Addressing Domain Mismatch in ASR for Akuzipik Language Documentation

Summer Chambers[1], Sylvia L.R. Woodrose Schwartz[1], Matthew C. Kelley[1], Lane Woodrose Schwartz[2]

[1]George Mason University, [2]University of Alaska Fairbanks

Correspondence: schamb3@gmu.edu

Abstract

The use of ASR models in endangered language documentation has grown in popularity given the bottleneck of manual speech transcription. Meta's Massively Multilingual Speech (MMS) model is particularly popular for its extensibility to low-resource languages. However, it is mostly trained on read speech data from the Bible, meaning it may not perform well on other domains. We evaluated this model on data collected as part of a larger language documentation and revitalization project focused on Akuzipik, a polysynthetic Alaska Native language. We also finetuned and evaluated the model on a small (<1h) collection of speech. The original model performed well on a dataset that roughly matched the Bible training data in domain and writing style but struggled on a separate collection of spontaneous speech. Performance on spontaneous speech improved after finetuning on a sample of our full dataset, and error rates reduced less dramatically after finetuning only on read speech. Both finetuning scenarios show promise for future model improvement, especially considering the relative ease of collecting read speech data. This experiment confirms the challenge of transcribing spontaneous speech with the MMS ASR model but provides hope for improving model performance for language documentation purposes, even with scarce data.

1 Introduction

Automatic speech recognition (ASR) models can speed up the transcription of spoken language—a key step for data collection in language documentation. Manual speech transcription is a huge bottleneck for many language documentation projects, since natural language data must traditionally be transcribed in order to analyze the language itself and archive grammatical materials (Bird, 2021; Shi et al., 2021; Liang and Levow, 2025). Shi et al. (2021) note that manually transcribing one hour of speech can take up to 50 hours, even for a native speaker of that language. Tools like ASR models have the potential to relieve this bottleneck by assisting with transcription, promoting more efficient data collection and annotation. High-performance ASR models can also further language revitalization and reclamation efforts when integrated into language-learning applications and assistive or convenience-based technologies.

Training functional ASR models can be quite a challenge for languages or varieties considered to be "low-resource" (this set comprises the vast majority of languages spoken on Earth; see Joshi et al., 2020). That said, large multilingual ASR models pre-trained on vast amounts of data have become popular for their extensibility to new languages without requiring much labeled speech data in the target language. Models like wav2vec 2.0 (Baevski et al., 2020), XLS-R (Babu et al., 2022), and Whisper (Radford et al., 2023) are trained on hundreds of thousands of hours of speech data from between 50 to 150 languages. Meta increased the number of pre-training languages to 1,406 in their Massively Multilingual Speech (MMS) model (Pratap et al., 2024) by scraping online readings of the Bible, resulting in a capable base model and 1,107 language-specific "adapters" (lightweight modules that have been trained to perform well on one specific language in conjunction with the base model).

As part of a much broader effort related to the documentation and revitalization of Akuzipik —an Indigenous Alaskan language—in this paper, we evaluate the existing MMS Akuzipik adapter model on a new dataset with the goal of ascertaining the model's potential usefulness towards that project. As of writing this paper, no one has evaluated the performance of an ASR model on Akuzipik using data that does not come from the same domain it was trained on: recordings of the New Testament. We compare the model's performance on spontaneous and read speech and explore appropriate fine-

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 93–103

tuning strategies. After examining the most common kinds of ASR errors, we discuss approaches for improving the model and overall transcription pipeline for better efficacy in future language documentation and revitalization projects.

1.1 Akuzipik Language Documentation

Akuzipik (ISO 639-3: ess) is spoken on Sivuqaq (St. Lawrence Island) in Alaska and on the Chukotka Peninsula of Russia (Koonooka et al., 2021). A member of the Inuit-Yupik-Unangan language family, Akuzipik is an endangered language, with fewer than 1000 living speakers (Schreiner et al., 2022). Like others in its language family, Akuzipik is described as having polysynthetic morphology, associated with long, multi-morphemic words and a correspondingly large set of vocabulary (Hunt et al., 2023).

Despite the historical and ongoing decline in use of the Akuzipik language, documentation and revitalization/reclamation efforts are underway within the community on Sivuqaq, particularly in recent years with the establishment of a language revitalization committee. Linguists external to the community are also involved in the efforts, several of which involve developing digital resources for Akuzipik in conjunction with native speakers and community members (Hunt et al., 2023).

One of the near-term goals of the language documentation effort is to transcribe and digitize a collection of oral stories told by elders, which have cultural significance to the community. Such stories can be transcribed by hand, but using an ASR model—at least as a first pass—could make the process faster and less tedious for the native speakers who have the skills to transcribe such stories (Bird, 2021). Though most research suggests that correcting automatically-generated transcripts is faster than transcribing speech by hand, there is some disagreement over the conditions in which this is true (Gaur et al., 2016; Ma et al., 2024). We return to this topic in the discussion section.

