Choosing an ASR model for Dënë Sųłné: Navigating polysynthesis and unstandardized orthography

Olga Kriukova[1], Antti Arppe[2], Olga Lovick[1],

[1]University of Saskatchewan, [2]University of Alberta

Correspondence: olga.kriukova@usask.ca

Abstract

While several pre-trained multilingual models are actively used for fine-tuning on under-resourced and endangered languages, it remains unclear which architectures perform better and what factors explain their varying performance across languages. Although this question may be less pressing for languages with adequate resources, it is critical for endangered language communities, where the time and funding available to experiment with multiple model options is usually severely limited(Jimerson et al., 2023). We compare the performance of two ASR architectures, Wav2Vec2 and Whisper, on a Dënë Sųłné dataset. This language and dataset present several challenges common to under-resourced and endangered languages: unstandardized orthography, variation in pronunciation, and phonological and morphosyntactic structures that differ from the major languages represented in the multilingual datasets used for pre-training large ASR models. Although Wav2Vec2 reportedly outperforms Whisper in low-resource settings (see e.g., Coto-Solano et al., 2024; Nahabwe et al., 2025; Williams et al., 2023), our study shows that Whisper yields significantly better results on the Dënë Sųłné dataset. These findings suggest that model performance may depend not only on architecture, dataset size, or typological features of language, but also on dataset-specific characteristics. In our case, Whisper showed better adaptability to a dataset with inconsistent spelling and pronunciation. Further verification across similarly inconsistent datasets is required to assess the generalizability of this result.

1 Introduction

Automatic Speech Recognition (ASR) is an important technology for under-resourced and endangered languages in many respects (Jimerson and Prud'hommeaux, 2018; Prud'hommeaux et al., 2021). With reliable ASR technologies, language communities can create or expand their written language corpora via ASR-assisted transcription (Ćavar et al., 2016; Lane and Bird, 2021; Zhang et al., 2022), which in turn may assist further in language documentation (cf. Amith et al., 2021; Liu et al., 2022), the creation of educational materials (Prud'hommeaux et al., 2021), and the development of other NLP tools (Zhang et al., 2022).

Modern pre-trained multilingual models promise to provide accurate speech recognition even for languages with very small (1–2 hour) corpora (cf. Babu et al., 2021; Baevski et al., 2020; Meta Research, 2020). However, despite real progress in this area, accurate ASR for many under-resourced and endangered languages is still far from reality. The sizes of the pre-trained ASR model and corpus are not the only factors determining ASR success. Under-resourced languages often face one or more of the following challenges: 1) high-quality recordings are rare, with many languages having only fieldwork-quality audio (Ćavar et al., 2016; Liang and Levow, 2025; Wisniewski et al., 2020); 2) consistent transcriptions do not exist due to the lack of a standard orthography or the presence of competing standards (cf. Xie and Anastasopoulos, 2023); 3) recordings may be collected across different dialects with varying pronunciations (cf. Nigmatulina et al., 2020); or 4) recordings come from only one speaker (cf. Jimerson et al., 2023)).

Beyond data quality issues, many under-resourced languages are typologically different from the major languages represented in the training data of large pre-trained models (Jimerson et al., 2023; Wisniewski et al., 2020). They may have different phonological, morphosyntactic, and orthographic features that these models could not learn during pretraining. Additionally, many endangered languages feature sentence- and word-level code-switching with a dominant regional language (Guillaume et al., 2022), which requires ASR systems to capture two languages at once. While some ma-

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 15–25

jor languages also have features that are challenging for ASR—such as tones in Chinese and Vietnamese, or poor sound-to-letter correspondence in English and French—these problems are often resolved thanks to the availability of large language corpora. For the majority of endangered languages, this solution is not available.

Given these challenges, researchers working with endangered and acutely under-resourced languages must consider many factors before developing ASR for these languages. One key decision is the selection of a pre-trained model. Two main pre-trained model families frequently compared in this field are Wav2Vec2 and Whisper. Several studies have sought to determine whether one ASR model outperforms the others in resource-constrained settings (Jimerson et al., 2023; Nahabwe et al., 2025). However, as we show in Section 2, there is no clear leader in this area, and only limited explanations exist for why one model may work better for one language than another.

In this study, we explore which of these two model architectures performs better on Dënë Sųłné. This language presents many challenges, not only for ASR but for natural language processing in general (see Section 1.1). By examining Wav2Vec2 and Whisper performance on Dënë Sųłné, we aim to contribute to the growing discussion of model choice in under-resourced settings—a particularly important discussion given that language communities do not always have access to the computational resources needed to experiment with multiple ASR architectures (Jimerson et al., 2023).

1.1 Dënë Sųłné

This study focuses on Dënë Sųłné (ISO 639-3: chp; Glottolog: chip1261), a member of the Dene (Athabaskan) language family. It is an endangered Indigenous language spoken in several Canadian provinces and territories (Cook, 2004) by approximately 10,000 speakers (Statistics Canada, 2021). Our data for this study comes from speakers in the sister communities of Clearwater River Dene Nation and La Loche (Saskatchewan, Canada).

