Q Lam, Dung D Nguyen, Dat T Nguyen, Han K Lam, Thuc H Cai, Suong N Hoang, and Hao D Do. 2022. Instance-based transfer learning approach for Vietnamese speech synthesis with very low resource. In Future of Information and Communication Conference, pages 148–164. Springer.

Mark R Laws. 2003. Speech data analysis for diphone construction of a Maori online text-to-speech synthesizer. In SIP, pages 103–108. Citeseer.

Liliia Makashova. 2021. Speech synthesis and recognition for a low-resource languages - connecting TTS and ASR for mutual benefit.

Oscar Mayor, Jordi Bonada, and Jordi Janer. 2011. Audio transformation technologies applied to video games. In Audio Engineering Society Conference: 41st International Conference: Audio for Games. Audio Engineering Society.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In Interspeech, volume 2017, pages 498–502.

Ministry of Finance and Economic Management, Government of the Cook Islands. 2021. Census 2021: Key findings. https://www.mfem.gov.ck/statistics/census-and-surveys/census/267-census-2021.

Tan Dat Nguyen, Quang Tuong Lam, Duc Hao Do, Huu Thuc Cai, Hoang Suong Nguyen, Thanh Hung Vo, and Duc Dung Nguyen. 2022. A linguistic-based transfer learning approach for low-resource Bahnar text-to-speech. In 2022 9th NAFOSTED Conference on Information and Computer Science (NICS), pages 148–153. IEEE.

Sally Akevai Nicholas. 2012. Te Vairanga Tuatua o te Te Reo Māori o te Pae Tonga: Cook Islands Māori (Southern dialects).

Sally Akevai Nicholas and Rolando Coto-Solano. 2019. Glottal variation, teacher training and language revitalization in the Cook Islands. In Proceedings of the 19th International Congress of Phonetic Sciences, University of Melbourne, Australia, pages 3602–3606.

Sally Akevai Te Namu Nicholas. 2018. Language contexts: Te Reo Māori o te Pae Tonga o te Kuki Airani also known as Southern Cook Islands Māori. Language Documentation and Description, 15:64.

Oqaasileriffik. 2020. Oqaasileriffik - Martha.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022. Requirements and motivations of low-resource speech synthesis for language revitalization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7346–7359.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. arXiv preprint arXiv:2305.13516.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558.

Marc Schröder, Marcela Charfuelan, Sathish Pammi, and Ingmar Steiner. 2011. Open source voice creation toolkit for the MARY TTS platform. In 12th Annual Conference of the International Speech Communication Association-Interspeech 2011, pages 3253–3256. ISCA.

Isabella Shields, Catherine I Watson, Peter J Keegan, Rebekah Berriman, and Jesin James. 2019. Te Reo Māori voice for TTS. In Language Technologies, volume 4.

Richard Sproat and Chilin Shih. 1997. Subject: A Navajo language text-to-speech synthesizer work project no. 311402-2226 file case 60011.

Statistics New Zealand. 2018. 2018 Census totals by topic – national highlights (updated). (StatsNZWebsite).

Te Hiku Media. 2023. Kaitiakitanga License. https://github.com/TeHikuMedia/Kaitiakitanga-License.

Te Reo Irirangi o Te Hiku o Te Ika. 2017. kōreromāori.io. https://koreromaori.io/.

Tao Tu, Yuan-Jui Chen, Cheng-chieh Yeh, and Hung-Yi Lee. 2019. End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. arXiv preprint arXiv:1904.06508.

Alfonso Medina Urrea, José Abel Herrera Camacho, and Maribel Alvarado Garcıa. 2009. Towards the speech synthesis of Raramuri: a unit selection approach based on unsupervised extraction of suffix sequences. Advances in Computational Linguistics, page 243.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2802–2812.

Rodolfo Zevallos. 2022. Text-to-speech data augmentation for low resource speech recognition. arXiv preprint arXiv:2204.00291.

Rodolfo Zevallos, Nuria Bel, Guillermo Cámbara, Mireia Farrús, and Jordi Luque. 2022. Data augmentation for low-resource Quechua ASR improvement. arXiv preprint arXiv:2207.06872.

Appendix 1: Hyperparameters for MaryTTS

Module SnackVoiceQualityProcessor:

```
fftSize 512
frameLength 0.005
lpcOrder 12
maxPitch 400
minPitch 60
numFormants 4
samplingRate 16000
windowLength 0.025
```

Module LabelPauseDeleter:

```
pauseDurationThreshold 100
```

Module HMMVoiceConfigure:

```
fftLen 512
frameLen 400
frameShift 80
freqWarp 0
gamma 0
lf0BandWidth 1
lnGain 1
lowerF0 110
mgcBandWidth 35
mgcOrder 34
normalize 1
numIterations 5
numState 5
numTestFiles 5
questionsNum 001
sampfreq 16000
strBandWidth 5
strOrder 5
upperF0 280
version 1
windowType 1
```

Appendix 2: Hyperparameters for FastSpeech2

Training hyperparameters:

```
optimizer:
batch_size:  16
```

```
betas: [0.9, 0.98]
eps: 0.000000001
weight_decay: 0.0
grad_clip_thresh: 1.0
grad_acc_step: 1
warm_up_step: 4000
anneal_steps: [300000, 400000,
500000]
anneal_rate: 0.3

step:
total_step: 300000
log_step: 100
synth_step: 1000
val_step: 1000
save_step: 100000
```

Preprocessing hyperparameters:

```
preprocessing:
val_size: 58

text:
text_cleaners: ["basic_cleaners"]
language: "cim"
use_spe_features: false
spe_feature_dim: 36

audio:
sampling_rate: 22050
max_wav_value: 32767.0

stft:
filter_length: 1024
hop_length: 256
win_length: 1024

mel:
n_mel_channels: 80
mel_fmin: 0
mel_fmax: 8000

pitch:
feature: "phoneme_level"
normalization: True

energy:
feature: "phoneme_level"
normalization: True

speaker:
embedding: "none"
pretrained_path: ""
```

Model hyperparameters

```
transformer:
encoder_layer: 4
encoder_head: 2
encoder_hidden: 256
decoder_layer: 4
decoder_head: 2
decoder_hidden: 256
conv_filter_size: 1024
conv_kernel_size: [9, 1]
encoder_dropout: 0.2
decoder_dropout: 0.2
spe_features: false
spe_feature_dim: 36
depthwise_convolutions: true

variance_predictor:
filter_size: 256
kernel_size: 3
dropout: 0.5
use_energy_predictor: true

variance_embedding:
pitch_quantization: "linear"
energy_quantization: "linear"
n_bins: 256

use_postnet: True

use_spe_loss: False

multi_speaker:
use_multi_speaker: False
embedding_type: "one-hot"

locations:
variance_adaptor: False
encoder: False

multilingual: False

max_seq_len: 1000

vocoder:
model: "HiFi-GAN"

speaker: "universal"
```