Towards low-resource text-to-speech generation for Indigenous Pacific Northwest languages: The case of Haida *X̲aad Kíl*[*]

Morris Alper
Independent Scholar

Samopriya Basu
Simon Fraser University

Nathan Bennett
Future Ancestors Alliance

Ryan Kessler
X̲aadas Kíl K̲uyáas Foundation

S. Verlaine Ravana
X̲aadas Kíl K̲uyáas Foundation

Wendy F. *K'ah Skáahluwaa* Todd
X̲aadas Kíl K̲uyáas Foundation & University of Alaska Southeast–Juneau

Abstract: We present a prototype text-to-speech (TTS) system for Haida (*X̲aad Kíl*), a critically endangered language isolate spoken in Alaska and British Columbia. The system addresses challenges in language revitalization by providing a pedagogical tool for pronunciation and listening practice when access to Elders is limited. We explore technical difficulties of developing TTS for low-resource languages with complex phonological systems, such as those found in many Pacific Northwest (PNW) languages, while navigating ethical concerns regarding data consent. Preliminary results reveal a trade-off between phonetic accuracy, audio quality, and voice anonymization, highlighting the need for tools that support complex phonologies without compromising linguistic fidelity. We also propose a community-based evaluation framework using the Indigenous Traditional Knowledge Framework (ITKF). This work demonstrates the potential for ethically-informed TTS technology to support language revitalization in Indigenous context, while underscoring the need for continued research and development of voice technologies for low-resource languages.

Keywords: Text-to-speech; Low-resource languages; Indigenous languages; Haida; Language revitalization; Data consent

1 Introduction

Indigenous communities across North America are reclaiming their ancestral languages as an act of resistance to colonial assimilation policies and as an affirmation of cultural identity and sovereignty. In addition, research (Harding et al. 2025) has shown that Indigenous language reclamation has observable positive effects on community health and well-being. However, language revitalization efforts are hindered by limited access to fluent Elders and high-quality audio resources essential for mastering pronunciation and developing speaking proficiency. For Haida (X̲aad Kíl), a language isolate spoken in Southeast Alaska and Haida Gwaii (off the coast of British Columbia), this challenge is particularly acute. The language is critically endangered (UNESCO

[*] Háw'aa, we thank Áljuhl (Erma Lawrence) for her dedication to preserving the Haida language which made this project possible, Jordan Lachler for critical work on documenting Haida, Yakov Kolani for useful feedback, and Anthony Webster and Mark Turin for suggesting articles on misuse of AI in the Indigenous languages sphere. Errors, if any, are our responsibility alone.
Contact info: morrisalp@gmail.com, basusamapriya@gmail.com, nathan@futureancestorsalliance.org, wftoddsmythe@alaska.edu

2010), with fluent speakers numbering in the single digits. Access to Elders and practice opportunities is further constrained by geographic isolation and limited communication infrastructure in rural and remote communities. Additionally, Haida and other Pacific Northwest languages have complex sound systems posing a significant challenge for learners, for whom pronunciation is a barrier to achieving basic proficiency. To address this gap, culturally-informed Text-to-Speech (TTS) systems show promise as a supplementary resource, enabling learners to practice pronunciation and build fluency beyond the direct mentorship of Elders.

From a technical standpoint, developing TTS for Haida poses challenges that are compounded by technical barriers inherent to low-resource languages. Even many languages spoken by millions (e.g. Javanese) remain digitally under-resourced, hindering naive data-driven approaches. Indigenous North American languages face additional obstacles, as their extreme typological differences from Eurasian languages frequently mismatch existing assumptions and architectures in Natural Language Processing (NLP). The languages of the Pacific Northwest, a well-known language area or *Sprachbund* (Beck 2000; Thomason 2015), present particularly acute challenges for voice technologies due to their complex consonantal inventories, featuring distinctions foreign to English (e.g. uvular and pharyngeal places of articulation, lateral fricatives, contrastive glottalization, and contrastive tone). Conversely, these phonetic complexities are precisely what make TTS technology critically needed as a pedagogical tool for these languages.

This paper introduces a prototype text-to-speech (TTS) system developed specifically for the Haida language, designed as a community-informed pedagogical tool. The system addresses the fundamental challenge of training effective models with severely limited audio data while navigating complex ethical considerations around data use, consent, and cultural protocols. Haida culture has established values relating to ownership of oral narratives and songs as well as community consensus-building, and we use these cultural principles to inform our approach. In building our system, we explore the trade-off between speaker anonymity, audio quality, and phonetic accuracy, due to ethical concerns surrounding the use of TTS systems with voice sourced posthumously. We believe these considerations are important for the development of ethical Artificial Intelligence (AI), particularly in Indigenous language contexts.

The remainder of this paper is structured as follows: Section 2 contextualizes our work within low-resource NLP and Indigenous language TTS research while addressing ethical considerations around AI in Indigenous language contexts. Section 3 details our data sources, model architecture, and engineering contributions. Section 4 presents evaluation results and a planned framework for future community-oriented evaluation. Section 5 presents current limitations of our approach, and concludes with directions for future research and potential for expansion to other Pacific Northwest languages in collaboration with interested communities.

2 Background, Related Work and the Ethics of AI in the Indigenous Languages Sphere

We proceed to place our work in context by discussing TTS systems for low-resource languages (2.1), Indigenous language technologies (2.2) and relevant ethical and technical challenges, proposed ethical frameworks for approaching these issues (2.3).