Many features of Dënë Sųłné are known to complicate the ASR development, especially in the low resource-settings. It is a polysynthetic language with highly productive verbal morphology, a large phoneme set (35 consonants and 6 vowels), and an unstandardized orthography.

As a heavily prefixing polysynthetic language, Dënë Sųłné exhibits significant complexity in its verbal paradigms (Cook, 2004, p. 91). Verbs participate in highly productive derivational processes which, combined with inflectional paradigms, can generate hundreds or thousands of surface realizations from a single root (cf. Arppe et al., 2017; Lovick et al., 2018). In practice, this productivity significantly amplifies the out-of-vocabulary problem (cf. Abate et al., 2020), while the tight fusion of some morphemes complicates the ability of ASR models to learn meaningful subword units. In addition to this richness, verbs in Dënë Sųłné exibit age-variation that increases the number of observable forms even further.

On top of this morphological richness, Dënë Sųłné marks both nasality and high tone on all six vowels, and both contrasts may be phonemic (e.g. ***yą*** /ya/ ‘sky’ vs. ***yą́*** /yá/ ‘lice’ (Cook, 2004, 6); *thɪ̨yį* ’I stand’, *thį̨yį* ’you stand’ (Elford and Elford, 1998, 293). However, since the orthography is not fully standardized (see Kriukova et al., 2026b for more details) and many speakers have not received formal literacy instruction, transcription tends to be perception-based, with individual variation in pronunciation adding a further source of inconsistency. Nasality and tone markers are consequently the primary site of spelling variation, with a single syllable often appearing in two to four written forms (e.g. ***hots’ı̨***, ***hots’í***, ***hóts’ɪ*** for ’from there’). Combined with the morphological richness described above, this orthographic instability substantially inflates the number of unique tokens in a corpus, compounding data sparsity.

In order to make the corpus more suitable to be used as training data, we first standardized the most frequent types and some closed word classes (see Kriukova et al., 2026b). Though incomplete, this partial standardization significantly improved automatic transcription performance (see Kriukova et al., 2026a).

1.2 The ASR architectures

At the time of this writing, the two main pre-trained multilingual ASR architectures used in low-resource settings are Wav2Vec2 and Whisper. Wav2Vec2 (Baevski et al., 2020), developed by Meta AI, is a self-supervised encoder-only model that learns from unlabeled raw audio and is fine-tuned for ASR using CTC decoding. Notable variants include Wav2Vec2-XLS-R, trained on 128 languages at up to 2B parameters (Babu et al., 2021), and Wav2Vec2-BERT, which operates on mel spectrograms rather than raw waveforms (Seamless

Communication et al., 2023).

Whisper (Radford et al., 2023), developed by OpenAI, is a weakly-supervised encoder-decoder model trained end-to-end on large-scale labeled multilingual data, enabling strong zero-shot generalization. It comes in several sizes (tiny, base, small, medium, and large) all of which differ in the number of parameters they have. For instance, Whisper-medium has 769M parameters, while Whisper-large has twice as many.

The key difference between these two model families lies in their training paradigms—Wav2Vec2 uses self-supervised pretraining on unlabeled data followed by supervised fine-tuning. In contrast, Whisper relies exclusively on large-scale supervised learning. Architecturally, Wav2Vec2 employs an encoder-only structure with CTC decoding, whereas Whisper uses an encoder-decoder framework with autoregressive token generation.

2 Literature Review

Numerous studies have examined the efficacy of Whisper and Wav2Vec2 in low-resource ASR settings. Jimerson et al. (2023) compared the two architectures across eleven typologically diverse languages (with training data varying from 19 minutes to 17 hours) and found no consistently superior model. Performance appeared to be influenced by typological features such as phoneset size and type of morphology: Whisper tended to perform better on languages with larger phonesets and polysynthetic morphology, while Wav2Vec2 showed advantages on isolating languages. Dataset characteristics (e.g., audio quality, source type) were also identified as possible contributing factors.

Nahabwe et al (2025), whose study benchmarked the models on African languages, found that Whisper outperforms Wav2Vec2-BERT only in very low-resource conditions (1–10 hours), possibly due to its encoder-decoder architecture and the composition of its pretraining data. Moreover, they found that supplementation of Wav2Vec2-type models by a language model improved performance at 10–50 hours but caused degradation on larger datasets. Notably, Nahabwe et al.'s results for Wolof favored Wav2Vec2-XLS-R—the opposite of Jimerson et al.'s (2023) findings—further illustrating how dataset-specific factors can influence model comparisons.

Several language-specific studies have demonstrated Wav2Vec2's superiority over Whisper: in Bangla (Ridoy et al., 2025), Maltese (Williams et al., 2023), and two Chibchan languages, Bribri and Cabécar (Coto-Solano et al., 2024). The latter two are particularly notable given their very small datasets (143 and 54 minutes, respectively). Importantly, both languages also present additional challenges—tonal and nasal orthographic features, dialectal variation in Bribri, and non-standardized orthography in Cabécar.

