A Preliminary Plains Cree Speech Synthesizer

Atticus Harrigan galvin@ualberta.ca
Timothy Mills timills@ualberta.ca
Antti Arppe arppe@ualberta.ca

Alberta Language Technology Laboratory
University of Alberta
Edmonton, Alberta

Abstract

This paper discusses the development and evaluation of a Speech Synthesizer for Plains Cree, an Algonquian language of North America. Synthesis is achieved using Simple4All and evaluation was performed using a modified Cluster Identification, Semantically Unpredictable Sentence, and a basic dichotomized judgment task. Resulting synthesis was not well received; however, observations regarding the process of speech synthesis evaluation in North American indigenous communities were made: chiefly, that tolerance for variation is often much lower in these communities than for majority languages. The evaluator did not recognize grammatically consistent but semantically nonsense strings as licit language. As a result, monosyllabic clusters and semantically unpredictable sentences proved not the most appropriate evaluate tools. Alternative evaluation methods are discussed.

1 Introduction

While majority languages such as English provide ample data for the creation and training of speech recognition, corpus annotation, and general language models, under-resourced languages are in a unique position to benefit greatly from such technologies. Speech Synthesizers, mobile keyboards, in-browser reading guides and smart dictionaries all provide invaluable tools to help aid language learners in gaining proficiency in languages where speaker numbers are falling. With the ubiquity of speech synthesis systems in public transit, emergency broadcast systems, and (most notably) mobile phones, under-resourced language communities often clamour for such technology. In addition to the positive social implications of having your language associated with technological innovation, speech synthesis systems provide a very real benefit in endangered and under-resourced language communities: while it is unfeasible for elders and remaining fluent speakers to detail every possible word and story, an ideal speech synthesizer allows for a learner to hear, on demand, any word, phrase, sentence, or passage. Despite this obvious benefit, endangered language groups rarely have such technology available, especially in the context of North American Indigenous languages.

This paper details the development of an early synthesizer for Plains Cree, an Indigenous language of Canada, and an evaluation of the resulting system. Through synthesis via the Simple4All suite of programs, this paper documents the pros and cons of such a system, and investigates how a speech synthesizer can be evaluated in the context of North American Indigenous languages.

2 Plains Cree Phonology

Plains Cree is a polysynthetic language of the Algonquian family spoken mainly in western Canada. With a recorded speaker count of nearly 35,000[1], Plains Cree is classified as a stage 5 language (or "developing") according to the Expanded Intergenerational Transmission Disruption Scale (Ethnologue, 2016).

Much of the literature on the language has focused on its morphosyntax. The phonetics and phonology have been less well described. Indeed, only five of the 50 pages of Wolfart's (1996) sketch of the language deal with what can be categorized as phonetics or phonology, and only four pages of his earlier, more comprehensive, grammar were dedicated to the topic (Wolfart, 1973). In the former, which is the more phonologically inclined, Wolfart identifies the phoneme inventory

[1] Although this is the current cited number of speakers, it is likely that this is an optimistic count of speakers. The true number is likely several thousand speakers lower, though reliable demographics are difficult to obtain.

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 1 Papers, pages 64–73,
Honolulu, Hawai'i, USA, February 26–27, 2019.

of Plains Cree as containing eight consonants, two semi-vowels/glides, and seven vowels (three short and four long).[2]

Although Plains Cree vowels are distinguished by length, it is also the case that the long vowels are qualitatively different from the short vowels (with the short vowels being less peripheral than their long counterparts (Harrigan and Tucker, 2015; Muehlbauer, 2012)). Further, the issue of quality is exacerbated by the language's system of stress, which is poorly understood. Consonants show similar, though better understood, variation. Stops are generally voiceless word initially, geminated when between two short vowels, and show free-variation between voiced and voiceless realization otherwise (Wolfart, 1996). The affricate /ts/ and fricative /s/ appear in seemingly free variation with [tʃ] and [ʃ], respectively (Wolfart, 1996). Plains Cree also possesses diphthongs, though only as allophonic realizations of vowels adjacent to glides. When either long or short /a/, /o/, or /e/ occur before /j/ the diphthongs [aɪ], [oɪ], [eɪ] occur (respectively). When these come before /w/, the diphthongs [aʊ], [oʊ], and [eʊ] are realized (respectively). When short or long /i/ occurs before /j/, it is lengthened; when short /i/ comes before /w/, [o] is realized, while long /i:/ in the same position produces the diphthong [iʊ]

Possible syllables are described by Wolfart (1996, 431) as having an optional onset composed of a consonant (optionally followed by a /w/ if the consonant is stop, an affricate, or /m/), a vowel nucleus, followed by an optional coda composed of a single consonant or either an /s/ or /h/ followed by a stop/affricate. According to Wolfart (1996, 431) the vast majority of Cree syllables have no coda.

