Computational Linguistics in the Netherlands Journal 14 (2025) 343-364 Submitted 12/2024; Published 07/2025

Exploring the use of pre-trained ASR models for automatic assessment of children's oral reading

Bram Groenhof* BRAM.GROENHOF@RU.NL
Wieke Harmsen* WIEKE.HARMSEN@RU.NL
Helmer Strik* HELMER.STRIK@RU.NL

* *Centre for Language Studies, Radboud University, the Netherlands*

Abstract

Dutch children's reading skills have been declining consistently for many years. Oral reading fluency, a combination of decoding skills and word recognition skills, is a fundamental pre-requisite for one's reading competence. Children's oral reading fluency is often tested through oral word reading tasks, which are time-consuming to carry out as teachers have to administer the tests in a one-on-one setting, in which they have to indicate the word reading correctness on-the-fly. One possible way of alleviating this workload is to use automatic speech recognition (ASR) to aid in the assessment process. A key concern is that many ASR models struggle with children's speech.

We explored the performance of two pre-trained ASR models: Wav2Vec2.0-CGN and Faster-Whisper-v2. We had them carry out correctness judgement on an oral word reading task, using data from the Children's Oral Reading Corpus (CHOREC). This corpus contains oral reading data of word lists from native Dutch-speaking primary school children aged 6-12 from Flanders. We compared the results of the ASR models to those of assessors in CHOREC by using the agreement metrics specificity, recall, accuracy, F1-score, and MCC as agreement metrics. We then used two different methods to improve the baseline results, by post-correcting ASR model correctness judgements using manually defined error categories.

We found that allowing a deviation from the prompt by one error category obtained the best results for the overall metrics. Faster-Whisper-v2 (accuracy = .89; F1-score = .58; MCC = .54) outperformed Wav2Vec2.0 (accuracy = .70; F1-score = .39; MCC = .38). The MCC values show that both ASR models had mild agreement with assessors. We expected the accuracy levels for both models to be lower than the lowest assessor inter-rater accuracy level (.86), but Faster-Whisper-v2 performed better than expected (.89). However, one should be careful in interpreting this result, since the high accuracy scores are partially due to the imbalanced dataset.

We conclude that the performance of standard pre-trained ASR models is promising, but given the current quality of the procedure caution should be exercised in its use. Future research could aim to improve the performance of the whole procedure by e.g. using methods like fine-tuning and validation, and through collaborative research with teachers.

Keywords: ASR, Wav2Vec2.0, Whisper, children's speech, oral reading.

1. Introduction

In the Netherlands, reading comprehension skills of children have been declining for many years (OECD 2023). One prerequisite skill for attaining reading comprehension skills is referred to as decoding skills. Decoding skills have long been noted as important skills to be prolific at for one to become a competent reader (Kendeou et al. 2009, Perfetti and Hogaboam 1975).

Two of the most known tests that aim to test the decoding skills of children are the Klepel-R made by Pearson (van den Bosch et al. 2019, Weijnman 2013) and the three-minute-exam (In Dutch: Drie Minuten Toets, DMT) made by Cito (Cito B.V. 2017). Both of these exams require children to do oral word reading tasks, through the reading of word lists. The DMT is the most widely used example of this in the Netherlands. Dutch children aged 6-12 take this test at least once every academic year. They are assessed using this test until they reach the maximum level or until they finish primary school. Because of its ubiquity in the Netherlands, we will use the DMT

to help illustrate what a typical oral word reading task looks like, and what issues are prevalent in the assessment procedure of it.

Van Til et al. (2018) provide an overview of the administration and assessment process of the DMT. Children receive different word lists that vary in difficulty. They are asked to read all of the words out loud as quickly and as correctly as possible. The test is administered and instantaneously marked by teachers manually using pre-made marking sheets. This is done by hand using pen and paper. If the word is read correctly, it is not marked. If the word is read incorrectly, the teacher makes a note of this. If a child skips a word or gets stuck on a word for five seconds, the word is marked as skipped. In the case of a child getting stuck on a word, the teacher whispers the correct reading (Cito B.V. 2017). The teacher has to conduct the DMT with every child individually, which is time-consuming. Moreover, the feedback that the pupils receive does not go deeper than the overall score, i.e. the number of words the child read correctly in three minutes.