The Wav2Vec2-XLS-R model was also employed for Tsúūtínà, a Dene language closely related to Dënë Sųłiné. With a training dataset of just under 7 hours, the model achieved a CER of 14.5% (C. Cox, personal communication, January 3, 2026), which is an excellent result for such a phonologically and morphologically complex, under-resourced language. Importantly, the training dataset followed a single orthographic convention, and the majority of the data came from one male speaker recorded under optimal conditions (C. Cox, personal communication, January 3, 2026).

Additionally, rather than choosing alternative architectures, some researchers have focused on attempting to finetune existing models more efficiently. LoRA-based fine-tuning has shown particular promise for Whisper-large in low-resource settings (Acharya et al., 2025; Ghimire et al., 2024; Simmons, 2025), though Y. Liu et al. (2024) found that vanilla fine-tuning with bottom-layer freezing can be comparably effective. The generalizability of these findings to languages absent from Whisper's pretraining data remains uncertain.

This study aims to determine the optimal ASR architecture and training conditions for the Dënë Sųłiné dataset we work with.

3 Methodology

3.1 Dataset

The dataset for this study comprises 22,203 utterances. The total length of the corpus is 15 hours and 3 minutes. The dataset is compiled by integrating data from three sources. The recordings made during the Talking Dene project served as the principal corpus (2020-2024; PI: Olga Lovick), supplemented by additional recordings collected by Kriukova for the present study and verb paradigm elicitations recorded by Willems (2025) with a single speaker. All the recordings, except verb paradigms, represent spontaneous speech. All 28 speakers whose recordings are used in this study

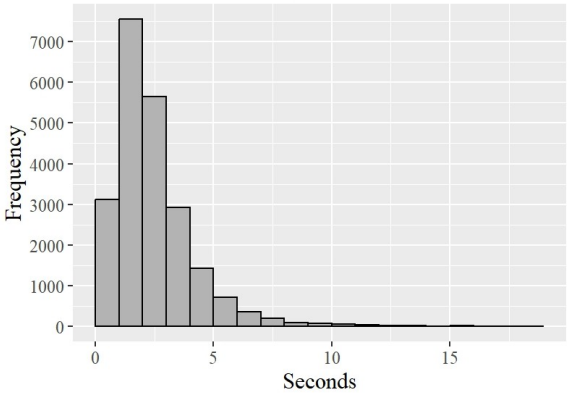


Figure 1: Duration of clips and their frequency.

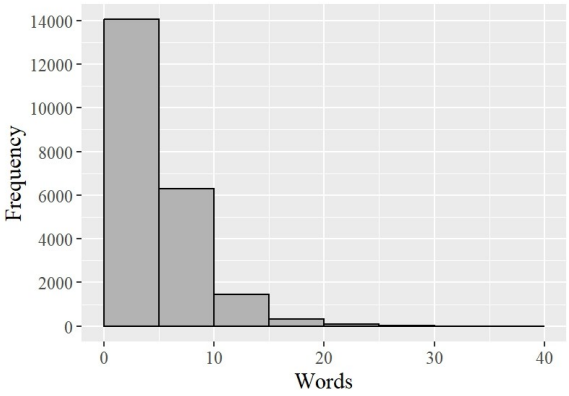


Figure 2: Number of words per utterance.

provided informed consent for the use of their data in model training. The information about the duration of the clips extracted from the recordings and their transcriptions is outlined in Figures 1 and 2.

Since the dataset is not fully orthographically standardized, evaluating the model on a random subset may yield unreliable results. Therefore, we tested our models on a dedicated testing set of 100 utterances. Although it is very small compared to the full dataset, each utterance in it was manually reviewed by Lovick to ensure it represents the transcription quality we aim for. This set is designed to evenly sample speakers represented in the training dataset across genders and ages. Within these constraints, utterances were selected at random, and code-switched utterances were not excluded, as code-switching is a natural and frequent feature of the speakers' language use. This ensures the test set reflects the realistic range of input the ASR system would encounter in practical deployment.

To quantify the extent of orthographic inconsistency in the original transcriptions, we treated them as a noisy baseline and compared them against the standardized corrected versions prepared by Lovick. This yielded a WER of 84.8% and a CER of 31%, where substitutions dominated (67.7%), reflecting the prevalence of non-standard spellings in the originals (not a measure of ASR performance). Insertions accounted for 16.6%, representing missing content that required addition—primarily the separation of fused forms into separate words (e.g., *yísoha* → *yísë o ghą*). Deletions were minimal at 0.5%.

3.2 ASR models

We fine-tuned Wav2Vec2-XLS-R-300M and Wav2Vec2-BERT, using HuggingFace guides.[1] The training scripts for Wav2Vec2-based models are published on GitHub[2]. During fine-tuning, we encountered training instabilities with a subset of 491 training pairs specific to these models. As this issue was discovered in the course of the experiments rather than anticipated by design, we discuss it in detail in the results section.

Among the Whisper models, we fine-tuned Whisper-medium and Whisper-large, following a HuggingFace tutorial.[3]. We experimented with several fine-tuning strategies to address the risk of overfitting when training Whisper-large on a small dataset. First, we applied vanilla fine-tuning, updating all model parameters. We then employed Low-Rank Adaptation (LoRA), which freezes the pretrained weights and introduces small trainable adapter matrices into the attention layers, significantly reducing the number of trainable parameters. We tested two LoRA configurations with varying rank (16; 64) and target modules (q_proj, v_proj; q_proj, v_proj, k_proj, out_proj, fc1, fc2). Additionally, we experimented with freezing the encoder and fine-tuning only the decoder, reducing trainable parameters by approximately half. We evaluated all approaches and selected the best-performing version of fine-tuned Whisper-large. The adapted fine-tuning scripts for Whisper models are also published on the same GitHub.