1.2 Spontaneous Speech

These oral stories fall under the broad category of "spontaneous" (natural/unplanned) speech as opposed to "read" speech which is read aloud from a book or other textual source material. This distinction is important for a few reasons. First, since spontaneous speech involves more natural productions of spoken language, it may be the most appropriate form of data for language documentation. Tucker and Mukai (2023) note that spontaneous speech displays much more variation than read speech. While this makes it ideal for analyzing sociolinguistic variation and capturing unique features of a language community such as discourse patterns and oral tradition, it also makes spontaneous speech much more difficult for ASR models to get right (Liang and Levow, 2025).

As Nakamura et al. (2008) explain, spontaneous speech is both acoustically and linguistically different from read speech. They attribute reduced ASR accuracy for spontaneous speech to its "spectral reduction" (blurred acoustic distinctions). In general, spontaneous speech tends to be faster and contains more self-corrections, filler words, partial words, hesitations, repetitions, and reductions. Tucker and Mukai (2023) highlight speech reductions as a major difference from read speech. While the transcribed data in this experiment and the oral stories likely to require transcription are primarily monologic, Evain et al. (2024) show that more conversational and casual forms of spontaneous speech like multi-speaker dialogues among friends or family result in even higher ASR error rates.

1.3 Meta's ASR Model and Akuzipik Adapter

The MMS base model and its 1,107 language-specific adapters are publicly available for download and adaptation. The MMS model has been found to be one of the best choices for transcribing low-resource languages with very small amounts of labeled speech data (Mainzinger, 2024; Liang and Levow, 2025). Through multilingual pre-training, the base model captures basic acoustic principles of human speech, though its success still varies significantly based on the target language variety and its degree of representation in the training data. The adapters, which are generally trained on only one language each, capture the acoustic principles of a specific language's sound system.

This architecture is advantageous in that MMS's language-specific adapters are much smaller than the full model itself, which makes training or fine-tuning an adapter doable even on a less-than-super computer. Le Ferrand et al. (2024) note that among the languages the model was trained on are some of the first polysynthetic languages represented in multilingual models, which are morphologically rare around the world but not uncommon among the Indigenous languages of the Americas. Le Ferrand et al. (2024) saw very good results from a similar XLS-R model they trained on Akuzipik New

Testament data but did not evaluate the model's performance on other domains.

The MMS Akuzipik adapter was trained on the entire text of the Akuzipik translation of the New Testament, paired with 33 hours of speech read by five native speakers on Sivuqaq. The text is precise and formal, consisting of historical and religious content. Aside from the difference in speech format, the domain of the text itself is quite different from that of the speech we want to transcribe; these oral stories may include a mix of formal and informal speech and will contain reference to more modern topics, as well as neologisms or borrowings. For this reason, we chose to evaluate the MMS Akuzipik model on a selection of data collected for various language documentation tasks, with the expectation that an ASR model trained on only one domain (the New Testament) is likely to need adaptation to perform well on new domains.

2 Methods

2.1 Data

While not included in XLS-R or Whisper, Akuzipik was one of the 1,406 languages for which Meta scraped Bible recordings to produce their Massively Multilingual Speech (MMS) dataset (Pratap et al., 2024). Several websites such as bible.com provide downloadable links to two 33-hour collections of speech recordings of the Akuzipik New Testament, with and without music overlaid, split by chapter and book. There are five speakers (three women and two men) included in that data. Aside from the New Testament data, little to no labeled Akuzipik speech data is publicly available. See the section on ethical considerations for a discussion of some of the many concerns surrounding the curation of the MMS dataset.

For this project, speech recordings produced during language documentation fieldwork were used to evaluate and finetune the MMS Akuzipik adapter model. The data collection process—which occurred between 2023 and 2025 on Sivuqaq—was approved by the first author's institutional review board. Each native speaker involved in the effort signed informed consent forms and was compensated for their time. Since some speakers prefer to remain anonymous, numerical aliases are used.[1] See Schreiner et al. (2022) for more details on how the authors approach this kind of fieldwork in-person as well as from a distance.[2]

One data source includes readings of very short sentences or single words with an average duration of 4 seconds each. These were collected by the second author in 2023 for the purpose of being integrated into the existing online Akuzipik dictionary. A second source of data includes spontaneous speech in the form of a story about the speaker's childhood. This story was broadly elicited by the second author in 2024 as part of a project analyzing Akuzipik syntax and semantics. The final set of data used in this project includes readings of a short fable called "The North Wind and the Sun" which was translated into Akuzipik by native speakers. Recordings of this particular fable are often used in phonetic documentation, which is the purpose for which a group of linguists including the first author elicited these data in 2025. The duration of all three sets together is 42.5 minutes. See Table 1 for a more detailed breakdown of the data.