While not all oral word reading tasks use the exact process of the DMT, they do require administration and marking of read out word lists. Thus, it would be fruitful to look into possible ways to improve the assessment process of these types of tasks. This could free up much time for teachers to focus on other aspects of reading, or even for them to look at the results in more detail so that they can provide more feedback to students beyond the single score.

Automatic speech recognition models (ASR models) are potential tools that teachers could use to help in the assessment of oral word reading tasks. ASR models have been improving for many years, but progress for atypical speech has been lacking (Ngueajio and Washington 2022). Children's speech is a form of atypical speech, as it differs from regular adult speech in, for example, acoustic variability (Jain et al. 2023). In recent years, the fact that many ASR models tend to struggle with children's speech has been noted clearly (Feng et al. 2024, Jain et al. 2023, Yeung and Alwan 2018). For children's oral reading ability specifically, interest in using ASR models as tools for recognition and assessment of children's speech for languages other than English has become more prominent (Harmsen et al. 2023, Klebanov et al. 2020, Loukina et al. 2017, Mich et al. 2020, Molenaar et al. 2023, Piton et al. 2023).

This paper will add to this body of knowledge, by focusing on children's speech in the context of oral word reading tasks and the possible use of different pre-trained ASR models in the assessment process of these types of exams. Furthermore, most studies that use ASR models apply them in context-heavy scenarios, such as sentences or stories. Our focus will be on word lists, which is a much more novel and understudied application of ASR models.

Our aim is to examine the feasibility of utilizing commonly used pre-trained ASR models as tools to support teachers in the assessment of children's oral word reading performance in exams that require children to read aloud word lists, akin to the DMT. We use Wav2Vec2.0 (Baevski et al. 2020) and Whisper (Radford et al. 2023) to automatically assess children's oral reading performance. Whisper was chosen because it is considered State-of-the-art (SOTA) at the moment of writing, Wav2Vec2.0 was chosen because of its common usage, though it was considered to be SOTA when it was initially released.

The performance of the ASR models will be measured through agreement metrics, using regular assessor judgements as the ground truth. Additionally, we aim to improve the results by allowing leniency in the ASR models' judgements. Our goal is to find out whether it is possible for these ASR models to automatically assess the oral reading skills of children, so that they can be utilized by teachers to save time and gather more diagnostic information.

2. Background

2.1 Decoding Skills in Oral Word Reading Tasks

We briefly introduced decoding skills and its importance as a prerequisite for reading fluency in the introduction. Nevertheless, it is important to concretely define what decoding skills are to get a better view of what skills oral word reading tasks measure.

Van Til et al. (2018) define oral word reading skills as one's ability to recognize written words quickly and correctly. This is taught to children in roughly two phases. First, the child is taught that words consist of graphemes and that each of them represents a sound in spoken language. This is known as the alphabetical principle: the relationships between graphemes and sounds. In addition, children start to develop their phonemic awareness in this phase. This allows them to read words orally, because phonemic awareness refers to the understanding that spoken words are constructed using phonemes. By the end of the first phase, children are able to read simple words that follow a simple consonant-vowel-consonant structure. In the second phase, children increase the speed at which they are able to read and the phonemic awareness is extended to more complex words. In this phase, children start to develop phonemic proficiency on top of phonemic awareness; they learn how to bend and manipulate the individual sounds to pronounce them more naturally in words. This is done through, for example, co-articulation (Bell, 2023). As children read more, they generally develop their reading skills to the point where they understand written language as much as they do spoken language by the end of primary school (Wentink 1997).

Thus, based on these two phases, there are two essential skills children require to be able to do well on an oral word reading task: decoding skills and word recognition skills. Decoding skills to one's ability to map graphemes (letters) to phonemes (sounds). Word recognition skills refer to the ability to find the meaning (semantics) of the read word. Of course, children already know many words from spoken language before they learn how to read, which aids in word recognition (Van Til et al., 2018). Naturally, a child reads faster the better they are at these processes.

