Neural Text-to-Speech for Myaamia: Speech Synthesis for an Indigenous Algonquian Language

Anita Baral[1] **John Femiani**[1] **Hunter Lockwood**[2]
Daniela Inclezan[1] **Balaram Bhandari**[2]

[1]Department of Computer Science and Software Engineering, Miami University, USA
[2]Myaamia Center, Miami University, USA

barala@miamioh.edu, femianjc@miamioh.edu, lockwoht@miamioh.edu,
inclezd@miamioh.edu, bhandab@miamioh.edu

Abstract

We present the first neural text-to-speech (TTS) implementation for Myaamia (Miami-Illinois), an Indigenous Algonquian language of North America. Developed in collaboration with the Myaamia Center at Miami University, our approach upholds principles of data sovereignty. Using 14,358 utterances (10.4 hours total, 8.18 hours for training) from seven speakers, we train and evaluate FastSpeech, Glow-TTS, and VITS, assessing synthesis quality through objective (MCD, F0 RMSE, duration RMSE) and subjective (expert evaluation) metrics. VITS outperforms other models in spectral and prosodic accuracy, but challenges remain in phonetic precision and prosody modeling. Our results confirm the feasibility of neural TTS for Myaamia, with direct implications for language learning and revitalization. This work offers a replicable framework for other low-resource Indigenous languages while ensuring ethical, linguistic data governance.

1 Introduction

Since the 1990s, the global decline of Indigenous languages has driven revitalization efforts, and technology plays an increasingly important role in preserving linguistic and cultural heritage (Bird, 2020). The Myaamia (Miami-Illinois) language, traditionally spoken in the southern Great Lakes region of North America, became dormant after the last first-language speakers passed away in the mid-20th century following forced relocation. In recent decades, the Miami Tribe of Oklahoma has led systematic revitalization efforts, using linguistic expertise, education, and digital resources to reclaim the language (Baldwin et al., 2016).

Central to this work is the Myaamia Center at Miami University, which supports Myaamia language and cultural revitalization through research, education, and community partnerships. The Center developed and maintains the Indigenous Languages Digital Archive (ILDA), a web-based platform that brings together written and audio language materials to support archives-based language reclamation (Baldwin et al., 2016). Complementing ILDA, the Šaapohkaayoni community education portal provides Myaamia community members with access to self-directed learning modules regardless of geographic location. While these resources provide strong support for language learning, access to spoken language remains limited. An estimated 95% of the available audio recordings originate from just two individuals, which creates a bottleneck for learners seeking to develop listening and pronunciation skills. Neural TTS systems offer a promising way to address this gap by generating natural-sounding speech, reducing the burden on the few available speakers and expanding access to spoken language resources for learners (Brinklow, 2021). This study presents the first neural TTS system for Myaamia, developed in collaboration with the Myaamia Center at Miami University.

Our contributions to low-resource speech synthesis include:

1. We develop and evaluate the first neural TTS system for Myaamia, leveraging well-established neural TTS models (FastSpeech (Ren et al., 2019), Glow-TTS (Kim et al., 2020), and VITS: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (Kim et al., 2021)) trained on 8 hours and 18 minutes of speech data.

2. We establish a performance benchmark for Myaamia TTS, assessing spectral (MCD), prosodic (F0 RMSE), and temporal (duration RMSE) accuracy alongside subjective evaluations by Myaamia linguists.

3. We uphold Indigenous Data Sovereignty principles, with all linguistic data remaining un-

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 1–10

der the ownership and governance of the Myaamia Center.

2 Related Work

The development of speech technology for endangered and Indigenous languages sits at the intersection of technological innovation and language preservation (Bird, 2020; Kraljevski et al., 2024; Pine et al., 2022).

