Automatic Transcription for Estonian Children's Speech

Anonymous Author
Affiliation / Address line 1
email@domain

Anonymouser Author
Affiliation / Address line 1
email@domain

Anonymousest Author
Affiliation / Address line 1
email@domain

Abstract

We evaluate the impact of recent improvements in Automatic Speech Recognition (ASR) on transcribing Estonian children's speech. Our research focuses on fine-tuning large ASR models with a 10-hour Estonian children's speech dataset to create accurate transcriptions. Our results show that large pre-trained models hold great potential when fine-tuned first with a more substantial Estonian adult speech corpus and then further trained with children's speech.

1 Introduction

Automatic Speech Recognition (ASR) continues to face challenges in accurately transcribing children's speech. Research efforts are underway to adapt adult ASR models to better handle the unique pronunciation variations and limited vocabulary that are characteristic of children's speech (Thienpondt and Demuynck, 2022; Dutta et al., 2022). These adaptations are necessary due to the limitations of current ASR systems, which often lack adequate representation of children's speech and struggle to generalize to new examples.

Recent advancements in ASR technology, including the use of large transformer-based models and unsupervised pre-training techniques, have resulted in improved performance for adult speech recognition, with the ability to train on a diverse range of data without human annotations (Baevski et al., 2020; Radford et al., 2022; Hsu et al., 2021). These models demonstrate greater robustness and generalization compared to previous systems. However, the effectiveness of these advanced ASR models for children's speech, especially in low-resource languages like Estonian, remains untested.

In this paper, we are investigating two multilingual speech models - Facebook's Wav2Vec2-XLS-R (Babu et al., 2021) and OpenAI's Whisper (Radford et al., 2022) - as potential starting points for building an ASR system transcribing Estonian children's speech. Our objective is to determine the potential of these models in creating low-effort ASR systems for children speaking a low-resource language like Estonian, for which there are no ASR systems for children's speech.

To accomplish this, we fine-tune the XLS-R and Whisper models from scratch using children's speech data. We also fine-tune pre-existing models for the Estonian language with additional children's speech recordings. Furthermore, we compare the quality of the ASR system by evaluating a pre-made Estonian ASR system provided by Microsoft Azure and exploring its fine-tuning capabilities.

Our research indicates that XLS-R models and Whisper models can serve as effective starting points for building an ASR system using only 10 hours of children's speech. However, for optimal performance, these models should first be fine-tuned with Estonian adult speech. We achieve the best word error rate of around 15 using an XLS-R model that was fine-tuned with Estonian ASR datasets and further trained with children's speech. Furthermore, our results show that the Azure speech-to-text model performs similarly to the Estonian XLS-R model but not as well as the fine-tuned public models.

In the next sections, we describe which data we used for evaluation and training, which models we used and how we fine-tuned these and last but not least we present and analyse the results.

2 Dataset and evaluation

The Children ASR dataset used in this work consists of speech recordings from 53 children aged 6 to 13. The data was collected by the Children's

Clinic of Tartu University Hospital and contains a mix of both boys and girls speaking about various topics such as answering questions, describing pictures, talking about their family and friends, and more. The dataset is divided into three subsets - test, dev, and train - with no overlap in speakers or texts.

The test set contains all age and gender groups and has a total recording duration of 278 minutes (approximately 4.6 hours). The development set is missing some speakers and has a total recording duration of 182 minutes (approximately 3 hours). The training set is also missing some speakers and has a total recording duration of 613 minutes (approximately 10 hours). A breakdown of the total recording duration for the test set by age and gender of the speakers is shown in Table 1.

Age	Girls (min)	Boys (min)	Total (min)
6	17	21	38
7	14	16	30
8	17	14	31
9	22	18	40
10	15	17	32
11	20	17	37
12	16	22	38
13	19	13	32
Total	140	138	278

Table 1: Total recording duration in minutes for the Estonian children ASR test set, broken down by age and gender of the speakers.

The children in the dataset speak about a wide range of topics, covering everything from answering questions and describing pictures to discussing their family and friends. They also include recordings of children reading fairytales, reciting poems, and saying specific sentences. The utterances in the dataset vary in their level of spontaneity - some are unscripted expressions of thoughts, while others feature children reading.

We evaluate the performance of our speech recognition models using the standard measure of word error rate (WER). This involves converting all text to lowercase and removing punctuation but not standardizing different spelling variations. Our reference transcriptions reflect the pronunciation of children, including any errors they may make. However, the line between correct and incorrect pronunciation is often blurry and some children's speech can be difficult to comprehend. We do not consider the ambiguity in human transcriptions and simply compare the models' output to our reference transcription, which could lead to increased WERs.

3 Models and training

We are using both public large speech models and private black box speech service. In the case of public models, we also searched for models already fine-tuned with Estonian speech data. We fine-tune the selection of these models with the children's speech dataset mentioned in the last section.