Numerous theoretical models of reading exist that try to represent the role of oral word reading skills, but there is no consensus on the individual importance of either decoding or word recognition skills. Van Til et al. (2018) points out that there is not a single perfect theoretical model, as human behavior is always different from a model representation. They mention that the focus should be on what theoretical models have in common. Three such theoretical models of reading are the Dual-Route Cascaded model (DRC; Coltheart et al., 2001), the triangle model (Harm and Seidenberg 2004), and the Connectionist Dual Process model of reading aloud (CDP++; Perry et al. (2013)). All three theoretical models are mentioned in Van Til et al. (2018) and they are cited as the main computational models for reading (Castles et al. 2018). We will introduce these models very briefly. We will focus on their commonalities to help us define oral reading fluency, to clarify the construct that we try to measure using the ASR models.

The DRC-model states that the process of reading a word happens through one of two routes: phonological or lexical. The mental lexicon plays a big role in this model as well. The lexicon is an internal system where important information about words is stored; including orthographic, phonological, and semantic information. When you read using the phonological route, you first decode each letter of the word that you read. Then, using the phonological and semantic information in the lexicon, the word and its meaning are recognized. When you read using the lexical route, the orthographic information in the lexicon activates all information at once without the need for decoding (Coltheart et al. 2001).

The triangle model uses processing layers that can become active when a word is read. Each of these layers has an in- and output layer. In this model, hidden units are represented by smaller layers. These facilitate more complex connections between the larger layers (semantics, phonology, and orthography). According to this model, part of learning how to read a certain word is to know how much each processing layer should weigh in for specific situations (Chang et al. 2020, Harm and Seidenberg 2004).

The CDP++ model describes a division of labor between lexical and non-lexical processes within a neural network. The model works by using two different routes: a direct route and a route using a hidden layer (Perry et al. 2013).

It is impossible to argue for the support of one of these theoretical models over the other, as numerous studies have shown advantages and disadvantages for each (Perry et al. 2013, Rapcsak et al. 2007, Seidenberg 2005, Woollams et al. 2007). However, there are commonalities which can help us identify the process of oral reading as well as what makes someone successful at oral reading.

Following this, we can define oral reading skills using the same important processes that we mentioned before: decoding skills and word recognition skills An important note is that word recognition is only possible if the word is stored in the mental lexicon (Castles et al. 2018). All three described theoretical models predict that oral reading goes faster and more correctly the more familiar someone is with the letters, clusters of letters, or full words. In the models, this is exemplified through the strength of the representations and connections of its parts (Van Til et al. 2018).

Put together, successfully performing an oral word reading task requires a combination of decoding skills and word recognition skills. A child will do well if they are prolific at mapping graphemes to phonemes successfully in combination with having strong mental representations of the words they are required to read. The more familiar the child is with a certain word or part thereof, the stronger it is established in the lexicon, the more rapid and correctly it can be obtained. A combination of these skills will show in a child's ability to recognise and read aloud words quickly and correctly. We will use the term oral reading fluency to refer to a combination of decoding skills and word recognition skills from this point onward.

2.2 The Current Study

The main research question for the current study is as follows:

> To what extent can commonly used pre-trained ASR models be incorporated to assess oral reading tests made by children automatically?

The relevancy of this question is embedded in the trends of Dutch children's reading skills and ASR models described above. ASR models have been improving for many years, and despite issues with correctly identifying children's speech, their potential as tools for educators cannot be understated (Cleuren et al. 2008, Klebanov et al. 2020). We explore the possibilities of applying commonly used pre-trained ASR models as tools in oral word reading tasks. We develop and utilize a pipeline that generates and uses these ASR models' transcriptions to judge whether children have read words correctly or not (correctness judgements). In doing so, we can assess the validity of these judgements by making a comparison between ASR models' correctness judgements and those of human assessors. Note that we do not intend to test whether ASR models could replace the teacher. We do not advocate for the replacement of teachers and assessors by ASR models in the correctness judgements of oral word reading tasks, but for their use as tools. If the ASR models perform well, the program can be improved upon iteratively so that teachers can use these models to aid them in the assessment process.

2.2.1 ASR Model Selection

Previous studies have described that children's speech is problematic for many ASR models to process correctly, because children's speech is considered atypical when compared to the native adult speech. Children's speech is typically more varied than adult speech and most ASR models are trained on little to no children's speech at all since this data is scarce (Cleuren et al. 2008, Jain et al. 2023). This brings limitations to the possibility of using ASR models to judge children's oral reading ability.

