# Requirements and Motivations of Low-Resource Speech Synthesis for Language Revitalization

# Requirements and Motivations of Low-Resource Speech Synthesis for Language Revitalization

**Aidan Pine**[1] aidan.pine@nrc.ca
**Dan Wells**[2] dan.wells@ed.ac.uk
**Nathan Thanyehténhas Brinklow**[3] nathan.brinklow@queensu.ca
**Patrick Littell**[1] patrick.littell@nrc.ca
**Korin Richmond**[2] korin.richmond@ed.ac.uk

## Abstract

This paper describes the motivation and development of speech synthesis systems for the purposes of language revitalization. By building speech synthesis systems for three Indigenous languages spoken in Canada, Kanien'kéha, Gitksan & SENĆOŦEN, we re-evaluate the question of how much data is required to build low-resource speech synthesis systems featuring state-of-the-art neural models. For example, preliminary results with English data show that a FastSpeech2 model trained with 1 hour of training data can produce speech with comparable naturalness to a Tacotron2 model trained with 10 hours of data. Finally, we motivate future research in evaluation and classroom integration in the field of speech synthesis for language revitalization.

## 1 Introduction

There are approximately 70 Indigenous languages spoken in Canada, from 10 distinct language families (Rice, 2008). As a consequence of the residential school system and other policies of cultural suppression, the majority of these languages now have fewer than 500 fluent speakers remaining, most of them elderly. Despite this, interest from students and parents in Indigenous language education continues to grow (Statistics Canada, 2016); we have heard from teachers that they are overwhelmed with interest from potential students, and the growing trend towards online education means many students who have not previously had access to language classes now do.

Supporting these growing cohorts of students comes with unique challenges for languages with few fluent first-language speakers. A particular concern of teachers is to provide their students with opportunities to hear the language outside of class. Text-to-speech synthesis technology (TTS) shows potential for supplementing text-based language learning tools with audio in the event that the domain is too large to be recorded directly, or as an interim solution pending recordings from first-language speakers.

Development of TTS systems in this context faces several challenges. Most notable is the usual assumption that neural speech synthesis models require at least tens of hours of audio recordings with corresponding text transcripts to be trained adequately. Such a data requirement is far beyond what is available for the languages we are concerned with, and is difficult to meet given the limited time of the relatively small number of speakers of these languages. The limited availability of Indigenous language speakers also hinders the subjective evaluation methods often used in TTS studies, where naturalness of synthetic speech samples is judged by speakers of the language in question.

In this paper, we re-evaluate some of these challenges for applying TTS in the low-resource context of language revitalization. We build TTS systems for three Indigenous languages of Canada, with training data ranging from 25 minutes to 3.5 hours, and confirm that we can produce acceptable speech as judged by language teachers and learners. Outputs from these systems could be suitable for use in some classroom applications, for example a speaking verb conjugator.

## 2 Background

### 2.1 Language Revitalization

It is no secret that the majority of the world's languages are in crisis, and in many cases this crisis is even more urgent than conservation biologists' dire predictions for flora and fauna (Sutherland, 2003). However, the 'doom and gloom' rhetoric that often follows endangered languages over-represents vulnerability and under-represents

[1]National Research Council Canada
[2]University of Edinburgh
[3]Queen's University

the enduring strength of Indigenous communities who have refused to stop speaking their languages despite over a century of colonial policies against their use (Pine and Turin, 2017). Continuing to speak Indigenous languages is often seen as a political act of anti-colonial resistance. As such, the goals of any given language revitalization effort extend far beyond memorizing verb paradigms to broader goals of nationhood and self-determination (Pitawanakwat, 2009; McCarty, 2018). Language revitalization programs can also have immediate and important impacts on factors including community health and wellness (Whalen et al., 2016; Oster et al., 2014).

There is a growing international consensus on the importance of linguistic diversity, from the Truth & Reconciliation Commission of Canada (TRC) report in 2015 which issued nine calls to action related to language, to 2019 being declared an International Year of Indigenous Languages by the UN, and 2022-2032 being declared an International Decade of Indigenous Languages. From 1996 to 2016, the number of speakers of Indigenous languages increased by 8% (Statistics Canada, 2016). These efforts have been successful despite a lack of support from digital technologies. While opportunities may exist for technology to assist and support language revitalization efforts, these technologies must be developed in a way that does not further marginalize communities (Brinklow et al., 2019; Bird, 2020).

### 2.2 Why TTS for Language Revitalization?

Our interest in speech synthesis for language revitalization was sparked during user evaluations of Kawennón:nis (lit. 'it makes words'), a Kanien'kéha verb conjugator (Kazantseva et al., 2018) developed in collaboration between the National Research Council Canada and the Onkwawenna Kentyohkwa adult immersion program in Six Nations of the Grand River in Ontario, Canada. Kawennón:nis models a pedagogically-important subset of verb conjugations in XFST (Beesley and Karttunen, 2003), and currently produces 247,450 unique conjugations. The pronominal system is largely responsible for much of this productivity, since in transitive paradigms, agent/patient pairs are fused, as illustrated in Figure 1.

In user evaluations of Kawennón:nis, students often asked whether it was possible to add audio to the tool, to model the pronunciation of unfamiliar words. Assuming a rate of 200 forms/hr for 4 hours per day, 5 days per week, this would take a teacher out of the classroom for approximately a year. Considering Kawennón:nis is anticipated to have over 1,000,000 unique forms by the time the grammar modelling work is finished, recording audio manually becomes infeasible.

(1) ***Se***nòn:wes
**you.to.it**-like-habitual
'**You** like **it**.'

(2) ***Take***nòn:wes
**you.to.me**-like-habitual
'**You** like **me**.'

Figure 1: An example of fusional morphology of agent/patient pairs in Kanien'kéha transitive verb paradigms (from Kazantseva et al., 2018)

The research question that then emerged was 'what is the smallest amount of data needed in order to generate audio for all verb forms in Kawennón:nis'. Beyond Kawennón:nis, we anticipate that there are many similar language revitalization projects that would want to add supplementary audio to other text-based pedagogical tools.

### 2.3 Speech Synthesis

The last few years have shown an explosion in research into purely neural network-based approaches to speech synthesis (Tan et al., 2021). Similar to their HMM/GMM predecessors, neural pipelines typically consist of both a network predicting the acoustic properties of a sequence of text and a vocoder. The feature prediction network must be trained using parallel speech/text data where the input is typically a sequence of characters or phones that make up an utterance, and the output is a sequence of fixed-width frames of acoustic features. In most cases the predictions from the TTS model are log Mel-spectral features and a vocoder is used to generate the waveform from these acoustic features.