2.2 Model Evaluation and Finetuning

The first step of our experiment was to run the full dataset through the MMS model with the Akuzipik adapter. The model and adapter used are available through Huggingface and were downloaded locally. Each of the model's predicted transcriptions was then evaluated against its gold-standard transcription for that sound file.

In ASR evaluation, word error rate (WER) is the most popular metric, but character error rate (CER) is not uncommonly used. WER is discussed with reference to overall model usefulness later, but CER was our preferred metric for a few reasons. CER provides a less "harsh" evaluation of transcriptions, particularly for certain languages with large vocabularies or long words. For instance, agglutinative and polysynthetic languages tend to have longer words made of many morphemes, so they would be mathematically punished more severely than a morphologically isolating language would for the same number of overall errors with WER (Le Ferrand et al., 2024). K et al. (2025) argue that CER is a "better" metric overall in that it is more closely correlated than WER is with human judgment of transcription errors.

After evaluating the model off-the-shelf on this

[1]The following speakers wished to be identified by name: 1: Petuwak Christopher Koonooka, 2: Apa John Apangalook, 3: Amaghalek Beulah Nowpakahok.

[2]Although the data and models are not made public to respect the privacy and data sovereignty of the Yupik people, code for this project is available at: https://github.com/SaintLawrenceIslandYupik/ComputEL2026

Dataset	Duration per utterance	Duration	Speaker Aliases	Speakers in Bible Data	Recording Device Used
Read Phrases	≈ 4 s	26 min	[4, 5, 6]	[]	[Zoom, phone]
Spontaneous	≈ 10 s	12 min	[1]	[1]	[Zoom]
Read Fable	≈ 12 s	4 min	[1, 2, 3]	[1, 3]	[Zoom]

Table 1: Datasets used for model evaluation and finetuning. Professional-quality Zoom brand recorders or mobile phones used.

new data, we then finetuned the Akuzipik adapter. Finetuning the adapter is much faster/easier than finetuning the entire MMS base model, which would require significant computing resources and much more data. Finetuning was done using a sample of the 42.5 minutes of all labeled data for training, development, and testing sets—see Table 2 for a detailed breakdown.

Allocation of each sound file and transcription from the three source datasets into training, development, and testing sets was random, except for ensuring that the same sentence only showed up in one of the three splits if it was associated with multiple sound files (recorded by multiple speakers or on multiple recording devices). The smallest source dataset—read speech from a translation of Aesop's fable "The North Wind and the Sun"—was so small that we decided only to include it in the test set, not train or dev. The other two source datasets are split into all three train/dev/test sets.

The finetuning process closely followed a Huggingface blog post (Von Platen, 2023), which details how to finetune an MMS adapter on a small amount of data. Sound files were all in WAV format, resampled to 16000 Hz, and converted to a single channel when necessary. Aside from the actual data and language-specific files like "vocabulary", which is character-based for the MMS model, the only changes made to the code in that blog post were switching WER to CER as the evaluation metric and reducing the batch size from 32 to 4 due to GPU memory constraints. Otherwise, default Huggingface training arguments for Wav2Vec2CTC models were used. The model trained for 4 epochs, which took ≈ 4 hours on a single laptop's NVIDIA GTX 1650 GPU. The model iteration with lowest CER on the dev set was chosen as the best finetuned model.

After finetuning, we looked at the overall improvement in CER on the 12-minute test set. Leaving one source dataset out of the train and dev sets entirely allowed us to observe the degree of overfitting to the small train set. We then looked at the types of errors that were most common before and after finetuning.

3 Results

3.1 Original Model Evaluation

After evaluating the predictions of the original model and adapter, the mean CER on the full dataset was 15.6%. See Figure 1a for a breakdown of CER by speaker and source dataset.

As expected, the model performs worst on the "Spontaneous Story" source dataset (mean CER: 18.9%), but it doesn't do much better on the "Read Phrases" source dataset (mean CER: 14.8%). This could be because the speaker from the "Spontaneous Story" appeared in the original model's training dataset (the New Testament recordings) while none of the speakers in the "Read Phrases" set did. The model performs extremely well on the "Read Fable" source dataset (mean CER: 1.8%), even for the speaker whose voice did not appear in the Bible data. This impressive performance could possibly be due to the similarity in domains of Aesop's fable—a formal and antiquated story translated from Ancient Greek—and the New Testament text itself, which was translated from an English version. The format of the "Read Fable" also contrasts from the "Read Phrases" in that each audio file corresponds to a reasonably long sentence rather than a short phrase or one-word command.