For public models, we use two multilingual ones: Facebook's XLS-R and OpenAI's Whisper (Radford et al., 2022). XLS-R model is trained with speech modelling objective, not ASR but it can be fine-tuned to ASR with Connectionist Temporal Classification (CTC) (Graves et al., 2006) algorithm. The Whisper on the other hand is a multipurpose model that contains both transformer encoder and decoder blocks and has been trained on several speech-processing tasks, like multilingual speech recognition, speech translation and voice activity detection (Radford et al., 2022).

The available XLS-R models have 300 million, 1 billion and 2 billion parameters, we are using the two smaller ones in this work. The Whisper model comes in six different sizes; we are using medium and large-v2 since the Estonian error rates for other ones are relatively high. There is one Estonian-specific fine-tuned model available for the 300 million parameter version, trained with over 700 hours of Estonian speech data (Alumäe and Olev, 2022). There are several Estonian Whisper models available in HuggingFace but these are trained with fewer data examples. We are using the best available medium and large-v2 ones.[1][2]

We use standard fine-tuning procedures. For training XLS-R-based ASR models from scratch, we use the learning rate of 3e-4, a 400-step warmup and train the models for 60 epochs with children's speech dataset, which is less than 4000 steps. When further fine-tuning the Estonian XLS-R model with children's speech, we use the learning rate of 2e-5 and 200 warmup steps. We fine-tune all the Whisper models with warmup 10%

[1] https://huggingface.co/agnesluhtaru/whisper-medium-et-ERR2020

[2] https://huggingface.co/agnesluhtaru/whisper-large-et-ERR2020-v2

of the steps and learning rate 1e-05. When fine-tuning the out-of-the-box Whisper models, we train these for 5000 steps or atound 40 epochs and when fine-tuned models already trained with Estonian adult speech, we train the large model for 2000 steps or over 16 epochs and medium model for 1000 steps or eight epochs.

For the private model, we use Microsoft Azure Speech service's speech-to-text[3], which requires an Azure subscription and a Speech resource. The transcription services can be accessed by making REST requests.

Microsoft Azure offers the option to fine-tune the model with custom datasets. This process involves uploading data to train the models, followed by deploying the trained models. Since audio-based fine-tuning is not available for Estonian, we use text-based tuning for our work with the texts from the children's speech dataset.

4 Results

In this section, we describe the results of all the models based on Facebook's XLS-R, OpenAI'S Whisper and Microsoft Azure speech-to-text.

4.1 XLS-R

Table 2 shows the word error rate (WER) scores of fine-tuned Estonian XLS-R models using only 10 hours of Estonian children's speech data, the fine-tuned Estonian model (Alumäe and Olev, 2022) and Estonian model further trained with children's speech. We can see that the limited amount of data for fine-tuning XLS-R from scratch results in a high WER of over 30 for both models with 300 million and one billion parameters. Training an ASR model using only 10 hours of speech data can be challenging, especially when the speech is for a low-resource language and children.

The results show that the pre-trained Estonian ASR model has a WER of around 20, while further fine-tuning the model with children's speech data leads to even better results, with a WER of less than 15. Based on the lower WER score for fine-tuned one billion parameter model, we can suggest that a larger model fine-tuned with Estonian data first and then further trained on children's speech could lead to even better results.

The results indicate that fine-tuning the Estonian ASR model using children's speech data improves performance across all age groups (refer to Figure 1). Younger speakers tend to have a higher word error rate (WER) than older speakers, although this relationship is not always straightforward. There are some exceptions, such as the recognition performance for 13-year-olds being worse than that of younger age groups. This highlights that speaker variability plays a role in the WER results. Nevertheless, the fine-tuning of the ASR model using children's speech data reduces the differences in recognition performance across age groups, resulting in improved overall performance.

Model	Test	Dev
xls-r-300M-children	36.3	34.58
xls-r-1B-children	30.89	31.06
xls-r-300M-et	20.62	19.15
xls-r-300M-et-children	15.31	14.30

Table 2: Comparison of WER scores for Facebook's Wav2Vec2 XLS-R (Babu et al., 2021) based models fine-tuned with only Estonian children's speech, only Estonian adult speech (Alumäe and Olev, 2022) and first fine-tuned to Estonian and further trained with children's speech.

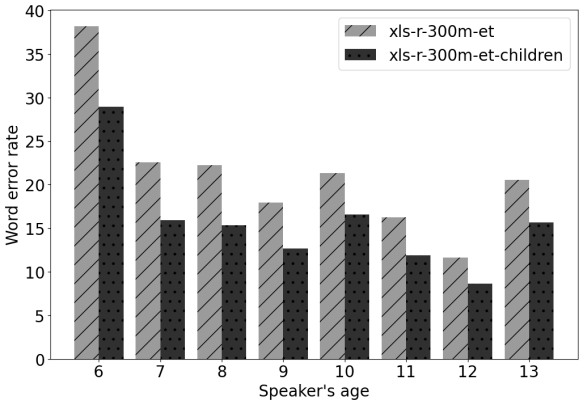

Figure 1: Performance comparison of Estonian XLS-R ASR and children's speech fine-tuned models across age groups.