Much of the previous work on low resource speech synthesis has focused on transfer learning; that is, 'pre-training' a network using data from a language that has more data, and then 'fine-tuning' using data from the low-resource language. One of the problems with this approach is that the input space often differs between languages. As the

inputs to these systems are sequences of characters or phones, and as these sequences are typically one-hot encoded, it can be difficult to devise a principled method for transferring weights from the source language network to the target if there is a difference between the character or phone inventories of the two languages. Various strategies have emerged for normalizing the input space. For example, Demirsahin et al. (2018) propose a unified inventory for regional multilingual training of South Asian languages, while Tu et al. (2019) compare various methods to create mappings between source and target input spaces. Another proposal is to normalize the input space between source and target languages by replacing one-hot encodings of text with multi-hot phonological feature encodings (Gutkin et al., 2018; Wells and Richmond, 2021).

### 2.4 Speech Synthesis for Indigenous Languages in Canada

There is extremely little published work on speech synthesis for Indigenous languages in Canada (and North America generally). A statistical parametric speech synthesizer using Simple4All was recently developed for Plains Cree (Harrigan et al., 2019; Clark, 2014). Although it was unpublished, two highschool students[1] created a statistical parametric speech synthesizer for Kanien'kéha by adapting eSpeak (Duddington and Dunn, 2007). We know of no other attempts to create speech synthesis systems for Indigenous languages in Canada. Elsewhere in North America, a Tacotron2 system has been built for Cherokee (Conrad, 2020), and some early work on concatenative systems for Navajo was discussed in a technical report (Whitman et al., 1997), as well as on Rarámuri (Urrea et al., 2009).

## 3 Indigenous Language Data

Although the term 'low resource' is used to describe a wide swath of languages, most Indigenous languages in Canada would be considered 'low-resource' in multiple senses of the word, having both a low amount of available data (annotated or unannotated), and a relatively low number of speakers. Most Indigenous languages lack transcribed audio corpora, and fewer still have such data recorded in a studio context. Due to the limited number of speakers, creating these resources is non-trivial: there are limited amounts of text from which a speaker could read, and there are few people available who are sufficiently literate in the languages to transcribe recorded audio. Re-focusing speakers' limited time to these tasks presents a significant opportunity cost; they are often already over-worked and over-burdened in under-funded and under-resourced language teaching projects.

As mentioned in §2.1, language technology projects that aim to assist language revitalization and reclamation efforts must be centered around the primary goals of those efforts and ensure that the means of developing the technology do not distract or work against the broader sociopolitical goals. A primary stress point for many natural language processing projects involving Indigenous communities surrounds issues of data sovereignty. It is important that communities direct the development of these tools, and maintain control, ownership, and distribution rights for their data, as well as for the resulting speech synthesis models (Keegan, 2019; Brinklow, 2021). In keeping with this, the datasets described in this paper are not being released publicly at this time.

To test the feasibility of developing speech synthesis systems for Indigenous languages, we trained models for three unrelated Indigenous languages, Kanien'kéha (§3.1), Gitksan (§3.2), and SENĆOŦEN (§3.3).

### 3.1 Kanien'kéha

Kanien'kéha[2] (a.k.a. Mohawk) is an Iroquoian language spoken by roughly 2,350 people in southern Ontario, Quebec, and northern New York state (Statistics Canada, 2016). In 1979 the first immersion school of any Indigenous language in Canada was opened for Kanien'kéha, and many other very successful programs have been started since, including the Onkwawenna Kentyohkwa adult immersion program in 1999 (Gomashie, 2019).

In the late 1990s, a team of five Kanien'kéha translators worked with the Canadian Bible Society to translate and record parts of the Bible; one of the speakers on these recordings, Satewas, is still living. Translation runs in Satewas's family, with his great-grandfather also working on Bible translations in the 19th century. Later, a team of four speakers and learners, including this paper's third author, aligned the text and audio at the utterance

[1] https://wiki.laptop.org/go/Instructions_for_implementing_a_new_language_%22voice%22_for_Speak_on_the_XO

[2] As there are different variations of spelling, we use the spelling used in the communities of Kahnawà:ke and Kahnesetà:ke throughout this paper

level using Praat (Boersma and van Heuven, 2001) and ELAN (Brugman and Russel, 2004).

While a total of 24 hours of audio were recorded, members of the Kanien'kéha-speaking community told us it would be inappropriate to use the voices of speakers who had passed away, leaving only recordings of Satewas's voice. Using a GMM-based speaker ID system (Kumar, 2017), we removed utterances by these speakers, then removed utterances that were outliers in duration (less than 0.4s or greater than 11s) and speaking rate (less than 4 phones per second or greater than 15), recordings with an unknown phase effect present, and utterances containing non-Kanien'kéha characters (e.g. proper names like 'Euphrades'). Handling utterances with non-Kanien'kéha characters would have required grapheme-to-phoneme prediction capable of dealing with multilingual text and code-switching which we did not have available. The resulting speech corpus comprised 3.46 hours of speech.

### 3.2 Gitksan

Gitksan[3] is one of four languages belonging to the Tsimshianic language family spoken along the Skeena river and its surrounding tributaries in the area colonially known as northern British Columbia. Traditional Gitksan territory spans some 33,000 square kilometers and is home to almost 10,000 people, with approximately 10% of the population continuing to speak the language fluently (First Peoples' Cultural Council, 2018).

As there were no studio-quality recordings of the Gitksan language publicly available, and as an intermediate speaker of the language, the first author recorded a sample set himself. In total, he recorded 35.46 minutes of audio reading isolated sentences from published and unpublished stories (Forbes et al., 2017).

### 3.3 SENĆOŦEN

The SENĆOŦEN language is spoken by the W̱SÁNEĆ people on the southern part of the island colonially known as Vancouver Island. It belongs to the Coastal branch of the Salish language family. The W̱SÁNEĆ community runs a world-famous language revitalization program[4], and uses an orthography developed by the late SENĆOŦEN speaker and W̱SÁNEĆ elder Dave Elliott. While the community of approximately 3,500 has fewer than 10 fluent speakers, there are hundreds of learners, many of whom have been enrolled in years of immersion education in the language (First Peoples' Cultural Council, 2018).