2.1 TTS for Low-Resource Languages

Mature voice technologies for high-resource languages such as English are typically trained on thousands of hours of transcribed audio, exemplified by the ~700K hours of data used for the popular Whisper automatic speech recognition model (Radford et al. 2023). Unfortunately, these

approaches are not feasible for languages with limited digital footprint. However, the promise of TTS systems for language education has attracted significant interest in their application to low-resource languages and Indigenous languages in particular (Pine et al. 2025). Recent advances have made it possible to develop functional TTS systems with much smaller datasets, in some cases containing as little as half an hour of transcribed audio.

Existing approaches to low-resource TTS may use various strategies to increase performance on a limited data budget. These include methods such as multilingual training and transfer learning strategies (Lux et al. 2022; Wang et al. 2025), data augmentation (Byambadorj et al. 2021), and the use of models with greater known data efficiency (Pine et al. 2025). Our approach similarly achieves usable results with very limited data. We leave technical enhancements such as transfer learning and data augmentation to future work, while contributing a novel examination of the trade-off between speaker anonymity and phonetic accuracy, a particularly important consideration for endangered languages with few remaining speakers.

2.2 Indigenous Language Technology: Opportunities, Challenges and Risks

TTS and other language technologies offer significant potential for Indigenous language revitalization efforts. However, as AI becomes a more prominent tool for language preservation, these systems can provide new privacy and ethical concerns that have been highlighted around consent, governance, and resource sharing, particularly when projects are initiated by institutions or developers outside of Indigenous communities (Yun-Pu Tu 2025), thereby raising concerns around data exploitation issues that Indigenous communities fear and often resist.

One major concern is the use of publicly available language materials such as YouTube recordings, textbooks, or archived audio without explicit permission. Public access is often mistaken for consent to use. Yet, for many Indigenous Nations, language and cultural knowledge is passed down through specific protocols that govern who can hear, use, and share them. Disregarding these practices leads to real harm, especially for communities who are still healing from cultural loss. For instance, Sámi language communities have objected to the development of natural language processing tools without engagement or consent, raising concerns about what the Sámi Council has called "digital colonialism" (Bird 2020). Similar concerns have been raised about the scraping of online data in Indigenous Polynesian languages like te reo Māori and ʻōlelo Hawaiʻi by major technology firms such as OpenAI to train AI models on that are put on open access without permission from the respective speaker communities (Mahelona et al. 2023).

Beyond consent issues, recent research has also highlighted risks to speaker privacy. Speech synthesis models trained on real voices can retain biometric features, sometimes making it possible to identify individuals in the training data — especially within small language communities (Huang et al. 2024). This privacy risk is particularly acute for Indigenous languages, where speaker communities are often small and closely-knit, making individual identification more likely and potentially more harmful.

A further concern that has arisen recently and likely to get worse, unfortunately, is the AI-generation of entire fake textbooks and lexica purportedly of Indigenous languages. These are largely hallucinated by Large Language Models, presumably, through training on scarce materials available in the languages they claim to represent, but are mostly incomprehensible in any language. Often, these same books claim authorship by language experts, both Indigenous and otherwise, without any real involvement or authorization from them, and are intended only to profit off of people's genuine desire to learn their heritage languages. Such publications have been

generated for Abenaki, Kanien'kéha (Mohawk), Diné bizaad (Navajo) and Anishinaabemowin (Ojibwe), among others (Becking 2024; Glorieux-Stryckman 2024).

These concerns around consent, privacy, and community governance directly inform our approach to developing TTS for Haida, where the small speaker community amplifies both the potential benefits and risks of voice technology.

2.3 Ethical Frameworks for Indigenous Language AI

In response to these challenges, advocates and researchers working with Indigenous data have developed comprehensive frameworks for ethical language model development. These frameworks emphasize several core principles that must guide language technology projects involving Indigenous languages:

- **Free, prior, and informed consent** requires that communities have full knowledge of proposed data use and modeling approaches before any work begins, as well as potential risks and long-term implications of resulting AI models.
- **Community-led governance** ensures that Indigenous communities maintain control over their data, models, and usage guidelines throughout the development process and beyond.
- **Cultural protocols** acknowledge that Indigenous communities have established traditions and guidelines for determining what linguistic and cultural material is appropriate for digitization and public release.
- **Transparent benefit-sharing** requires clear agreements about how any advantages gained from such development will be distributed between researchers, technologists, and language communities.

Without implementing these safeguards and working with community-appointed Indigenous representatives, even well-meaning AI projects risk reinforcing the same extractive patterns that language revitalization efforts are meant to challenge.

As we move forward with this early version of a Haida TTS model, we have tried to follow the values and principles that many Indigenous communities have called for. This isn't a finished product, and will not be complete without more community input. But we do see it as a way to start a conversation — one rooted in respect and care about what might be possible and what should be protected. Language work isn't just about recovering words, after all. It's about relationships, trust, and honoring the intentions of those who carried these teachings before us. We are approaching this project carefully, with the goal of doing things differently not repeating the extractive practices others have justified in the name of innovation. We want this to be a path that opens doors but only the ones the community chooses to open.

