On the possibility of using pre-trained ASR-models to assess oral reading exams automatically

Bram Groenhof, Wieke Harmsen, Helmer Strik
Radboud University

Dutch children's reading skills are in decline. One way oral reading skills are measured among primary school students in the Netherlands is the three-minute-exam ('Drie Minuten Toets', DMT). The DMT is time-consuming because teachers must administer it one-on-one, marking word reading correctness in real time. One possible way of alleviating this workload is to use automatic speech recognition (ASR) to aid in the assessment process. Many ASR models struggle with children's speech, but since the DMT only needs a binary correct/incorrect judgment for every word, perfect transcription isn't necessary. Additionally, the ASR-transcriptions can be analysed to obtain diagnostic information about a child's oral reading performance. This information can be utilized by teachers to instruct students based on what type of words or sounds individual students struggle with.

We explored the performance of two state-of-the-art (SOTA) pre-trained ASR-models: Wav2vec2.0 and Whisper. We had them carry out assessments on oral reading word tasks, similar to the DMT, using data from the Children's Oral Reading Corpus (CHOREC). Word lists were read by native Dutch-speaking primary school children aged 6-12 from Flanders and marked by assessors. The judgements of ASR-models and assessors were compared using accuracy, F1-score, and Matthews Correlation Coefficient (MCC) as agreement metrics. Two methods to improve the baseline results were applied: rule-based and similarity-based (using standardized Levenshtein distances).

We found that rule-based improvements obtained the best results for the overall metrics. It involves aspects such as (de)voicing and short versus long vowels. Whisper (accuracy = .54; F1-score = .58; MCC = .54) outperformed wav2vec2.0 (accuracy = .69; F1-score = .39; MCC = .37). The MCC values show that both ASR-models showed mild correlations with assessors.

We conclude that the performance of pre-trained ASR-models, especially Whisper, are promising. We are currently expanding this line of research using recordings of real DMTs. Using the rule-based improvement method, we aim to obtain more detailed diagnostic information from the DMT (e.g., which phonetic aspects the children struggle with).