A text-to-speech synthesis system for Border Lakes Ojibwe

Christopher Hammerly, Sonja Fougère
Department of Linguistics
University of British Columbia
Vancouver, Canada
chris.hammerly@ubc.ca
sonjaf16@student.ubc.ca

Giancarlo Sierra, Scott Parkhill
Harrison Porteous, Chad Quinn
CultureFoundry
Victoria, Canada
giancarlosierra, scottparkhill
harrison, chadquinn
@culturefoundrystudios.com

Abstract

This paper describes the development of a text-to-speech synthesis system for Border Lakes Ojibwe, which is being deployed within a web-based language learning platform. We discuss our approach to community engagement, recording and editing transcribed sets of utterances for model training, the technical implementation of the speech synthesis model itself, how the system is being used by teachers and learners within the web-based platform, strategies for future extensions of this type of work to other Indigenous voices, dialects and languages, and possibilities for applications in additional educational contexts and beyond.

1 Introduction

Ojibwe (known by speakers as Anishinaabemowin) is an Indigenous language of the Algonquian family consisting of a diverse set of mutually intelligible varieties spoken throughout large swaths of what is colonially known as Canada (through much of Ontario and westward to Manitoba) and the United States (from Michigan to outlying communities in Montana). In part as a consequence of the colonial policies of the United States and Canada such as the residential school system, which explicitly aimed to decrease the use of Indigenous languages and disrupt the passing of cultural knowledge, many Ojibwe speaking communities have seen a decline in the number of fluent speakers and in children learning the language from a young age.

Despite this difficult context, there are an estimated 28,130 speakers of the Ojibwe language within Canada (O'Donnell and Anderson, 2016), and robust revitalization efforts all across Ojibwe country. We present one component of our team's ongoing work to build technologies that can be used as tools for language revitalization: a text-to-speech (TTS) synthesis system, which is being deployed through a web-based platform for language learning that currently has over 3,000 active users. We describe an initial project that has built these tools for use by language instructors and learners in school and community settings within the Treaty #3 lands of Northwestern Ontario, where the Border Lakes variety of the Southwestern Ojibwe dialect group is spoken (Valentine, 1994). We especially focus on our process for creating training data for Indigenous speech synthesis systems.

2 Background

2.1 Positionality and community engagement

The project was initiated by the Seven Generations Education Institute in Fort Frances, Ontario as part of their *Anishinaabemodaa* "Waking up Ojibwe" language initiative, and has been conducted in collaboration with a team of researchers at the University of British Columbia, the Halifax-based language revitalization organization SayItFirst, and the Victoria-based educational start-up CultureFoundry. We include positionality statements from each member of the team who has worked directly on the technical side of the TTS:

- Hammerly is of mixed Anishinaabe and Norwegian-American descent and a member of the White Earth Nation in Minnesota, who currently works as a professor of linguistics.
- Frazier is of Acadian descent, hailing from rural Nova Scotia, and works as a graduate student in linguistics.
- Sierra is an Ecuadorian mestizo with Indigenous Latin American and European ancestry who now resides in British Columbia and works as a software engineer.
- Parkhill is a white Canadian and works as a software developer.
- Porteous is of European-Canadian descent with multidisciplinary interests, working in

programming and software development for language revitalization efforts.

- Quinn is of European-Canadian descent with a background in computational intelligence, and is currently leading CultureFoundry.

This work emerged in early 2020 from long-standing collaborations between members of our team and individuals and organizations within Treaty #3 lands. The Seven Generations Education Institute, whose leadership has continually guided this project, is an Indigenous owned and controlled community post-secondary institution operated by a board that represents ten First Nations in the Rainy River area. They initially approached their long-time partner in educational resource creation SayItFirst about building an online educational platform for Ojibwe. These organizations then partnered with CultureFoundry to create the platform and related tools, including the TTS system presented here. In August 2020, CultureFoundry approached Hammerly to work on the project. Hammerly has worked with elders and language teachers in the Rainy River region since 2016 in his capacity as a PhD Student in linguistics, as a professor of linguistics, and as a program assistant at an adult language immersion camp in Northern Minnesota. While he is part of the wider Ojibwe community, he has no known direct family connections to Treaty #3 communities.