All models for this study were trained and tested on Plato, a high-performance computing cluster at the University of Saskatchewan. Average training time for the Wav2Vec2-based models ranged from 2 to 8 hours, and for Whisper models, from 10 to 40 hours, depending on the number of epochs.

[1] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2, https://huggingface.co/blog/fine-tune-w2v2-bert

[2] https://github.com/HelgaKr/DS-ASR

[3] https://huggingface.co/blog/fine-tune-whisper

3.3 Language model

Since Wav2Vec2 models perform better when supplemented by a language model (Baevski et al., 2020; Jimerson et al., 2023), we also trained one for our experiments. Our n-gram language model was trained on the same corpus we used for ASR training, excluding the paradigm recordings (4.4% of the full dataset), since they are concatenated and do not represent valid utterances. To train the model, we used the KenLM package (Heafield, 2011). Text preprocessing matched the Wav2Vec2 training pipeline: Unicode NFC normalization, removal of special characters, and lowercasing. The script we used to train this language model is published on GitHub[4].

4 Results

In this study, we experimented with two Whisper models and two Wav2Vec2-based models. Additionally, we tested all Wav2Vec2 models with and without a language model. Our findings are summarized in Figure 3 and demonstrate that Whisper models outperformed Wav2Vec2-based ones in all cases. Whisper-medium showed the best WER among all the models at 60.7%. Whisper-large delivered the best CER of 34%; however, the difference in CER between large and medium Whisper was negligible (see Figure 3). Given the minimal CER difference and shorter training time, we consider Whisper-medium to be the best-performing model among those we tested in this study. To verify that this performance gap is not not due the small test set, we conducted paired bootstrap resampling (10,000 iterations) on the WER scores of Whisper-medium and the best Wav2Vec2-based model, yielding a 95% confidence interval of [1.21%, 14.68%] — excluding zero and confirming that the difference is statistically significant.

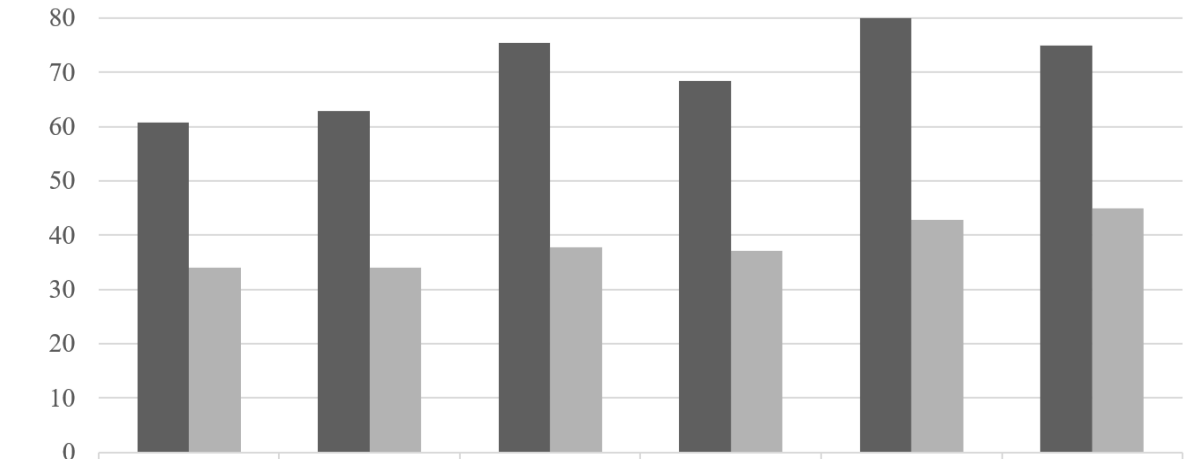
Counter to our expectations, LoRA did not improve Whisper-large performance in our settings. Both LoRA configurations resulted in higher error rates (WER 89% and 84.5%, CER 57% and 54.2%, correspondingly) compared to the vanilla fine-tuned model, likely due to insufficient adapter capacity for a domain substantially different from the pretraining data. Freezing the encoder while fine-tuning the decoder also did not lead to improvements. Ultimately, vanilla fine-tuning of Whisper-large provided the best result for this model.

[4] https://github.com/HelgaKr/DS-ASR

The breakdown of substitutions, deletions, and insertions (macro-averaged) made by the Whisper-medium and Wav2Vec2-BERT with LM (see Table 1) reveals that Whisper produced substantially more complete transcriptions. Macro-averaging was chosen to ensure that shorter utterances, which are common in conversational speech, contributed equally to the evaluation rather than being dominated by longer utterances. The results show that Wav2Vec2 deleted 71% more words (13.7% vs 8.0%) and 78% more characters (17.4% vs 9.8%) than Whisper. In contrast, Whisper exhibited higher insertion rates—48% more at the word level (6.8% vs 4.6%) and 90% more at the character level (7.6% vs 4.0%).

