Automatic speech recognition systems for applications with child users

Author: Ing. Richard Ševc

Supervisor: doc. Ing. Stanislav Ondáš, PhD.

Motivation

Automatic Speech Recognition (ASR) technology has made significant strides in recent years, becoming a key component of Natural Language Processing (NLP). ASR systems are widely used in various applications, from voice-controlled devices to automatic transcription services. However, most research in ASR has focused on adult speech, which is generally more consistent in pitch, pronunciation, and speech patterns. In contrast, children's speech, which is characterized by higher pitch, variability, and less precise articulation, remains underexplored especially in languages like Slovak.

Contributions

This diploma thesis tackles the unique challenges of recognizing Slovak children's speech. The goal is to adapt ASR models like wav2vec 2.0 and its multilingual version, XLS-R, to better process the specific traits of children's speech. The thesis also explores data augmentation methods, proven to improve ASR accuracy in similar cases.

A major contribution is expanding the Slovak children's speech dataset, which originally included recordings from the TV show "Táraninky", by adding new data from the radio show "Rozhlasové Leporelo". This enriched the dataset with more diverse speech patterns and intonations, creating a more robust resource for ASR training.

The experimental part of the thesis applies and compares various training strategies using the expanded dataset to improve ASR model performance. The best-performing model was integrated into a web application that allows users to transcribe Slovak children's speech by recording or uploading audio. This aims to enhance ASR technology for Slovak, where resources are limited.

Results

Training on the expanded dataset significantly improved the model's performance. Additionally, the trained models were used to correct manual transcriptions from previous studies, further refining the dataset. The final model was evaluated using this corrected test set, showing a noticeable improvement in accuracy.

The model's performance was further boosted by integrating a language model developed at the Technical University of Košice (TUKE). The fine-tuned XLS-R model achieved a Word Error Rate (WER) of 26.48% without a language model and 16.54% with one, demonstrating that adding new data still provides significant performance gains and further expansions could continue to improve the model. The best-performing model, when paired with this language model, was deployed in a web application built using the Flask framework. This application offers a simple interface for users to transcribe audio files or live recordings into text. The web application was dockerized and deployed on Google Cloud, where it was thoroughly tested.