While our work was researcher-initiated rather than community-led, we do count among us scholars from within the community. We have sought to implement the principles described above through ongoing family consent and feedback. The recordings used in this project were created by and with Áljuhl (Erma Lawrence) and later compiled on haidalanguage.org by Jordan Lachler. We obtained oral permission from Áljuhl's immediate family to use her recordings for educational purposes and kept them informed throughout. We are also in the process of reaching out to Dr. J. Lachler, the site administrator, for his express permission to use these materials only in a limited, non-commercial research context. This highlights a gap in established standards for consent and governance when archival resources are applied to language technology. This is an area where more

community guidance is needed. For this reason, we are not releasing the model publicly at this stage and see this work as an early step that should be shaped by ongoing conversation with the Haida community. Ultimately, all data and tools developed through this project are for the Haida people. It is vital to recognize their right through sovereignty and self-determination to decide what should be shared, protected, or used to guide the next generation.

3 Present Contribution

Below we present our prototype of a TTS system for Haida based on very limited data, including findings on the trade-off between audio quality, phonetic accuracy, and anonymity entailed by current methods. We proceed to discuss (3.1) existing recorded Haida language data, (3.2) the data source used for our proposed system, (3.3) the preprocessing applied to this data, including optional steps to clean audio and change speaker voice, (3.4) details of the TTS model used and its implementation, and (3.5) the user interface we have designed for our prototype.

3.1 Recorded Data for Haida

Recorded data for Haida suitable for TTS training faces significant limitations due to issues of data quantity, quality, and ethical considerations regarding data usage. The bulk of Alaskan Haida recordings were collected between the 1970's and early 2000's in casual or home environments, resulting in acoustic noise, tape deterioration, and quality degradation from accidental re-recording over existing materials. While limited higher-quality studio recordings exist through the work of linguists and community members, these remain privately held and unavailable for public use. The scarcity of publicly accessible Haida recordings, combined with the specialized linguistic expertise and knowledge required for accurate orthographic transcription, creates substantial barriers to TTS development. Given these complexities, this preliminary work serves as a technological demonstration aimed at building the foundation of ethical community consensus regarding the feasibility of TTS for the Haida language while honoring Haida protocols. This process is essential before engaging with more extensive, less publicly available Haida audio archives that hold great linguistic value but require community guidance, permissions, and oversight for responsible use.

3.2 Data-source used

Data currently used to train our model is sourced exclusively from recordings by fluent Haida speaker Áljuhl (Erma Lawrence), compiled and edited by linguist Jordan Lachler. These recordings, consisting of approximately 30 minutes of audio in total, were recorded in 1974, followed by a small set of additional phrases in 2003. These are publicly available on Jordan Lachler's website haidalanguage.org, and are parallel to Lawrence's two books *Alaskan Haida Phrasebook for Beginners* (Volume 1, *Kasaan Dialect*) and *Kíilang Sk̲'at'áa* ("*Learning Your Language*"). We have explicitly received consent to use these materials, as discussed below.

The recordings are in .mp3 format. Each file is approximately 4–10 seconds long, containing a recording of Lawrence saying a short phrase twice or three times in succession. These are accompanied by textual transcriptions in the classic Alaskan Haida orthography. An excerpt from the website is shown below.

Hlk'yáawdaalw uu íijang. This is a broom.
Hlk'yáawdaalwaay í'waan-gang. The broom is big.
Hlk'yáawdaalwaay gigwáay sgíidang. The broom handle is red.

K̲'ust'áan-gyaa uu íijang. This is a crab.
K̲'ust'áan-gyaa uu táaw 'láa íijang. Crabs are good food.
K̲'ust'áan-gyaa uu isdéi díi guláagang. I like to get crabs.

Figure 1: Screenshot from haidalanguage.org

The following shows a single such audio file loaded in the Audacity media player of the phrase *Hláa uu hlgánggulaang* ("I am working"), repeated twice in succession by Áljuhl (Erma Lawrence). These recordings have static background noise and audible reverberation, as they were captured in a home environment with standard recording equipment (rather than a studio environment) — we later discuss the implications of this and mitigation strategies.

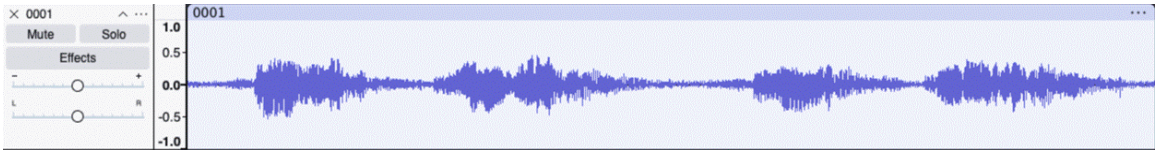

Figure 2: A typical audio-sample from haidalanguage.org

We automatically extract the orthographic transcriptions accompanying each audio file by processing the HTML of the pages containing these phrase lists. The classic Haida orthography uses the underlined letters ⟨g̲ x̲⟩ to represent the epiglottal consonants /ʔ͡ʜ ʜ/, and on these pages, they are displayed using HTML underline functionality (rather than using existing Unicode characters). Our parsing code includes a functionality to detect HTML underlines and convert them into these characters.