Additionally, since missing or added nasality and tone markers lead to considerable spelling variation in the corpus, but do not always reflect lexical distinctions, we checked how many deletions and insertions involved these diacritic symbols. The analysis showed that tone and nasal marking accounted for approximately 17–20% of character-level errors in both Whisper-medium and Wav2Vec2-BERT with LM. However, the error profiles differed: Wav2Vec2 deleted 38% more tone marks (51 vs. 37) and 21% more nasal marks (17 vs. 14) than Whisper, while Whisper inserted nearly three times more tone marks (31 vs. 11). If these errors are excluded, effective CER drops from 34.1% to approximately 28% for Whisper and from 37.2% to approximately 31% for Wav2Vec2. This suggests that both models perform better at the character level than raw CER indicates. At the word level, diacritic-only errors—where the base word is correct but tone or nasal marking differs—accounted for only 7.1% of Whisper's word errors and 4.0% of Wav2Vec2's. Consequently, WER is less inflated by diacritic issues than CER, and the majority of word-level errors (>90%) reflect genuine base-word misrecognitions or multiple spelling errors.

Moving beyond quantitative metrics to examine the outputs of the best-performing model from each architecture, we observed that certain sentences were transcribed accurately by both models or with only minor mistakes (Example 1a). Mostly, such sentences contained high-frequency vocabulary. However, a notable divergence between the models emerges in other utterances. Since Wav2Vec2 operates at the character level, it frequently generates non-words that are phonetically close to the target forms, such as *bëcjënętdı* in Ex-



	Whisper-med	Whisper-lrg	Wav2Vec2 Bert	Wav2Vec2 Bert_LM	Wav2Vec2 XLS-R	Wav2Vec2 XLS-R_LM
WER	60.7	62.9	75.4	68.4	80	75
CER	34.1	34	37.8	37.2	42.9	45

Figure 3: WER and CER comparison for all fine-tuned models.

Level	Metric	Whisper-medium	Wav2Vec2-BERT_LM
Word	Substitution	45.87%	50.10%
	Deletion	8.03%	13.72%
	Insertion	6.84%	4.61%
Character	Substitution	16.84%	15.76%
	Deletion	9.79%	17.43%
	Insertion	7.63%	4.01%

Table 1: Comparison of substitutions, deletions, and insertions made by the models.

ample 1b. Whisper, by contrast, operates at the subword and word levels, and consequently tends to mainly produce attested lexical items that best match the acoustic input, such as *bëch'ánıdílë* in Example 1b. As a result, Wav2Vec2 outputs often resemble phonetic transcriptions (Example 1c).

We also found that Wav2Vec2 transcriptions of code-switched utterances exhibited clear signs of catastrophic forgetting with respect to English. Despite applying parameter freezing strategies during fine-tuning, performance on English did not improve. Only the integration of a language model led to modest gains. Given that English recognition was not a top priority for our purposes—where Dënë Sųłıné remained the primary target—we did not pursue further optimization of the Wav2Vec2-based pipeline. Notably, Whisper did not exhibit this behaviour (Example 1d).

Additionally, during this study, we observed that the two model architectures behaved differently on our dataset. Although both accept audio recordings in the standard ASR format (16,000 Hz, mono), Wav2Vec2 exhibited training instabilities with certain audio-transcription pairs in our corpus. Specifically, we observed sudden loss spikes followed by collapse to zero when the model encountered utterances with sparse transcriptions (fewer than 10 characters) or fast speech rates. Consequently, we adapted our fine-tuning scripts to exclude these problematic pairs from the training data. In contrast, Whisper models handled all files in our dataset without issue. In total, the Wav2Vec2-based models were fine-tuned on 491 fewer files than the Whisper models.

5 Discussion and Conclusions

5.1 Models' performance and language features

Our study found that the Whisper architecture provided more accurate speech recognition for Dënë Sųłıné. This finding aligns with Jimerson et al. (2023), who showed that Whisper models perform better for languages with large phoneme invento-

(a)	**Ground truth:**	*west la loche nı́ są́ hhamá nı́ west la loche nádhër ú*
	Whisper-medium:	*west la loche nı́ są́ hhamá nı́ west la loche nadhër ú*
	Wav2Vec2-BERT with LM:	*west la loche nı́ są́ hamá nı́ west la loche nádhër ó*
	Translation:	'It was in West La Loche. My mom was living in West La Loche.'
(b)	**Ground truth:**	*há bëch'ánëdíla hájá*
	Whisper-medium:	*hą* ***bëch'ánıdílë*** *hájá*
	Wav2Vec2-BERT with LM:	***bëcjënętdı*** *lı já*
	Translation:	'Okay, and you don't like it anymore?'
(c)	**Ground truth:**	*cause kǫ́t'u náts'ëdé sëba darıtɬʔə́dh kǫ́t'u nëstı dúé á*
	Wav2Vec2-BERT with LM:	*cause kǫ́t'u nesedé sëba darıtl'ë kǫ́t'u nestee dúé*
	Wav2Vec2-BERT with LM:	*cause kǫ́t'u néts'ëdë* ***sëbąderëdléh*** *kǫ́t'u nestı dúé*
	Translation:	'Cause if everyone is up, and it is loud, I can't sleep like that.'
(d)	**Ground truth:**	*small baby horə́sʔı sëkóę́ yísë o ghą atú́ łá*
	Wav2Vec2-BERT with LM:	*small baby horə́sʔı sëkóę́ yísı ha la*
	Wav2Vec2-BERT with LM:	*small* ***bebe*** *horə́sʔı sëkóę́ yísı ha hu łá*
	Wav2Vec2-BERT w/o LM:	***smal bebey*** *horə́sʔı sëkóę́ yísıhą hu łá*
	Translation:	'I want a small baby for my house.'