As there were no studio-quality recordings of the SENĆOŦEN language publicly available, we recorded 25.92 minutes of the language with PENÁĆ David Underwood reading two stories originally spoken by elder Chris Paul.

## 4 Research Questions

Given the motivation and context for language revitalization-based speech synthesis, a number of research questions follow. Namely, how much data is required in order to build a system of reasonable pedagogical quality? How do we evaluate such a system? And, how is the resulting system best integrated into the classroom? In §4.1, we discuss the difficulty of evaluating TTS systems in low-resource settings. We then discuss preliminary results for English and Indigenous language TTS which show that acceptable speech quality can be achieved with much less training data than usually considered for neural speech synthesis (§4.2). Finally, we suggest possible directions for pedagogical integration in section §4.4.

### 4.1 Low-Resource Evaluation

One of the most significant challenges in researching speech synthesis for languages with few speakers is evaluating the models. For some Indigenous languages in Canada, the total number of speakers of the language is less than the number typically required for statistical significance in a listening test (Wester et al., 2015). While the number of speakers in these conditions is sub-optimal for statistical analysis, we have been told by the communities we work with that the positive assessment of a few widely respected and community-engaged language speakers would be practically sufficient to assess the pedagogical value of speech models in language revitalization contexts. For the experiments described in this paper, we ran listening tests for both Kanien'kéha and Gitksan with speakers, teachers, and learners, but were not able to run any such tests for SENĆOŦEN due to very few speakers with already busy schedules.

While some objective metrics do exist, such as

[3] We use Lonnie Hindle and Bruce Rigsby's spelling of the language, which, with the use of 'k' and 'a' is a blend of upriver (gigeenix) and downriver (gyets) dialects

[4] https://wsanecschoolboard.ca/sencoten-language/

Mel cepstral distortion (MCD, Kubichek, 1993), we do not believe they should be considered reliable proxies for listening tests. Future research on speech synthesis for languages with few speakers should prioritize efficient and effective means of evaluating results.

In many cases, including in the experiment described in §4.2, artificial data constraints can be placed on a language with more data, like English, to simulate a low-resource scenario. While this technique can be insightful and it is tempting to draw universal conclusions, English is linguistically very different from many of the other languages spoken in the world. Accordingly, we should be cautious not to assume that results from these types of experiments will necessarily transfer or extend to genuinely low-resource languages.

### 4.2 How much data do you really need?

The first question to answer is whether our Indigenous language corpora ranging from 25 minutes to 3.46 hours of speech are sufficient for building neural speech synthesizers. Due to the prominence of Tacotron2 (Shen et al., 2018), it seems that many people have assumed that the data requirements for training *any* neural speech synthesizer of similar quality must be the same as the requirements for this particular model. As a result, some researchers still choose to implement either concatenative or HMM/GMM-based statistical parametric speech synthesis systems in low-resource situations based on the assumption that a "sufficiently large corpus [for neural TTS] is unavailable" (James et al., 2020, p. 298). We argue that attention-based models such as Tacotron2 should not be used as a benchmark for data requirements among all neural TTS methods, as they are notoriously difficult to train and unnecessarily inflate training data requirements.

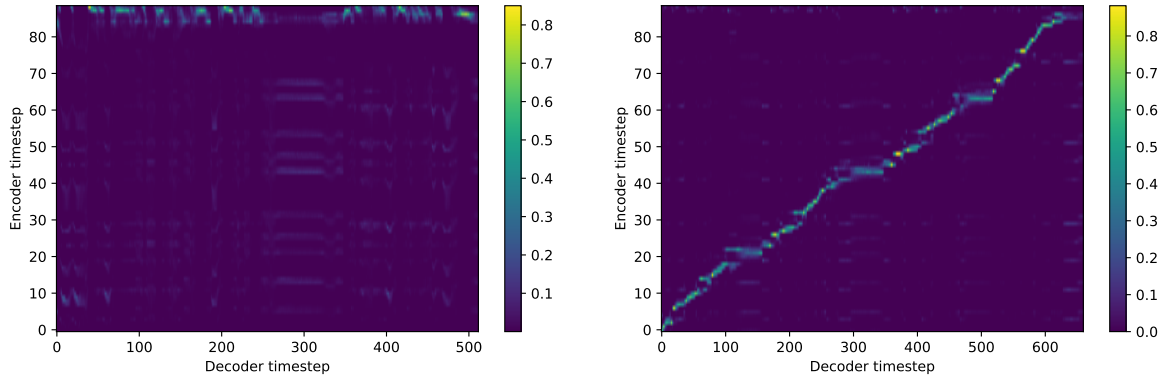
#### 4.2.1 Replacing attention-based weak duration models

Tacotron2 is an autoregressive model, meaning it predicts the speech parameters $\hat{y}_t$ from both the input sequence of text $x$ and the previous speech parameters $y_1, ..., y_{t-1}$. Typically, the model is trained with 'teacher-forcing', where the autoregressive frame $y_{t-1}$ passed as input for predicting $\hat{y}_t$ is taken from the ground truth acoustic features and not the prediction network's output from the previous frame $\hat{y}_{t-1}$. As discussed by Liu et al. (2019), such a system might learn to copy the teacher forcing input or disregard the text entirely, which could still optimize Tacotron2's root mean square error function over predicted acoustic features, but result in an untrained or degenerate attention network which is unable to properly generalize to new inputs at inference time when the teacher forcing input is unavailable. Attention failures represent a characteristic class of errors for models such as Tacotron2, for example skipping or repeating words from the input text (Valentini-Botinhao and King, 2021).

There have been many proposals to improve training of the attention network, for example by guiding the attention or using a CTC loss function to respect the monotonic alignment between text inputs and speech outputs (Tachibana et al., 2018; Liu et al., 2019; Zheng et al., 2019; Gölge, 2020). As noted by Liu et al. (2019), increasing the so-called 'reduction factor' – which applies dropout to the autoregressive frames – can also help the model learn to rely more on the attention network than the teacher forcing inputs, but possibly at the risk of compromising synthesis quality.