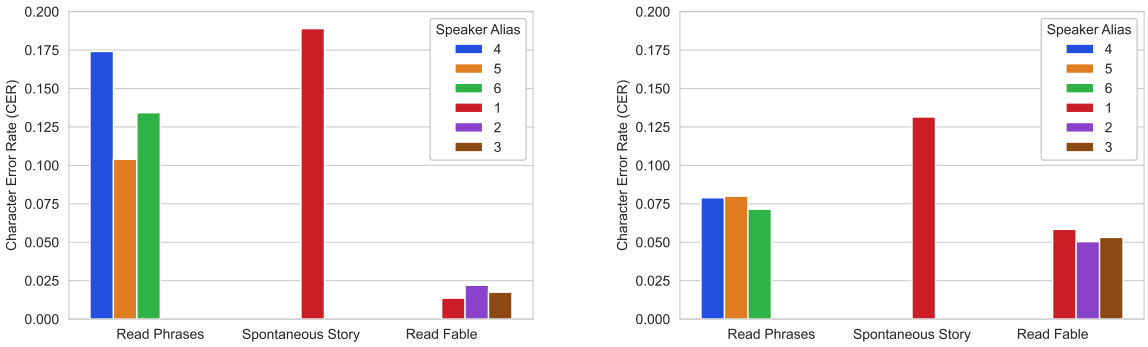
3.2 Finetuned Model Evaluation

After comparing the predictions of the original model on the smaller test set to those generated by the finetuned model, we see a 5.5-point reduction in mean CER on the finetuned transcripts, moving from 14.3% down to 8.8%. See Figure 1b for a visual representation of those results.

Following these evaluations, a new question of interest was identified. Since spontaneous speech

Split	Datasets	Duration	# Files	# Sentences	Speakers
Train	[Read Phrases, Spontaneous]	27 min	408	196	[4, 5, 6, 1]
Dev	[Read Phrases, Spontaneous]	3 min	54	27	[4, 5, 6, 1]
Test	[Read Phrases, Spontaneous, Read Fable]	12 min	151	68	[4, 5, 6, 1, 2, 3]

Table 2: Breakdown of data used in finetuning. For each of the training, development, and testing sets, we show total duration in minutes, number of sound files, number of unique sentences, and speakers included.



(a) Original MMS Akuzipik model on full dataset

(b) Finetuned Akuzipik model on test dataset

Figure 1: Mean CER for original model and finetuned model by dataset and speaker. Note that the voices of both Speakers 1 and 3 were included in the New Testament data used to train the MMS base model and Akuzipik adapter. In (a), observe that only Speaker 1 appears in more than one dataset, with drastically differing CER values between the two. In (b), it is clear that while the finetuned model does much better on the read phrases and spontaneous story datasets, it does worse on the read fable dataset, which was left out of the finetuning data, likely indicating overfitting to the training set.

data is much harder to transcribe, and therefore less populous in the datasets we've collected so far, it would be beneficial to know if finetuning solely on read speech can yield similar improvements on spontaneous speech. We performed a brief post-hoc experiment to test this by again finetuning the original adapter after reassigning the three source datasets into different train, dev, and test sets, this time including only the read speech data in the train set, and splitting the spontaneous speech alone into dev and test—shown in Table 3.

The results of this post-hoc experiment are that CER dropped slightly from 20.6% to 17.3% on the purely-spontaneous test set. This may indicate that model performance on spontaneous speech (the target domain/format) can still be improved when finetuning with only read speech (the easier format to collect as labeled data). See Figure 2 for a visualization of all finetuning experiments and CER improvement.

The pattern of improvement for the datasets included and deterioration for the one not included in the finetuning data likely indicates overfitting to the small (27 minute) training set. That said, the worsened CER of the "Read Fable" set is still objectively low at $\approx 5\%$, and since we are most interested in improving the model's performance on spontaneous speech, this CER increase may not necessarily be of great concern. It does serve as a sanity check for very extreme overfitting. One advantage of finetuning only the adapter model is that catastrophic forgetting—one potential consequence related to overfitting in which the model essentially "forgets" what it learned during earlier training—is unlikely given the frozen weights of the pre-trained base model (Fazel et al., 2021; Eeckt and hamme, 2023).

3.3 Error Analysis on Original Test Dataset

The following error analysis was performed on the test dataset from the first finetuning experiment, since it was judged to better represent the variety of data included in all three source datasets, whereas the test dataset from the second finetuning experi-

	Datasets	Duration	# Files	# Sentences	Speakers
Train	[Read Phrases, Read Fable]	30 min	418	96	[4, 5, 6, 1, 2, 3]
Dev	[Spontaneous Story]	3 min	42	42	[1]
Test	[Spontaneous Story]	9.5 min	153	153	[1]

Table 3: Breakdown of train, dev, and test sets in post-hoc read-speech-only finetuning experiment. Note that overall proportions and values of duration between train, dev, and test sets are kept as similar as possible to the previous finetuning experiment. This allowed us to use the same hyperparameters and training arguments as before.