Example 1: Transcriptions produced by the models.

ries. One of the languages in their study, Hupa, belongs to the same language family as Dënë Sųłiné and, similarly, achieved better results with Whisper, further supporting this pattern. Additionally, two polysynthetic languages in Jimerson et al.'s study achieved better WER than both Wav2Vec2 and Wav2Vec2 with language models: Hupa and Seneca. In our study, we obtained similar results, with both Whisper models outperforming Wav2Vec2-based models with a language model. These results may indicate that Whisper performs better with polysynthetic languages and those with large phoneme inventories. Nevertheless, a study on ASR development for Tsúūtínà (Cox, 2023; Rodríguez and Cox, 2023) achieved great recognition results (CER 14.5%) using a Wav2Vec2-XLS-R model (without a language model), with a smaller dataset (C. Cox, personal communication, January 3, 2026). Although Cox did not directly compare Wav2Vec2 with Whisper during the development, the fact that a closely related language with an almost identical phoneme inventory, morphological characteristics, and smaller dataset size produced such different outcomes warrants further investigation. One possible explanation is that the Tsúūtínà dataset has consistent spelling, good-quality recordings, and mostly represents the speech of a single speaker (C. Cox, personal communication, January 3, 2026), which significantly reduced variation in pronunciation. In contrast, our dataset contains substantial spelling and pronunciation variability on top of the fieldwork quality recordings, which may explain why all Wav2Vec2-based models underperformed in our case.

Our results also run contrary to numerous studies that have found Wav2Vec2 to outperform Whisper in low-resource settings (cf. Coto-Solano et al., 2024; Ridoy et al., 2025; Williams et al., 2023), or on datasets larger than 10h (Nahabwe et al., 2025). These studies attributed Wav2Vec2's success to its architecture. However, based on our findings and those of Jimerson et al. (2023), we suggest that the relative performance of these model families in low-resource settings may depend less on their architecture and language dataset size, and more on language features and dataset characteristics, or on dataset consistency in particular. It should be noted, however, that a direct architectural comparison in our study is complicated by the fact that Wav2Vec2-based models were trained on 491 fewer utterances than Whisper models due to training instabilities encountered during fine-tuning (see Section 4 for details), and this should be taken into account when interpreting the performance gap.

5.2 The role of dataset consistency

Our corpus, despite being relatively large by under-resourced standards, demonstrates that size alone does not guarantee better WER and CER. Variation in pronunciation or spelling poses a significant challenge in low-resource contexts because even a relatively large dataset may contain enough inconsistency to hinder effective learning. We therefore suggest that the performance gap we observed between the two architectures is likely related to the overall inconsistency of our dataset, and that Whisper may be more adaptable under such conditions.

This finding has broader implications. Inconsistent datasets may seem like a niche problem in ASR, as researchers typically strive to use the "cleanest" data for training. However, such inconsistency is not uncommon in under-resourced language contexts (for examples, see Jones & Mooney 2017) and may even discourage researchers from attempting machine learning on datasets they perceive as less than "ideal". While it is possible to standardize some datasets to some degree, complete standardization can be difficult and time-consuming for various reasons (Hinton, 2014; Jones and Mooney, 2017). It is therefore important to understand which ASR models can better adapt to such conditions, enabling the development of ASR systems even in the absence of standardization. Our study suggests that Whisper performs better under such conditions. However, since spelling consistency is rarely reported for low-resource ASR training datasets—especially in comparative studies—it remains difficult to generalize how Whisper and Wav2Vec2 compare in performance across languages with inconsistent or unstandardized orthography. Further research on speech recognition for such languages is needed to verify this hypothesis.

5.3 Support of the ASR-assisted transcription

Since ASR models for under-resourced languages are frequently developed to support ASR-assisted transcription, it was essential to evaluate the relative suitability of Whisper-medium and Wav2Vec2-BERT with the LM for this task. In ASR-assisted workflows, deletions impose a greater correction burden than insertions: missing words require transcribers to re-listen to the audio and reconstruct content, whereas hallucinated words are typically salient and can be easily removed. Whisper's lower deletion rate (8.0% vs. 13.7% at the word level) yields more complete initial drafts, reducing the need for time-consuming gap-filling. This difference likely reflects the models' architectures: Wav2Vec2's CTC-based approach ties output directly to audio frames and tends to return blanks when uncertain, whereas Whisper's seq2seq decoder is biased toward generating complete transcriptions.