2.1 Neural TTS Systems

Neural text-to-speech synthesis has advanced rapidly, from WaveNet's (van den Oord et al., 2016) neural waveform generation to Tacotron's (Wang et al., 2017) end-to-end approach. Subsequent architectures addressed efficiency and data constraints: FastSpeech (Ren et al., 2019) and FastSpeech 2 (Ren et al., 2021) introduced non-autoregressive generation, while Glow-TTS (Kim et al., 2020) applied flow-based modeling for efficient training. VITS (Kim et al., 2021) combined variational autoencoders with adversarial training in a fully end-to-end framework. Since then, further advances have emerged, including neural codec language models such as VALL-E (Wang et al., 2023), diffusion-based approaches like NaturalSpeech 2 (Shen et al., 2023), style-based models such as StyleTTS 2 (Li et al., 2023) and flow-matching methods like F5-TTS (Chen et al., 2025). However, these systems typically require substantially larger datasets and computational resources, and are not yet widely supported in open-source toolkits for low-resource language development. We implement and evaluate FastSpeech, Glow-TTS, and VITS for Myaamia, selecting these architectures for their proven effectiveness in low-resource settings and their availability in the Coqui TTS framework.

2.2 Low-Resource and Indigenous Language Speech Synthesis

Modern neural approaches have reduced the data requirements of TTS systems, making them increasingly viable for language preservation and education (Xu et al., 2020). However, challenges persist, including data scarcity, limited native speaker evaluation, and language-specific phonological complexities (Gumma et al., 2024; Hammerly et al., 2023). Earlier work includes rule-based synthesis for Navajo (Whitman et al., 1997), while more recent neural efforts include high-quality TTS for Kanyen'kéha (Mohawk), Plains Cree, SENĆOTEN (Pine et al., 2022, 2025), Võro (Rätsep and Fishel, 2023), Border Lakes Ojibwe (Hammerly et al., 2023), Mundari (Gumma et al., 2024), a multilingual system for Ojibwe, Mi'kmaq, and Maliseet (Wang et al., 2025), and Shipibo-Konibo (Menendez and Gomez, 2025). Our work builds on this foundation, extending neural TTS to Myaamia and contributing the first benchmark for this language.

3 Methodology

Figure 1 illustrates the Myaamia TTS pipeline, covering data preprocessing, model training and inference using FastSpeech (Ren et al., 2019), Glow-TTS (Kim et al., 2020), and VITS (Kim et al., 2021), and evaluation of synthesized speech.

3.1 Dataset: Myaamia-TTS

The dataset consists of 14,358 utterances (approximately 10.4 hours) of Myaamiaataweenki (Myaamia language) recordings created at the Myaamia Center since 2010. The recordings were collected in a controlled environment with minimal ambient noise and feature seven speakers, two of whom contributed approximately 95% of the recordings. Myaamiaataweenki employs a phonemic writing system derived from Americanist phonetic notation. The writing system has 20 basic symbols: 4 vowels, each of which can be short or long (a, aa, e, ee, i, ii, o, oo), and 12 consonants (p, t, k, c, s, š, h, m, n, l, w, y). Potential phonological challenges include preaspirated consonants (hp, ht, hk, hs, hš, hc), articulated with a brief puff of air before the consonant, and a set of vowel-devoicing rules. Still, the phoneme inventory overlaps heavily with English. Table 1 presents a representative subset of phonetic symbols along with example words where the symbol's sound is highlighted in blue.

The dataset consists of text sequences ranging from 3 to 163 characters. Because Myaamia employs a phonemic orthography with high grapheme-to-phoneme correspondence, we train our TTS systems at the character level, where each input character closely approximates a phoneme-level representation. This approach establishes a baseline for Myaamia TTS and allows us to empirically identify the specific cases where the character-phoneme mapping diverges, as discussed in Sections 4.3 and 5. Transcripts are drawn directly from the ILDA database, where they were authored and curated by Myaamia Center linguists using the same orthography taught to learners; no additional

