Fine-Tuning ASR models for Very Low-Resource Languages: A Study on Mvskoke

Julia Mainzinger and **Gina-Anne Levow**
University of Washington
jmainz,levow@uw.edu

Abstract

Recent advancements in multilingual models for automatic speech recognition (ASR) have been able to achieve a high accuracy for languages with extremely limited resources. This study examines ASR modeling for the Mvskoke language, an indigenous language of America. The parameter efficiency of adapter training is contrasted with training entire models, and it is demonstrated how performance varies with different amounts of data. Additionally, the models are evaluated with trigram language model decoding, and the outputs are compared across different types of speech recordings. Results show that training an adapter is both parameter efficient and gives higher accuracy for a relatively small amount of data.

1 Introduction

Endangered languages are often overlooked in research on speech technology and other NLP applications. Research obstacles include data scarcity and the effort it takes to collect new data, as well as funding and a perceived limited impact on small speech communities. However, these technologies can be hugely beneficial to assisting community-led language revitalization efforts and are worthy of the effort it takes, if it is done with consideration and care for the speech community.

Automatic Speech Recognition (ASR) technology can help speed up transcription and documentation work, as well as be a stepping stone to other applications such as spoken term detection, which can help in identifying certain topics or key information contained in recordings. Other useful applications for the speech community are speech-to-text input and automatic subtitling. These applications can be helpful in encouraging use of the language and promoting language education.

ASR is a relatively mature technology when applied to high-resource languages (Baevski et al., 2020). But it is only more recent advancements such as model size and multilinguality that have enabled comparable accuracy for resource-constrained situations (Pratap et al., 2023). This work focuses on the evaluation and analysis of two highly multilingual speech models when trained for Mvskoke, a language indigenous to the southeastern United States (Martin and Mauldin, 2000).

1.1 The Mvskoke Language

The Mvskoke language is spoken by members of the Muscogee (Creek) Nation and Seminole Nation in Oklahoma, and members of the Seminole tribe of Florida. It is estimated that less than 300 first-language speakers remain, and nearly all are over the age of 60[1]. Recent years have seen an interest among tribal members to revitalize the language, which has led to several new initiatives such as a Master-Apprentice Program at the College of the Muskogee Nation, and new educational and preservation resources being created and collected by the Language Program at the Muscogee Creek Nation tribal government. ASR can assist in some of these efforts.

The language is synthetic and agglutinative, with a traditional orthography of 20 latin letters (Martin, 2011; Frye, 2020). The orthography is relatively transparent and allows for spelling variations. The advantage of a transparent orthography is that transcriptions can remain relatively close to the speech signal. The disadvantage is that the error rates can appear higher since spelling may vary between model predictions and reference transcriptions.

1.2 ASR for Low-Resource Languages

HMM-based and E2E can achieve usable results on very low resource languages, without large pre-trained multilingual models. An ASR system for

[1]This estimate is from personal communication with a member of Ekvn-Yefolecv, a community of Mvskoke people.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 76–82

Yoloxóchitl Mixtec compares HMM and end-to-end (E2E) encoder/decoder and finds E2E performed best, with a WER of 16.0%. This model has been incorporated into documentation workflow (Shi et al., 2021; Amith et al., 2021). Jimerson et al. (2023) show that an HMM-neural hybrid trained from scratch can outperform pre-trained neural networks for some languages, but is worse for others. This shows that there is no clear choice for system architecture, and that choice of architecture may in fact be dependent on the features of the language.

1.3 Fine-tuning Pre-trained Models

Fine-tuning a pre-trained model is a common approach for low-resource settings. An ASR model for Cherokee using a fine-tuned XLSR-53 has a WER of 64% (Zhang et al., 2022). A fully-convolutional neural network (CNN) for Seneca sees improvement from transfer learning from English (Thai et al., 2020). In their paper on endangered languages of Nepal, Meelen et al. (2024) demonstrates an effective ASR pipeline using XLSR-53 and shows the relationship between dataset size and model performance. For the current work, we choose to fine-tune multilingual transformer models due to the ease of implementation (Pratap et al., 2023).