Additionally, the distribution of deletions and insertions related to nasality and tone markers revealed in our analysis further supports Whisper's better suitability for ASR-assisted transcription of Dënë Sųłıné. Diacritic errors require less correction effort than base-word errors: the intended word remains immediately recognizable, requiring only a minor character-level edit rather than retyping the entire word. Notably, a higher proportion of Whisper's substitution errors were diacritic-only mistakes (9.2% vs. 5.5%), where the base word is correct and only the tone or nasal marking needs adjustment. Wav2Vec2, by contrast, produced more errors that required more editing or full-word replacement. Moreover, Whisper tends to preserve more diacritic symbols than Wav2Vec2. For a language like Dënë Sųłıné, these two factors can make a workflow of ASR-assisted transcription easier.

5.4 Practical considerations

During our experiments with the models, we found a practical disadvantage of Wav2Vec2: not all recordings were suitable for it. While Whisper processed all recordings without issue, Wav2Vec2 became unstable when encountering recordings with a high character-to-frame ratio. This suggests that Wav2Vec2's architecture may struggle with extreme ratios that are unavoidable in some languages or recording environments, whereas Whisper's architecture handles such mismatches well.

Nevertheless, Wav2Vec2 has one important advantage: it trains significantly faster. For communities and researchers without access to free computing infrastructure, such as university resources, Wav2Vec2 may be a more affordable option. For instance, in our case, Wav2Vec2-BERT supplemented with a language model showed results not much worse than those of both Whisper models. Therefore, in situations when training resources are limited, resorting to Wav2Vec2 should not result in a significant loss in transcription quality.

Given that the Wav2Vec2 architecture operates at the character level, we expected it to handle Dënë Sųłıné, with its rich derivational morphology,

more effectively, avoiding the OOV problem entirely. This, however, did not prove to be the case. Nevertheless, we do not rule out the possibility that if fine-tuned on a larger, more standardized corpus, Wav2Vec2-based models may outperform Whisper models. For now, however, Whisper is the clear choice for our dataset.

Limitations

This study compares Wav2Vec2 and Whisper on a single dataset drawn from two communities, characterized by high orthographic inconsistency. Further experiments across a broader range of languages are needed to determine whether Whisper's advantage is consistent in such contexts, or whether it is specific to cases where unstandardized orthography co-occurs with large phoneme inventories and polysynthetic morphology. Future work should prioritize datasets that share similar characteristics to enable more generalizable conclusions.

Additionally, the architectural comparison is not perfectly controlled: as noted in Sections 3 and 4, training instabilities led to Wav2Vec2 being fine-tuned on 491 fewer utterances than Whisper. While this was an emergent issue rather than a deliberate design choice, it should be taken into account when interpreting the performance gap between the two architectures.

Ethical considerations

This study was approved by the University of Saskatchewan Board of Ethics (Beh-REB-4918). All speech data used in this study was used with explicit consent from speakers. Participating communities were informed about the results of this study and were involved in the testing and evaluation of the fine-tuned Whisper-medium model. The dataset and models cannot be made publicly available until the Clearwater River and La Loche communities decide whether and how they want to distribute them.

Acknowledgments

We are grateful to the Clearwater River and La Loche (SK, Canada) Dene communities for the opportunity to work with their language. We especially want to thank the research assistants from the Clearwater River for their help in the data collection and transcription for this study: Trina Lemaigre and Chastity Sylvestre. Moreover, we want to thank all participants, whose recordings were used for the training of the Automatic Speech Recognition model (some referred to by pseudonym): Rebecca Dene, Teresa Dene, Mitchell Guetre, Gerald E. Haineault, Brenda Herman, Rhonda Herman, Sharon Kennedy, Alison Lemaigre, Andrea Lemaigre, Antoinelle Lemaigre, Edainya Lemaigre, Jeannie Lemaigre, Jennifer Lemaigre, Johnny Lemaigre, Mikki Lemaigre, Miranda Lemaigre, Randall Lemaigre, Taitlyn Lemaigre, Taylon Lemaigre, Tina Lemaigre, Trina Lemaigre, Tyanne Lemaigre, Doreen Moise, Ernie Piche, Heather Piche, Ursula Piche, and Jeff Toulejour. We also want to thank Nial Willems for his help with verb-paradigm checking and for providing his checked transcriptions and recordings for this study. This study was funded by the SSHRC Partnership Grant 895-2019-1012 "21st Century Tools for Indigenous Languages".

References

Solomon Teferra Abate, Martha Yifiru Tachbelie, and Tanja Schultz. 2020. Multilingual acoustic and language modeling for Ethio-Semitic languages. In *Proc. Interspeech 2020*, pages 1047–1051.

Priyobroto Acharya, Soham Chaudhuri, Sayan Das, Dipanjan Saha, and Dipankar Das. 2025. JUNLP@LT-EDI-2025: Efficient Low-Rank Adaptation of Whisper for Inclusive Tamil Speech Recognition Targeting Vulnerable Populations. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 17–25, Naples, Italy. Unior Press.