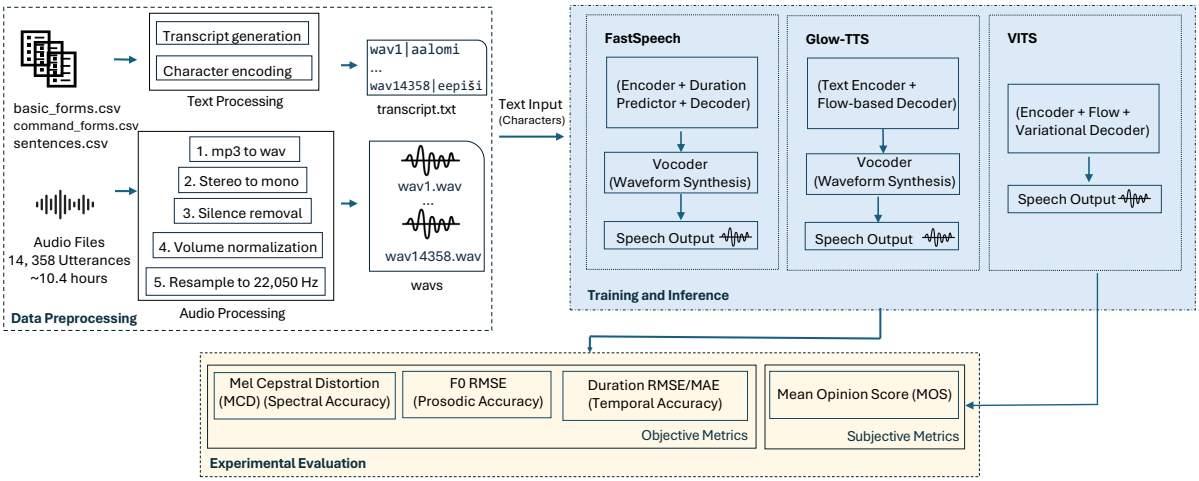

Figure 1: Overview of the text-to-speech (TTS) pipeline showing data preprocessing, training, and inference architectures (FastSpeech, Glow-TTS, and VITS), and evaluation metrics.

Table 1: Myaamia Phonetic Symbols and Examples. A representative subset is shown; the full inventory includes 20 symbols (see Section 3).

Symbol	Myaamia Example (Target Sound)	English Approximation (Similar Sound)
a	**a**ya	p**a**pa
ee	n**ee**we	b**ay**
p	aa**p**ooši	**p**ot
k	**k**iinte	**k**ey
hp	paa**hp**ilo	no English equivalent
ht	ee**ht**eeki	no English equivalent
nk	ii**nk**a	li**ng**er
nt	kii**nt**e	ti**nd**er

phonemic re-transcription was applied. Audio–transcript pairings and the exported character inventory were verified by Myaamia Center linguists before training, and no sub-utterance alignment was performed. Each model learns alignment internally from utterance-level pairs.

The majority of sequences fall within the shorter range, as shown in Table 2. Taken from dictionary recordings, the data falls into three major categories: basic expressions, command forms (imperatives), and example sentences. Below are examples from each category, all derived from the stem **ayaa-**, which means ‘go to a place’ in English.

- Basic form: **iiyaayani** (‘you go’)
- Command form: **ayaataawi** (‘let’s go!’)
- Example sentence: **nipwaantiikaaninkiši iiyaayaani** (‘I am going to school’)

Audio for the dataset was somewhat variable; files included three different sample rates: 44100 Hz (69.6%), 22050 Hz (29.8%), and 48000 Hz (0.6%). Recordings varied in duration from 0.52 to 20.06 seconds, with an average length of 2.61 seconds. Because two speakers contribute roughly 95% of the recordings, the synthesized voice, pronunciation, and prosody primarily reflect those two speakers rather than the broader community; we address this imbalance as a data limitation in Section 5.

Table 2: Distribution of Utterance Lengths in Characters

Character Count	Number of Utterances	Percentage	Total Characters
3-14	8,559	59.6%	96,117
15-25	5,131	35.7%	89,882
26-37	524	3.6%	15,651
38-83	140	1.0%	6,446
84-163	4	0.0%	511
Total	14,358	100.0%	208,607

3.2 Data Preparation and Standardization

The audio data consisted of stereo recordings in MP3 format. The corresponding transcripts were maintained in three database tables (*basic_forms*, *command_forms*, and *sentences*) within the Myaamia ILDA database and exported to CSV format. The preprocessing pipeline established a mapping between audio files and transcripts using unique identifiers from the database tables, generating a transcript.txt file containing paired audio filenames and transcription entries *(e.g., wav1|aahkohkimilo)*.

