Automatic Speech Recognition for Icelandic

Making use of the cross-lingual potential of Facebook AI‘s wav2vec2-large-xlsr-53 pre-trained model for training Icelandic ASR models

Guðjón Kristjánsson **Study No: 202202801 – GK**

School of Communication and Culture, Aarhus University

Data Science, Prediction and Forecasting Exam, 02/06-2023

Abstract

The purpose of this study was to fine-tune three distinct Automatic Speech Recognition models on the Facebook AI‘s wav2vec2-large-xlsr-53 pre-trained model using three Icelandic datasets. These datasets were Samrómur, Málrómur and Althingis‘ Parliamentary Speeches. The fine-tuned models were evaluated using WER-score, a standard for large vocabulary continuous speech recognition tasks such as these. The hypothesis was that these models would all yield low WER-scores due to their size and quality. The final results varied greatly, with the model fine-tuned using the Samrómur dataset performing significantly better than the other two models. The hypothesis was therefore not supported since only one model turned out to perform well. Further research is encouraged, especially on the Samrómur dataset due how well the model fine-tuned using it did.

Table of contents

Introduction

In today’s world, the exponential evolution of technology is becoming increasingly clearer. New innovative systems keep getting released and have substantially altered the way we go about our daily lives. The sector of computational linguistics has perhaps seen the most dramatic leap in the last few years. One subsector of computational linguistics that has recently become very noticeable is speech recognition.

1.1 Automatic Speech Recognition

Speech recognition systems such as Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) are transformative technologies that have proven very useful to people and businesses. TTS systems convert written text into speech which can greatly enhance the quality of life for people with visual or reading problems. ASR systems, however, accept speech and transform it into text. These systems have also proven to be very beneficial, especially for people with hearing or writing disabilities. Both these systems therefore provide support and additional communication tools for people who need them. Both systems have also proven to be useful and reduce costs in industries where transcriptions or translations are needed (Katyal et al., 2014). As previously mentioned, ASR models are designed to accept speech and transform it into text. This process begins with creating a list of possible texts and then the most probable text is selected. Generally, ASR models have an acoustic front-end that processes the speech and extracts useful features. This extraction process involves many methods such as principal component analysis, linear discriminant analysis, linear predictive coding and dynamic feature extraction, among others. Following the feature extraction, the model moves into the processing phase, where the acoustic model and a language model are used to generate a proposed transformation of the spoken language. While the acoustic model is based on statistical representations of distinct phonetic units of a word, the language model contextualizes the likelihood of words belonging to a certain language. The decoder then tries to find the most probable sequence of words to match the spoken language, usually creating a list of possible transcriptions called the n-best list (Filippidou & Moussiades, 2020).

1.2 Word Error Rate

Different ways can be used to measure the performance of an ASR model but when it comes to large vocabulary continuous speech recognition (LVCSR), the standard evaluation metric used is Word Error Rate (WER). To use the WER for evaluation, the requirements are at least two hours of transcribed test data so that it can yield reliable results (Ali & Renals, 2020). The formula for the WER is defined as:

$$WER = (S + D + I) / N_1 = (S + D + I) / (H + S + D)$$

where S stands for total number of placements, D stands for the total number of deletions, I stands for the total number of entries, N_1 stands for the total number of reference words and finally H stands for the total number of successes (Filippidou & Moussiades, 2020). The WER-score is usually on the scale of 0% to 100%, meaning that if a model gets 0% it got every word right and if it gets 100% it misinterpreted every word. One limitation of WER is however that there is a chance to get a WER-score over 100%. The way this can happen is if the ASR model generates more than one incorrect word for each input word. If you then compare two models and one has a WER-score of 100% and the other has a WER-score of 120%, one might think that the latter model is worse. That is, however, not the case as both models are completely useless. For example, if the first model is given the word "door" and it guesses the word is "more", it is incorrect. If the second model is given the same word and it guesses the word "more", but it also guesses the word "lore", it is also incorrect. The latter model is however not more wrong, they are both equally wrong as neither model guessed the correct word (Morris et al., 2004).

1.3 Facebook AI's wav2vec2-large-xlsr-53 pre-trained model

The wav2vec2-large-xlsr-53 pre-trained model, developed by Facebook AI, is a cross-lingual ASR model. It is an extension of the wav2vec 2.0 model, also developed by Facebook AI, but has been trained on 53 languages. Three datasets were used when training the model which were the CommonVoice dataset, the BABEL dataset and the Multilingual LibriSpeech. These three datasets together consist of over 52 thousand hours of speech data. The wav2vec2-large-xlsr53 model performs very well on all languages and notably well on languages other than English. The similarity of languages used in pre-training and fine-tuning likely plays an important role there. The architecture of the model works in a way where the feature encoder takes raw audio data and transforms it into latent speech representations. It then uses a Transformer to help the model understand the meaning or context behind these latent speech representations. The model then implements quantization which simplifies the representations so they're easier for the model to analyze (Conneau et al., 2020).