During the summer of 2020 before work on the project began, members of SayItFirst approached community elders on behalf of our team in the traditional way with an offering of *asemaa* (tobacco) and a gift, and asked for guidance regarding whether such a project should be undertaken. After positive feedback during this consultation, we started work on the project, and have continued to consult with these elders on an ongoing basis as they record sentences for model training. For example, one concern that has been raised as we have presented this work in academic settings (including by a reviewer of this paper) is whether we have placed restrictions on the generation of obscene or inappropriate speech. Based on our consultations, we have not placed any restrictions on what can be generated. The idea conveyed to us is that all parts of the language have their place and should be able to be represented, and that there is trust and respect within the community that will guide how these tools are used. We also note that ownership of all recordings associated with this project are retained by the individuals who are contributing their voices, participants are paid for the time they spend creating recordings, and the participants have the right to pull out of the project (e.g. to stop making recordings or have existing recordings used within the project) at any time.

2.2 Related work on speech synthesis

While there is a rich history of research and development on speech synthesis for Western and other so-called majority languages (e.g. Taylor, 2009), speech synthesis for Indigenous languages, including text-to-speech systems, is still a significantly under-developed area of research. One of the most significant challenges is many Indigenous languages are also "low resource", in the sense that there are no existing data sets such as transcribed corpora of speech that can be leveraged for training, and many active speakers are aging and face significant time pressures to engage with a wide range of cultural and linguistic revitalization efforts.

We are aware of a number of recent efforts to develop speech synthesis for Indigenous languages of North America: Pine et al. (2022), who developed such systems for Kanien'kéha (also known as Mohawk; Iroquoian), Gitksan (Tsimshianic), and SENĆOŦEN (Coast Salish) with data sets ranging from 25 minutes to 3.5 hours, comparing Tacotron2 (Shen et al., 2018) and FastSpeech2 (Ren et al., 2021) models; Harrigan et al. (2019), who present a speech synthesizer for Plains Cree (Central Algonquian) based on 2.5 hours of speech and the Simple4All system (Simple4All, 2011–2014); Lütkebohle (2020) who describes a TTS system with Tacotron2 for Cherokee (exact number of hours unclear, but less than 3); Duddington and Dunn (2007) for Mohawk using eSpeak; and Whitman et al. (1997) for Navajo (details unknown).

3 Building utterance data sets

We recorded new sets of utterances to form the training data for our model. We chose this path for a number of reasons. First, to our knowledge, there are no existing large-scale data sets of high-quality transcribed audio for any dialect of Ojibwe. Second, a major priority set out by our community partners is to create a speech synthesis system that reflect the particular way that members of Border Lakes communities speak, including capturing micro-variation that exists between speakers and

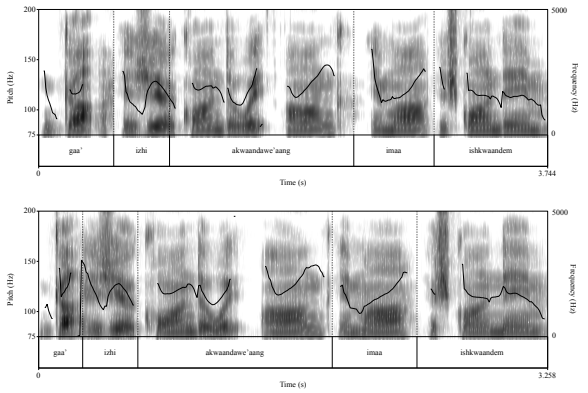

Figure 1: Spectrogram and pitch-track comparison of a speaker audio (top) versus synthesized audio (bottom) of the same string of text.

sub-varieties within Treaty #3 lands. Finally, this gives greater control over the quality of the TTS output, with the input being tailored to its specific application for educational contexts.