3.3 Data Preprocessing

3.3.1 Processing Textual Metadata

Our metadata consists of a LJSpeech-formatted table containing audio filenames and ground-truth (GT) IPA transcripts, as shown in Figure 3 below.

```
wavs/el/el-what_are_you_drying_then.wav|gˈuus tɬ'aa dˈaŋ xilˈaadaaŋ? gˈuus tɬ'aa dˈaŋ xilˈaadaaŋ? gˈuus tɬ'aa dˈaŋ xilˈaadaaŋ?
wavs/el/el-I_am_drying_black_seaweed.wav|sʔˈiw uu ɬ xilˈaadaaŋ. sʔˈiw uu ɬ xilˈaadaaŋ. sʔˈiw uu ɬ xilˈaadaaŋ.
wavs/00/0029.wav|gˈuusgjaa uu ˈiidʒaŋ? gˈuusgjaa uu ˈiidʒaŋ?
wavs/00/0058.wav|tˈaan-gjaa uu ˈiidʒaŋ. tˈaan-gjaa uu ˈiidʒaŋ.
wavs/00/0059.wav|quŋˈaaj uu ˈiidʒaŋ. quŋˈaaj uu ˈiidʒaŋ.
wavs/00/0060.wav|k'aajɬt'ˈaagjaa uu ˈiidʒaŋ. k'aajɬt'ˈaagjaa uu ˈiidʒaŋ.
wavs/00/0061.wav|xuujˈaaj uu ˈiidʒaŋ. xuujˈaaj uu ˈiidʒaŋ.
wavs/00/0062.wav|k'ˈaawgjaa uu ˈiidʒaŋ. k'ˈaawgjaa uu ˈiidʒaŋ.
wavs/00/0063.wav|sq'ˈin-gjaa uu ˈiidʒaŋ. sq'ˈin-gjaa uu ˈiidʒaŋ.
```

Figure 3: Sample metadata

For simplicity, we include the repeated text in transcripts (rather than splitting them into separate audio segments). Empirically we find that trained models generalize to synthesize non-repeating segments at inference time. As seen in the table above, the GT transcripts are in IPA

rather than using the original Haida orthography, produced with a light-weight grapheme-to-phoneme (G2P) function. This has a number of potential advantages:

- Haida has been written with a variety of orthographies (Enrico 2003; Swanton 1905, 1908). Rather than giving preferential treatment to a single orthography, we envision our TTS model being used with the user's preferred orthography as input, with G2P conversion implemented to convert this input to underlying IPA.
- As we initialize from a pretrained checkpoint, using standard IPA may make training more efficient by better matching the grapheme-audio matching learned during pretraining.
- Certain transcription conventions may affect the learning process. For example, we choose to denote high tone with the IPA vertical arrow character (ˈ) before the relevant vowel (Enrico 2003; Lawrence 1977). This contrasts with the use of an acute accent over vowels (á é í ó ú) where, if represented as single composed Unicode characters (rather than base vowel + combining accent), the model must learn to interpret each high-tone vowel as a separate segment.

Note that our IPA transcription uses various conventions specific to us, and is not a narrow phonetic transcription. In particular:

- We use an IPA arrow (ˈ, normally used for a tonal upstep) to denote high tone, as discussed above.
- We use "voiced" IPA symbols (e.g. ⟨d g⟩) to represent unaspirated segments ([t k]) and "unvoiced" symbols (e.g. ⟨t k⟩) to represent aspirated segments ([t^h k^h]) in positions where these contrast, as this matches the surface orthography.
- We use the epiglottal symbols ⟨ʔ ʜ⟩ to represent the segments written ⟨g̲ x̲⟩ in the classic orthography. These have various realizations depending on speaker and dialect, but we treat these as arbitrary symbols matching our speaker's realization of these segments.

Some of these choices may affect the model's learning process, and we envision future work testing this systematically, as we discuss in the Future Work section (5.3).

3.3.2 Processing Audio Data

We convert all audio files from .mp3 to .wav format and resample to 22.05 kHz using ffmpeg, as this format is required for training the downstream TTS model. Training on this audio data directly presents two issues: as we will show, its outputs reflect the noisy quality of the raw recordings (static background noise and audible reverberation), and it produces a voice that sounds similar to the original speaker. Therefore, we also attempt a two-step mitigation strategy using state-of-the-art (SOTA) voice models. We first clean each audio recording using Adobe Speech Enhance. Then, we select a small subset of (~50) recordings and apply Eleven Labs Voice Changer to produce clean recordings in another voice ("Mark – Natural Conversations"), adding these recordings to the metadata list used for training. As the latter in particular often significantly distorts the phonemes in the audio, we manually review the transcripts of these added recordings and adjust them to reflect

the audio, even if it is no longer valid Haida. For example, when *sg̱íw* "laver" is distorted to **spíw* (which does not exist in Haida), we use ⟨p⟩ in the IPA transcript. In this case, we also add an additional column to the metadata file with speaker ID — 0 for recordings in Áljuhl's voice, and 1 for recording with the modified voice — as this is used for multi-speaker training. We will show that these strategies introduce phonetic distortion, producing a trade-off between phonetic accuracy, audio quality, and anonymity.

3.4 TTS Model and Implementation Details

We adopt the VITS architecture (Kim et al. 2021) as maintained in the Piper open-source project as our base TTS model. This lightweight neural TTS architecture is popular for low-latency and edge computing applications such as smart homes and assistive technology. We selected this architecture as a lightweight baseline for several reasons: it is convenient for prototyping, easy to train and test on our hardware, allows for fast inference, and can run locally on standard CPU-only hardware, which is not practical for larger TTS models. Additionally, the Piper library and community support provide practical advantages over more experimental, research-oriented TTS frameworks.

We use the Piper-medium variant with approximately 20 million parameters, initializing from a pretrained English checkpoint and fine-tuning on our paired audio-transcript data. For training on raw audio, we employ single-speaker training mode, while for cleaned audio with voice changing, we use Piper's limited multi-speaker capability by training in multi-speaker mode. In all settings, we train for approximately 12 hours with batch size 32, using default Piper training hyperparameters. We export our trained model to ONNX format for inference with Piper's ONNX inference script.