FastSpeech2 (Ren et al., 2021), and similar systems like FastPitch (Łańcucki, 2021), present an alternative to Tacotron2-type attentive, autoregressive systems with similar listening test results and without the characteristic errors related to attention. Instead of modelling duration using attention, they include an explicit duration prediction module trained on phone duration targets extracted from the training data. For the original FastSpeech, target phone durations derived from the attention weights of a pre-trained Tacotron2 system were used to provide phone durations (Ren et al., 2019). In low-resource settings, however, there might not be sufficient data to train an initial Tacotron2 in the target language in the first place. For FastSpeech2, phone duration targets are instead extracted using the Montreal Forced Aligner (MFA, McAuliffe et al., 2017), trained on the same data as used for TTS model training. We have found MFA can provide suitable alignments for our target languages, even with alignment models being trained on only limited data.

Faster convergence of text-acoustic feature alignments has been found to speed up overall encoder-decoder TTS model training, as stable alignments provide a solid foundation for further training of the decoder. Badlani et al. (2021) show this by adding a jointly-learned alignment framework to a Tacotron2 architecture, reducing time




(a) 5 Hr LJ Corpus Subset

(b) 10 Hr LJ Corpus Subset

Figure 2: Visualization of Tacotron2 Attention Network Weights extracted after 100k steps trained on the LJ corpus. The weights of the attention network should be diagonal and monotonic as seen in subfigure (b). Subfigure (a) shows that the network trained on a 5 hour subset of the LJ corpus results in a degenerate attention network.

to convergence. In contrast, they found that replacing MFA duration targets in FastSpeech2 training offers no benefit – forced alignment targets already provide enough information for more time-efficient training compared to an attention-based Tacotron2 system. Relieving the burden of learning an internal alignment model also opens the door to more data-efficient training. For example, Perez-Gonzalez-de-Martos et al. (2021) submitted a non-attentive model trained from forced alignments to the Blizzard Challenge 2021, where their system was found to be among the most natural and intelligible in subjective listening tests despite only using 5 hours of speech; all other submitted systems included often significant amounts of additional training data (up to 100 hours total).

#### 4.2.2 Experimental Comparison of Data Requirements for Neural TTS

To investigate the effects of differing amounts of data on the attention network, and in preparation for training systems with our limited Indigenous language data sets, we trained five Tacotron2 models on incremental partitions of the LJ Speech corpus of American English (Ito and Johnson, 2017). We used the NVIDIA implementation[5] with default hyperparameters apart from a reduced batch size of 32 to fit the memory capacity of our GPU resources. We artificially constrained the training data such that the first model saw only the first hour of data from the shuffled corpus, the second model that same first hour plus another two hours (3 total) etc., so that the five models were trained on 1, 3, 5, 10 and 24 (full corpus) hours of speech. The models were trained for 100k steps and, as seen in Figure 2, using up to 5 hours of data the attention mechanism does not learn properly, resulting in degenerate outputs.

For comparison, we trained seven FastSpeech2 models with batch size 16 for 200k steps on 15 and 30 minute, 1, 3, 5, 10 and 24 hour incremental partitions of LJ Speech. Our model[6] is based on an open-source implementation (Chien, 2021), which adds learnable speaker embeddings and a decoder postnet to the original model, as well as predicting pitch and energy values at the phone rather than frame level. We also added learnable language embeddings for supplementary experiments in cross-lingual fine-tuning; while not reported in this paper, we refer the interested reader to Pine (2021) for discussion of these experiments. Motivated by concerns of efficiency in model training and inference, and the possibility of overfitting a large model to limited amounts of data, we further modified the base architecture to match the LightSpeech model presented in Luo et al. (2021). We removed the energy adaptor, replaced the convolutional layers in the encoder, decoder and remaining variance predictors with depthwise separable convolutions (Kaiser et al., 2018) and matched encoder and decoder convolutional kernel sizes with Luo et al. (2021). This reduced the number of model parameters from 35M[7] to 11.6M without noticeable change in voice quality and sped up train-

[5] https://github.com/NVIDIA/tacotron2

[6] https://github.com/roedoejet/FastSpeech2

[7] In the implementation of Chien (2021); the original FastSpeech2 is slightly smaller at 27M parameters.

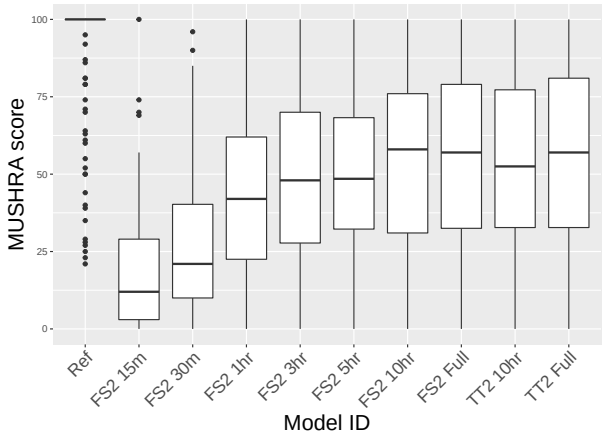



Figure 3: Box plot of survey data from MUSHRA questions comparing Tacotron2 (TT2) and FastSpeech2 (FS2) models with constrained amounts of training data. 'Ref' refers to reference recordings of natural speech.

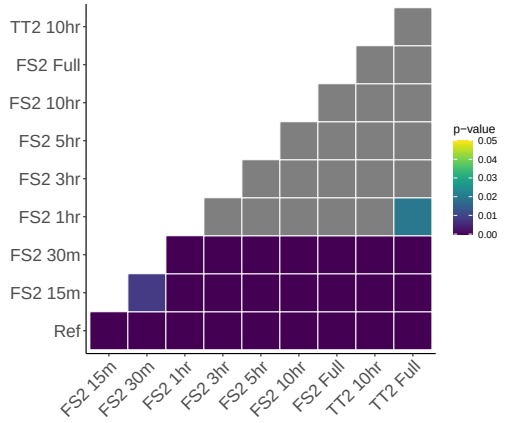



Figure 4: Pairwise Bonferroni-corrected Wilcoxon signed rank tests between each pair of voices. Cells correspond to the significance of the result of the pairwise test between the model on the y-axis and the model on the x-axis. Darker cells show stronger significance; grey cells did not show a significant difference in listening test results. FS2 refers to models built with FastSpeech2, TT2 refers to models built with Tacotron2, and 'Ref' to reference recordings. Samples available at https://roedoejet.github.io/msc_listening_tests_data/

ing by 33% on GPU or 64% on CPU. For additional discussion of the accessibility benefits of these changes with respect to Indigenous language communities, see Appendix A.