4.2 Whisper

The performance of the out-of-the-box Whisper models on the children's dataset (see Table 3) is comparable to the scores reported by Radford et al. (2022) on the Estonian Common Voice 9 Ardila et al. (2020). All models have a WER of at least 35. So, although we can use Whisper without fine-

[3] https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/speech-to-text

tuning, it does not transcribe Estonian speech well and therefore does not give great transcriptions for Estonian children's speech as well.

When only fine-tuning the model with 10 hours of children's speech, we can already get better results, with large-v2 model WER around 20, which is significantly better than using a model fine-tuned with Estonian speech that showed much better performance for XLS-R. Although we don't have reasonable evidence for saying that XLS-R models are better because the Whisper models are not optimal.

Model	Test	Dev
Whisper-medium	46.11	43.21
Whisper-large-v2	36.01	35.06
Whisper-medium-children	25.08	24.29
Whisper-large-v2-children	20.38	20.58
Whisper-medium-et	28.78	26.83
Whisper-large-v2-et	29.2	28.13
Whisper-medium-et-children	18.66	17.49
Whisper-large-v2-et-children	16.02	15.73

Table 3: Comparison of WER scores for OpenAI Whisper (Radford et al., 2022) models and Whisper models fine-tuned with only Estonian children's speech, only Estonian adult speech and first fine-tuned to Estonian and further trained with children's speech.

Despite using the Estonian Whisper models fine-tuned with fewer audio text pairs than the XLS-R model, when trained further with children's speech, the large model achieved similar WER as the double fine-tuned smaller XLS-R model.

4.3 Azure

The results from our evaluation of the children's speech dataset show that the out-of-the-box Azure speech-to-text model performs similarly or better than the fine-tuned Estonian XLS-R model (Alumäe and Olev, 2022). As indicated in Table 4, the Microsoft Azure speech-to-text scores are around 20 or below.

Model	Test	Dev
Microsoft Azure	18.93	20.18
Azure text-tuned	20.31	21.21

Table 4: WER scores for Microsoft Azure speech-to text and it's custom text-tuned version.

However, the experiment also shows that text-tuning is not the best approach for this particular dataset. The dataset mostly contains simpler vocabulary and not much terminology, most likely leading to quick overfitting with text-tuning. Currently, text-tuning is the only option available for the Estonian language, but it might not be the best use case for children's speech datasets.

5 Discussion

Our experiments show that children's speech recognition continues to be a tricky problem but big speech models are looking promising. It is possible to build an ASR system for Estonian children's speech without any bells and whistles using only 10 hours of data and get output that is decent and might be good enough for use in chatbots. However, when it comes to six-year-olds, whose speech is difficult to understand even for the human ear, the system is still struggling.

We evaluate different models and it appears that both OpenAI's Whisper and Facebook's XLS-R are viable options for developing a speech recognition model for Estonian children's speech. The current best word error rate is around 15 with XLS-R. However, it remains unclear if this pre-trained model is optimal for children's speech or if a lower error rate could be achieved with Whisper after fine-tuning with a similar amount of Estonian adult speech. Additionally, we do not obtain comparable results with the Azure service, as it does not permit fine-tuning with audio data.

Our findings suggest that the results could be improved by using a larger XLS-R model as the base or by fine-tuning Whisper models with more data. Additionally, we do not use a separate language model, which is possible with both Whisper and XLS-R models and could potentially enhance the performance of these models.

6 Conclusion

We test the performance of two speech recognition models, XLS-R and Whisper, on transcribing Estonian children's speech. We fine-tune the models with children's speech data and compared them to an off-the-shelf system from Microsoft Azure. Both models fine-tuned with children's speech, outperform Microsoft Azure, which does not allow fine-tuning with audio for Estonian, and are promising for children's ASR system.

References

Tanel Alumäe and Aivo Olev. 2022. Estonian speech recognition and transcription editing service. volume 10, page 409–421.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. https://aclanthology.org/2020.lrec-1.520 Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. https://doi.org/10.48550/ARXIV.2111.09296 Xls-r: Self-supervised cross-lingual speech representation learning at scale.

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.

Satwik Dutta, Sarah Anne Tao, Jacob C. Reyna, Rebecca Elizabeth Hacker, Dwight W. Irvin, Jay F. Buzhardt, and John H.L. Hansen. 2022. https://doi.org/10.21437/Interspeech.2022-555 Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments. In *Proc. Interspeech 2022*, pages 4322–4326.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. https://doi.org/10.1145/1143844.1143891 Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. https://doi.org/10.1109/TASLP.2021.3122291 Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. https://doi.org/10.48550/ARXIV.2212.04356 Robust speech recognition via large-scale weak supervision.

Jenthe Thienpondt and Kris Demuynck. 2022. https://doi.org/10.21437/Interspeech.2022-10964 Transfer Learning for Robust Low-Resource Children's Speech ASR with Transformers and Source-Filter Warping. In *Proc. Interspeech 2022*, pages 2213–2217.