1.4 Icelandic language technology

In the last decade, there has been tremendous progress in the development of language technologies. This progress is, however, dependent on each language. The development and advancement of language technology systems is very resource demanding and costly. This can prove especially challenging for smaller countries where their language is spoken by few. Icelandic is an example of this since only about 350.000 people speak Icelandic. The Icelandic government, however, decided in 2017 to take on this challenge and fund a five-year plan to enhance and develop language technologies for Icelandic. The plans' ultimate goal was to make sure that Icelandic will be available on language technology platforms. To achieve this, the focus was on developing resources for text and speech systems and to make them available online for others to use. This plan was set in to motion in October of 2019, after two years of preparation (Nikulásdóttir et al., 2020).

1.5 The Icelandic datasets

For the current project, three Icelandic datasets were used. Those datasets were Málrómur, Samrómur and Althingis' Parliamentary Speeches.

The Málrómur dataset consists of 152 hours of speech data recorded from 563 participants, between 2011-2012. Those 152 hours are then split up into training, validation and test sets. The training set contains 119 hours and 3 minutes of speech, the validation set contains 3 hours and 22 minutes and the test set contains 13 hours and 41 minutes. The speech data was all manually checked by evaluators to determine whether the text included only the utterances the participants had read. The evaluation process included four evaluators and two phases. The datasets included 127.286 recorded segments, of which 108.568 were kept in, the rest was considered not good enough (Steingrímsson et al., 2017).

The Samrómur dataset consists of almost 146 hours of speech data, collected from 16th of October until 22nd of November in 2019. Those 146 hours are then split into training, validation and test sets. The training set contains 114 hours and 34 minutes, the validation set contains 15 hours and 16 minutes and the test set contains 15 hours and 51 minutes. The Samrómur dataset is an ongoing project, aimed at creating the largest speech corpus for Icelandic. The latest version of the dataset was released in March 2020. Due to good marketing efforts and media coverage, the data collection was a lot quicker than previously expected. In total, 45.000 utterances were collected, with almost perfect gender balance. For next release, the data collectors hope to include speech data from children as well and non-native Icelandic speakers (Mollberg et al., 2020).

The Althingis' Parliamentary Speeches dataset consists of 542 hours of speech data from 196 speakers, collected from 2005 until 2016. Those 542 are then split into training, validation and test sets. The training set contains 514 hours and 29 minutes, the validation set contains 14 hours and 2 minutes and the test set contains 13 hours and 52 minutes. The training set contains 192 speakers, that being 104 males and 88 females. The validation and test sets contain only speech data from 2016 which contains 59 speakers, that being 29 males and 30 females. The Althingis' Parliamentary Speeches currently the largest Icelandic dataset meant for fine-tuning ASR systems. Although some challenges were faced when creating this dataset like noisy background for instance, the creators deem this dataset very well suited for fine-tuning speech recognition systems (Helgadóttir et al., 2017).

1.6 The aim of the study

The purpose of the present study is to make use of Facebook's wav2vec2-large-xlsr-53 (Conneau et al., 2020) exceptional cross-lingual capabilities and the quality and size of the Icelandic datasets. Three separate ASR models will be fine-tuned using the wav2vec2-large-xlsr-53 pre-trained model (Conneau et al., 2020) and the previously mentioned Icelandic datasets – Málrómur (Steingrímsson et al., 2017), Samrómur (Mollberg et al., 2020) and Althingis' Parliamentary Speeches (Helgadóttir et al., 2017). Fine-tuning three different models offers the possibility of comparing them to see which performs the best. To be able to evaluate each models' performance and compare the models to each other, the Word Error Rate (WER) will be used. The hypothesis for the present study is that due to the quality and size of the datasets, all models will perform well and yield a low WER-score.