3.5 User Interface

We create a user interface for using our prototype TTS system, designed with the Gradio library for machine learning model prototyping (Fig. 4). While not currently publicly released, a snapshot is provided below, with an explanation as to how it works and is used.

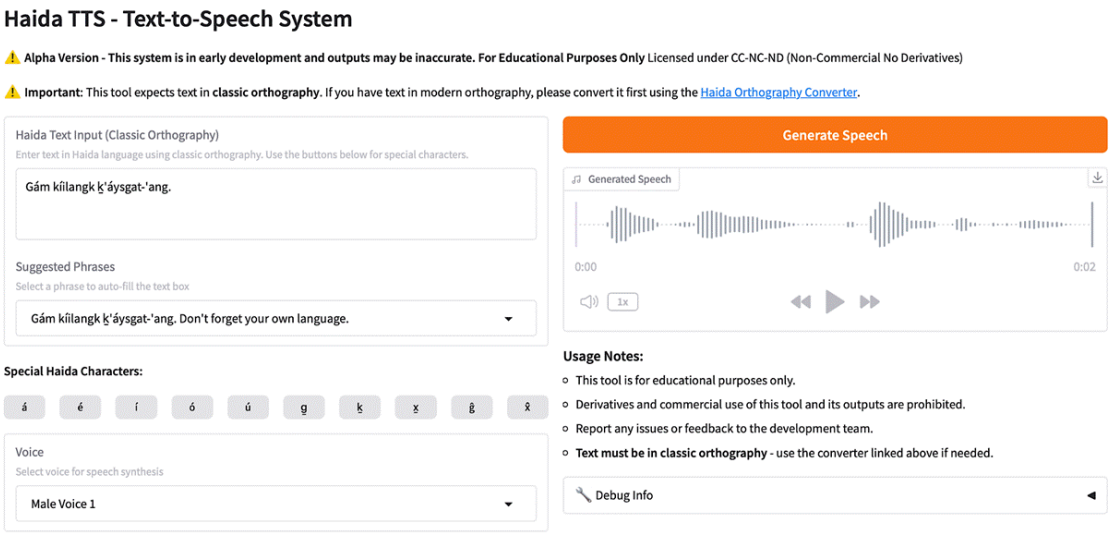

Figure 4: Prototype Haida TTS interface

The user may input Haida text into the text box in the upper-left corner. They are also provided with a drop-down list of suggested phrases; when one is selected, it is automatically filled in the input text box. As the classic Haida orthography uses various special characters, these are provided as buttons which automatically paste them in the input area for users who may not have their keyboard configured to input them directly.

Underneath, the user may select a voice. Our prototype currently only supports a single extra voice (either Áljuhl's voice for the variant trained on raw audio, or the male voice with voice changing applied for the variant trained on cleaned and modified audio), but we envision a future system allowing for multiple voice selections to cover male and female voices and diverse ages and styles.

When the user presses "Generate Speech", the TTS model generates audio for the input text which may be played and downloaded with the audio widget on the right-hand side. The low-latency Piper model typically takes a fraction of a second to generate this audio (running on CPU-only hardware).

Under the hood, this demo performs G2P conversion on the input text, to convert it into IPA which can be processed by the TTS generation model. This can be viewed by clicking on the "Debug Info" drop-down at the bottom-right. We also link to an orthography conversion tool, currently a work in progress, which will enable users to convert between the classic Alaskan Haida orthography expected by our system, and other orthographies such as that of the dictionaries of Jordan Lachler (2010) and John Enrico (2005) which may differ significantly from the system used here.

4 Results and Evaluation

Below, we first present our initial, manual evaluation of TTS outputs in Section 4.1, comparing our model trained on raw audio and on clean and voice-changed data. In Section 4.2, we then lay out a comprehensive community-based evaluation framework, which we intend to employ as we develop future iterations of our TTS system.

4.1 Preliminary Model Evaluation

We conduct a preliminary manual evaluation of our TTS models to assess their initial performance before developing a comprehensive community-based evaluation framework. This preliminary assessment is performed by a trained linguist Haida expert co-author to establish baseline performance metrics and identify key technical challenges that must be addressed. We perform a quantitative evaluation over a set of fourteen sentences selected randomly from the sentences in Lawrence (2010) and graded on a 1–5 scale for several phonetic and prosodic criteria intended to cover aspects of Haida most distinct from larger Eurasian languages. These scores are tabulated in Table 1, and results per sentence are graphed in Fig. 5 below.

Table 1: Average scores per metric used for assessment

Metric	Raw audio	Anonymized audio
Vocal timbre	4.57/5 | 91.43%	2.07/5 | 41.42%
Prosodic naturality	4.07/5 | 81.43%	2.93/5 | 58.57%
Pitch naturality	4.71/5 | 94.29%	3.5/5 | 70.00%
Lateral accuracy	4.5/5 | 88.00%	2.2/5 | 44.00%
Uvular accuracy	5/5 | 100%	2.33/5 | 46.67%
Affricate accuracy	4.56/5 | 91.11%	2.67/5 | 53.33%
Epiglottal accuracy	3.83/5 | 76.66%	3.17/5 | 63.33%
Ejective accuracy	4.5/5 | 90.00%	2.13/5 | 42.50%
Output consistency	4.29/5 | 85.71%	4.29/5 | 85.71%
Overall performance	**88.73%**	**56.17%**

These results match our qualitative observations, highlighting significant differences between training approaches. Output trained solely on raw audio of Áljuhl (Erma Lawrence) with no filters to clean or disguise her voice performed the best, generating very clear and understandable Haida although with poor audio quality reflecting that of the recordings used during training. We find it to be comprehensible and expect issues such as unnatural prosody to improve incrementally with more data from additional speakers.