Jonathan D. Amith, Jiatong Shi, and Rey Castillo García. 2021. End-to-end automatic speech recognition: Its impact on the workflow in documenting Yoloxóchitl Mixtec. In *Proceedings of the 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 64–80.

Antti Arppe, Christopher Cox, Mans Hulden, Jordan Lachler, Sjur .N. Moshagen, Miikka Silfverberg, and Trond Trosterud. 2017. Computational modeling of verbs in Dene languages: The case of Tsuut'ina. In *Working papers in Athabascan Linguistics ("Red Book" series)*, Fairbanks. Alaska Native Language Center.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint*. ArXiv:2111.09296 [cs].

Alexei Baevski, Henri Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *34th Conference on Neural Information Processing Systems*, pages 12449–12460, Vancouver, Canada.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint*. ArXiv:2312.05187 [cs].

Eung-Do Cook. 2004. *A grammar of Dëne Sųłiné (Chipewyan)*. Number 17 in Algonquian and Iroquoian Linguistics. University of Manitoba, Winnipeg.

Rolando Coto-Solano, Tai Wan Kim, Alexander Jones, and Sharid Loáiciga. 2024. Multilingual Models for ASR in Chibchan Languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8521–8535, Mexico City, Mexico. Association for Computational Linguistics.

Christopher Cox. 2023. XLS-R-ELAN: An implementation of XLS-R automatic speech recognition as a recognizer for ELAN.

Leon Elford and Marjorie Elford. 1998. *Dene (Chipewyan) Dictionary*. Northern Canada Mission Distributions.

Rupak Raj Ghimire, Prakash Poudyal, and Bal Krishna Bal. 2024. Improving on the Limitations of the ASR Model in Low-Resourced Environments Using Parameter-Efficient Fine-Tuning. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 408–415, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Leanne Hinton. 2014. Orthography wars. In M Cahill and Keren Rice, editors, *Developing orthographies for unwritten languages*, pages 139–168. SIL International.

Robbie Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the right ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 1008–1016.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for Documenting Acutely Under-Resourced Indigenous Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mari C. Jones and Damien Mooney. 2017. Creating orthographies for endangered languages. In Mari C. Jones and Damien Mooney, editors, *Creating orthographies for endangered languages*, pages 1–35. Cambridge University Press.

Olga Kriukova, Antti Arppe, and Olga Lovick. 2026a. Data-centric approach to low-resource ASR model performance improvement: The case of Dënë Sųłiné. *(Submitted)*.

Olga Kriukova, Gabrielle Fontaine, Alison Lemaigre, Dagmar Jung, Antti Arppe, and Olga Lovick. 2026b. Using automatic speech recognition to assist with standardization of Dënë Sųłiné transcripts. *(Submitted)*.

William Lane and Steven Bird. 2021. Local Word Discovery for Interactive Transcription. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2058–2067, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages. *arXiv preprint*. ArXiv:2506.17459 [cs].

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of Whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2022. Enhancing documentation of Hupa with automatic speech recognition. In *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 187–192.

Olga Lovick, Christopher Cox, Miikka Silfverberg, and Antti Arppe. 2018. A computational architecture for the morphology of Upper Tanana. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Meta Research. 2020. Wav2vec 2.0: Learning the structure of speech from raw audio.

Alvin Nahabwe, Sulaiman Kagumire, Denis Musinguzi, Bruno Beijuka, Jonah Mubuuke Kyagaba, Peter Nabende, Andrew Katumba, and Joyce Nakatumba-Nabende. 2025. Benchmarking Automatic Speech Recognition Models for African Languages. *arXiv preprint*. ArXiv:2512.10968 [cs] version: 1.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for Non-standardised Languages with Dialectal Variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation and Conservation*, 15:491–513.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Md Sazzadul Islam Ridoy, Sumi Akter, and Md Aminur Rahman. 2025. Adaptability of ASR models on low-resource language: A comparative study of Whisper and Wav2Vec-BERT on Bangla. *arXiv preprint*. ArXiv:2507.01931 [cs] version: 1.

Lorena M Rodríguez and Christopher Cox. 2023. Speech-to-text recognition for multilingual spoken data in language documentation. In *Proceedings of the 6th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 117–123.

Mark Simmons. 2025. Data augmentation for low-resource bilingual ASR from Tira linguistic elicitation using Whisper. In *Proceedings of the 8th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 155–161, Honolulu, HI, USA.

Statistics Canada. 2021. Mother tongue by geography, 2021 Census.

Nial Austen Willems. 2025. The ts'ë- passive in Dëne Sųtłıné. Master's thesis, University of Saskatchewan, Saskatoon, Canada.

Aiden Williams, Andrea DeMarco, and Claudia Borg. 2023. The applicability of Wav2Vec2 and Whisper for low-resource Maltese ASR. In *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 39–43. SIGUL. Accepted: 2024-09-19T06:26:48Z.

Guillaume Wisniewski, Séverine Guillaume, and Alexis Michaud. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 306–315, Marseille, France. European Language Resources association.

Ruoyu Xie and Antonios Anastasopoulos. 2023. Noisy parallel data alignment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1501–1513, Croatia.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced Aligner, ASR. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 4004–4011, Slovenia.