Methods

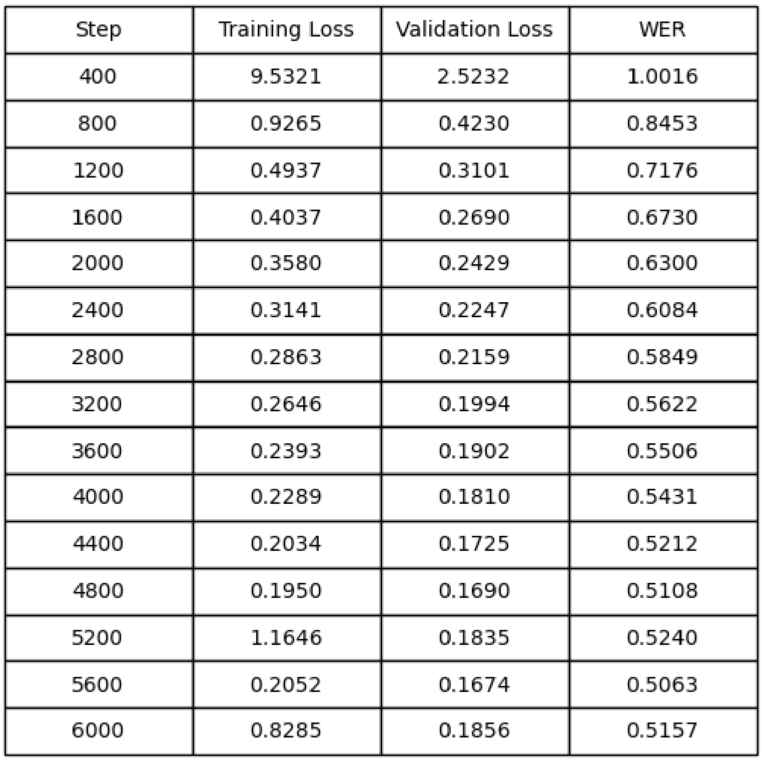
In the following chapter I will introduce methods and measurements used for the present study. Three separate ASR models will be fine-tuned on the pre-trained wav2vec2-large-xlsr-53 model (Conneau et al., 2020) using three Icelandic datasets – Málrómur (Steingrímsson et al., 2017), Samrómur (Mollberg et al., 2020) and Althingis' Parliamentary Speeches (Helgadóttir et al., 2017), which will be loaded directly from Huggingface. The models will be trained using Python3 (Van Rossum & Drake, 2009) with a 32 vCPU, 196GB memory on UCloud. Python packages that will be imported and used are Datasets (Lhoest et al., 2021), Transformer (Wolf et al., 2020), LibROSA (McFee et al., 2015), JiWER (Morris et al., 2004), IpyWidgets which is a part of the Jupyter notebook (Kluyver et al., 2016) and PyTorch (Paszke et al., 2019). The pre-trained model will be imported using the “AutoModelForCTC” from the Transformer (Wolf et al., 2020) library. Hyperparameters will be implemented with attention, hidden and layer dropout being set to 0.1, feature projection dropout set to 0, time mask probability set to 0.05 and CTC loss reduction set to mean. Lastly, the padding token ID is set to match the padding token ID of the processor’s tokenizer and the vocabulary size is set to match the length of the processor’s tokenizer. Training arguments will be set using the “TrainingArguments” from the Transformer (Wolf et al., 2020) library. For the training arguments, batch size is set to 32, gradient accumulation steps set to 2, number of training epochs set to 6, gradient checkpointing set to true and learning rate set to 3e-4. The models training process will then be evaluated and saved every 400 steps, with maximum number of saved checkpoints set to 2. Lastly, the use of FP16 precision will be set to false as no GPU was for this project. The training process will then be done by the Trainer from the Transformer (Wolf et al., 2020) library with the training being done on the training dataset and the evaluation done on the test dataset. WER-score (Morris et al., 2004) will be used to evaluate the models performance every 400 steps. Lastly, after training is done, each model will be tested on their corresponding validation dataset which was unused during the training process. The code used is built upon a public notebook from GitHub on how to build a code to train an ASR model (*Notebooks*, 2020/2023).

Results

The results of training the models vary so each model’s results will be reported on and described separately. The model trained using the Málrómur dataset (Steingrímsson et al., 2017) was named malromur_asr_model, the model trained using the Samrómur dataset (Mollberg et al., 2020) was named samromur_asr_model and finally the model trained using the Althingis’ Parliamentary Speeches (Helgadóttir et al., 2017) dataset was named althingi_asr_model.

1.1 The results for the malromur_asr_model

The training of the malromur_asr_model took 95 hours, 11 minutes and 51 seconds in total. The process and the results of this training can be seen beneath in table 1 and figure 1.



Step	Training Loss	Validation Loss	WER
400	9.5321	2.5232	1.0016
800	0.9265	0.4230	0.8453
1200	0.4937	0.3101	0.7176
1600	0.4037	0.2690	0.6730
2000	0.3580	0.2429	0.6300
2400	0.3141	0.2247	0.6084
2800	0.2863	0.2159	0.5849
3200	0.2646	0.1994	0.5622
3600	0.2393	0.1902	0.5506
4000	0.2289	0.1810	0.5431
4400	0.2034	0.1725	0.5212
4800	0.1950	0.1690	0.5108
5200	1.1646	0.1835	0.5240
5600	0.2052	0.1674	0.5063
6000	0.8285	0.1856	0.5157

Table 1. The results of the training for the malromur_asr_model.