Plains Cree has a standardized orthography referred to as the Standard Roman Orthography (SRO) The standard is best codified by Okimâsis and Wolvengrey (2008) and makes use of a phonemic representation, ignoring allophonic variation such as the stop voicing described above. The standard also offers morphophonological information; for example, although the third-person singular *apiw* ('S/he sits') is pronounced /apo/, it is not written as such, so as to allow the reader to apply morphological rules such as inflecting for the imperative by removing the third-person morpheme, {-w}, and adding {-tân}. Were the third-person form written <apow> or <apo>, one might incorrectly expect the imperative form to be /apota:n/, rather than the correct /apɪta:n/. Vowels in the SRO are marked for length by using either a circumflex or macron over the vowel, and <e> is always to be written as long. When diphthongs are allophonically present, the underlying phonemes are used, as in the example of [apo] being written as <apiw>. The orthography is mostly shallow: the seven consonants (/p, t, k, ts, s, h, m, n/) are represented by single graphemes (< p, t, k , c, s, h, m, n>); short vowels (/a, i, o/) are written without any diacritic (<a, i, o>), while long vowels (/a:, i:, o:, e:/) are written with a circumflex or macron (<â, î, ô, ê>)

While this standard has been codified in Okimâsis and Wolvengrey (2008), it is not universally (or even largely) adopted. While major publications such as Masuskapoe (2010), Minde (1997), or those put out by provincial/federal departments, are written in the SRO, many publications and communications (especially those informal) are done using nonstandard conventions. In the Maskwacîs Dictionary of Cree Words(1997), for example, <ch> is used in place of <c>, <h> is used (in addition to its regular phonemic representation) essentially to impart that a vowel is different than the expected English pronunciation, and vowel length is not identified. The Alberta Elder's Cree Dictionary (LeClaire et al., 1998) does mark vowel length, though <e> is always written without length marking.

These orthographic variations, as well as the paucity of data provide challenges for language technology.

3 Speech Synthesis

Broadly speaking, speech synthesis comes in two main forms: *parametric* and *concatenative*. Parametric synthesis aims to recreate the particular rules and restrictions that exist to manipulate a sound wave as in speech (Jurafsky and Martin, 2009, 249). Concatenative synthesis, rather than focusing on recreating the parameters that produce speech, concerns itself with stitching together pre-existing segments to create a desired utterance (Jurafsky and Martin, 2009, 250).

The contemporary focus of speech synthesis is so-called *text to speech* (TTS) (Jurafsky and Martin, 2009, 249), wherein input text is transformed

[2]There is only one /e/ phoneme in the language and it is generally considered long for historical reasons. Initially, Plains Cree did have a short /e/, but this eventually converged with long /i/ (Bloomfield, 1946, 1).

into sound. In this process, text is phonemically represented and then synthesized into a waveform (Jurafsky and Martin, 2009, 250). Phonemic representation is accomplished though various means: dictionaries of pronunciation offer up a transcription of common words and sometimes even names, though these are almost guaranteed to *not* contain every single word a synthesizer could expect to encounter (Jurafsky and Martin, 2009). As a result, other methods such as "letter-to-sound" rules (which aims to provide a phonemic representation for a grapheme given a context) have also been employed (Black et al., 1998). Other units of representation have also been considered, such as those smaller than the phoneme (Prahallad et al., 2006). In general, most contemporary TTS systems generate a phonemic representation through machine learning (Jurafsky and Martin, 2009, 260). Black et al. (1998), for example, instituted a system wherein a computer is fed set of graphemes and a set of allowable pronunciations for each of those graphemes, along with the probabilities of a particular grapheme-to-phoneme pairing. More recent techniques such as those described by Mamiya et al. (2013) instead take a subset of speech and text data that are manually aligned and, using machine learning algorithms, learn the most likely phonemic representation of each grapheme given the context. In any case, grapheme-to-phoneme alignment results in a system that is able to produce sound for a given text sequence. A speech synthesis system will also provide intonational information to best reflect naturalistic speech (Jurafsky and Martin, 2009, 264).