#### 4.2.3 Results

We conducted a short (10-15 minute) listening test to compare the two Tacotron2 models that trained properly (10h, full) against the seven FastSpeech2 models. We recruited 30 participants through Prolific, and presented each with four MUSHRA-style questions where they were asked to rank the 9 voices along with a hidden natural speech reference (ITU-R, 2003). MUSHRA-style questions were used as a practical way to evaluate this large number of models.

While it only took 30 minutes to recruit 30 participants using Prolific, the quality of responses was quite varied. We rejected two outright as they seemingly did not listen to the stimuli and left the same rankings for every voice. Even still, there was a lot of variation in responses from the remaining participants, as seen in Figure 3. We tested for significant differences between pairs of voices using Bonferroni-corrected Wilcoxon signed rank tests. Pairwise test results are summarized in the heat map of their p-values in Figure 4.

In the results from the pairwise analysis, we can see that natural speech is rated as significantly more natural than all synthetic speech samples. Naturalness ratings for the FastSpeech2 voices trained on 15m and 30m of data are significantly lower than all other voices, and significantly different from each other. The results for the remaining voices, while showing consistent improvements in naturalness ratings as more data is added (as shown in Figure 3), are not significantly different from each other. This is a relevant and important finding for low-resource speech synthesis because it shows that a FastSpeech2 voice built with 3 hours of data can achieve subjective naturalness ratings which are not significantly different from a Tacotron2 voice built with 24 hours of data. Similarly, the results of the listening test for our FastSpeech2 voice built with 1 hour of data are not significantly different from our Tacotron2 voice built with 10 hours of data. Additionally, while all the FastSpeech2 voices were intelligible, all Tacotron2 models trained with less than 10 hours of data produced unintelligible speech.

### 4.3 Indigenous Language Experiments

Despite the difficulty in evaluation (§4.1), we built and evaluated a number of TTS systems for the Indigenous languages described in §3. We had a baseline concatenative model available for Kanien'kéha that we had previously built using Festival and Multisyn (Taylor et al., 1998; Clark et al., 2007). Additionally, we trained cold-start FastSpeech2 models for each language, as well as models fine-tuned for 25k steps from a multilin-

gual, multispeaker FastSpeech2 model pre-trained on a combination of VCTK (Yamagishi et al., 2019), Kanien'kéha and Gitksan recordings. A rule-based mapping from orthography to pronunciation form was developed for each language using the 'g2p' Python library in order to perform alignment and synthesis at the phone-level instead of character-level (Pine et al., Under Review).

#### 4.3.1 Results

We carried out listening test evaluations of Gitksan and Kanien'kéha models. Participants were recruited by contacting teachers, learners and linguists with at least some familiarity with the languages.

For the Kanien'kéha listening test, 6 participants were asked to answer 20 A/B questions comparing synthesized utterances from the various models. We used A/B tests for more targeted comparisons between different systems, namely cold-start vs. fine-tuned and neural vs. concatenative. Results showed that 72.2% of A/B responses from participants preferred our FastSpeech2 model over our baseline concatenative model. In addition, 81.7% of A/B responses from participants preferred the cold-start to the model fine-tuned on the multi-speaker, multi-lingual model, suggesting that the transfer learning approach discussed in §2.3 might not be necessary for models with explicit durations such as FastSpeech2 since they are relieved of the burden to learn an implicit model of duration through attention from limited data.

For the Gitksan listening test, we did not build a concatenative model as with Kanien'kéha and so we were not comparing different models, but rather just gathering opinions on the quality of the cold-start FastSpeech2 model. Accordingly, 10 MOS-style questions were presented to 12 participants for both natural utterances and samples from our FastSpeech2 model. The model received a $3.56 \pm 0.26$ MOS compared with a MOS for the reference recordings of $4.63 \pm 0.19$ as shown in Figure 5. While both Kanien'kéha and Gitksan results seem to corroborate our belief that these models should be of reasonable quality despite limited training data, it is difficult to make any conclusive statement given the low number of eligible participants available for evaluation.

As the main goal of our efforts here is to eventually integrate our speech synthesis systems into a pedagogical setting, we also asked the 18 people who participated across Kanien'kéha and Gitksan listening tests directly whether they approved of the synthesis quality. As seen in Figure 6, participant responses were generally positive; full responses are reported in Appendix B.

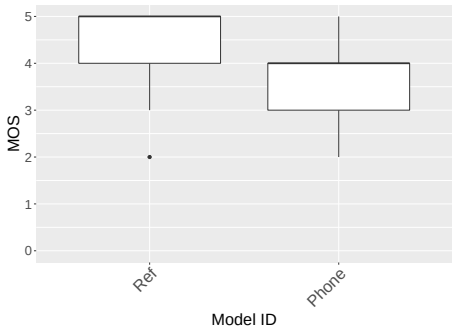
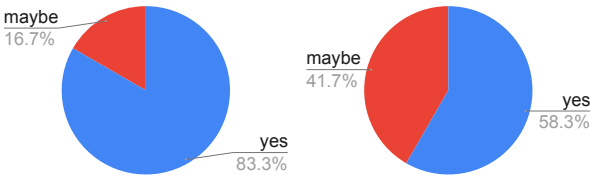



Figure 5: Box plot of MOS results for Gitksan listening test. 'Ref' is the reference voice and 'Phone' is the phone-based FastSpeech2 neural model. Variable results for the reference voice are likely due to the natural speech recordings coming from a non-native speaker.




(a) Kanien'kéha (b) Gitksan

Figure 6: Responses from qualitative survey asking participants "Would you be comfortable with any of the voices you heard being played online, say for a digital dictionary or verb conjugator if no other recording existed?". No participants responded "no".

### 4.4 Integrating TTS in the Classroom

Satisfying the goal of adding supplementary audio to a reference tool like Kawennón:nis can be straightforwardly implemented by linking entries in the verb conjugator to pre-generated audio for the domain from a static server. This implementation also limits the potential of out of domain utterances that might be deemed inappropriate, which is an ethical concern in communities with low numbers of speakers where the identity of the 'model' speaker is easily determined.