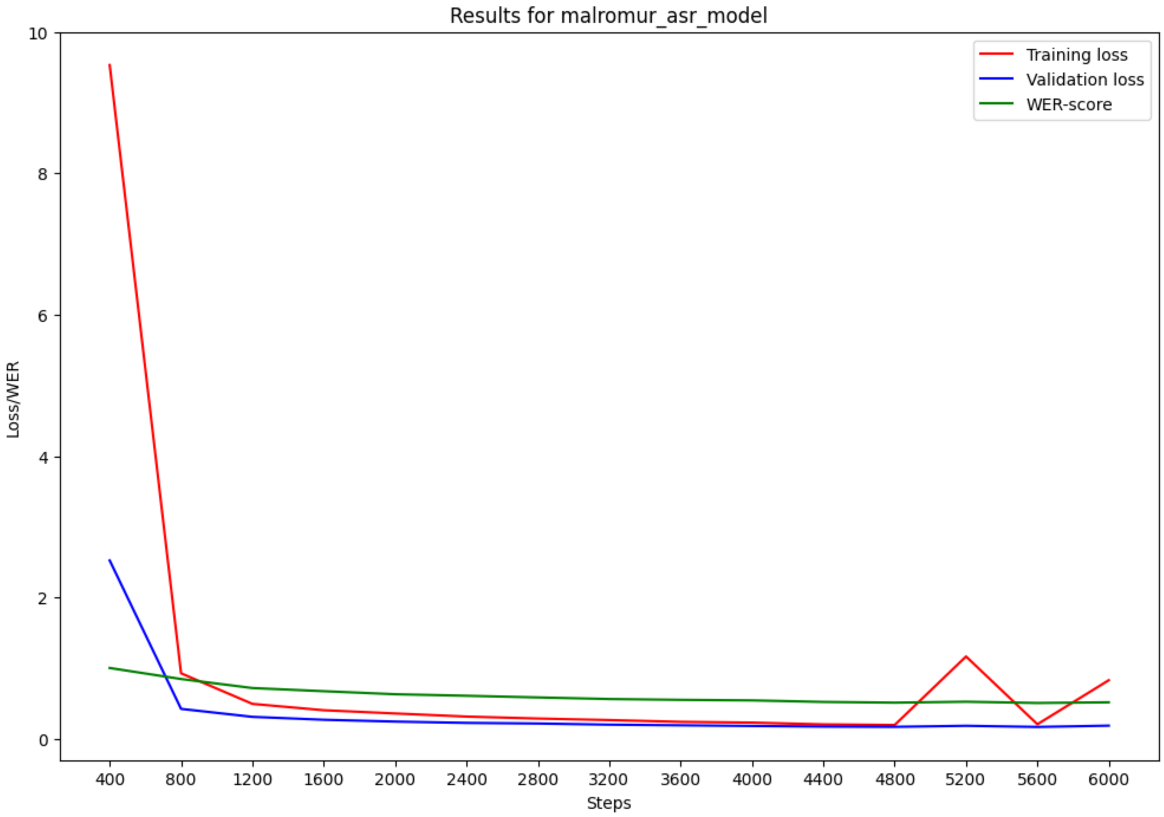
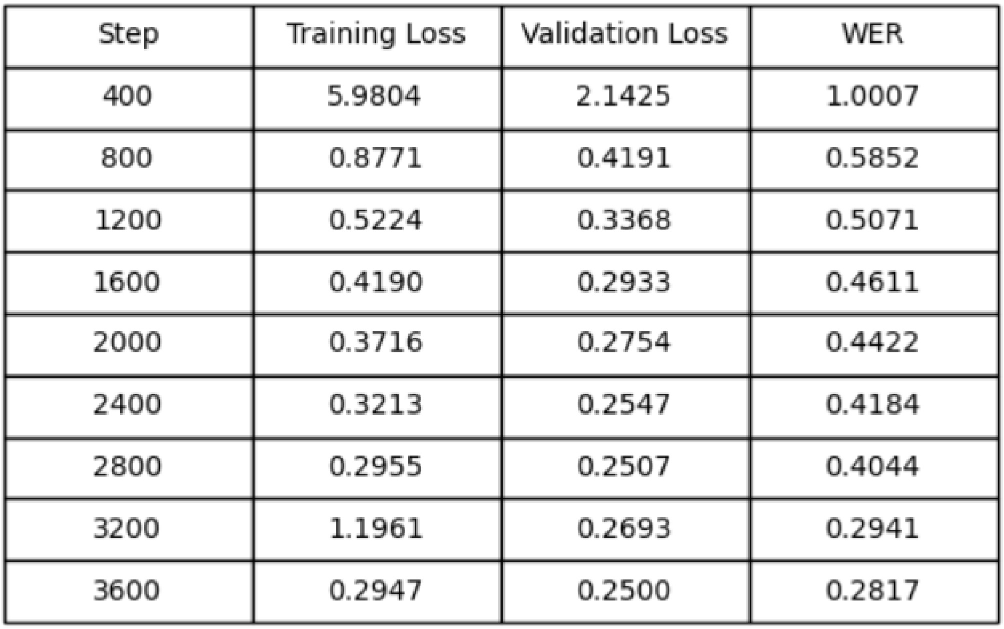


Figure 1. Visual representation of the results of the training for the malromur_asr_model.

In table 1, the training and validation loss and WER-score for each 400 steps can be seen. In figure 1, a visual representation of the training and validation loss and WER-score for each 400 steps can also be seen. The total number of optimization steps was 6258, with the final model being saved at 6000 steps. The connection to the UCloud computer was lost just before the 4800-step mark and the training had to be resumed from the 4800-step checkpoint. This resulted in a spike of training loss and a slight increase in validation loss and WER-score between the 4800-step mark and the 5200-step mark. The connection was lost again just before the 6000-step mark and the training had to be resumed from the 5600-step checkpoint. This again resulted in a spike of training loss and slight increase in validation loss and WER-score between the 5600-step mark and the 6000-step mark. The model was then tested afterwards on the validation dataset which yielded a WER-score of 0.5176.