While there have been speech synthesis efforts for minority language (Duddington et al., n.d), little focus has been paid to North American languages. Although there appears to be a Mohawk voice available for the eSpeak open-source software project (Duddington et al., n.d), the only published account of an Indigenous language synthesizer seems to be a 1997 technical report detailing a basic fully-concatenative synthesizer for the Navajo language (Whitman et al., 1997). In this instance, the authors compiled a list of all possible diphones (two-phoneme pairs) and had a Navajo speaker read these in a list (Whitman et al., 1997, 4). These diphones were then manually segmented, concatenated, and adjusted for tone (as Navajo is a tonal language) (Whitman et al., 1997). According to the authors, although the system was small and lacked much of the data one might prefer when building a speech synthesizer, the concatenative method they used managed to produce an intelligible synthesizer (Whitman et al., 1997, 13). Other than this effort, it appears that speech synthesis for North American languages has been largely non-existent. The reason for this is likely due to the lack of resources in these languages. Languages of North America may lack even a grammar, though many will have a variety of recordings of important stories or conversations (Arppe et al., 2016). Few languages of North America have a standard and well established written tradition. As a result, speech synthesis development is necessarily difficult for these languages. Plains Cree, as one of the most widely spoken indigenous languages of North America with roughly 20,000 speakers (Harrigan et al., 2017), provides relatively large amounts of standardized text for a North American language (Arppe et al., forthc.). The sources range from biblical texts (Mason and Mason, 2000), to narratives (Vandall and Douquette, 1987), and even interviews between fluent speakers (Masuskapoe, 2010). The texts used for TTS synthesis in this paper are described in Section 4.

4 Materials

4.1 Training Data

This study uses two varieties of materials: training data and a TTS toolchain. Training data comes in the form of the biblical texts of Psalms from Canadian Bible Society (2005). These texts, narrated by Dolores Sand, are accompanied by transcriptions in the SRO (Canadian Bible Society, 2005). The audio files are uncompressed, stereo audio with a 16 bit depth, 44,100 sampling frequency. Not all Psalms were available as an audio recording, though 52 files totaling 2 hours 24 minutes and 50 seconds in length were available as training data. Because the toolchain discussed below requires files with only a single channel for input, the Psalm recordings were converted into mono-channel files using the *SoX* utility (Bagwell, 1998–2013). Finally, the computation was completed on a virtual server with an Intel 2.67 GHz Xenon X5650 processor and 16 GB of RAM.

4.2 Tool Chain

Simple4All (Simple4All, 2011–2014) aims to develop lightly or unsupervised speech recognition/synthesis (Simple4All, 2011–2014). Two of the major outputs of the project are *ALISA*, a lightly supervised alignment system (Simple4All, 2014), and *Ossian*, a front-end synthesizer (Simple4All, 2013-2014). ALISA aligns based on ten minutes of pre-training data which has been manually segmented (Stan et al., 2013). The system learns and then attempts to align the rest of the training data and text transcript provided to it (Stan et al., 2013) (as detailed later, ALISA alignment was not particularly successful, and so hand alignment was conducted.). Resulting from alignment is a set of utterance sound files and text files with the respective orthographic representation. These files are then fed directly to into Ossian which uses these data to train itself and produces a synthesized voice. Ossian itself is a 'collection of Python code for building text-to-speech (TTS) systems, with an emphasis on easing research into building TTS systems with minimal expert supervision' (Simple4All, 2013-2014).

5 Evaluation

Two iterations of the synthesizer were created. The first iteration was built on alignment from ALISA. In order to evaluate the synthesized voice, a combination of various common metrics was used. Both functional (i.e. intelligibility of the system (Hinterleitner, 2017, 24)) and judgment (i.e. how pleasant the system is to use) tests were implemented. Ideally, at least a dozen participants would be used, but due to the limited number of people who speak and are literate in the language, and being mindful of the need to not bias speakers for future non-pilot level tests, one speaker was deemed appropriate for each iteration of the synthesis[3].

For functional level analysis, a modification of the Cluster Identification Task (Jekosch, 1992) was used. In this modified task, basic V, CV, and CVC syllables (*not* words) were randomly presented (at least twice, but as often as the participant requests) after which the participant was asked to write down what they heard. Although complex onsets and codas exist in Plains Cree, they are certainly less frequent, as discussed above. Although each possible syllable would ideally be presented at least once, the number of evaluations would total nearly 2000 items, a task too onerous for a single session and participant. In order to acclimate the participants, three test scenarios using English syllables from a native English speaker (/skwi/, /cle/, and /ram/) were run.

In addition to the Cluster Identification Task, the participant was asked to take part in a modified Semantically Unpredictable Sentence task (Benoît et al., 1996). In this task the participant was asked to listen to sentences that, while morphosyntactically correct, were semantically unlikely such as *ê-mowat sêhkêw*, 'You eat the car', where *ê-mowat* licenses any grammatically animate noun like *sêhkêw*, 'car,' but is much more likely to refer to food than a vehicle.

After listening to the stimuli at least twice, the participant was asked to write down what they heard. This test produces a situation wherein speakers are less able to rely on context for discrimination (Hinterleitner, 2017) while making use of real words rather than just monosyllables. A total of 5 semantically regular sentences were also presented so as to investigate how well our system works. As we would expect a greater level of comprehension in these sentences, any other result would indicate very poor performance.