We are currently working with three speakers of Border Lakes Ojibwe—one highly fluent heritage speaker who primarily learned Ojibwe as an adult, but received significant exposure as a child, and two elders who learned Ojibwe from birth, one male and one female, each from a different community within the region. This will ultimately result in three distinct speech synthesis systems for the Border Lakes variety. The overall plan is to record a large set of training utterances with the heritage speaker (20,000 utterances) and a smaller set of utterances with the two elders (5,000 utterances) in order to minimize the strain on time for the elders. We then plan to use the heritage speaker model as a base, using the two elder data sets to fine-tune new voices. At present, we are still in the early stages of utterance recording with the two elders who are contributing their voices, therefore we focus here on the process for the heritage speaker model, where we have recorded over 15,000 utterances.

3.1 Utterance corpus

We entertained two possible approaches to creating a set of utterances: (i) source text from existing stories and books in Ojibwe for speakers to record, or (ii) record new stories and sentences and transcribe them into text. Given the high labor demands associated with transcription, and the current absence of automated speech recognition software for Ojibwe, which could streamline the transcription process, we pursued the former option: to record existing texts to build our training data set.

We first sourced all books we could find in Ojibwe, restricting our set to those identified with the broader Southwestern Ojibwe dialect group. All books were written using the "double vowel" orthography system based in the latin alphabet (originally devised by Charles Fiero), which is the standard system used across the Southwestern group and thus an important component of model training within this context.

We then digitized the collected books and used optical character recognition (OCR) software. We then converted the Ojibwe portions of each volume to plain text, and used a combination of regular expression-based automation and by-hand checking by an RA familiar with the Ojibwe language to ensure that all OCR errors were corrected in the plain text versions (e.g. 'm' being parsed as 'ln'). Finally, we created "split" versions of the texts that consisted of a single utterance per line. Splitting was done automatically as a first pass using sentence-final characters (i.e. '.', '?', and '!'). Since longer sentences are harder for speakers to fluently record and potentially decrease the quality of training due to increased issues at the forced-alignment stage, we isolated sentences that were longer than 170 characters and split them by hand, ensuring that breaks were made at natural points such as commas or other major clausal boundaries.

3.2 Utterance recording

For our recording hardware, we used a Rode NT-1 KIT cardioid condenser microphone with a Focusrite Scarlett 2i2 USB audio interface and a Microsoft Surface laptop. We additionally included a set of portable acoustic panels to improve isolation. The biggest challenge for the recording process was to create high-quality audio outside of a studio environment. Use of a studio was not possible for a number of reasons, with the primary issue being that the nearest studio was at least a 90 minute drive from our speakers. We therefore opted for an at-home and portable recording set-up that struck a balance between accessibility, comfort, and quality. We further instructed speakers to find a quiet room away from appliances or environmental noise.

We used SpeechRecorder (Draxler and Jänsch, 2004) as our recording software, which provides a user-friendly interface for prompting and recording a single high-quality audio file for each individual utterance. We split our utterance set into