1.4 Adapters

Houlsby et al. (2019) introduced adapter modules, which allow fine-tuning pretrained models by adding only a few trainable parameters per task rather than training all of the existing parameters. The recent Massive Multilingual Speech (MMS) models include adapters that are trainable for certain tasks such as ASR, and have been shown to be more memory efficient and yield better performance for low-resource languages (Pratap et al., 2023).

1.5 Language Model Decoding

Utilizing a language model (LM) can be helpful because often text data can be more easily gathered than audio data. This is true in the case of Mvskoke. (Jimerson et al., 2023) demonstrate that using a language model always increases accuracy, but the gains are minimal in comparison with other factors such as model architecture. On the other hand, Orken et al. (2020) show that ASR for two agglutinative languages, Turkish and Tatar, see a marked improvement from use of a language model. In this work, we investigate the performance of the multilingual models with and without LM decoding.

2 Data

The texts and recordings used in these experiments primarily come from language documentation work conducted over the last few decades. Two documentation books, by Haas et al. (2015) and Gouge et al. (2004) are collections of stories, historical letters, and other cultural documents. A portion of these texts were recorded in a studio setting by two female speakers. In order to incorporate male speakers and spontaneous speech, a small segment of the New Testament was selected, as well as a few short sections of recorded interviews.

Splits. Train and development sets are split 90/10 at run-time. Two evaluation sets are kept separate from the training set. "Eval (clean)" is read speech from the same documentation sources as the training set, and "eval (other)" is noisier speech, consisting of one overlapping male speaker and one held-out female speaker. In the transcripts for all the audio data, there are a total of 6,840 utterances and 19,154 words, for an average of 2.8 words per utterance. The train and "eval (other)" sets include both read and spontaneous speech, while the "eval (clean)" set is only read speech. Other features of the datasets are shown in Table 1.

Language Model. The text data for the language model (LM) includes the two books above as well as the transcriptions from a series of interviews conducted by the Pumvhakv School in 2015. For these experiments, the interview recordings are not used for training due to noise including nature sounds, speech errors, and singing, but the transcriptions provide valuable vocabulary. The texts and transcriptions of the evaluation set were excluded from the text training data. The text corpus used for LM training has 118,021 words and 27,795 unique words.

At this time, the dataset will not be publicly released due to copyright constraints of the source material. Currently, the Muscogee (Creek) Nation is working to consolidate data and establish language resource policies. However, much of the source of the data can be viewed on the Muskogee Documentation Project website [2].

[2] https://muskogee.pages.wm.edu/

	train+dev	eval clean	eval other
Total Length	4.1h	21m	27.6m
Avg. Length	2.6s	2.5s	2.5s
F Speakers	2	2	1
M Speakers	2	0	1

Table 1: Prepared audio datasets. Train and development sets are split 90/10 at run-time, and the evaluation sets are held out for testing. Evaluation sets are partitioned into clean and noisy speech.

3 Methodology

The goal of this work is to evaluate the effectiveness of fine-tuning an adapter for a large multilingual model. This is one state-of-the-art path for ASR that requires less manual work than other methods such as an HMM, and generally requires less data due to the existing pre-trained acoustic knowledge of the multilingual models. Additionally, other aspects that are evaluated are how much data is required and whether or not a language model can improve results.

3.1 Models

This study evaluates models introduced by Meta's Massively Multilingual Speech (MMS) project (Pratap et al., 2023). MMS models are speech representation models with a wav2vec2.0 architecture that are pre-trained on unlabeled data from 1,406 languages (Baevski et al., 2020; Pratap et al., 2023). The base models are available in 300 million and 1 billion parameter versions. Of particular interest in this study is the MMS-1B-l1107, a model that was fine-tuned for ASR from the MMS-1B base model (Pratap et al., 2023). This model features an adapter with 2 million parameters on top of the base 1 billion parameters, based off of a method introduced by Houlsby et al. (2019). The adapter layers allow the large multilingual acoustic knowledge to be fine-tuned for a new language in a computationally efficient way.