To assess the pleasantness of synthesis, a set of scales with opposing metrics was created. Scales where end points fall beyond metric markings were used (see Figures 1 and 2 for an example). This was done so as to avoid the tendency for participants to not rate at the ends of scales and the tendency for individuals to have difficulty in distinguishing a stimuli that is either terrible or very good (Hinterleitner, 2017). The scales were comprised of the following pairs: Natural vs. Unnatural, Pleasant vs. Unpleasant, Quick vs. Slow, and overall Good vs Bad. Judgments were elicited for two utterances: one a an excerpt from the training data, and one a synthesis of an utterance pulled from a corpus (Arppe et al., forthc.; Ahenakew, 2000; Bear et al., 1992; Kâ-Nîpitêhtêw, 1998; Masuskapoe, 2010; Minde, 1997; Vandall and Douquette, 1987; Whitecalf, 1993). The use of non-

[3]The second iteration actually contained two evaluations by the same participant due to the initial evaluation tasks containing non standard spellings in the Semanticaly Unpredictable Sentence and Judgment tasks. Only these tasks were re-administered (with new stimuli verified for orthographic consistency). The first evaluation of the second synthesizer was substantially similar to the second evaluation and so will not be detailed in this paper.

synthesized data was used as a point of comparison. The identities of these stimuli were not made known to the participant. Because the participant was a former language instructor, general comments regarding the usefulness of the synthesizer were invited.

6 Results

Although the methodologies described above were carefully considered for the task of evaluation, major modifications had to be made once the evaluation was actually undertaken. The first iteration of the synthesizer proved difficult for evaluation, with the participant barely able to complete most of the tasks. According to the participant, syntheses were unintelligible, unpleasant, too quick, and overall bad. Listening to the syntheses through headphones was so jarring that evaluation was done through basic laptop speakers. Judgment assessment of longer utterances was extremely difficult for the participant, and stimuli were deemed to be too long. The cause of this poor synthesis is likely due to the fact that ALISA managed to align only 24 minutes of nearly 2.5 hours of training data.

To respond to this issue, the entirety of the training data was aligned by hand. This second iteration of the synthesizer was substantially more natural, intelligible, and pleasant. As the first participant's schedule was quite busy, a second participant was recruited. This participant was a male, middle-aged former Cree-language teacher who is proficient in the SRO. The following section details the evaluation of the second iteration of the speech synthesizer. All stimuli were presented in a randomized order.

6.1 Modified Cluster Identification Task

So as not to exhaust the participant, 70 clusters were presented. Of the tested clusters, 21 were identified correctly in their entirety. Another 16 were classified as minor errors (accepted variation in the orthograhy) such as 6 instances of <ê> being written as <î> (likely the result of the two phonemes overlapping in vowel space (Harrigan and Tucker, 2015) as well as perhaps the influence of English orthography) and 7 long vowels being written as short vowels following by an <h> (a common non-standard way of indicating that a vowel is long). These two types make up the majority (13/16) of minor errors. Together, minor errors and correctly identified clusters make up the majority of responses (53%). There were 11 items where the onset was misheard, with the majority of these (6) being <c> misheard as <s>. Given that <c> represents /ts/, this is not wholly surprising. Remaining error types were smaller in their tokens: 3 clusters had a vowel identified correctly, but the participant missed the onset entirely; 4 vowels were identified with the correct quality but the wrong length; 6 clusters showed the wrong quality but the correct length; and 4 clusters were heard as incorrect vowel with incorrect lengths. See Table 6.1 for a full list of stimuli and results. A highlighted row indicates a sentence where the participant's transcription deviated from the input. Boldface letters indicate where the deviance occurred.

6.2 Semantically Unpredictable Sentences

Semantically Unpredictable sentences showed similar results. All sentences where the participant's transcriptions varied from the SRO input were semantically unpredictable. The differences in transcription were nearly always restricted to differences in vowel length. In the one case where vowel quality differed (*kikî-sîkinik minôs*), the error was an instance of input <î> being written by the participant as <ê>. As mentioned previously, this variation is unsurprising due to overlapping vowel spaces. Table 6.2 summarizes the evaluation of the SUS task. Those sentences preceded by a hash-mark are semantically unpredictable, while those without are semantically predictable. As above, a highlighted row indicates a sentence where the participant's transcription deviated from the input, and boldface letters indicate where the deviance occurred.