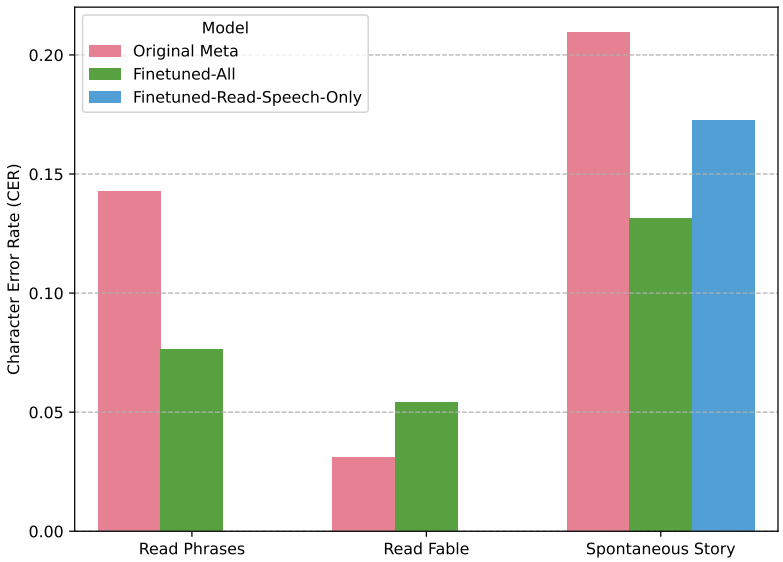


Figure 2: Mean CER by model. The "Original Meta" model is the off-the-shelf MMS model with Akuzipik adapter. "Finetuned-All" represents the model and adapter finetuned on subsets of the read phrases and spontaneous story data. "Finetuned-Read-Speech-Only" represents the model and adapter finetuned only on the read phrases and read fable datasets. Recall that this model was trained on the entirety of those two datasets and can therefore only be evaluated on the spontaneous story. Note that the Finetuned-All model shows improvement in CER for the read phrases and spontaneous story datasets, which were included in the finetuning data. It also shows worsening on the read fable dataset, which was excluded from the finetuning data. The Finetuned-All model shows good improvement on the spontaneous story data, while the Finetuned-Read-Speech-Only model shows moderate improvement.

ment included only one speaker and source dataset. We first look at the types of errors that were most frequent for both the original and finetuned models.

Note that Akuzipik has bilabial, apical, velar, and uvular voiceless stops, as well as a wide variety of voiced, voiceless, rounded, and unrounded fricatives and nasals. Also note that the single characters /a/, /i/, and /u/ indicate a short vowel, whereas a long vowel (/aa/, /ii/, or /uu/) is represented by two characters. Akuzipik also has a central vowel represented as /e/, but it only ever appears in the short form. See Chen (2023) for more on the phonological inventory of Akuzipik.

By far, the most common type of error we observed was a mistranscription of vowel length. In our test dataset, 77.5% of the model's transcriptions included at least one error related to vowel length (the vowel predicted was short when it should have been long, or vice versa). For the finetuned model, this percentage dropped slightly to 69.5% of the test dataset. See Table 4 for a real example of such an error.

In their evaluation of the MMS model on Mocho'—a Mayan language with a vowel length distinction—Liang and Levow (2025) also observed that vowel length was particularly difficult for the model to capture. It could be that the architecture of the ASR model itself is not optimal

for capturing supra-segmental features like vowel length. However, variation in both pronunciation and spelling of long vowels in Akuzipik may be the bigger culprit.

While Hunt et al. (2020) show that orthographically long vowels in Akuzipik are, in general, phonetically longer in duration, native speakers do not always agree on vowel length as it is represented orthographically, resulting in frequent spelling variation. This could contribute to the model's difficulty in transcribing vowel length correctly, if gold-standard transcriptions of the same sound are inconsistent. Due to frequent spelling variation, speakers may also consider these kinds of errors relatively minor or easy to correct, though more research is to confirm this.

The next most frequent error made by either model was an insertion, deletion, or substitution of a vowel not related to a vowel length error. 34.4% of transcriptions in the test set produced by the original model contained at least one vowel error unrelated to length, compared to 27.2% produced by the finetuned model. See Table 5 for an example.

Another kind of error commonly observed involved a word boundary issue. Specifically, the model inserted or deleted a space where it should not have, either breaking up a word into two or more, or combining two or more words into one. A word boundary error was present in 20.5% of the test set as transcribed by the original model and in 15.9% of the test set as transcribed by the finetuned model. See Table 6 for an example.