In order to evaluate MMS in comparison with its predecessors, we also train XLSR-53, a popular choice for low-resource ASR. XLSR-53 has the same wav2vec2.0 architecture and is pre-trained on 53 languages with 300 million parameters (Conneau et al., 2020). In order to compare a similarly-sized MMS model, we also train MMS-300M (Pratap et al., 2023). MMS-1B is not included for this experiment due to memory constraints of the hardware used.

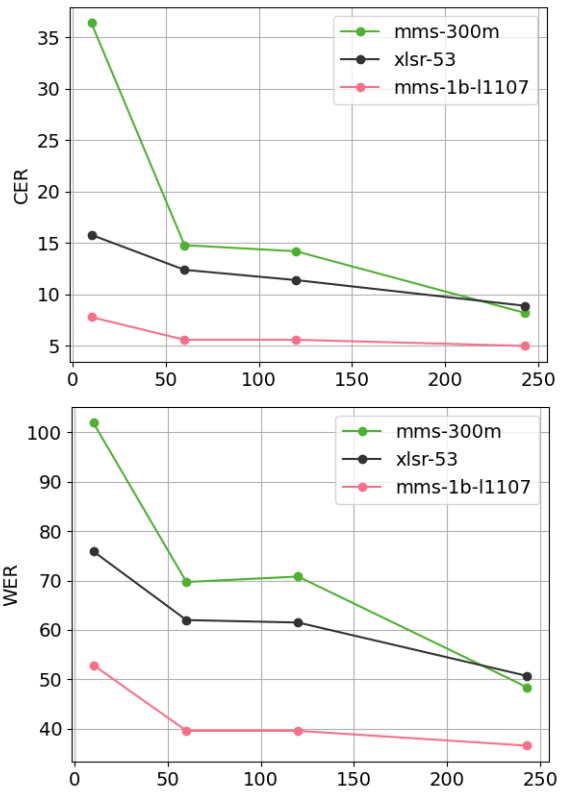

Figure 1: Word error rate and character error rate for each model given the length of training data in minutes.

MMS-1B-l1107 was chosen over MMS-1B-all based off of a simple empirical test in which the former performed better, the details of which can be found in Appendix A.

3.2 Implementation

Implementation follows the steps detailed by Patrick von Platen to fine-tune the MMS adapter using Huggingface Transformers[3] (Wolf et al., 2019). For MMS-1B-l1107, the base model is frozen and only the adapter layer is trained. For the other two models, the entire model weights are trained. The data is split into sets of 10, 60, 120, and 243 minutes. Early stopping criteria ends training before overfitting. More hyperparameters are detailed in Appendix A. The best model is saved with the lowest character error rate (CER), and then evaluated on the clean and noisy evaluation sets.

The language model is a trigram model trained with KenLM (Heafield, 2011). This LM is then used in a CTC decoder after the models are trained.

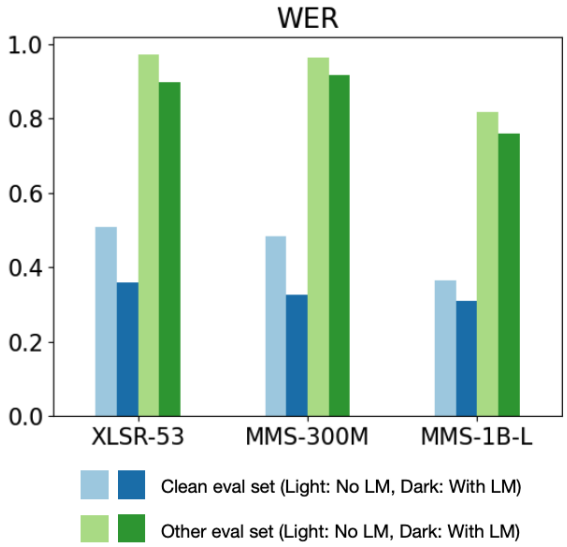

Figure 2: Word error rate on evaluation sets when decoding with a trigram language model, for each model trained on 243 minutes of audio data. MMS-1B-L is the MMS-1B-L1107 model.