6.3 Judgment Tasks

Impressionistic judgment of the synthesizer shows that the system performed worse than an actual native-speaker production. Figure 1 represent the evaluation of the synthesizer, and Figure 2 the representation of the natural utterance. Unlike the first iteration of the synthesizer (which was almost entirely rated as negatively as possible on all scales), the second synthesizer was rated as somewhat *unnatural, unpleasant* and *bad*, but not drastically so. The synthesizer was rated as only slightly slower than the middle ground between *quick* and *slow*. The naturally produced stimuli

Stimulus	Heard	Stimulus	Heard
kit	kit	**y**it	it
tit	kit	wiit	wiit
coot	ot	mo**o**t	moht
sat	sat	wot	wot
woot	rot	n**ii**t	neet
maat	maat	s**ii**t	seet
miit	miit	s**oo**t	sat
wit	wit	hot	hot
taat	taat	sit	sit
hat	hat	seet	seet
keet	kit	naat	naat
kot	pat	**m**ot	not
kiit	keet	p**o**t	pat
pat	at	y**ee**t	yit
teet	NA	sot	sot
cot	sot	**c**it	sit
wat	wat	paat	paat
n**oo**t	not	h**ee**t	hiht
kaat	kaat	to**o**t	toht
not	not	**c**eet	**s**iht
h**aa**t	hat	**hii**t	neht
yaat	yaat	pit	pit
yot	eewt	kat	kat
ciit	seet	p**ii**t	peet
koot	pat	mat	mat
t**ii**t	teet	y**a**t	yeet
peet	peet	tot	tot
ho**o**t	hoht	**c**at	set
neet	neet	hit	ahiht
weet	weet	nit	nit
nat	nat	po**o**t	poht
tat	tat	**caa**t	saht
yiit	eet	wa**a**t	waht
yoot	eewt	m**ee**t	miht
mit	mit	s**aa**t	seht

Table 1: Cluster Identification Results

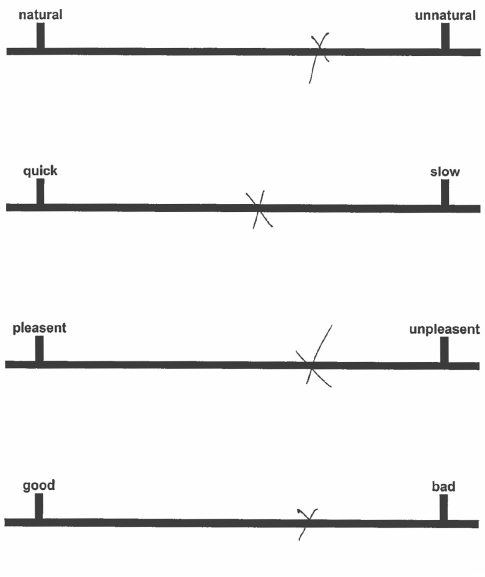

Figure 1: Synthesized Voice Judgment Task Evaluation

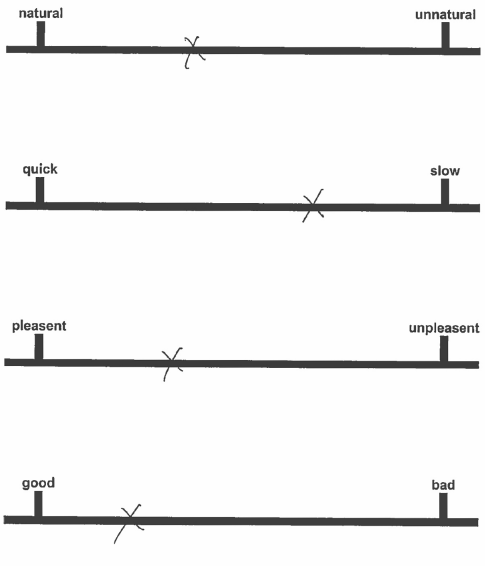

Figure 2: Naturally Produced Voice Judgment Task Evaluation

showed almost the opposite pattern, being modestly *natural*, *pleasant*, *good*, and *slow*.

Interestingly, the naturally produced utterance was not judged to be maximally *natural, pleasant,* or *good*. This is likely due to the fact that the utterance was not of quick or conversational speech, but rather a *performed* recording of biblical text.

7 Discussion

Although a synthesizer was developed, alignment through ALISA was unsuccessful, with just 24 minutes of roughly 2.5 hours correctly aligned (19%). According to the developer, Adriana Stan, despite attempting to create an alignment system for a wide variety of languages, polysynthetic languages such as Plains Cree were not considered or tested; as a result, ALISA's parameters dispreferer very long words, instead assuming an error in processing (p.c. Adriana Stan, Jan 31, 2018). Although ALISA allowed for some adjustment in this respect, changing the average acceptable length for a word to any extent did not produce any significant increase in alignment (either by lowering or raising this threshold). Further, ALISA