In contrast, output trained on cleaned audio and with voice changing applied was noticeably less phonetically accurate, exhibiting unrepresentative pronunciations of lateral obstruents, dorsal stops (especially fronting in /Cj/ environments so that velars sound coronal, and uvulars velar), inconsistency in rendering epiglottal and ejective sounds, and numerous prosodic and tonal deviations from the Haida language. Thus, while this synthesized speech has better overall audio quality, it is not usable as-is for the purpose of Haida language pedagogy.

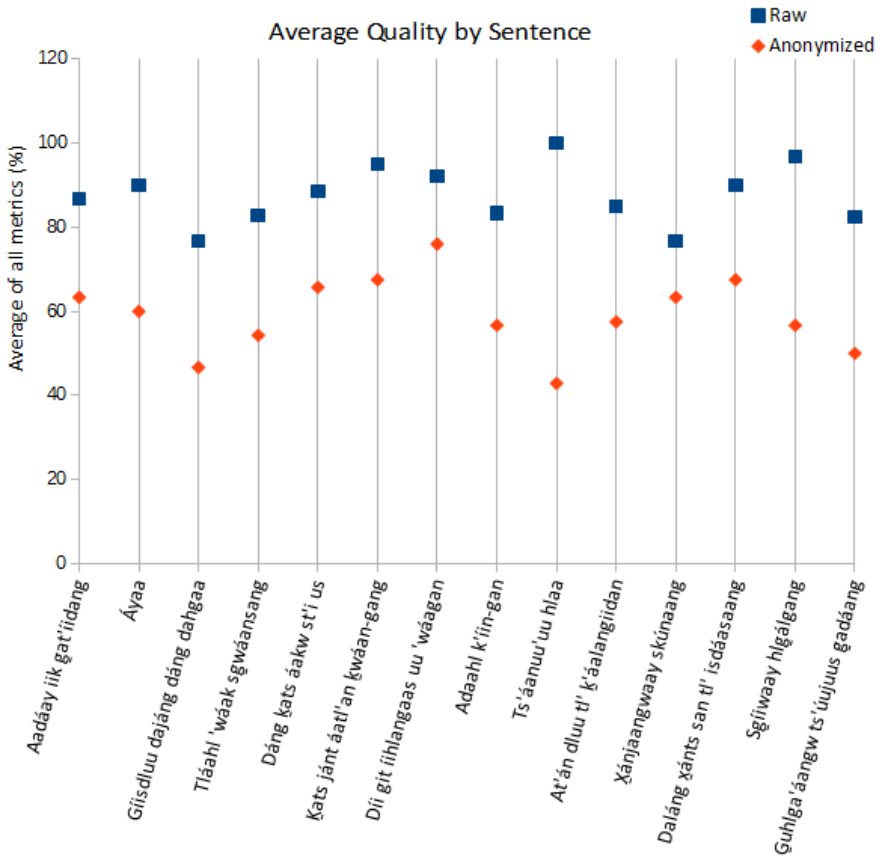

Figure 5: Quality of sentences as determined by the average of the qualitative metrics assessed for each of them

This preliminary evaluation highlights a critical technical challenge: we may train on original voice recordings that preserve acoustic fidelity but retain speaker identity, or we may clean and modify voices but introduce acoustic distortions that compromise linguistic accuracy. Bridging this gap will require further research and community input on acceptable trade-offs between voice anonymization and linguistic fidelity.

4.2 Future Evaluation Framework

Building on these preliminary findings, we plan to implement a comprehensive evaluation framework that blends case study and ethnographic design, incorporating both qualitative and quantitative data using community-based participatory research principles and the Culturally Responsive Indigenous Evaluation (CRIE) model. This approach emphasizes respect, protection, and ethical use of materials while reducing social and cultural risks (CEMA 2015; Waapalaneexkweew & Dodge-Francis 2018).

The evaluation will be conducted in alignment with Haida worldviews, ethical behaviors, and cultural methodologies, implementing practices of informed consent (LaFrance & Nichols 2010; LaFramboise & Place 1983). Data collection and feedback will follow guidelines established using the Indigenous Traditional Knowledge Framework (ITKF), which is based on the guiding principles that: (1) Indigenous knowledge and language provides important insights into the natural

environment, (2) Indigenous knowledge and language offers valuable perspectives on natural and social phenomena through a cultural lens, and (3) Indigenous knowledge and language belongs to the tribal nation, which maintains authority and control over the knowledge such that permission is required to collect, analyze, and disseminate findings (CEMA 2015; Waapalaneexkweew & Dodge-Francis 2018). Evaluation will blend case study and ethnographic design and include both qualitative, semi-quantitative, and quantitative data. Formative assessment will be ongoing and allow for the project to gain real-time knowledge of efficacy, to make adjustments rapidly, and to ensure that project goals are being met.

We propose to engage three key groups within the community:

- **Beginner learners**: Assessing comprehensibility and learning utility.
- **Potential teachers**: Evaluating pedagogical effectiveness and cultural appropriateness.
- **Expert speakers**: Determining linguistic accuracy and cultural authenticity.