1.2 The results for the samromur_asr_model

The training of the samromur_asr_model took 56 hours, 41 minutes and 43 seconds in total. The process and the results of this training can be seen beneath in table 2 and figure 2.



Step	Training Loss	Validation Loss	WER
400	5.9804	2.1425	1.0007
800	0.8771	0.4191	0.5852
1200	0.5224	0.3368	0.5071
1600	0.4190	0.2933	0.4611
2000	0.3716	0.2754	0.4422
2400	0.3213	0.2547	0.4184
2800	0.2955	0.2507	0.4044
3200	1.1961	0.2693	0.2941
3600	0.2947	0.2500	0.2817

Table 2. The results of the training for the samromur_asr_model.

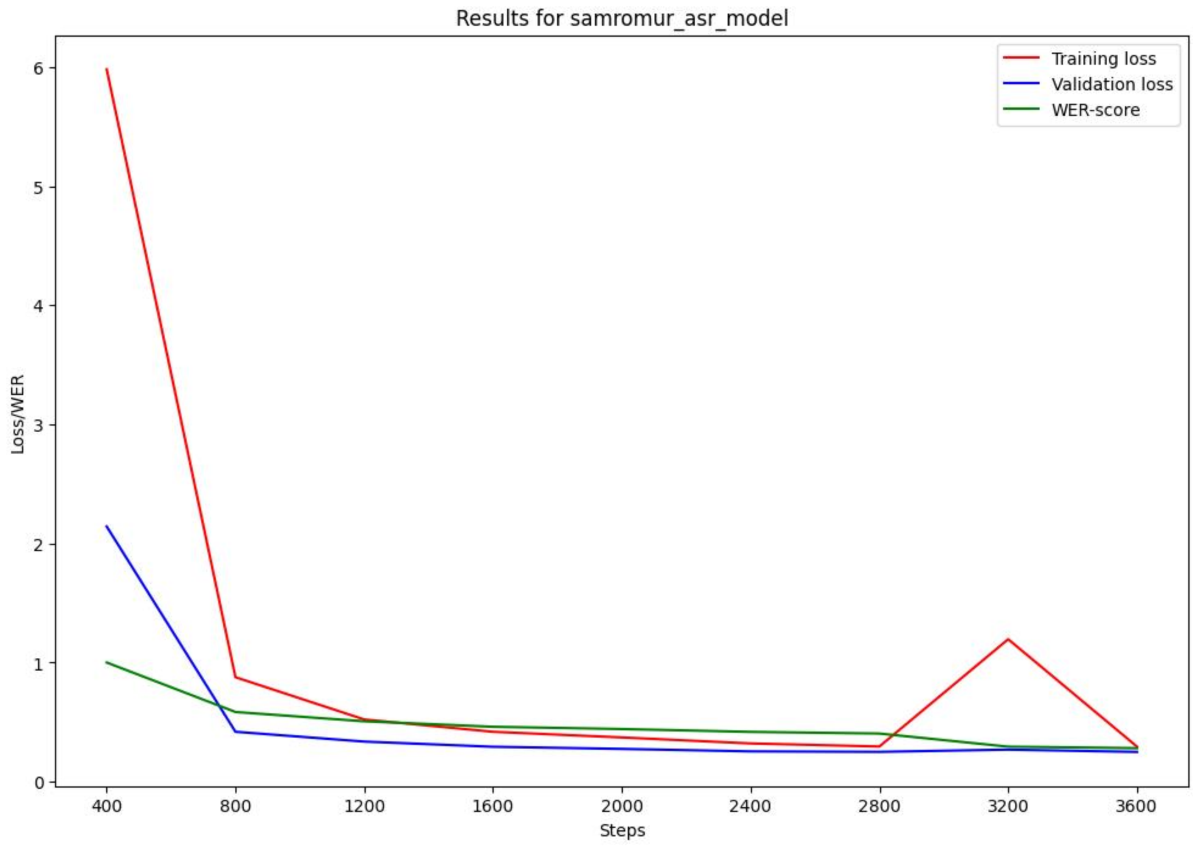
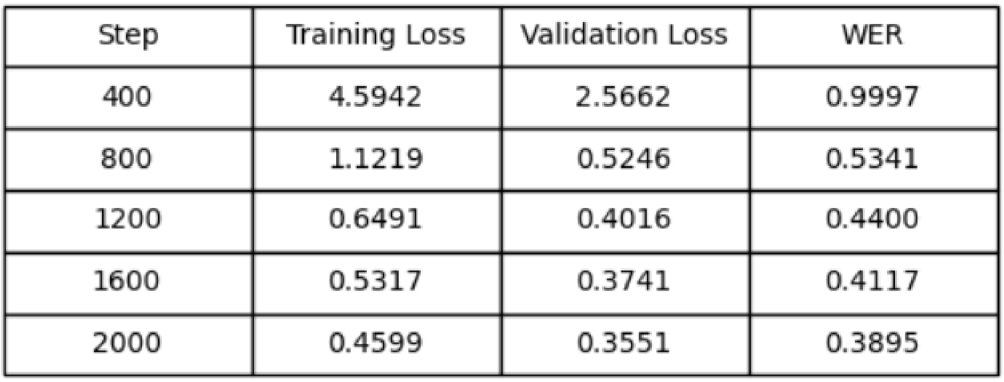


Figure 2. Visual representation of the results of the training for the samromur_asr_model.