Stereo recordings were converted to mono format, and silence removal was applied using a threshold of -40 dB for segments exceeding 0.5 seconds. Volume normalization was performed to maintain consistent amplitude levels, and all audio was resampled to 22,050 Hz. Mono conversion was applied because the recordings are single-speaker and contain no meaningful stereo information. The 22,050 Hz target matches the Coqui TTS framework default and avoids upsampling the 29.8% of files already recorded at that rate, while retaining the full frequency range relevant to intelligible speech. Character encoding was applied to the transcript data to properly represent the special characters š and Š (IPA /ʃ/) in the Myaamia orthography. The standardized dataset was partitioned using an 80/20 split ratio, yielding 11,486 utterances for training and 2,872 for testing. The split was performed by random utterance-level sampling without explicit speaker or character stratification; given the size of the test set, the speaker distribution and character inventory in the training and test partitions are expected to closely mirror those of the full corpus. Total duration of the training data was 8 hours and 18 minutes, while the evaluation dataset was 2 hours and 4 minutes of audio. During training, the Coqui TTS framework automatically reserved a subset of the training partition for validation using its default *eval_split_size* of 0.01 (1%), yielding approximately 115 validation utterances per model. All models were trained for 1000 epochs, and convergence was monitored through training and validation loss curves.

3.3 Speech Synthesis For Myaamia

We explored three TTS architectures of increasing complexity: FastSpeech (Ren et al., 2019), Glow-TTS (Kim et al., 2020), and VITS (Kim et al., 2021). The text processing pipeline supported all Myaamia characters, including the special characters š and Š, along with punctuation and numbers. Our study began with FastSpeech's parallel sequence generation, followed by Glow-TTS's efficient training and robust voice conversion via normalizing flows, and concluded with VITS's end-to-end variational autoencoder (VAE) and adversarial training. VITS has shown promise for low-resource languages, with recent applications in African languages (Ogun et al., 2024), Mundari (Gumma et al., 2024), and Ojibwe (Hammerly et al., 2023). All models were trained on an NVIDIA A30 GPU (24GB) with a batch size of 32 for 1000 epochs. VITS used mixed precision (fp16) with an AdamW optimizer (lr = 0.001) and a text encoder featuring six layers, two attention heads, and 768 FFN channels. Its loss function prioritized mel loss (45.0) over KL, generator, discriminator, and duration losses. Glow-TTS used an RAdam optimizer (lr = 0.001) with a NoamLR scheduler (4,000 warmup steps) and a relative positional transformer encoder. FastSpeech applied the Adam optimizer (lr = 0.0001) with NoamLR and FFTransformers for both encoder and decoder. Hyperparameters follow the Coqui TTS default configurations for each architecture; no systematic hyperparameter search was performed.

The Real-Time Factor (RTF) measures processing time relative to audio duration, with values below 1.0 indicating real-time capability. To reflect realistic deployment conditions, RTF values were measured using CPU inference across all 2,872 test samples. FastSpeech achieved the best efficiency among the three models (RTF: 7.43 ± 3.00) due to its feed-forward architecture. Glow-TTS had moderate speed (RTF: 8.59 ± 3.29) as normalizing flows added computational overhead. VITS was the slowest (RTF: 12.71 ± 4.53) due to its complex VAE-GAN architecture, though this trade-off was justified by superior output quality (see Section 4). While none of the models achieve real-time CPU inference, all remain practical for offline generation in educational contexts. Training times followed a similar pattern: FastSpeech (27.69h), Glow-TTS (33.87h), and VITS (60.10h).

4 Experimental Evaluation

Our evaluation used 2,872 test samples (20% of the total dataset) to assess all three TTS architectures through objective and subjective metrics.

4.1 Objective Evaluation

We evaluated synthesis quality using three objective metrics. For all metrics, we extracted features from both reference and synthesized speech, applied dynamic time warping (DTW) for temporal alignment, and computed error measures between corresponding frames. Mel Cepstral Distortion (MCD) (Kubichek, 1993) assessed spectral quality by computing the difference between mel-frequency cepstral coefficients (MFCCs), excluding the 0th coefficient (Vasilijevic and Petrinović, 2011). Lower MCD values indicate greater spectral similarity. F0 RMSE (Tsanas et al., 2014) quantified prosodic accuracy by measuring pitch contour deviation between synthesized and reference speech (Luo et al., 2016). Duration RMSE (Henter et al., 2017) evaluated temporal accuracy by comparing segment lengths between synthesized and reference speech. Table 3 summarizes the results.