These community members will be drawn upon to iteratively evaluate our models with the following protocol:

- **Test Preparation**: Hold out a small proportion (tentatively 10%) of recordings as a test set. These are not used for training the TTS model.
- **Comparative Assessment**: Present both original-voice and anonymized versions to community evaluators.
- **Subjective Feedback**: Collect naturalness and clarity ratings (1–5 scale) alongside qualitative feedback.
- **A/B Testing**: Conduct preference comparisons between outputs of different models.
- **Interview-based assessment:** Collect feedback from these community members about their possible uses of such a TTS model. This includes rating its potential relevance, advantages and drawbacks for their particular use cases, and potential cultural or ethical considerations.

We will use this feedback to iteratively refine our model, ensuring that technical development proceeds in accordance with community values and linguistic preservation goals. The results will also inform decisions about whether to release synthetic audio with original speaker voices or only anonymized versions, pending appropriate community permissions.

5 Discussion

We conclude with some discussion on the limitations of our application thus far in Section 5.1, its potential applications in pedagogy and revitalization in 5.2, and finally, in Section 5.3, directions for future collaborative research and development.

5.1 Limitations

Our current model faces several significant technical challenges that constrain its immediate applicability. The primary limitation is a fundamental trade-off between phonetic accuracy and voice anonymization: using raw audio preserves relative phonetic accuracy but retains the original

speaker's voice characteristics, while cleaned audio with voice masking produces less accurate linguistic output. In addition, both model versions show reduced accuracy when synthesizing voice from very short (e.g. single-word) inputs, relative to longer sentences. Future work may investigate promising work-arounds such as generating multiple repeated instances of a word in sequence and selecting the middle generation, or more fundamental solutions to this issue.

Additional limitations stem from our constrained training data. The model, unsurprisingly, struggles with rare phonemes, particularly those represented by ⟨ĝ⟩ and ⟨x̂⟩ (only found in loanwords) which never appear in the limited training data, and shows bias due to the skewed distribution of forms in the training data, such as a tendency to over-use present tense verb endings (ending in *-ng*) at the expense of past tense verbs (ending in *-n*). These issues reflect the broader challenge of working with limited linguistic resources for endangered languages.

The model's evaluation remains preliminary and somewhat subjective, conducted only by a linguist rather than community members or intended users. This limited assessment scope prevents comprehensive understanding of the model's practical utility and cultural appropriateness.

Finally, any pedagogical or community application must await explicit community consensus regarding appropriate use, as the technology raises important questions about voice consent, cultural protocol, and the ethics of synthetic speech generation from archival recording.

5.2 Potential Applications

With appropriate community authorization, we foresee this TTS technology addressing several critical needs in Haida language education. A key area that this may address is the current imbalance between the slightly larger set of written materials and the near-total lack of spoken audio. Learning tools such as dictionaries and phrasebooks could be digitized and enhanced with hyperlinked audio, which is largely infeasible to obtain via recording Elders, though that is certainly the best option. This could be scaled up to create full audiobooks, or to narrate archival texts documented since the 1970's by the Society for the Preservation of the Haida Language. This would especially benefit auditory learners who struggle to learn from written materials alone. In addition, our method could be used to generate audio for culturally-important transcribed narratives such as those recorded by Swanton (1905; 1908), which have not been spoken aloud for several generations due to factors such as the outdated orthography in which they were written.

Our method could also enable various interactive applications, which are particularly needed for threatened languages like Haida for which opportunities for speaking practice remain extremely limited. It could be integrated into language learning applications to provide spoken examples for vocabulary and sentence practice, as is common for high-resource languages. As an ambitious future direction, our TTS model could be integrated with Haida-adapted language models to power conversational AI for language learning.

5.3 Future Work

We foresee future work iterating on and improving our Haida TTS model from various angles. Our immediate priority is expanding the scope of our extremely limited training data while ensuring its high quality. To obtain additional data, we must engage directly with the community to identify existing recordings, establish ownership and permissions, determine ethical usage parameters, and coordinate transcription efforts. Quality improvement requires thorough manual cleaning and verification of all materials. We believe that increasing data size and coverage of speaker styles and phonological phenomena is fundamental to improving the quality of synthesized Haida speech.

This could also enable extending the model to additional dialects of Haida (Massett, Skidegate), which have their own distinctive phonetic traits.

Regarding technical improvements to our method, further research is needed to resolve our observed trade-off between overall audio quality, voice anonymization, and linguistic accuracy. We envision enhanced models for removing noise from audio and changing speaker identity with less distortion of the characteristic sounds of Haida or other low-resource languages. In addition, we intend to test different model architectures and to explore accuracy-optimized variants that may be slower but generate more precise output, particularly valuable for applications like online speaking dictionaries. We will also test how different grapheme-to-phoneme conversions affect results and evaluate whether our current IPA conventions should be modified to better capture Haida phonology for TTS applications.

Recent research on multi-language training for typologically similar Indigenous languages (Wang et al. 2025) suggests potential for collaborative Pacific Northwest language TTS development, though this requires careful consideration of cross-community data consent protocols. All development will proceed under community guidance, with the Haida people as intended beneficiaries and decision-makers regarding appropriate applications and implementation.

References

Beck, D. 2000. Grammatical convergence and the genesis of diversity in the Northwest Coast Sprachbund. *Anthropological Linguistics*, 42: pp. 147–213.