In table 2, the training and validation loss and WER-score for each 400 steps can be seen. In figure 2, a visual representation of the training and validation loss and WER-score for each 400 steps can also be seen. The total number of optimization steps was 3876, with the final model being saved at 3600 steps. The connection to the UCloud computer was lost just before the 3200-step mark. The training then had to resume from the 2800-step checkpoint. This resulted in a spike of training loss and a slight increase in validation loss between the 2800-step mark and the 3200-step mark. Despite this, the model kept improving as the WER-score dropped significantly. The model was then tested afterwards on the validation dataset which yielded a WER-score of 0.2546.

1.3 The results for the althingi_asr_model

The training of althingi_asr_model dataset took 29 hours, 36 minutes and 11 seconds in total. The process and the results of this training can be seen beneath in table 3 and figure 3.



Step	Training Loss	Validation Loss	WER
400	4.5942	2.5662	0.9997
800	1.1219	0.5246	0.5341
1200	0.6491	0.4016	0.4400
1600	0.5317	0.3741	0.4117
2000	0.4599	0.3551	0.3895

Table 3. The results of training for the althingi_asr_model.

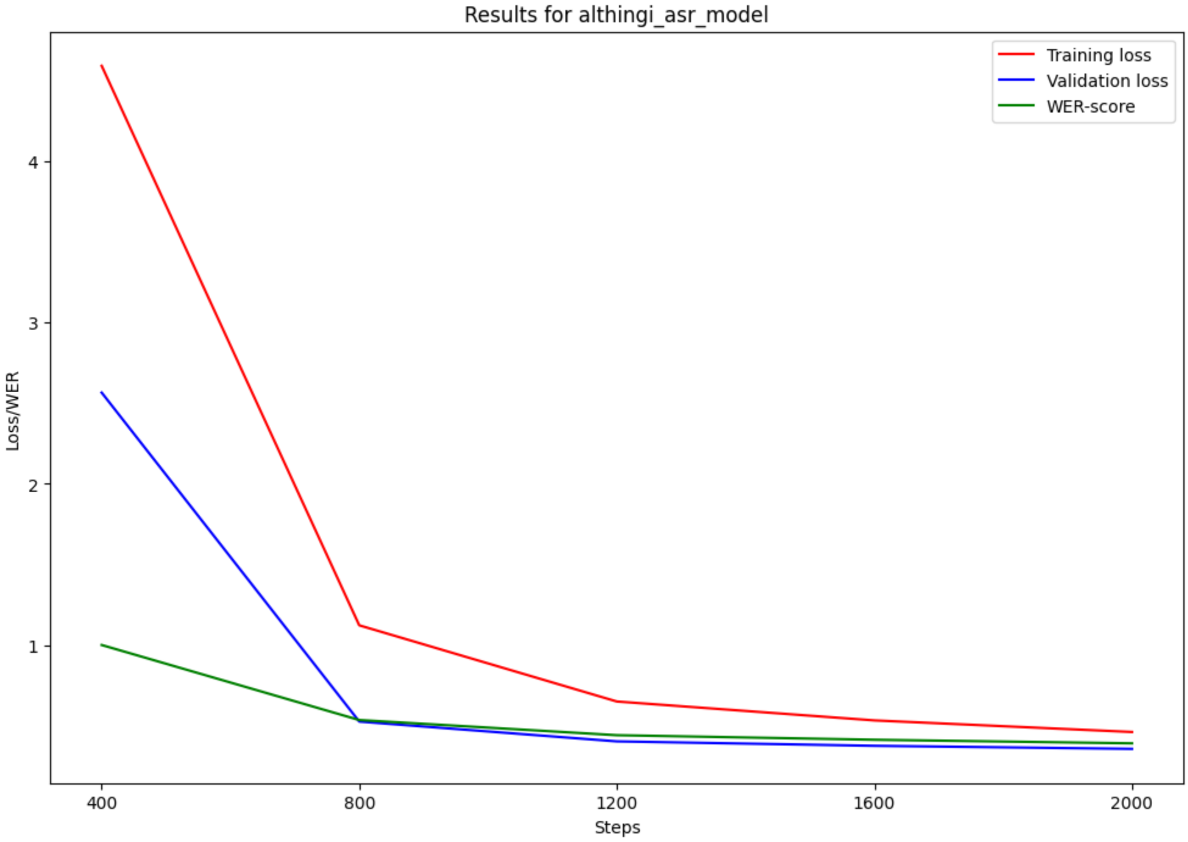


Figure 3. Visual representation of the results of the training for the althingi_asr_model.