Notably, errors involving consonants (missing, extra, or incorrect consonants) were relatively uncommon for both models. Only 13.2% of transcriptions from the original model and 8.6% from the finetuned model contained consonant deletions, insertions, or substitutions. Table 7 shows an example.

4 Discussion

The improvement in CER observed after finetuning on less than 30 minutes of training data indicates that even small amounts of data can make a big difference when adapting to a new domain. An overall average CER of less than 10% means that more than 90% of characters are transcribed correctly, which is promising, and may be "good enough" to speed up the process of human transcription, but this has yet to be confirmed.

We calculate the average WER after finetuning

Model	Transcript	Gloss; CER
Gold Standard	tagituguuq	'It was foggy'
Original Model	tagitu**u**q	0.10
Finetuned Model	tagitu**u**q	0.10

Table 4: Example of an error in vowel length. For each of the following tables, all incorrect characters are shown in red, and the incorrect character of interest is underlined in red.

Model	Transcript	Gloss; CER
Gold Standard	qelaneghllaak	'I was so anxious to get there'
Original Model	q**a**la**a**neghll**a**k	0.23
Finetuned Model	q**a**laneghll**a**k	0.15

Table 5: Example of an incorrect vowel. Model confusion between vowels /a/ and /e/ was unsurprisingly common, as several Akuzipik words display spelling variations with /e/ or /a/ substituted for one another.

Model	Transcript	Gloss; CER
Gold Standard	ligikamken	'I understand you'
Original Model	li**i**gi_kamken	0.20
Finetuned Model	ligikamken	0.00

Table 6: Example of an incorrect word boundary—the unnecessary space is underlined in red.

Model	Transcript	Gloss; CER
Gold Standard	utaqiigi	'wait'
Original Model	uta**a**qii**v**i	0.25
Finetuned Model	utaq**i**gi	0.12

Table 7: Example of an incorrect consonant. Substitution of /v/ for /g/ by the original model is less unusual than it might appear, as /g/ is pronounced as a fricative in Akuzipik. Therefore, /v/ and /g/ differ only in place of articulation, not manner.

to be slightly under 60%, and according to results from Gaur et al. (2016), a model with that high of a WER may not actually speed up the process of human transcription through manual correction. On the other hand, results from Ma et al. (2024) suggest that a model with 60% WER would still be viable for speeding up transcription. Clearly, further research is needed to understand if a model with low CER but high WER can speed up transcription, and to understand how this may differ when the transcribers and correctors are native speakers or non-native linguists.

We also see that read speech alone can be used to improve the model's performance on spontaneous speech, though not quite as dramatically. While including spontaneous speech in training/finetuning data is preferred, this means that read speech is still likely to be beneficial, which is advantageous, since read speech is much easier to transcribe and collect as a kind of pre-labeled data. This could be useful for further improving performance on the oral stories of current relevance to language documentation efforts. However, as Evain et al. (2024) discuss, other kinds of spontaneous speech such as multi-speaker conversations are likely to elicit much higher error rates, so further finetuning/adaptation would be necessary for future application to that kind of domain.

The small increase in CER on the "Read Fable" dataset held out from training confirms that some overfitting is likely when finetuning on such a small set of data. While finetuning with more data is usually better, we should prioritize collecting data from domains that are most important for our use case. We should also incorporate as much variety as possible in terms of speakers, recording devices, environments, speech registers, code-mixing and borrowings, etc. if we wish to perform well in a variety of scenarios. Data augmentation methods can be employed to add artificial noise to data, making a model more robust to these kinds of variations, but incorporating "real-life" noise is likely to be most effective (Lakshminarayanan and Prud'hommeaux, 2024).

As mentioned, including more data in the finetuning set is likely to improve the model further. While several hours of Akuzipik audio exist, the vast majority are either recordings of full elicitation sessions in a mix of English and Akuzipik that need significant preprocessing, or spontaneous speech that has yet to be transcribed. Speaker diarization and forced alignment tools may be useful for processing read speech data, but it may also be possible to noisily transcribe some of the existing spontaneous speech data with the current finetuned model. For further iteration, those "noisy" predicted transcriptions could be fed into the finetuning set to improve the model.

In future experiments, we see potential for the existing Akuzipik spell-checker and morphological parser/dictionary to detect and correct some ASR errors before passing off the transcripts for human correction. The orthographic spell-checker (Schwartz and Chen, 2017) detects "impossible" character sequences in Akuzipik text. For instance, two vowels of different quality appearing in sequence (e.g. "ia") would be flagged as it does not occur in Akuzipik orthography. Though it does not rely on a lexicon of valid Akuzipik words and is therefore limited in the kinds of errors it can detect, this tool has previously been useful in automatically detecting optical character recognition (OCR) errors during text digitization, so it has potential to do the same for ASR transcripts.