4 Results

The MMS-1B-l1107 performed best overall, with best results of 37% word error rate (WER) and 5% character error rate (CER). Results are shown in Figure 1.

Data size effects. Interestingly, the XLSR-53 performed better than the MMS-300M on smaller amounts of data. However, more data (4 hours) improves the MMS-300M to a point that surpasses XLSR-53. The reason for this is unclear. One explanation could be due to the fact that the former trained longer. Early stopping criteria ended training around 10-13 epochs for all models except the MMS-300M at 243 minutes, which took longer to converge and trained for 23 epochs. Further experimentation is needed to determine if this trend continues to hold for more data. Table 2 shows resulting error rates for each model.

MMS vs XLSR. Other papers have shown that XLSR-53 outperforms MMS in some situations, such as Uralic languages and Arabic, both of which have tens of thousands of hours of training data available (Mihajlik et al., 2023; Younis and Mohammad, 2023). Mvskoke on the other hand only has a few hours of data, possibly making MMS the better candidate. This is consistent with the findings of the original authors of MMS, that higher-resource languages show some degradation in MMS compared with previous models that cover fewer languages, but that most extremely low-resource languages benefit from the large amount of languages represented in MMS (Pratap et al., 2023).

The advantage that MMS-1B-l1107 presents is that it has been fine-tuned specifically for the task of ASR. Adding a new language-specific adapter for Mvskoke also means that only a small number of parameters need to be trained. Ultimately, fine-tuning the adapter only for the MMS-1B-l1107 is both more memory efficient and gives better performance.

Model	WER		CER	
	120	243	120	243
XLSR-53	62	51	11	9
XLSR-53 + LM	40	36	10	7
MMS-300M	71	48	14	8
MMS-300M + LM	43	33	10	6
MMS-1B-L	40	37	6	**5**
MMS-1B-L + LM	**34**	**31**	**5**	**5**

Table 2: Error rate percentages for different models with different data amounts in minutes, compared with language model (LM) decoding, on the eval (clean) set. MMS-1B-L is the MMS-1B-L1107 model.

LM Decoding. Language model (LM) decoding improves all of the models by several percentage points. The performance improvement is less for the better models, but even the best model (MMS-1B-l1107) improves slightly in WER. Figure 2 shows the decrease in error rate for each model with the LM. However, in the best model, the CER is not improved. Sometimes the language model breaks apart long out-of-vocabulary words into more common words, which degrades the transcription. For example, "vcvkvhoyvte hvmkat" ("one of the ones who had followed") is transcribed as "vcakkvhoyvte hvmkat" without an LM, which is phonetically similar, but is changed to "vcakv oketv hvmkat" by the LM, which is nonsensical. So although the WER goes down overall for the whole evaluation, some information may be lost. This may be dis-preferred for some applications such as spoken term detection (Le Ferrand et al., 2021). More example outputs are shown in Appendix B.

5 Conclusion and Future Work

This study shows that fine-tuning multilingual transformer models is an effective method for train-

[3] https://huggingface.co/blog/mms_adapters

ing ASR systems in low-resource language contexts. Fine-tuning the adapter for a 1 billion parameter model, MMS-1B-l1107, yields better results when compared to training entire models such as XSLR-53 and MMS-300M. However, the performance of such systems depends highly on the recording quality and type of speech. Although language modeling improves overall accuracy measures such as WER and CER, it can also degrade the output in some cases. Alternatives like subword or character-level modeling could offer a more effective approach, particularly for applications where fidelity to the original speech signal is preferred.