In table 3, the training and validation loss and WER-score for each 400 steps can be seen. In figure 3, a visual representation of the training and validation loss and WER-score for each 400 steps can also be seen. The total number of optimization steps was 2364, with the final model being saved at 2000 steps. The model was then tested afterwards on the validation dataset which yielded a WER-score of 0.3803.

1.4 Comparison between the models

The visual comparison between the models can be seen beneath in figure 4.

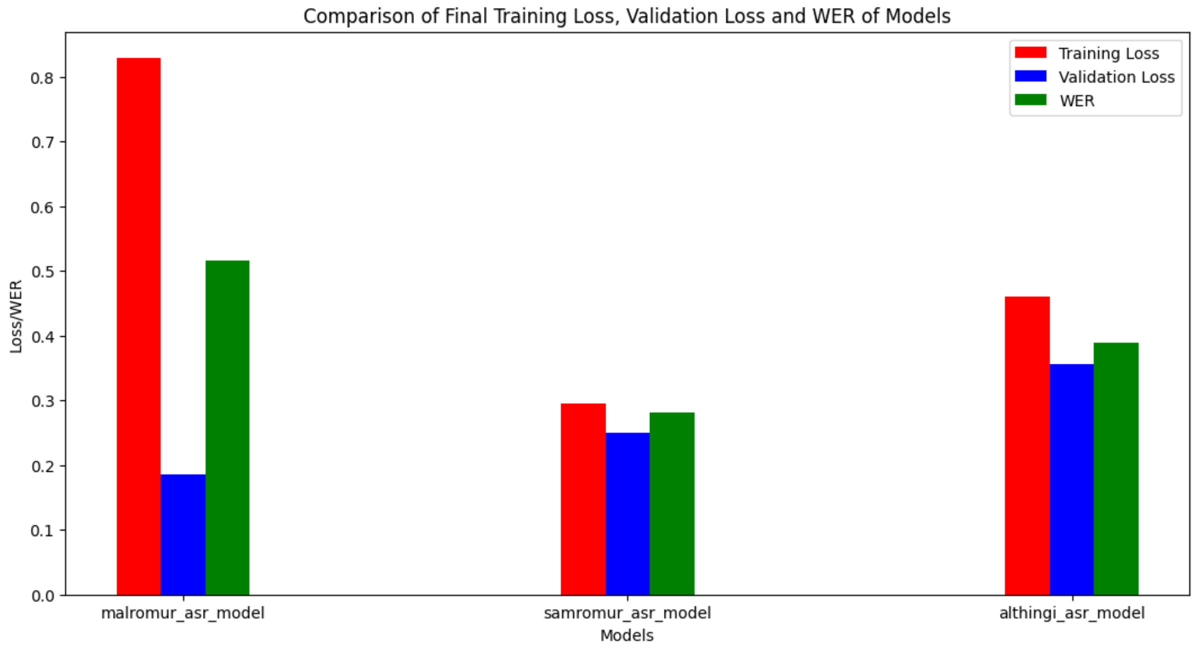


Figure 4. Visual comparison of the results of the training for the three models.

As can be seen, the samromur_asr_model performs considerably better than the other two models, especially better than the malromur_asr_model.

Discussion

This purpose of this study was to fine-tune three ASR models on the wav2vec2-large-xlsr-53 pre-trained model using three Icelandic datasets, designed for training ASR models, and then compare those models to see which one performed the best. During training, all the models kept improving with training loss, validation loss and WER-score consistently dropping for every model throughout. However, how the models turned out varies greatly. The malromur_asr_model performed significantly worst with a final WER-score of 0.5157. Furthermore, when tested on the validation dataset afterwards, the WER-score was even higher with a WER-score of 0.5176, which means the model got roughly every other word right. The althingi_asr_model came second with a WER-score of 0.3895 when tested on the last saved step. When tested on the validation dataset afterwards it yielded a WER-score of 0.3803. The model then performed much better than the malromur_asr_model but still unreliable. Lastly, the best performing model was samromur_asr_model with a WER-score of 0.2817 on the last saved step. Furthermore, when tested on the validation dataset afterwards, it yielded a WER-score of 0.2546. The final WER-score for samromur_asr_model is significantly lower than for the other two models and must be considered pretty good. The reason why the model does so much better than the other two models is, however, unknown. The dataset used to train samromur_asr_model is smaller than the other two datasets and all models were trained for the same number of epochs. However, the size of the dataset isn't everything and the quality of the data could partly explain this difference. For instance, the Althingi Parliamentary Speeches dataset (Helgadóttir et al., 2017) had issues with noisy background and includes data recorded over a much larger period of time, which could impact the models performance. The Málrómur (Steingrímsson et al., 2017) and Samrómur (Mollberg et al., 2020) datasets both include a large number of speakers. However, the data for the Samrómur dataset (Mollberg et al., 2020) was collected more recently and after the Icelandic government had set it's five-year plan in motion. In fact, the plan was set in motion in the same month as the speech data was recorded for the dataset after two years of preparation (Nikulásdóttir et al., 2020) in 2019. This could mean that the speech data was recorded with much better equipment and in a much better environment.