However, the ability to synthesize novel utterances could be pedagogically useful. Students often come into contact with words or sentences which do not have audio, and teachers often have to prepare new thematic word lists or vocabulary lessons that could benefit from a more general purpose speech synthesis solution. In those cases,

with community and speaker input, we might consider what controls would be necessary for the users of this technology. One potential solution is the variance adaptor architecture present in FastSpeech2, allowing for phone-level control of duration, pitch and energy; an engaging demonstration of a graphical user interface for the corresponding controls in a FastPitch model is also available.[8] We would like to focus further efforts on designing a user interface for speech synthesis systems that satisfies ethical concerns while prioritizing language pedagogy as the fundamental use case.

In addition to fine-grained prosodic controls, we would like to explore the synthesis of hyperarticulated speech, as often used by language teachers when modelling pronunciation of unfamiliar words or sounds for students. This style of speech typically involves adjustment beyond the parameters of pitch, duration and energy, and is characterized by more careful enunciation of individual phones than is found in normal speech. This problem has parallels to the synthesis of Lombard speech (Hu et al., 2021), as used to improve intelligibility by speakers who find themselves in noisy environments.

## 5 Conclusion

In this paper, we presented the first neural speech synthesis systems for Indigenous languages spoken in Canada. Subjective listening tests showed encouraging results for the naturalness and acceptability of voices for two languages, Kanien'kéha and Gitksan, despite limited training data availability (3.5 hours and 35 minutes, respectively). More extensive evaluation on English shows that the FastSpeech2 architecture can produce speech with similar quality to a Tacotron2 system using a fraction of the amount of speech usually considered for neural speech synthesis. Notably, a FastSpeech2 voice trained on 1 hour of English speech achieved subjective naturalness ratings not significantly different from a Tacotron2 voice using 10 hours of data, while a 3-hour FastSpeech2 system showed no significant difference from a 24-hour Tacotron2 voice.

We attribute these results to the fact that FastSpeech2 learns input token durations from forced alignments, rather than jointly learning to align linguistic inputs to acoustic features alongside the acoustic feature prediction task as in attention-based architectures such as Tacotron2. Given forced alignments of sufficient quality, which we found to be achievable even by training a Montreal Forced Aligner model only on our limited Indigenous language training data, this makes for more data-efficient training of neural TTS systems than has generally been explored in previous work. These findings show great promise for future work in low-resource TTS for language revitalization, especially as they come from systems trained from scratch on such limited data, rather than pre-training on a high-resource language and subsequent fine-tuning on limited target language data.

[8] https://fastpitch.github.io/


## Acknowledgements

We would like to gratefully acknowledge the many people who worked to record the audio for the speech synthesis systems described in this project. In particular, Satewas Harvey Gabriel, and PENÁĆ David Underwood.

Much of the text and experimentation related to this paper was submitted as partial fulfillment of the first author's M.Sc. dissertation at the University of Edinburgh (Pine, 2021).

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.


## References


Valerie Alia. 2009. *The New Media Nation: Indigenous Peoples and Global Communication*, ned - new edition, 1 edition. Berghahn Books.

Rohan Badlani, Adrian Łancucki, Kevin J. Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2021. One TTS Alignment To Rule Them All. *arXiv:2108.10447*.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519.

Paul Boersma and Vincent van Heuven. 2001. Speak and unSpeak with PRAAT. *Glot International*, 5(9/10):341–347.

Nathan Thanyehténhas Brinklow. 2021. Indigenous language technologies: Anti-colonial oases in a colonizing (digital) world. *WINHEC: International Journal of Indigenous Education Scholarship*, 16(1):239–266.

Nathan Thanyehténhas Brinklow, Patrick Littell, Delaney Lothian, Aidan Pine, and Heather Souter. 2019. Indigenous Language Technologies & Language Reclamation in Canada. *Proceedings of the 1st International Conference on Language Technologies for All*, pages 402–406.

Hennie Brugman and Albert Russel. 2004. Annotating Multi-media/Multi-modal Resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Chung-Ming Chien. 2021. ming024/FastSpeech2. https://github.com/ming024/FastSpeech2. Original-date: 2020-06-25T13:57:53Z.

Robert AJ Clark. 2014. Simple4all. In *Proc. Interspeech 2014*, pages 1502–1503.

Robert AJ Clark, Korin Richmond, and Simon King. 2007. Multisyn: Open-domain unit selection for the festival speech synthesis system. *Speech Communication*, 49(4):317–330.

Michael Conrad. 2020. Tacotron2 and Cherokee TTS. https://www.cherokeelessons.com/content/tacotron2-and-cherokee-tts/.

Isin Demirsahin, Martin Jansche, and Alexander Gutkin. 2018. A unified phonological representation of South Asian languages for multilingual text-to-speech. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 80–84.

Jonathan Duddington and Reece Dunn. 2007. eSpeak: Speech Synthesizer. http://espeak.sourceforge.net/.

EPA. 2019. Emissions & generation resource integrated database (eGRID). https://www.epa.gov/egrid.

First Peoples' Cultural Council. 2018. Report on the status of B.C. https://fpcc.ca/resource/fpcc-report-of-the-status-of-b-c-first-nations-languages-2018/.

Clarissa Forbes, Henry Davis, Michael Schwan, and Gitksan Research Lab. 2017. Three Gitksan Texts. *Papers for the International Conference on Salish and Neighbouring Languages*, 52:47–89.

Eren Gölge. 2020. Solving Attention Problems of TTS models with Double Decoder Consistency. https://erogol.com/solving-attention-problems-of-tts-models-with-double-decoder-consistency/.

Grace A. Gomashie. 2019. Kanien'keha / Mohawk Indigenous language revitalisation efforts in Canada. *McGill Journal of Education / Revue des sciences de l'éducation de McGill*, 54(1):151–171.

Alexander Gutkin, Martin Jansche, and Tatiana Merkulova. 2018. FonBund: A Library for Combining Cross-lingual Phonological Segment Data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2236–2240. European Language Resources Association (ELRA).

Atticus Harrigan, Antti Arppe, and Timothy Mills. 2019. A Preliminary Plains Cree Speech Synthesizer. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 64–73, Honolulu. Association for Computational Linguistics.

Qiong Hu, Tobias Bleisch, Petko Petkov, Tuomo Raitio, Erik Marchi, and Varun Lakshminarasimhan. 2021. Whispered and Lombard Neural Speech Synthesis. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 454–461.

Innu-Atikamekw-Anishnabeg Coalition. 2020. Export of Canadian Hydropower to the United States - First Nations in Québec and Labrador Unite to Oppose Hydro-Québec Project.