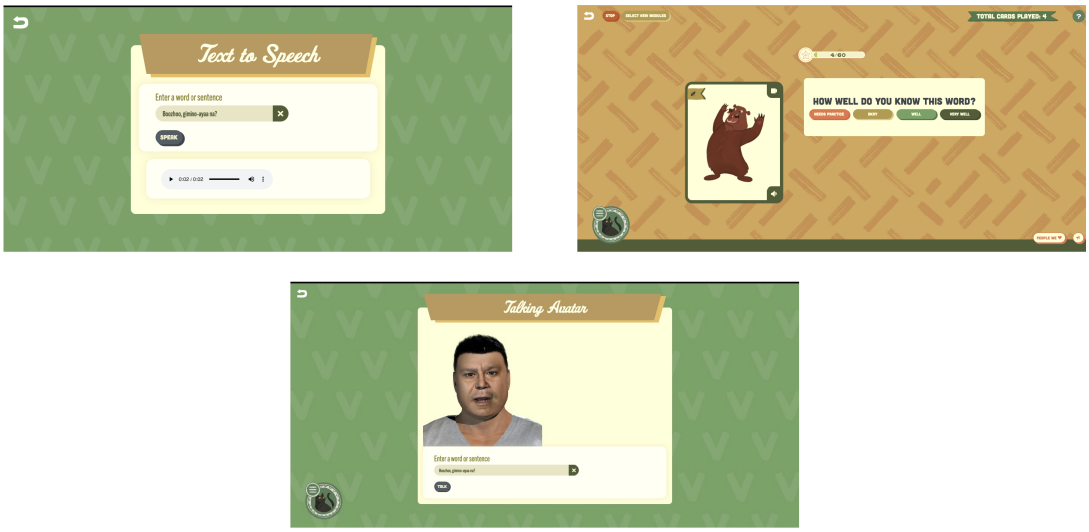

Figure 2: Integration of text-to-speech system as standalone component (upper left), in the flashcard activity (upper right), and with talking avatar (bottom).

projects of 300-500 utterances, and speakers were instructed to talk as if they were speaking to a learner of Ojibwe, and to re-record any utterances with pauses, disfluencies, or missing words. Furthermore, our speakers could edit text within the application if they noticed a typo, wanted to change a word to better fit their dialect, or change spelling to better match the conventions of the community. This was a critical feature, as the texts represented a range of sub-varieties within the Southwestern group. These corrected and adapted versions were used as the transcriptions for the model training. To ensure all portions of the audio were captured, we built-in a pre-recording delay of 1,000 ms and a post-recording delay of 500 ms. All audio was recorded as a .wav file on a single channel at a rate of 44,100 Hz and a 16 bit depth.

3.3 Utterance editing

There were two main facets of the raw recordings that required editing prior to use for model training: (i) all utterances were preceded and followed by silence (see previous section for reasoning), and (ii) despite much care being taken by our speakers, there was occasional background noise from fans and other home appliances. We created a python-based pipeline using the packages *librosa* (McFee et al., 2015), *numpy* (Harris et al., 2020), and *noisereduce* to automate the editing of these recordings to trim silences from the beginning and end of each utterance (leaving all pauses and silences that occurred during an utterance), to reduce noise via spectral gating and with a low-pass filter, normalize loudness across utterances from different sessions, and re-sample to a rate of 22,050 Hz, the rate recommended for training our model.

4 Speech synthesis model

For the text-to-speech, we adopted VITS (Kim et al., 2021), a parallel end-to-end model, trained using the Coqui AI TTS library (Gölge, 2021). Our current data training set includes 11,265 unique utterances recorded by our heritage speaker collaborator, which was comprised of 54,866 words and 556,455 characters. The utterances had a mean length of 4.87 words (SD = 2.56, Range = [1, 22]) and 49.39 characters (SD = 25.06, Range = [3, 170]). All utterances include the edited audio file and associated transcription in the double vowel orthography. After editing, the utterance set used to train the model had a total duration of 696.93 minutes or 11.62 hours. The mean duration of each utterance was 3.87 seconds (SD = 2.04 , Range = [.279, 16.02]). We trained the model from scratch on an array of GPUs consisting of 4x NVIDIA RTX A6000s, 400GB of RAM, 48GB of VRAM per GPU, with 56 virtual CPUs. Total training time was 92.6 hours with a batch size of 64 for 80k steps with otherwise standard model parameters.