VITS achieved the lowest mean MCD (15.22) and F0 RMSE (17.50), indicating better spectral and pitch replication than other models. Glow-TTS showed intermediate performance (MCD: 17.91, F0 RMSE: 20.00), while FastSpeech had the largest deviations (MCD: 19.82, F0 RMSE: 30.51). All three models performed similarly in duration metrics, with RMSE values ranging from 0.29 to 0.31, suggesting comparable speech timing capabilities. However, MCD and F0 RMSE values were higher than those typically reported for well-resourced TTS systems (Kominek et al., 2008), likely reflecting the constraints of low-resource training. Despite VITS achieving the best results among the three models, further refinements are needed to enhance phonetic precision and natural prosody. Figure 2 presents a spectrogram comparison of the utterance *kiihkikaateeŝwilo (‘amputate my foot!’)* across all three models and the original recording, illustrating the spectral differences reflected in the objective metrics.

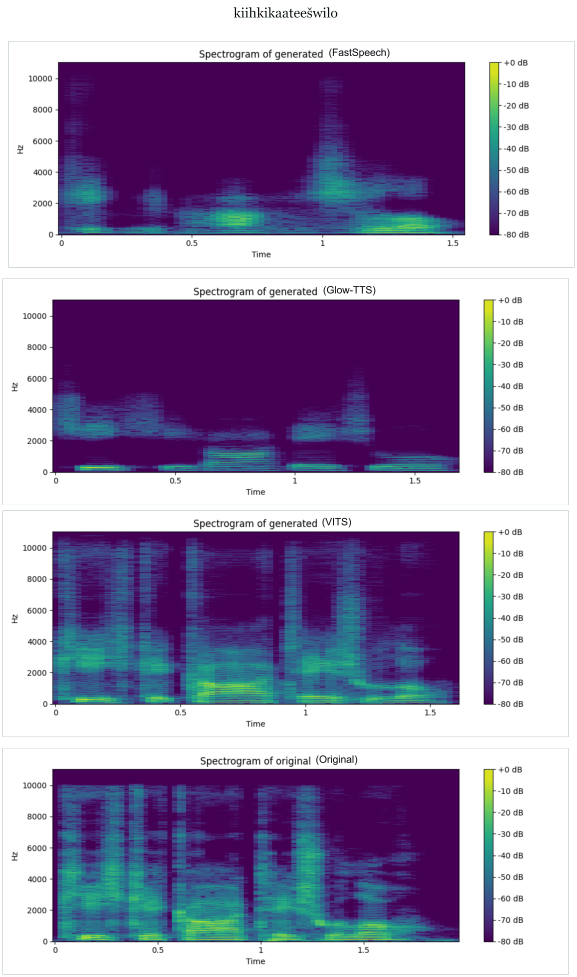

Figure 2: Spectrograms of the utterance *kiihkikaateeŝwilo* generated by FastSpeech, Glow-TTS, VITS, and the original reference audio.

4.2 Subjective Evaluation

To evaluate perceived quality and identify areas for improvement, we conducted a targeted subjective evaluation with four Myaamia experts using audio synthesized by the VITS model. Two of the experts were primary contributors to the recordings used to train the model, while the other two are linguists specializing in Myaamia. The Myaamia speaker community is extremely small, and only a limited number of additional speakers are available. Moreover, several evaluation criteria, including judgments on phonemic contrasts, prosodic alignment, and phonetic accuracy, require specialized linguistic expertise beyond conversational fluency, which further constrained participant selection. Future work will expand the evaluation to include additional community members, as discussed in Section 5. The study adhered to ethical research guidelines, as detailed in Section 6. Participants rated 20 audio samples: 10 evaluating general perceptual measures (naturalness, intelligibility, and overall quality) and 10 focusing on language-specific features (intonation, rhythm, and linguistic accuracy). Each set contained an equal number of synthesized and original recordings, with stratified randomization employed to minimize bias. Ratings were provided on a 5-point Likert scale, where 1 indicated “poor” quality (least natural, least intelligible, or least accurate), and 5 indicated “excellent” quality

Table 3: Objective Quality Metrics for FastSpeech, Glow-TTS, and VITS Text-to-Speech Models. The best score is represented in bold.

Metric	FastSpeech	Glow-TTS	VITS
MCD ↓	19.82 ± 2.33	17.91 ± 2.19	**15.22 ± 2.66**
F0 RMSE ↓	30.51 ± 29.93	20.00 ± 23.73	**17.50 ± 20.69**
Duration RMSE ↓	0.31 ± 0.37	0.30 ± 0.36	**0.29 ± 0.40**

(most natural, most intelligible, or most accurate).