Keith Ito and Linda Johnson. 2017. The LJ speech dataset. https://keithito.com/LJ-Speech-Dataset/.

ITU-R. 2003. Recommendation ITU-R BS.1534-1 - Method for the subjective assessment of intermediate quality level of coding systems. Technical Report ITU-R BS.1534-1, International Telecommunication Union.

Jesin James, Isabella Shields, Rebekah Berriman, Peter Keegan, and Catherine Watson. 2020. Developing resources for te reo Māori text to speech synthesis system. In P. Sojka, I. Kopeček, K. Pala, and A. Horák, editors, *Text, Speech, and Dialogue*, pages 294–302.

Lukasz Kaiser, Aidan N. Gomez, and Francois Chollet. 2018. Depthwise Separable Convolutions for Neural Machine Translation. In *International Conference on Learning Representations*.

Anna Kazantseva, Owennatekha Brian Maracle, Ronkwe'tiyóhstha Josiah Maracle, and Aidan Pine. 2018. Kawennón:nis: the wordmaker for Kanyen'kéha. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 53–64, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Te Taka Keegan. 2019. Issues with Māori sovereignty over Māori language data. *Let The Languages Live 2019 Conference*.

R. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128 vol.1.

Abhijeet Kumar. 2017. Spoken Speaker Identification based on Gaussian Mixture Models : Python Implementation.

Adrian Łańcucki. 2021. Fastpitch: Parallel Text-to-Speech with Pitch Prediction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592.

A. Levasseur, S. Mercier-Blais, Y. T. Prairie, A. Tremblay, and C. Turpin. 2021. Improving the accuracy of electricity carbon footprint: Estimation of hydroelectric reservoir greenhouse gas emissions. *Renewable and Sustainable Energy Reviews*, 136:110433.

Peng Liu, Xixin Wu, Shiyin Kang, Guangzhi Li, Dan Su, and Dong Yu. 2019. Maximizing Mutual Information for Tacotron. *arXiv:1909.01145*.

Renqian Luo, Xu Tan, Rui Wang, Tao Qin, Jinzhu Li, Sheng Zhao, Enhong Chen, and Tie-Yan Liu. 2021. LightSpeech: Lightweight and Fast Text to Speech with Neural Architecture Search. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5699–5703.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pages 498–502. ISCA.

Teresa L McCarty. 2018. Community-based language planning: Perspectives from Indigenous language revitalization. In *The Routledge handbook of language revitalization*, pages 22–35. Routledge.

Richard Oster, Angela Grier, Rick Lightning, Maria Mayan, and Ellen Toth. 2014. Cultural continuity, traditional Indigenous language, and diabetes in Alberta first nations: a mixed methods study. *International journal for equity in health*, 13:92.

Alejandro Perez-Gonzalez-de-Martos, Albert Sanchis, and Alfons Juan. 2021. VRAIN-UPV MLLP's system for the Blizzard Challenge 2021. In *Blizzard Challenge 2021 Workshop*.

Aidan Pine. 2021. *Low Resource Speech Synthesis*. M.Sc. dissertation, University of Edinburgh.

Aidan Pine, Patrick Littell, Eric Joanis, David Huggins-Daines, Christopher Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delaisie Torkornoo, and Sabrina Yu. Under Review. $G_i2P_i$: Rule-based, index-preserving grapheme-to-phoneme transformations.

Aidan Pine and Mark Turin. 2017. Language Revitalization. *Oxford Research Encyclopedia of Linguistics*.

Brock Thorbjorn Pitawanakwat. 2009. *Anishinaabemodaa Pane Oodenang: a qualitative study of Anishinaabe language revitalization as self-determination in Manitoba and Ontario*. Ph.D. thesis, University of Victoria.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. In *Advances in Neural Information Processing Systems*, volume 32.

Keren Rice. 2008. Indigenous languages in Canada. In *The Canadian Encyclopedia*.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Statistics Canada. 2016. Census of population. `https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm`.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

William J. Sutherland. 2003. Parallel extinction risk and global distribution of languages and species. *Nature*, 423:276–279.

Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4784–4788.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A Survey on Neural Speech Synthesis. *arXiv:2106.15561*.

Paul Taylor, Alan W Black, and Richard Caley. 1998. The architecture of the Festival speech synthesis system. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, pages 147–152.

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. In *Interspeech 2019*, pages 2075–2079.

A. M. Urrea, José Abel Herrera Camacho, and Maribel Alvarado García. 2009. Towards the Speech Synthesis of Raramuri: A Unit Selection Approach based on Unsupervised Extraction of Suffix Sequences. *Research in Computing Science*, 41:243–256.

Cassia Valentini-Botinhao and Simon King. 2021. Detection and Analysis of Attention Errors in Sequence-to-Sequence Text-to-Speech. In *Interspeech 2021*, pages 2746–2750. ISCA.

Dan Wells and Korin Richmond. 2021. Cross-lingual Transfer of Phonological Features for Low-resource Speech Synthesis. In *Proc. 11th ISCA Speech Synthesis Workshop*, pages 160–165.

Mirjam Wester, Cassia Valentini-Botinhao, and Gustav Eje Henter. 2015. Are we using enough listeners? No!—an empirically-supported critique of Interspeech 2014 TTS evaluations. In *Interspeech 2015*, pages 3476–3480.

D. Whalen, Margaret Moss, and Daryl Baldwin. 2016. Healing through language: Positive physical health effects of indigenous language use. *F1000Research*, 5:852.

Robert Whitman, Richard Sproat, and Chilin Shih. 1997. *A Navajo Language Text-to-Speech Synthesizer*. AT&T Bell Laboratories.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). University of Edinburgh, The Centre for Speech Technology Research (CSTR).

Yibin Zheng, Xi Wang, Lei He, Shifeng Pan, Frank K. Soong, Zhengqi Wen, and Jianhua Tao. 2019. Forward-Backward Decoding for Regularizing End-to-End TTS. In *Interspeech 2019*, pages 1283–1287.

# A Compute, Accessibility, & Environmental Impact

For reasons of environmental impact and accessibility, reducing the amount of computation required for both training and inference is important for any neural speech synthesis system, particularly so for Indigenous languages.