The parser (Schwartz et al., 2019; Chen et al., 2020) performs a harsher check than the spell-checker, since it will flag any word that does not successfully parse to a known base form plus derivational and inflectional morphemes. The parser and dictionary are still undergoing development to add more spelling variations, base vocabulary items, and more complete sets of possible morpheme combinations. In the future, these tools could be adapted to detect and even correct common ASR transcription errors, such as those related to vowel length.

5 Conclusion

In this paper, we conclude that Meta's MMS ASR model for Akuzipik, after finetuning on an in-domain dataset, is likely to be helpful towards language documentation efforts. An average CER of 13% for monologic spontaneous speech (the target domain) suggests that a task like the transcription of elders' oral stories could be sped up by the use of the finetuned ASR model as a first-pass, though additional research is needed to confirm this. Finetuning has promising results for improving model performance on out-of-domain and spontaneous speech, in particular. Error analysis provides insight into what model outputs are most likely to be incorrect. In this case, most errors were related to vowel length, which Akuzipik speakers may con-

sider minor. While there is much left to research and attempt, this initial endeavor serves as a good starting point for improving speech technology for language documentation and revitalization in the Akuzipik-speaking community and beyond.

Limitations

The analyses presented in this paper have many limitations. One large limitation was the amount of data collected. While it proved large enough to improve the existing model through finetuning, some overfitting was observed, and the test dataset was not large enough to break down each variable of interest. A larger test dataset could allow for statistical modeling of the effects of each variable—speaker identity, speech type, audio recording device, etc.—on CER. That we were only able to include spontaneous speech from one speaker is a significant limitation that should be addressed in future iterations of the project.

Though very few ASR researchers report morpheme error rate (MER), the authors wanted to explore the appropriateness of that metric for Akuzipik as one that may be more easily comparable to WER, the dominant metric in ASR research. Unfortunately, proper morphological segmentation of the data wasn't feasible in the time frame. Existing interlinear glosses of Akuzipik sentences have been compiled by Chen (2023), but the current parser tool produces morpheme sequences that do not correspond to actual surface forms, partially due to the phonological changes associated with suffixation. Work on the parser to produce surface-form segmentations is underway and should eventually permit us to calculate MER as an alternative ASR error metric.

The MMS ASR pipeline contains a language model as the final layer of the neural network model itself. The language model provides statistical reasoning to favor certain character sequences over others, based on the text it was trained on—this text usually consists of the gold-standard transcriptions used to train the ASR model. In this paper, we were unfortunately unable to probe or otherwise explicitly adjust the language model associated with the Akuzipik adapter. However, with more time, it may be possible to tweak or replace the language model, which could significantly improve performance of the ASR model without having to collect and label more audio data. Specifically, language models may be able to predict orthographic patterns that are not necessarily highly salient in the acoustic signal of speech—perhaps this would reduce vowel length errors in the Akuzipik model.

Finally, we should note that a similar experiment on a language which was not included in the MMS dataset and which has no pre-trained adapter would likely not be nearly as successful. The relatively low CER observed in this evaluation is sure to be explained—at least partially—by the inclusion of 33 hours of Akuzipik speech in the original MMS dataset and adapter training data.

Ethical considerations

Meta trained the MMS model used in this paper on data that is technically "publicly available", but this does not assuage ethical concerns regarding violated Indigenous data sovereignty. Similar data scraping and ASR model training processes have been condemned by Indigenous researchers and community members (Keoni Mahelona et al., 2023). As Pine et al. (2025) point out, in conjunction with their ASR model and adapters, Meta trained text-to-speech (TTS) systems which model the likenesses of the speakers in the Bible recordings training material without obtaining permission from those speakers (or the publishers of that material). Pine et al. (2025) discuss in depth the significant ethical consideration required when training a TTS model in an Indigenous language community. Language in these communities often has a very high degree of cultural and traditional significance. This could mean that the idea of a computer-generated "speaker" of that language may be upsetting or unacceptable. The potential for generation of disrespectful or otherwise uncharacteristic speech may be especially problematic.

Though Geng et al. (2025) show promising results in augmenting ASR training data with ethically-developed TTS systems for Indigenous languages, a subjective evaluation in the Akuzipik-speaking community of TTS as a concept and in relation to the pre-existing model trained by Meta will be necessary before it is appropriate to proceed in the use or evaluation of TTS systems for Akuzipik. For this project, all ASR models were used and trained locally, which mitigates immediate concerns regarding language data security. In time, community members may also decide to distance themselves from all of Meta's speech technology in favor of independently and ethically developed software.