As discussed in more detail in §6, we have not yet systematically evaluated the results produced by our model. However, we present a side-by-side comparison of a test utterance recorded by our speaker, but not used to train the TTS, and the same string produced synthetically with the resulting TTS system in Figure 1. We have also

demonstrated the feature to numerous community groups, teachers, and elders, and this informal feedback has been uniformly positive, usually including a sense of disbelief and surprise that the voice has been generated synthetically rather than from real recordings based on the tone and accuracy of the pronunciations.

5 Deployment in web-based application

The TTS system is being deployed in various ways within the *Anishinaabemodaa* web-based language learning platform produced in collaboration with teams at the Seven Generations Education Institute, SayItFirst, CultureFoundry, and the University of British Columbia.

There are three primary ways that the TTS system is being deployed or planned to be deployed (Figure 2). First, there is a "text-to-speech" page in the application where users can input any word or sentence in Ojibwe and generate an audio file of synthesized speech. This audio can be sped up and slowed down within the application and downloaded for offline use.

Second, the model for the platform allows teachers to input and generate their own materials, including lists of vocabulary that students should learn. These can then be integrated into games, including a flashcard activity. When a new word is added, audio files of synthesized speech are automatically generated, which can then be accessed by learners by clicking on the audio icon on the flashcard. This allows teachers maximum flexibility to have instant multi-modal media for each new word added to their lessons.

Finally, we have integrated the TTS system into a beta version of an animated talking avatar, providing a multi-modal tool where learners can input a string of text and see and hear how that text is pronounced. The current version is based off a lip-syncing technology developed by Speech Graphics, but has the limitation of being tuned to an English library. Our team is currently undertaking development of a Unity-based 3D model that can be customised to the articulatory properties of Ojibwe.

6 Outlook and extensions

So far, our TTS model of Ojibwe based on the voice of a single heritage speaker has been applied for pedagogical purposes. Moving forward, we envision the developed system as a tool for increasing the accessibility of written materials in the Ojibwe language. For example, those who may have limited literacy in the language, or Ojibwe language users with visual impairments, can use the TTS system to generate audio for a given text to be able to engage with listening rather than reading.

Second, we are currently in the process of recording data sets with two elders who are L1 speakers of Border Lakes Ojibwe. With this, we have had success with the same pipeline based in Speech Recorder, which provides flexibility and independence for our collaborators. Data from these multiple speakers will be combined to jointly train a multi-speaker TTS model.

Finally, we are making plans for systematic evaluation of our model on a variety of dimensions. So far, we have collected informal and anecdotal feedback from our partners about the quality of synthesized speech being produced, which has been uniformly positive. We plan to use a mixture of objective measures such as Mel cepstral distortion (Weiss et al., 2021), subjective measures such as ABX testing (Choi et al., 2021), and objective measures such as eye tracking (Govender and King, 2018). In all cases, we will engage in a user-centered research program that focuses on evaluation relative to the context of use and the needs of the community (Wagner et al., 2019).

References

Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34:16251–16265.

Christoph Draxler and Klaus Jänsch. 2004. Speechrecorder: A universal platform independent multi-channel audio recording software. In *LREC*.

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

Eren Gölge. 2021. Coqui TTS. https://www.coqui.ai.

Avashna Govender and Simon King. 2018. Using pupillometry to measure the cognitive load of synthetic speech. *System*, 50(100):2018–1174.

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.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

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

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25.

Vivian O'Donnell and Thomas Anderson. 2016. The Aboriginal languages of First Nations people, Métis and Inuit. Technical report, Statistics Canada.

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.

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*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Simple4All. 2011–2014. Simple4all: developing automatic speech synthesis technology. https://simple4all.org/.

Paul Taylor. 2009. *Text-to-speech synthesis*. Cambridge university press.

Jerry Randolph Valentine. 1994. *Ojibwe dialect relationships*. The University of Texas at Austin.

Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Eva Szekely, Christina Tånnander, et al. 2019. Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*.

Ron J Weiss, RJ Skerry-Ryan, Eric Battenberg, Soroosh Mariooryad, and Diederik P Kingma. 2021. Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5679–5683. IEEE.

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