Becking, M. 2024. Fraudulent Anishinaabemowin resources a serious concern. Available at: https://anishinabeknews.ca/2024/10/22/fraudulent-anishinaabemowin-resources-a-serious-concern/ (Accessed on: 07/07/2025).

Bird, S. 2020. Decolonizing speech and language technology. In *Proceedings of COLING 2020*: pp. 3505–3513. Available at: https://aclanthology.org/2020.coling-main.313/ (Accessed on: 07/07/2025).

Byambadorj, Z., Nishimura, R., Ayush, A., Ohta, K, & Kitaoka, N. 2021. Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation. *EURASIP Journal on Audio, Speech, and Music Processing*. Available at: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-021-00225-4 (Accessed on: 07/07/2025).

CEMA Task Group. 2015. CEMA Indigenous traditional knowledge framework project.

Enrico, J. 2003. *Haida syntax*. (2 volumes). Lincoln, NE: University of Nebraska Press.

Enrico J. 2005. *Haida Dictionary: Skidegate, Masset, and Alaskan Dialects*. (2 volumes). Alaska Native Language Center and Sealaska Heritage Institute.

Glorieux-Stryckman, M. 2024. AI outrage: Error-riddled Indigenous language guides do real harm, advocates say. Available at: https://www.montrealgazette.com/news/article562709.html (Accessed on: 07/07/2025).

Harding, L., DeCaire, R., Ellis, U., Delaurier-Lyle, K., Schillo, J., & Turin, M. 2025. Language improves health and wellbeing in Indigenous communities: A scoping review. *Language and*

Health, 3(1), 100047. Available at: https://doi.org/10.1016/j.csl.2024.101723 (Accessed on: 07/07/2025).

Huang, W.-C., Wu, Y.-C., & Toda, T. 2024. Multi-speaker Text-to-speech training with speaker anonymized data. arXiv:2405.11767. Available at: https://arxiv.org/abs/2405.11767 (Accessed on: 07/07/2025).

Kim, J., Kong, J., & Son, J. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *International Conference on Machine Learning*. PMLR. Available at: https://arxiv.org/abs/2106.06103 (Accessed on: 07/07/2025).

Lachler, J. 2010. *Dictionary of Alaskan Haida*. Sealaska Heritage Institute.

LaFrance, J. & Nichols, R. 2010. Reframing Evaluation: Defining An Indigenous Evaluation Framework. *The Canadian Journal of Program Evaluation*, 23(2): pp. 13–31.

LaFromboise, T.D., & Plake, B.S. 1983. Towards meeting the research needs of American Indians. *Harvard Educational Review*, 53: pp. 45–51.

Lawrence, E. 1977. *Haida dictionary*. Society for the Preservation of Haida Language & Literature.

Lawrence, E. 2010. *Alaskan Haida Phrasebook*. Sealaska Heritage Institute.

Lux, F. et al. 2022. Low-resource multilingual and zero-shot multispeaker TTS. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*: pp. 741–751. Available at: https://aclanthology.org/2022.aacl-main.56/ (Accessed on: 07/07/2025).

Mahelona, K., Leoni, G., Duncan, S., & Thompson, M. 2023. OpenAI's Whisper is another case study in Colonisation. Available at: https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/ (Accessed on: 07/07/2025).

Pine, A., Cooper, E., Guzmán, D., Joanis, E., Kazantseva, A., Krekoski, R., Kuhn, R., Larkin, S., Littell, P., Lothian, D., Martin, A., Richmond, K., Tessier, M., Valentini-Botinhao, C., Wells, D., & Yamagishi, J. 2025. Speech Generation for Indigenous Language Education. *Computer Speech & Language*, 90: Article 101723. Available at: https://doi.org/10.1016/j.csl.2024.101723 (Accessed on: 07/07/2025).

Radford, A., Kim J.W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I.. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of ICML 2023*: pp. 28492–28518. Available at: https://dl.acm.org/doi/10.5555/3618408.3619590 (Accessed on: 07/07/2025).

Swanton, J.R. 1905. *Haida texts and myths, Skidegate dialect*. Washington: Government Printing Office.

Swanton, J.R. 1908. *Haida texts, Masset dialect*. Washington: Government Printing Office.

Thomason, S.G. 2015. The Pacific Northwest linguistic area: Historical perspectives. In Bowern, C. & Evans, B. (eds.). *The Routledge Handbook of Historical Linguistics*, 727–737. London: Routledge.

Mosley, C., & Nicholas, A. (for UNESCO). 2010. Atlas of the world's languages in danger. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000187026 (Accessed on: 07/07/2025).

Waapalaneexkweew (Bowman, N., Mohican/Lunaape) & Dodge-Francis, C. 2018. Culturally responsive Indigenous evaluation & tribal governments: Understanding the relationship. In F. Cram, K.A. Tibbetts, & J. LaFrance (eds.), *Indigenous Evaluation. New Directions for Evaluation*, 159: pp. 17–31.

Wang, S. et al. 2025. Developing multilingual speech synthesis system for Ojibwe, Mi'kmaq, and Maliseet. arXiv:2502.02703. Available at: https://arxiv.org/abs/2502.02703 (Accessed on: 07/07/2025).

Yun-Pu Tu, M. 2025. Building ethical AI requires shifting power to Indigenous communities — in data control, design, and governance. Available at: https://policyoptions.irpp.org/magazines/may-2025/ai-indigenous-data (Accessed on: 07/07/2025).