Input	Participant Transcription	English Gloss
# ê-pâhpisit sêhkêw	ê-p**a**hpisit sêhkêw	'The car laughs.'
# sîwinikan pimohtêw	sîwinikan pimohtêw	'Sugar walks.'
# ê-postiskawacik kinosêwak	ê-postiskawacik kinosêwak	'You put the fishes on.'
awâsis wâpahtam masinahikan	awâsis wâpahtam masinahikan	'A child sees a book'
ê-wiyâtikosit iyiniw	ê-wiyâtikosit iyiniw	'The Indigenous person is happy'
nâpêw ê-mîcisot	nâpêw ê-mîcisot	'The man is eating'
kinêpik nipâw	kinêpik nipâw	'A snake is sleeping'
#atim kiwî-saskamohitin	atim kiwî-saskam**ô**h**î**tin	'I am going to put a dog in your mouth.'
#kikî-sîkinik minôs	kikî-s**ê**kinik minôs	'The cat poured you.'
iskwêw pîkiskâtisiw	iskwêw pîkiskâtisiw	'A woman is sad.'

Table 2: Evaluated Sentences

seemed relatively successful for other agglutinating languages such as Finnish and Turkish with similarly long words; this suggests that the issue with alignment may be more complex than just word length.

The evaluation of hand-aligned data with Ossian synthesizer showed promising results. The second iteration of the synthesis was relatively well received by the second participant, who remarked that the speech synthesizer, while not perfect, was serviceable and represented an exciting opportunity for language learners. Despite this, multiple issues arose throughout the evaluation process. Most significant was the issue of orthographic proficiency. Because the tasks selected for this evaluation relied on written responses, only participants with strong literacy in SRO could be considered. This is especially problematic for Indigenous languages of Canada, as many of these varieties are historically oral languages. Few fluent speakers of Plains Cree actually possess the level of literacy needed for the evaluation tasks. This severely restricts who can participate in evaluation of the synthesizer. A side effect of this restriction was the scarce availability of a native speaker to review the stimuli (as one would prefer not to have reviewers act as participants in the study), leading to the several orthographic inconsistencies in the first evaluation for both iterations of the speech synthesizer. For the second iteration of the synthesizer, evaluation was re-administered for those items with orthographic inconsistencies. One solution to this issue is to eschew the need for writing. Instead of writing down what they hear, participants could be asked to provide word-to-word translations (insofar as possible) and to compare these with the intended meaning of the stimuli. Though this would not address the participants' ability to recognize particular graphemes, the lack of a need for literacy would allow for a signficantly larger number of participants than what would be available when requiring literacy in the SRO. Alternatively, participants could be asked to repeat what they have heard, though this would likely require a complete reworking of stimuli.

In building a synthesizer for Plains Cree, the full Simple4All toolchain proved unreliable. In addition to the issues faced in ALISA alignment for Plains Cree, documentation installation of the software provided multiple challenges with little documentation for support. Future endeavors should consider newer systems such as the Merlin project (Wu et al., 2016) which has been using Deep Neural Nets (DNN), a form of machine learning that seem to provide better results than the Hidden Markov Models (HMMs) used in the Simple4All project. Although no such comparison of DNN vs. other machine learning methods has been reported for synthesis of North American languages, DNN based systems such as Google's WaveNet report significantly better results than other techniques such as HMMs (van den Oord et al., 2016). If continuing to use Simple4All, the results of this evaluation suggest that one should not use ALISA for text alignment. In addition, researchers should consider the use of prebuilt aligners for languages with similar phoneme inventories. In the case of Plains Cree, given that the phoneme inventory is a subset of English's, one could consider the use of force aligners built for English with a modified dictionary specific to Plains Cree. This should be feasible in theory, though it remains to be seen whether it is useful

in practice.

In regards to evaluation, the SUS task might best be abandoned altogether in assessing future Plains Cree synthesizers. Results showed very little difference between semantically unpredictable sentences and semantically predictable ones. Both participants noted that the 'unpredictable' sentences were, in fact, ungrammatical: the line between unlikely and allowable is very thin. It would be interesting to repeat this task with other native speakers, as it may be a participant-specific attribute (though this study's participants had experience as second-language instructors and were likely far more familiar with listening, identifying and interpreting odd, infelicitous, and/or mispronounced utterances than the average speaker). If this is a tendency that holds across native speakers of Plains Cree, it may be worth investigating the factors influencing these attitudes (such as a tendency for speakers of the language to be either very novice or very fluent, perhaps leading to a lack of familiarity or tolerance of variation as seen in the SUS task). Assuming this attitude holds for the general population, it would be best to choose somewhat semantically predictable, but less frequent, stimuli (e.g. *I ate the zebra*, where *zebra* is more predictable than *car*, but less predictable than *rabbit*) or avoid the SUS task entirely. Of course, this means the confound of semantic predictability endures, though this might be addressed presenting words in isolation (accepting that this does not allow for sentence level prosody to be assessed).