A future direction would be to incorporate the ASR model into a keyword-spotting or sparse transcription system. The high error rates for noisy recordings in this study mean that manual transcription may still be faster than correcting ASR output. Sparse transcription can be helpful in situations where high ASR error rates lead to low-quality transcriptions (Bird, 2021). Transcribing only high-confidence words can be useful for indexing recordings and providing an overview of recorded content that can then be used for knowledge gathering.

6 Limitations

Due to the computational effort, each model was only trained once for each data amount (10, 60, 120, and 243 minutes). The datasets were shuffled randomly at runtime when selecting the splits, for example one 10 minute set is slightly different than another 10 minute set. This creates some variability in the results, and is not as robust as training the models multiple times and taken an average of performance.

This study also does not include the MMS-1B, the adapter-less version of the MMS-1B-l1107, because of the computational requirements of training such a large model. Because of this, conclusions cannot be made about the performance of an adapter model compared to a model with an equal amount of parameters. This study does not seek to fully evaluate adapter architecture, rather only to say that it is an effective method for this setting.

Finally, the transformer architecture was not evaluated alongside other architectures. In low-resource settings, model architecture can affect performance significantly, and no single architecture is best for every language (Jimerson et al., 2023). This study only evaluates the models stated here and their performance on the Mvskoke language.

7 Acknowledgements

We are grateful to the students of the Linguistics Undergraduate Research Apprenticeship Program (LURAP) for their assistance editing recordings - Liyana Alam, Bill Yu, Anika Pontis, Myranda Fraker, Cassie Hu, Luke Eriksen, Isabella Lufschanowski, and Kim Dang. Thank you to Dr. Jack Martin for his collaboration in providing recordings and generously allowing us to build from his decades of work on Mvskoke language documentation. Thank you to the Sam Noble Museum as well as the Language Program at the Muscogee (Creek) Nation for assistance with archiving and accessing recordings. Finally, we are indebted to tribal elders, teachers, and leaders who provided feedback as well as wisdom and insight. Mvto.

References

Jonathan D Amith, Jiatong Shi, and Rey Castillo Garcia. 2021. End-to-end automatic speech recognition: Its impact on the workflow for documenting yoloxóchitl mixtec. In *First Workshop on NLP for Indigenous Languages of the Americas. 11 June 2021. https://www. aclweb. org/anthology/2021. americasnlp-1.8. pdf.*

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition.

Melanie Frye. 2020. Improving mvskoke (creek) language learning outcomes: A frequency-base approach. *Thesis, University of Oklahoma*.

Earnest Gouge, Edited, Translated by Jack B. Martin, and Juanita McGirt. 2004. *Totkv Mocvse / New Fire: Creek Folktales*. Norman: University of Oklahoma Press.

Mary R. Haas, James H. Hill, Jack B. Martin, Margaret McKane Mauldin, and Juanita McGirt. 2015. *Creek (Muskogee) Texts*. University of California Publications.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016, Toronto, Canada. Association for Computational Linguistics.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2021. Phone based keyword spotting for transcribing very low resource languages. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, volume 19 of *Proceedings of the Australasian Language Technology Workshop*, pages 79–86. Australasian Language Technology Association. Publisher Copyright: © ALTA 2021. All rights reserved.; 19th Workshop of the Australasian Language Technology Association, ALTA 2021 ; Conference date: 08-12-2021 Through 10-12-2021.

Jack B. Martin. 2011. *A Grammar of Creek (Muskogee)*. University of Nebraska Press.

Jack B. Martin and Margaret McKane Mauldin. 2000. *A Dictionary of Creek/Muskogee*. University of Nebraska Press.

M Meelen, A O'Neill, and R Coto-Solano. 2024. End-to-end speech recognition for endangered languages of nepal.

Péter Mihajlik, Máté Kádár, Gergely Dobsinszki, Yan Meng, Meng Kedalai, Julian Linke, Tibor Fegyó, and Katalin Mady. 2023. What kind of multi- or cross-lingual pre-training is the most effective for a spontaneous, less-resourced asr task? In *2nd Annual Meeting of the ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL 2023)*.

