COMBINING MULTILINGUAL RESOURCES AND MODELS TO DEVELOP STATE-OF-THE-ART E2E ASR FOR SWEDISH

Lukas Mateju, Jan Nouza, Petr Cerva, Jindrich Zdansky, Frantisek Kynych
<frantisek.kynych@tul.cz>

In terms of automatic speech recognition (ASR), Swedish is considered among the group of languages with limited available resources. Currently, there are only a few hundred hours of publicly accessible training data, mostly consisting of read speech. In an effort to gather a more extensive and realistic dataset, we have undertaken an exploration of existing multilingual methodologies. Additionally, we propose two novel approaches that involve merging Swedish with previously established Norwegian data and models. These methods efficiently harvest spoken Swedish from diverse sources like broadcast, parliament, YouTube, and audiobook archives, significantly speeding up the process and enhancing the Swedish ASR system's performance. Our evaluations show that our system surpasses leading commercial cloud services.

Index Terms: end-to-end speech recognition, transfer learning, multilingual training, Scandinavian languages, Swedish

INTRODUCTION

Swedish (SWE), along with Norwegian (NO) and Danish, is part of the North-Germanic language group, with Swedish being the most widely spoken (9.2M people, while the others have around 5M each). ASR research for Scandinavian languages was active in the 1990s but saw limited progress in the following two decades due to speech resource limitations and complex linguistic characteristics.

Modern end-to-end (E2E) systems [1] have alleviated these issues, driving new research. Recent MSc theses and papers since 2020 have explored E2E approaches for Swedish ASR [2], showing promising results with a word error rate (WER) below 10%. However, performance drops with more realistic data, highlighting the need for larger, diverse training resources.

To collect more data, public sources like broadcast archives, parliamentary records, and YouTube can be utilized, along with audiobooks and corresponding e-books. Tools are needed to prepare audio and text data, detect matching parts, and split them for training, typically involving an ASR system to convert audio to text. For less-resourced languages, this iterative process can be slow, so methods leveraging existing data and models for other languages are explored.

This paper focuses on Swedish, exploring multilingual techniques like transfer learning and multilingual training, utilizing previously created Norwegian data and E2E models to gather over 1,000 hours of realistic training data, resulting in a high-performing ASR model for various applications.

METHODOLOGY

In this work, we employ the ESPNet platform [4] for our end-to-end model, combining connectionist temporal classification (CTC) and attention-based encoder-decoder (AED). The model consists of a shared encoder (conformer [5]), two decoders (CTC-based and attention-based).

The shared encoder includes two sub-sampling convolutional layers (3×3 kernel, stride 2) and 12 blocks with eight attention heads. The CTC decoder uses a linear layer to transform encoder output. The attention decoder is a transformer with six blocks, 512 attention dimensions, and 2,048 position-wise feed-forward layer units. The model has 136M parameters.

Transfer learning can initialize the encoder from a high-resource language model. We compare the non-initialized (mono) and initialized (init) approaches and explore four multilingual training variations:

1. Joint training combines datasets from multiple languages for monolingual training (136M parameters).
2. Joint-with-lid extends joint training with language identity guidance.
3. Multi-decoder uses a pair of language-specific decoders for each language (44M language-specific parameters, 92M shared).
4. Multi-layer employs language-specific CTC decoders and shared attention decoders (13M language-specific parameters, 126M shared).

Each training batch is language-specific, requiring language knowledge during training and decoding for the last three modifications.

Table 1: WER [%] of SWE models (using NO as a support language) trained with increasing amounts of data across all test datasets.

model	25 h	50 h	100 h	200 h	300 h	500 h	1000 h
mono SWE	66,4	39.8	22.8	16.1	13.8	12.0	10.8
init SWE from NO	47,7	28.9	19.2	**13.5**	12.3	**11.1**	**9.9**
joint SWE+NO	22,6	19.5	16.2	13.7	12.5	11.9	10.4
joint-with-lid SWE+NO	**22.1**	20.3	**16.0**	13.7	12.6	11.7	10.4
multi-decoder SWE+NO	28.4	21.1	17.0	**13.5**	**12.2**	11.6	10.3
multi-layer SWE+NO	24.0	**19.4**	17.1	**13.5**	**12.2**	11.8	10.1

RESULTS AND DISCUSSION

To thoroughly assess the architecture's suitability for bootstrapping a new language, we randomly sampled and trained all variants with varying data amounts (from 25 to 1,000 hours). Each data addition (e.g., 25 to 50 hours) builds upon the previous set (25 hours) to simulate effective bootstrapping.

The results in Table 1, weighted WER values across all test datasets, reveal a clear trend: multilingual training significantly aids in the early stages when target data is limited. The closely related language data fills gaps, with more than a 35% WER difference between mono- and any multi-lingual model using only 25 hours of the target language. Around 200 hours, performance starts to even out, and initialization from a related language (init) becomes a faster and slightly better option. The related language in multilingual training starts to cause more confusion. Among multilingual modifications, the multi-layer approach performs best, isolating only language-dependent model parts, but all modifications are beneficial. Even with 1,000 hours, a solely monolingual model cannot surpass the use of a related language.

Table 2: WER [%] comparison to commercial solutions.

test set	our final	MS Azure	Google Cloud
NST5h	**2.9**	5.5	24.9
CV	**5.9**	10.5	22.4
SVT	12.6	**10.8**	35.0
PAR	**7.3**	11.5	26.5
YTB	11.3	**10.1**	31.4
ABOOK	**3.9**	11.4	23.7
FLEURS	**12.4**	12.9	21.1
all	**8.0**	10.1	16.8

We trained the final model (on NVIDIA A40 GPU in 7 days) on the complete 1,226-hour training corpus using the best-performing technique (initialization by the Norwegian model). Results in Table 2 show the lowest WER values for read speech sets (NST5h, CV, and ABOOK), with the ABOOK content being challenging (8% unseen words). Furthermore, the FLEURS set's WER demonstrates the model's ability to generalize for unseen data.

CONCLUSIONS

In this work, we developed a hybrid CTC/AED-based E2E ASR model for Swedish. We combined freely available datasets with new multilingual techniques to harvest over 1,200 hours of training data, leveraging Norwegian as a closely related language. Multilingual training proved advantageous with limited data, while transfer learning reduced WER values significantly with enough data. The final initialized model, trained on the entire dataset, on average outperforms two commercial cloud services. We evaluated the model with a diverse testing dataset and provide access to the test sets and detailed logs, including ASR-to-reference word alignments, on our cloud platform.

ACKNOWLEDGEMENTS

This work was supported by the Student Grant Competition (SGS) at the Technical University of Liberec in 2023.

REFERENCES

[1] D. Wang, et al., “An overview of end-to-end automatic speech recognition,” Symmetry, vol. 11, no. 8, 2019.

[2] M. Malmsten, et al., “Hearing voices at the national library - a speech corpus and acoustic model for the Swedish language,” in Fonetik 2022, Stockholm, Sweden, 2022.

[3] R. Al-Ghezi, et al., “Self-supervised end-to-end ASR for low resource L2 Swedish,” in Interspeech 2021, Brno, Czechia, 2021.

[4] S. Watanabe, et al., “Espnet: End-to-end speech processing toolkit,” in Interspeech 2018, Hyderabad, India, 2018.

[5] A. Gulati, et al., “Conformer: Convolution-augmented transformer for speech recognition,” in Interspeech 2020, Shanghai, China, 2020.