Based on the feedback from the participant, reducing the number of syllables presented would be beneficial, though the details of how to do so remain unclear, especially considering that this study ignored a large portion of possible syllables. Random sampling could be used by selecting simply one type of each sound in each syllable position (e.g. ensuring there is tested at least one syllable starting with a stop and ending with a fricative). Spreading out the set over many participants, such that every syllable is evaluated the same number of times but not by each participant, is perhaps a better solution as it allows every syllable to be evaluated; in this case, one would have to analyze results via some sort of mixed-effects model (where *speaker* acts as a random effect) to account for the variation between speakers. Further, this method requires many participants, an inherent restriction in working with minority languages, especially those of North America. Finally, the first participant indicated that a few of the monosyllables, while unattested in dictionaries, were actually vulgarities. As this task was meant to assess only syllable intelligibility separately from word-level intelligibility, and due to their offensive nature, it is important that future studies remove these words or at least warn participants of their possible presence.

8 Conclusion

This paper presents one of the first parametric syntheses of an Indigenous language of Canada, using the Simple4All packages ALISA and Ossian. Based on roughly 2.5 hours of speech, this method of speech synthesis makes use of lightly supervised forced alignment to ease the workload required by the researcher. Although Simple4All has been used with a variety of languages (Simple4All, 2011–2014), the forced alignment was largely unsuccessful with the Plains Cree data. No conclusive reason could be found to account for this, though it may be that word length played a factor. In contrast, the results of the second synthesizer based on hand-aligned training data present promising results, with many of the stimuli being understood in their entirety. Although this second synthesizer was clearly identified as non-natural speech, its output was intelligible and relatively well received by the participant. Where the participant's transcription of stimuli deviated from the input, deviations generally concerned different vowel lengths. The results of this paper also indicate that careful consideration must be given to the evaluation frameworks, since those techniques that have become established and applied successfully for majority languages may not be suitable for Indigenous languages, at least for those in the Canadian context.

Acknowledgements

The authors of this paper would like to thank Dolores Sand, the speaker on whom the synthesis was based for here willingness to have her data used in this project. We would also like to thank Martti Vainio, Antti Suni, and Adriana Stan for not only their work in developing Simple4All, but their generosity in helping to troubleshoot throughout the synthesis. This research was supported by the Social Sciences and Humanities Research Council

of Canada (Grants 890-2013-0047 and 611-2016-0207), the University of Alberta Kule Institute for Advanced Study Research Cluster Grant, and a Social Sciences and Humanities Research Council of Canada Doctoral Fellowship. We would also like to thank Arok Wolvengrey and Jean Okimâsis for their help in reviewing stimuli, Arden Ogg for helping to facilitate our relationship with Dolores Sand, and community members of Maskwacîs, Alberta for their input throughout the process of developing the synthesizer. Finally, we would like to thank our anonymous participants for their expertise, patience, and participation.

References

Alice Ahenakew. 2000. âh-âyîtaw isi ê-kî-kiskêyihtahkik maskihkiy/They knew both sides of medicine: Cree tales of curing and cursing told by Alice Ahenakew, edited by HC Wolfart and F Ahenakew. *Publications of the Algonquian Text Society. Winnipeg: University of Manitoba Press*.

Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N Moshagen. 2016. Basic language resource kits for endangered languages: A case study of plains Cree. *CCURL*, pages 1–8.

Antti Arppe, Katherine Schmirler, Atticus G. Harrigan, and Arok Wolvengrey. forthc. A morphosyntactically tagged corpus for Plains Cree. In *Papers of the 49th Algonquian Conference*.

Chris Bagwell. 1998–2013. Sox – sound exchange, the swiss army knife of audio manipulation. http://sox.sourceforge.net/sox.html.

Glecia Bear, Minnie Fraser, Irene Calliou, Mary Wells, Alpha Lafond, and Rosa Longneck. 1992. *kôhkominawak otâcimowiniwâwa/Our Grandmothers' Lives: As Told In Their Own Words, edited by F Ahenakew and HC Wolfart*, volume 3. University of Regina Press.

Christian Benoît, Martine Grice, and Valérie Hazan. 1996. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, 18:381 – 392.

Alan W. Black, Kevin A. Lenzo, and Vincent Pagel. 1998. Issues in building general letter to sound rules. In *The Third ESCA/COCOSDA Workshop on Speech Synthesis, Blue Mountains, Australia, November 26-29, 1998*, pages 77–80.

Leonard Bloomfield. 1946. Algonquian. In *Linguistic structures of Native America*, volume 6, pages 85–129. Viking Fund Publications in Anthropology, New York.