Table 4 summarizes the subjective evaluation. Although naturalness and understandability scores are lower than those of the original recordings, they fall within the expected range given the limited data. Notably, the model achieves an intonation and rhythm score of 3.52 ± 0.81, indicating a strong capacity to capture essential prosodic features of Myaamia. However, to more closely approximate the fluidity and authenticity of native speech, further refinements in stress alignment and phonetic articulation are necessary.

As shown in Table 5, our system demonstrates competitive performance compared to other low-resource TTS models, despite being trained on a more limited dataset. Mundari TTS, trained on 24.76 hours of data, achieves a higher MOS for naturalness (3.69 ± 1.18) (Gumma et al., 2024), while our model scores 3.05 ± 0.89 with significantly less data. Similarly, Võro TTS uses 17 hours of training data, including 14 hours of Estonian, yet our model performs competitively despite relying solely on Myaamia data (Rätsep and Fishel, 2023).

4.3 Error Analysis

To identify specific phonological challenges in the synthesized speech, we analyzed qualitative feedback provided by expert evaluators on the VITS-generated samples. Each evaluator rated five synthesized utterances drawn from distinct subsets of the test data. The model exhibited difficulty with consonant distinctions specific to Myaamia. In the utterance *noonki šayiipaawe aalaankwiaani ('I'm tired this morning')*, one evaluator noted that the voiceless fricative *š* was realized as the affricate /tš/ (orthographic *c* in Myaamia), collapsing a phonemic contrast. In *maalami eelaamhsenki ('it is too windy')*, the same evaluator observed that the consonant cluster /mhs/ appeared to be reduced or realized closer to /nk/, suggesting the model may struggle to maintain multi-segment consonant sequences. In *weelaantaweeyani ('you are climbing')*, the lateral approximant /l/ was perceived as the labial-velar approximant /w/ by another evaluator.

Beyond consonant-specific issues, multiple evaluators noted reduced consonant clarity across their respective sample sets. Vowel length, which is phonemically contrastive in Myaamia (e.g., *a* vs. *aa*), was also affected. Two evaluators independently reported issues with vowel duration, including cases where vowels were perceived as longer than expected and instances of incorrect vowel length in some utterances (e.g., *iihia ('yes'), eetiliwatenki ('it is thin ice')*; *pinšiwa awiilawi meeneehwiki ('the cat food is gone')*). Given that the character-based input represents long vowels as doubled characters, this may reflect limitations in the duration model's handling of repeated graphemes. Evaluators also reported audible artifacts, including glitching, skipping, and unnatural segmentation at syllable boundaries. One evaluator identified specific positions within *aayaapweeyohsiaanki ('we take a walk')* where synthesis degraded, while another noted that the final syllable of *keekiipiinkweeholaci ('you blindfold him/her')* was choppy. These issues were particularly noticeable in longer, multi-word utterances. No comparable issues were reported in evaluations of the original recordings, suggesting these are model-specific limitations. These findings inform the directions for improvement discussed in Section 5.

5 Limitations and Future Work

While this work demonstrates the feasibility of neural TTS for Myaamia, several limitations remain.

Modeling Limitations. Our system uses character-based input rather than phoneme-based representations, which may have contributed to some of the phonetic issues identified in Section 4.3, particularly consonant conflation and vowel length errors. A hybrid character-phoneme approach, as explored for Mundari (Gumma et al., 2024), may improve pronunciation accuracy while maintaining flexibility. Additionally, the current system does not explicitly model prosodic features such as pitch, stress, or rhythm. Although the

Table 4: Subjective evaluation of synthesized Myaamia speech vs. original recordings, based on expert ratings. Mean Opinion Score (MOS) with standard deviations reported.

Metric	Generated Speech (MOS ↑ ± SD)	Original Speech (MOS ↑ ± SD)
Naturalness	3.05 ± 0.89	**4.79 ± 0.42**
Understandability	3.05 ± 1.10	4.55 ± 0.69
Overall Quality	3.20 ± 0.77	4.75 ± 0.44
Intonation/Rhythm	**3.52 ± 0.81**	4.52 ± 0.56
Linguistic Quality	3.40 ± 0.91	4.45 ± 0.56

Table 5: Naturalness comparison of TTS models across languages using Mean Opinion Score (MOS) and standard deviation (SD) on a 5-point Likert scale. Training data size (hours) included.