Acknowledgments

We thank the community on St. Lawrence Island, Alaska for their collaboration and support of this research. We give special thanks to the Akuzipik speakers who lent their time, voices, and linguistic expertise to this particular project, including Apa John Apangalook, Petuwak Christopher Koonooka, Amaghalek Beulah Nowpakahok, and others who prefer to remain anonymous.

References

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. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech 2022*, pages 2278–2282. ISCA.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Steven Bird. 2021. Sparse Transcription. *Computational Linguistics*, 46(4):713–744.

Emily Chen. 2023. *Modeling Saint Lawrence Island Yupik morphology to support revitalization*. Thesis, University of Illinois at Urbana-Champaign.

Emily Chen, Hyunji Hayley Park, and Lane Schwartz. 2020. Improved Finite-State Morphological Analysis for St. Lawrence Island Yupik Using Paradigm Function Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2676–2684, Marseille, France. European Language Resources Association.

Steven Vander Eeckt and Hugo Van hamme. 2023. Using Adapters to Overcome Catastrophic Forgetting in End-to-End Automatic Speech Recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ArXiv:2203.16082 [eess].

Solene Virginie Evain, Solange Rossato, and François Portet. 2024. Unraveling Spontaneous Speech Dimensions for Cross-Corpus ASR System Evaluation for French. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17165–17175, Torino, Italia. ELRA and ICCL.

Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. 2021. SynthASR: Unlocking Synthetic Data for Speech Recognition. In *Interspeech 2021*, pages 896–900, ISCA. ISCA.

Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*, W4A '16, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Mengzhe Geng, Patrick Littell, Aidan Pine, Penáć, Marc Tessier, and Roland Kuhn. 2025. Supporting SENĆOTĒN language documentation efforts with automatic speech recognition. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 29–39, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Benjamin Hunt, Harim Kwon, and Sylvia Schreiner. 2020. An acoustic analysis of St. Lawrence Island Yupik vowels. *The Journal of the Acoustical Society of America*, 148:2471–2471.

Benjamin Hunt, Lane Schwartz, Sylvia Schreiner, and Emily Chen. 2023. Community consultation and the development of an online Akuzipik-English dictionary. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–143, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Thennal D K, Jesin James, Deepa Padmini Gopinath, and Muhammed Ashraf K. 2025. Advocating Character Error Rate for Multilingual ASR Evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4926–4935, Albuquerque, New Mexico. Association for Computational Linguistics.

Keoni Mahelona, Gianna Leoni, Suzanne Duncan, and Miles Thompson. 2023. OpenAI's Whisper is another case study in Colonisation.

Christopher Petuwaq Koonooka, Sylvia L. R. Schreiner, Giulia Masella Soldati, Lane Schwartz, Benjamin Hunt, Preston Haas, Emily Chen, and Hyunji Hayley Park. 2021. Akuzipik/Yupik (St. Lawrence Island, Alaska, USA; Chukotka, Russia) - Language Snapshot. *Language Documentation and Description*, 20(0). Number: 0.

Vigneshwar Lakshminarayanan and Emily Prud'hommeaux. 2024. Exploring the impact of noise in low-resource ASR for Tamil. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 30–34, St. Julian's, Malta. Association for Computational Linguistics.

Eric Le Ferrand, Zoey Liu, Antti Arppe, and Emily Prud'hommeaux. 2024. Are modern neural ASR architectures robust for polysynthetic languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2953–2963, Miami, Florida, USA. 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. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37, Vienna, Austria. Association for Computational Linguistics.

Marcus Ma, Lelia Glass, and James Stanford. 2024. Introducing Bed Word: a new automated speech recognition tool for sociolinguistic interview transcription. *Linguistics Vanguard*, 10(1):641–653.

Julia Mainzinger. 2024. Fine-tuning ASR Models for Very Low-Resource Languages: A Study on Mvskoke.

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui. 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2):171–184.

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.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

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*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Sylvia L. R. Schreiner, Benjamin Hunt, Emily Chen, Preston Haas, and Ukaall Crystal Aningayou. 2022. Semantic fieldwork from a distance with speakers of Akuzipik. *Semantic fieldwork methods*, 4(2).

Lane Schwartz and Emily Chen. 2017. Liinnaqumalghiit: A web-based tool for addressing orthographic transparency in St. Lawrence Island/Central Siberian Yupik. *Language Documentation*, 11.

Lane Schwartz, Emily Chen, Benjamin Hunt, and Sylvia L.R. Schreiner. 2019. Bootstrapping a Neural Morphological Analyzer for St. Lawrence Island Yupik from a Finite-State Transducer. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 87–96, Honolulu. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yolóxochitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

Benjamin V. Tucker and Yoichi Mukai. 2023. *Spontaneous Speech*. Elements in Phonetics. Cambridge University Press, Cambridge.

Patrick Von Platen. 2023. Fine-Tune MMS Adapter Models for low-resource ASR.