Canadian Bible Society. 2005. *kihci-masinahikan nikamowina: âtiht kâ-nawasônamihk*. Toronto: Canadian Bible Society.

Jonathan Duddington, Martin Avison, Reece Dunn, and Valdis Vitolins. n.d. eSpeak speech synthesizer.

Ethnologue. 2016. Plains Cree.

Atticus Harrigan and Benjamin Tucker. 2015. Vowels spaces and reduction in Plains Cree. *Journal of the Canadian Acoustics Association*, 43(3).

Atticus G. Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27(4):565–598.

Florian Hinterleitner. 2017. *Quality of Synthetic Speech: Perceptual Dimensions, Influencing Factors, and Instrumental Assessment*, 1st edition. Springer Publishing Company, Incorporated.

Ute Jekosch. 1992. The cluster-identification test. In *CSLP-1992*.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Jim Kâ-Nîpitêhtêw. 1998. *Counselling Speeches of Jim Kâ-Nîpitêhtêw, edited by F Ahenakew and HC Wolfart*. University of Manitoba Press.

Nancy LeClaire, George Cardinal, Emily Hunter, and Earle G. Waugh. 1998. *Alberta Elders' Cree Dictionary: Alperta ohci kehtehayak nehiyaw otwestamâkewasinahikan*, 1st edition. Edmonton: University of Alberta Press.

Yoshitaka Mamiya, Junichi Yamagishi, Oliver Watts, Robert A.J. Clark, Simon King, and Adriana Stan. 2013. Lightly Supervised GMM VAD to use Audiobook for Speech Synthesiser. In *Proc. ICASSP*.

Maskwachees Cultural College. 1997. *Maskwacîs Dictionary of Cree Words/Nehiyaw Pîkiskwewinisa*. Maskwachees Cultural College.

William Mason and Sophia Mason. 2000. *Oski Testament.* Toronto: Canadian Bible Society.

Cecilia Masuskapoe. 2010. *piko kîkway ê-nakacihtât: kêkêk otâcimowina ê-nêhiyawastêki, edited by HC Wolfart and F Ahenakew*. Algonquian and Iroquoian Linguistics.

Emma Minde. 1997. *kwayask ê-kî-pê-kiskinowâpahtihicik/Their Example Showed Me the Way, edited by F Ahenakew and HC Wolfart*. Edmonton: University of Alberta Press.

Jeffrey Muehlbauer. 2012. Vowel Spaces in Plains Cree. *Journal of the International Phonetic Association*, 42(1):91–105.

Jean Okimâsis and Arok Wolvengrey. 2008. *How to Spell it in Cree.* Regina : Miywâsin Ink.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *Arxiv*.

Kishore Prahallad, Alan W. Black, and Mosur Ravishankar. 2006. Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 2006, Toulouse, France, May 14-19, 2006*, pages 853–856.

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

Simple4All. 2013-2014. Ossian speech synthesis toolkit. http://homepages.inf.ed.ac.uk/owatts/ossian/html/index.html.

Simple4All. 2014. Alisa: An automatic lightly supervised speech segmentation and alignment tool. http://simple4all.org/product/alisa/.

A. Stan, P. Bell, J. Yamagishi, and S. King. 2013. Lightly supervised discriminative training of grapheme models for improved sentence-level alignment of speech and text data. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, INTERSPEECH 2013 - 14th Annual Conference of the International Speech Communication Association, pages 1525–1529, (1)Communications Department, Technical University of Cluj-Napoca.

Peter Vandall and Joe Douquette. 1987. *wâskahikaniwiyiniw-âcimowina/Stories of the House People, edited by F Ahenakew*. University of Manitoba Press.

Sarah Whitecalf. 1993. kinêhiyâwiwininaw nêhiyawêwin/The Cree language is our identity: The La Ronge lectures of Sarah Whitecalf, edited and translated by HC Wolfart and F Ahenakew. *Publications of the Algonquian Text Society / Collection de la Société d'édition des textes algonquiennes. Winnipeg: University of Manitoba Press*.

Robert Whitman, Chilin Shih, and Richard Sproat. 1997. A navajo language text-to-speech synthesizer. Technical report, AT&T Bell Laboratories.

H. Christoph Wolfart. 1973. *Plains Cree: a grammatical study.* Transactions of the American Philosophical Society: new ser., v. 63, pt. 5. Philadelphia, American Philosophical Society, 1973.

H. Christoph Wolfart. 1996. Sketch of Cree, an Algonquian Language. In *Handbook of American Indians. Volume 17: Languages*, volume 17, pages 390–439. Smithsonian Institute, Washington.

Zhizheng Wu, Oliver Watts, and Simon King. 2016. Merlin: An open source neural network speech synthesis system. In *9th ISCA Speech Synthesis Workshop (2016)*, pages 218–223.