TTS Model (Language)	Training Data Size (Hours)	Naturalness (MOS ↑ ± SD)
Myaamia TTS (Ours)	8.18	3.05 ± 0.89
Võro TTS (Rätsep and Fishel, 2023)	17.00	3.62 ± 0.15
Mundari TTS (Gumma et al., 2024)	24.76	3.69 ± 1.18

VITS model achieved reasonable intonation scores (3.52 ± 0.81), evaluators noted audible artifacts in longer utterances (Section 4.3). Incorporating prosodic embedding layers and attention-based duration predictors may help address both issues (Henter et al., 2017).

Data and Deployment. The dataset imbalance noted in Section 3.1 limits generalization across speaker variation. Future work should prioritize expanding and diversifying the corpus through new recordings from additional speakers. Inference speed also remains a practical constraint for mobile or low-resource deployment; model compression techniques such as pruning or knowledge distillation could improve feasibility for integration into platforms like the Šaapohkaayoni Education Portal. Ongoing collaboration with the Myaamia Center remains essential to ensure that synthesized speech aligns with the linguistic expectations of speakers and learners (Baldwin et al., 2016; Brinklow, 2021).

Evaluation Limitations. The subjective evaluation was limited to four experts, and only the VITS model was assessed. While this reflects the small size of the Myaamia speaker community and the specialized linguistic expertise required for judgments on phonemic contrasts and prosodic alignment, it limits the generalizability of perceptual findings. Future evaluations should include additional community members at varying levels of Myaamia proficiency, to examine whether speakers at different proficiency levels perceive synthesized speech differently. Such evaluations should also assess all three models and report inter-rater agreement metrics to strengthen the evaluation methodology.

6 Ethical Considerations and Sovereignty

This research is grounded in the principles of Indigenous Data Sovereignty and aligns with the CARE Principles for Indigenous Data Governance (Carroll et al., 2020). The Myaamia Center maintains full ownership and governance over all linguistic and cultural data used in this study. The speech corpus was curated through the Indigenous Languages Digital Archive (ILDA), a platform purpose-built to support the reclamation goals of tribal communities, and all work was conducted under Miami University IRB protocol 05009e through a long-standing partnership with the Myaamia Center. Prior to the initiation of this research, the use of ILDA audio data for TTS development was approved by the director of the Myaamia Center, and consent was obtained from the primary speakers whose recordings comprise the dataset. All model training and evaluation were carried out on Miami University managed computing resources governed by the same institutional access controls as the ILDA database; no data was shared with third

parties outside this environment.

From its inception, this project has been developed in collaboration with the Myaamia Center, with center staff and affiliated experts providing guidance on data use, model development, and evaluation criteria. To ensure the resulting technology aligns with both cultural values and pedagogical needs, the subjective evaluation was carried out by linguistic experts and the primary speakers who contributed to the original recordings. Their direct participation means the synthesized speech is validated by the very individuals whose voices the models aim to represent. By centering this expertise at every stage, this work seeks to provide a sustainable, ethically governed tool in support of ongoing Myaamia language reclamation efforts.

7 Conclusion

This work presents the first neural text-to-speech system for the Myaamia language. We evaluated three architectures, FastSpeech, Glow-TTS, and VITS, trained on 8.18 hours of community-curated speech from the Indigenous Languages Digital Archive. VITS achieved the best performance across objective metrics (MCD: 15.22, F0 RMSE: 17.50), and subjective evaluation by Myaamia experts yielded encouraging results, particularly for intonation and rhythm (MOS: 3.52 ± 0.81). At the same time, error analysis revealed specific challenges in consonant distinctions, vowel length accuracy, and audible artifacts in longer utterances, indicating clear directions for future improvement.

By combining objective metrics with expert evaluation from original speakers and linguists, we established a performance baseline grounded in community standards. While further refinement is needed, particularly in phoneme-level modeling, prosodic control, and dataset diversification, this work provides a foundation for integration into community learning platforms such as the Šaapohkaayoni Education Portal. Our ongoing collaboration with the Myaamia Center ensures that development proceeds with appropriate oversight and accountability.