Mamyrbayev Orken, Keylan Alimhan, Bagashar Zhumazhanov, Tolganay Turdalykyzy, and Farida Gusmanova. 2020. *End-to-End Speech Recognition in Agglutinative Languages*, pages 391–401.

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. 2023. Scaling speech technology to 1,000+ languages.

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.

Bao Thai, Robert Jimerson, Raymond Ptucha, and Emily Prud'hommeaux. 2020. Fully convolutional ASR for less-resourced endangered languages. 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 126–130, Marseille, France. European Language Resources association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Hiba Adreese Younis and Yusra Faisal Mohammad. 2023. Arabic speech recognition based on self supervised learning. In *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*, pages 528–533.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can nlp help revitalize endangered languages? a case study and roadmap for the cherokee language.

A Training Details

Hyperparameters. The implementation for this experiment follows the guide by Patrick von Platen using HuggingFace transformers[4]. Hyperparameters were defined as follows:

- Learning rate = 1e-3
- Maximum epochs = 30
- Best model metric = CER
- Early stopping = 3
- Early stopping threshold = 0.003

Most models stopped training around 10-13 epochs, with the exception of the MMS-300M trained on the full dataset, which took longer to converge and stopped at 23 epochs.

MMS-1B-l1107 vs MMS-1B-all. MMS-1B-l1107 was chosen over MMS-1B-all for a few reasons. Both models are fine-tuned for ASR from the

[4] https://huggingface.co/blog/wav2vec2-with-ngram

base MMS-1B model using labeled data. MMS-1B-l1107 was fine-tuned on the MMS-lab set only, which is a collection of New Testament recordings in 1,107 languages (Pratap et al., 2023). MMS-1B-all includes more data, however the additional data is for a smaller subset of languages, many of which are higher-resourced. This may be detrimental for an extremely low-resource language. This hypothesis was tested somewhat empirically by training the adapters for both MMS-1B-l1107 and MMS-1B-all with 60 minutes of Mvskoke training data, and the MMS-1B-l1107 performed better (decrease of 8% WER and 1% CER on test set). Therefore this work continues with the MMS-1B-l1107 model.

B Example Output

Table 3 shows examples of outputs from the best model, MMS-1B-l1107 trained on the full data set. Example 1 shows output on a female speaker not present in the training data, speaking conversationally. The model misses a word boundary and the LM does not make any changes. However, the transcription is still true to the speech signal. In example 2, the language model (LM) substitutes a common alternative spelling for the same word, resulting in a higher error rate but is still a good transcription. Example 3 shows how the LM can in fact degrade transcription quality, when it attemps to break an out-of-vocabulary word into more common words. In this case, the output without LM decoding makes a closer transcription. The final example, example 4, shows that LMs can improve the transcription on familiar words.

Table 3: Examples of ASR outputs from MMS-1B-l1107.

1.	**Held-out female speaker**		
Eval (other)	*"'Wring its neck,' he told me."*		
Reference	nokfiyvs kihcen cvkihcen	CER	WER
No LM	nokfiyvskihcen cvkihcen	12	67
With LM	nokfiyvskihcen cvkihcen	12	67
2.	**Minor spelling changes**		
Eval (clean)	*"We don't want you. Go back," he was told*		
Reference	ceyacēkot os yefulkvs kihocen	CER	WER
No LM	ceyacēkot os yefulkvs kihocen	0	0
With LM	ceyac**e**kot os yefulkvs kihocen	3	25
3.	**LM degrades transcription**		
Eval (other)	*"one of the ones who had followed"*		
Reference	vcvkvhoyvte hvmkat	CER	WER
No LM	vcakkvhoyvte hvmkat	8	5
With LM	vcakv oketv hvmkat	32	100
4.	**LM improves transcription**		
Eval (other)	*"November"*		
Reference	ohrolopē eholē	CER	WER
No LM	orrolope v ehoflē	38	150
With LM	ohrolopē eholē	0	0