Limitations

The initial plan for this project was to merge four Icelandic datasets for ASR model training. However, when trying to merge the datasets, errors kept coming up that a solution was not found for. Therefore, the focus was shifted, and the plan was to train for separate models using those four distinct datasets. Another challenge was faced when one of those datasets was unable to load and was therefore left out. The remaining three datasets were then used for training and uploaded to Huggingface. However, the plan was then to use those pre-trained models for further testing by loading the validation datasets from the other models. For instance, for the samromur_asr_model, the plan was to load the validation datasets from the Althingis' Parliamentary Speeches dataset (Helgadóttir et al., 2017) and the Málrómur dataset (Steingrímsson et al., 2017) and use those to test the model even further and on new data. That was however not doable for this project as errors kept arising that were not solvable at the present time.

Conclusion

To conclude, the present study set out to train three distinct ASR models on three different datasets and see how they performed. The hypothesis for the study was that all models would perform well due to the size and quality of the datasets used for training. That was, however, not the case for all models. The malromur_asr_model performed very badly after training, managing only to get a WER-score of 0.5157 and therefore managing to get less than half of the words right. It cannot therefore be deemed very reliable. The althingi_asr_model performed much better but still only managed to achieve a WER-score of 0.3895 and therefore cannot also be deemed reliable. The samromur_asr_model came on top with a WER-score of 0.2817, which can be deemed somewhat reliable. The reason why the samromur_asr_model performed so much better could be due to the quality of the speech recordings in the Samrómur dataset (Mollberg et al., 2020). These speech recordings were made in the same month as the Icelandic government rolled out its five-year plan to enhance and develop the Icelandic language technology sector (Nikulásdóttir et al., 2020) in 2019. It is therefore likely that these speech recordings were of higher quality than those that were recorded before. The Samrómur dataset is an ongoing project with its latest release being in March 2020 (Mollberg et al., 2020). It will therefore be interesting to see the size and quality of its next release, whenever that will be. Further research is encouraged to see if these findings can be replicated or improved with different hyperparameters or other datasets. It is also encouraged to try fine-tuning these models further on other Icelandic datasets and see if they can be further improved.

References

Ali, A., & Renals, S. (2020). *Word Error Rate Estimation Without ASR Output: E-WER2*. https://doi.org/10.48550/ARXIV.2008.03403

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). *Unsupervised Cross-lingual Representation Learning for Speech Recognition*. https://doi.org/10.48550/ARXIV.2006.13979

Filippidou, F., & Moussiades, L. (2020). A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations* (Vol. 583, pp. 73–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-49161-1_7

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., & Guðnason, J. (2017). Building an ASR Corpus Using Althingi's Parliamentary Speeches. *Interspeech 2017*, 2163–2167. https://doi.org/10.21437/Interspeech.2017-903

Katyal, A., Kaur, A., & Gill, J. (2014). Automatic Speech Recognition: A Review. *International Journal of Engineering and Advanced Technology*, *Volume-3*(Issue-3), 71–74.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & team, J. development. (2016). Jupyter Notebooks? A publishing format for reproducible computational workflows. In F. Loizides & B. Scmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90). IOS Press. https://eprints.soton.ac.uk/403913/

Lhoest, Q., del Moral, A. V., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Scao, T. L., Sanh, V., Xu, C., Patry, N., … Wolf, T. (2021). *Datasets: A Community Library for Natural Language Processing*. https://doi.org/10.48550/ARXIV.2109.02846

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. *Proceedings of the 14th Python in Science Conference*, *8*.

Mollberg, D. E., Jónsson, Ó. H., Þorsteinsdóttir, S., Steingrímsson, S., Magnúsdóttir, E. H., & Gudnason, J. (2020). Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 3463–3467. https://aclanthology.org/2020.lrec-1.425

Morris, A. C., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. *Interspeech 2004*, 2765–2768. https://doi.org/10.21437/Interspeech.2004-668

Nikulásdóttir, A. B., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., & Steingrímsson, S. (2020). *Language Technology Programme for Icelandic 2019-2023*. https://doi.org/10.48550/ARXIV.2003.09244

Notebooks. (2023). [Jupyter Notebook]. Hugging Face. https://github.com/huggingface/notebooks/blob/main/examples/speech_recognition.ipynb (Original work published 2020)

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran

Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Steingrímsson, S., Guðnason, J., Helgadóttir, S., & Rögnvaldsson, E. (2017). Málrómur: A Manually Verified Corpus of Recorded Icelandic Speech. *Proceedings of the 21st Nordic Conference on Computational Linguistics*, 237–240. https://aclanthology.org/W17-0229

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Appendix

Link to the GitHub repository that contains all coding scripts:

https://github.com/gudjonkri20/DataScienceExam

Link to the Huggingface repository containing all the fine-tuned models:

https://huggingface.co/gudjonk93