References

Daryl Baldwin, David J. Costa, and Douglas Troy. 2016. Myaamiaataweenki eekincikoonihkiinki eeyoonki aapisaataweenki: A miami language digital tool for language reclamation. *Language Documentation and Conservation*, 10:394–410.

Steven Bird. 2020. Decolonising speech and language technology. In *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pages 3504–3519. Association for Computational Linguistics (ACL).

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

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):1–12.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *Preprint*, arXiv:2410.06885.

Varun Gumma, Rishav Hada, Aditya Yadavalli, Pamir Gogoi, Ishani Mondal, Vivek Seshadri, and Kalika Bali. 2024. MunTTS: A text-to-speech system for Mundari. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 76–82, St. Julians, Malta. Association for Computational Linguistics.

Christopher Hammerly, Sonja Fougère, Giancarlo Sierra, Scott Parkhill, Harrison Porteous, and Chad Quinn. 2023. A text-to-speech synthesis system for border lakes ojibwe. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 60–65. Association for Computational Linguistics.

Gustav Eje Henter, Srikanth Ronanki, Oliver Watts, and Simon King. 2017. Non-parametric duration modelling for speech synthesis with a joint model of acoustics and duration. In *Proceedings of Interspeech*, pages 1213–1217, Stockholm, Sweden. ISCA.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *Advances in Neural Information Processing Systems*, volume 33.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

John Kominek, Tanja Schultz, and Alan W. Black. 2008. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *Proceedings of*

the First Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2008), pages 63–68, Hanoi, Vietnam.

Ivan Kraljevski, Frank Duckhorn, Daniel Sobe, Constanze Tschoepe, and Matthias Wolff. 2024. Preserving language heritage through speech technology: The case of upper sorbian. In *Proceedings of the 26th International Conference on Speech and Computer (SPECOM 2024)*, pages 3–12, Belgrade, Serbia. Springer Nature.

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

Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Preprint*, arXiv:2306.07691.

Zhaojie Luo, Tetsuya Takiguchi, and Yasuo Ariki. 2016. Emotional voice conversion using neural networks with different temporal scales of f0 based on wavelet transform. In *Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW9)*, pages 238–243, Sunnyvale, CA, USA. ISCA.

Daniel Menendez and Hector Gomez. 2025. Text-to-speech system for low-resource languages: A case study in Shipibo-konibo (a Panoan language from Peru). In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 1–7, Albuquerque, New Mexico. Association for Computational Linguistics.

Sewade Ogun, Abraham T. Owodunni, Tobi Olatunji, Eniola Alese, Babatunde Oladimeji, Tejumade Afonja, Kayode Olaleye, Naome A. Etori, and Tosin Adewumi. 2024. 1000 african voices: Advancing inclusive multi-speaker multi-accent speech synthesis. *Preprint*, arXiv:2406.11727.

Aidan Pine, Erica Cooper, David Guzmán, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratékha' Martin, Korin Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells, and Junichi Yamagishi. 2025. Speech generation for indigenous language education. *Computer Speech & Language*, 90:101723.

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)*, volume 1, pages 7346–7359. Association for Computational Linguistics.

Liisa Rätsep and Mark Fishel. 2023. Neural text-to-speech synthesis for võro. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 723–727, Tórshavn, Faroe Islands. University of Tartu Library.

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 *9th International Conference on Learning Representations (ICLR)*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, volume 32.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *Preprint*, arXiv:2304.09116.

Athanasios Tsanas, Matías Zañartu, Max A. Little, Cynthia Fox, Lorraine O. Ramig, and Gari D. Clifford. 2014. Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering. *The Journal of the Acoustical Society of America*, 135(5):2885–2901.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *Preprint*, arXiv:1609.03499.

Antonio Vasilijevic and Davor Petrinović. 2011. Perceptual significance of cepstral distortion measures in digital speech processing. *Automatika*, 52:132–146.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Shenran Wang, Changbing Yang, Michael l Parkhill, Chad Quinn, Christopher Hammerly, and Jian Zhu. 2025. Developing multilingual speech synthesis system for Ojibwe, mi'kmaq, and maliseet. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 817–826, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. *Preprint*, arXiv:1703.10135.

Robert Whitman, Richard Sproat, and Chilin Shih. 1997. A navajo language text-to-speech synthesizer. Technical Report 11222-930830-13TM, AT&T Bell Laboratories.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. *Preprint*, arXiv:2008.03687.