## A.1 Accessibility, Training & Inference Speed

While language revitalization efforts are *mostly* encouraging about integrating new technologies into curriculum, there is a growing awareness of the potential harms. Beyond assessing the benefits and risks of introducing a new technology into language revitalization efforts, communities are concerned with the way the technology is researched and developed, as this process has the ability to empower or disempower language communities in equal measure (Alia, 2009; Brinklow et al., 2019). The current model for developing speech synthesis systems is not very equitable – models need to be run on GPUs by people with specialized training. For Indigenous communities to create speech synthesis tools for their languages, they should not be required to hand over their language data to a large government or corporate organization. A pre-training, fine-tuning pipeline could be attractive for this reason; communities could fine-tune their own models on a laptop if a multilingual/multi-speaker model were pre-trained on GPUs at a larger institution. Reducing the computational requirements for training and inference of these models could help ensure language communities have greater control over the process of the development of these systems, less dependence on governmental organizations or corporations, and more sovereignty over their data (Keegan, 2019).

Strubell et al. (2019) present an argument for equitable access to computational resources for NLP research; put another way, we might say that systems which require less compute are more accessible. Reducing the number of parameters in a neural TTS model should translate to increased efficiency, and might make the model less prone to overfitting when training on limited amounts of data. As discussed in §4.2.2, we modified the base implementation of FastSpeech2 from Chien (2021) closely following the lightweight alternative discovered through neural architecture search in Luo et al. (2021). These changes reduced the size of the model from Chien (2021) from 35M to 11.6M parameters, reduced the size of the stored model from 417 MB to 135 MB and significantly improved inference and train times as summarized in Table 1. We saw a 33% improvement in average batch processing times on the GPU during training, and 64% on the CPU, which may be even more relevant for Indigenous language communities with limited computational resources. During inference, we saw a 15% speed-up on GPU and 57% on CPU.

Results were timed by running the model for 300

| | | FastSpeech2 | Adapted System |
|---|---|---|---|
| Training | GPU | 90.52 ms ($\sigma$ 3.31) | 60.04 ms ($\sigma$ 1.70) |
| | CPU | 7561.50 ms ($\sigma$ 263.55) | 2720.88 ms ($\sigma$ 92.99) |
| Inference | GPU | 12.00 ms ($\sigma$ 0.30) | 10.23 ms ($\sigma$ 0.78) |
| | CPU | 138.73 ms ($\sigma$ 3.94) | 59.50 ms ($\sigma$ 1.85) |

Table 1: Mean and standard deviation of training and inference times for a single forward pass of baseline FastSpeech2 and adapted models.

repetitions and taking the mean. The GPU (Tesla V100-SXM2 16GB) was warmed up for 10 repetitions before timing started, and PyTorch's built-in GPU synchronization method was used to synchronize timing (which occurs on the CPU) with the training or inference running on the GPU. CPU tests were performed on an Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz with 4 cores and 16GB memory reserved. All timings used a batch size of 16.

### A.2 CO2 Consumption

Strubell et al. (2019) also argue that NLP researchers should have a responsibility to disclose the environmental footprint of their research, in order for the community to effectively evaluate any gains and to allow for a more equitable and reproducible field.

All experiments for this paper requiring a GPU were run on the Canadian General Purpose Science Cluster (GPSC) in Dorval, Quebec. Experiments were all run on single Tesla V100-SXM2 16GB GPUs. Strubell et al. (2019) provide the following equation for estimating $CO_2$ production:

$$p_t = \frac{1.58t(p_c + p_r + (g * p_g))}{1000} \quad (1)$$

where $t$ is time, $p_t$ is total power for training, $p_c$ is average draw of CPU sockets, $p_r$ is average DRAM memory draw, $g$ is the number of GPUs used in training and $p_g$ is the average draw from GPUs. In our case, we estimate $t$ to be equal to 1,541.98[9] after summing the time for experiments based on their log files, $p_c$ is 75 watts, $p_r$ is 6 watts, $g$ is 1, and $p_g$ is 250 watts, and the equation for grams of CO2 consumption is $CO_2 = 34.5p_t$ as the average carbon footprint of electricity distributed in Quebec is estimated at 34.5g CO2eq/kWh (Levasseur et al., 2021). This results in a total equivalent carbon consumption of 27,821.65 grams, roughly equivalent to driving a single passenger gas-powered vehicle for 110 kilometres according to the average rate of 404 grams/mile (EPA, 2019).

This is a comparatively low CO2 consumption for over 1500 GPU hours, largely due to the low CO2/kWh output of Quebec electricity when compared with the 2019 USA average of 400g CO2eq/kWh (EPA, 2019). However, CO2 equivalents are just a proxy for environmental impact and should not be understood to comprehensively account for social and environmental impact. Hydroelectric dam projects in Quebec, like the ones powering the GPSC have a sordid and complex history in the province. Innu Nation Grand Chief Mary Ann Nui spoke to this when she commented that "over the past 50 years, vast areas of our ancestral lands were destroyed by the Churchill Falls hydroelectric project, people lost their land, their livelihoods, their travel routes, and their personal belongings when the area where the project is located was flooded. Our ancestral burial sites are under water, our way of life was disrupted forever. Innu of Labrador weren't informed or consulted about that project" (Innu-Atikamekw-Anishnabeg Coalition, 2020).

[9]Note this estimate is based on the total number of hours spent running experiments from the M.Sc. dissertation this paper draws its experiments from. There were additional models trained for experiments that are not discussed in this paper. As such, this is a generous overestimation of $t$.

## B Qualitative Results

Question:

"Would you be comfortable with any of the voices you heard being played online, say for a digital dictionary or verb conjugator if no other recording existed?"

**Kanien'kéha responses**:

- Yes.
- yes
- Yes

- Out of the two voices I hear, the first was clearer to understand
- Yes, voices sounds really good!
- yes

**Gitksan responses**:

- yes
- Yes, but the ones that have the most whistling or buzzing would be annoying.
- maybe?? I think for a talking dictionary people do want to hear original pronunciations, but it could be a useful interim solution or a way to do short phrases!
- Yes
- Yes.
- Assuming there is a single control for the last section of the survey/test, then some of the synthesised voices actually sound really good and I would be comfortable hearing those in an online dictionary where audio didn't exist for a particular word or phrase.
- yes
- The ones with higher ratings for sure, some of the lower ratings were just about the sound quality because that hampered hearing the speech quality. So I may have confounded the results with that, but point remains that it is always good to try to avoid poor audio recordings for online dictionaries
- Maybe/yes
- only ones rated fair or above fair
- Absolutely yes
- yes, as long as they were identified as synthesized