While it is true that ASR models tend to struggle with judging correctness of oral word reading tasks performed by children, it does not insinuate that different assessors are always in agreement

about judgements. In their publication on the Children's Oral Reading Corpus (CHOREC), Cleuren et al. (2008) investigated how consistently the assessors agreed on judgements and found that across all participating schools the inter-rater agreement varied between 86.4% and 99.6%. Harmsen et al. (2023) also looked at inter-rater agreement of teachers assessing native Dutch children's oral readings of word lists. These word lists were taken from the Dutch automatic reading tutor (DART) corpus and were developed to be like those in the DMT. Assessors were instructed to assess children using DMT guidelines. They found moderate agreement between teachers, stating that "for around 40% of the words, less than 80% of the teachers agreed" (Harmsen et al. 2023). A main advantage of using ASR models for judgements is that it will make the same judgements consistently. For this to be successful however, the ASR model must be making these judgements correctly, or it is invalid.

As mentioned previously, ASR models struggle with transcribing children's speech correctly, causing difficulties in using them for correctness judgements. Despite this, there are many studies that show hopeful results for its capabilities; both for pre-trained and fine-tuned models. In their paper, Piton et al. (2023) explore the possibilities of commercially developed pre-trained ASR models (IBM Watson) to generate transcriptions for analysis of French and Italian children's speech. While they conclude that these ASR models themselves do not provide fine-grained analysis of children's speech themselves, they also speak positively of the possibilities for using the transcripts to classify children's speech as correct or incorrect.

If we turn to fine-tuned ASR models' performances, the results are much more optimistic. First, the previous findings for languages other than Dutch. Klebanov et al. (2020) created an app for children to use for oral reading using ASR. They state that the ASR transcriptions proved to be very useful when they scored the recordings of children, not needing orthographic transcriptions after validation the ASR model on external corpora only. Bernstein et al. (2017) developed an app using a hybrid-based ASR model for children. The purpose of this app was to explore the possibility of self-administered oral reading tests. They showed that children were able to self-administer the oral reading test quite well: the words correct per minute (WCPM) scores from automatic (ASR) assessments correlated highly with those of teachers. Mich et al. (2020) developed a web application for assessment of reading skills of Italian Children. They used a fine-tuned Kaldi model based on words that they knew the children were going to be assessed on. They conclude that teachers can use their system for assessment of children's oral reading skills. Finally, Jain et al. (2023) illustrates how the fine-tuning of models for children's speech specifically can improve an ASR models performance on recognizing children's speech. They showed that the performance of a Whisper-based ASR model's performance, which consisted of adult speech, would improve significantly when fine-tuning the model using children's speech data. They take special note of the improvements that were made when they included linguistically diverse correct and erroneous readings such as accented speech.

For Dutch, the results are similar. Molenaar et al. (2023) made use of four Kaldi-based models and two Whisper-based models to assess Dutch children's oral reading accuracy. They found that the best performing model was a Kaldi-based one that had a language model that contained both prompts and orthographic transcriptions. This would imply that here, the orthographic transcriptions are crucial. For DMT-like tasks specifically, Harmsen et al. (2023) evaluated the performance of three ASR models (one based on Kaldi and two based on Whisper) on child speech data from the DART corpus. They found that the ASR model based on Whisper performed best, meaning that it was the best at predicting a teacher majority vote; it performed the most similar to teachers. This best-performing model had two important characteristics. First, it was able to produce pseudo- and non-words. Second, the model was provided with the prompts for the correct word readings. These studies show the potential usefulness of ASR transcriptions for assessment of children's oral reading skills.

3. Method

3.1 CHOREC (Children's Oral Reading Corpus)

The data used for this paper came from the CHOREC corpus (Cleuren et al. 2008). We developed two pipelines for this paper in Python 3.11 (Foundation 2022). We included two pipelines, as they differ slightly based on which ASR model is used. Pre-processing, obtaining the baseline results, and doing both experiments can be replicated using these scripts. (Groenhof 2024a, Groenhof 2024b).

In total, the CHOREC corpus contains oral speech recordings of 400 Flemish children who speak Dutch as their native language. At the time of recording, the children were elementary school students attending either regular elementary schools (N = 274) or elementary schools for children with specific learning disabilities (N = 126). All children were between 6 and 12 years old.

All children in the CHOREC corpus performed oral word reading tasks, making it suitable for our research purposes. However, not all the participant data is relevant for our research. As we discussed in the background section, we define oral reading fluency as a combination of decoding skills and word recognition skills. CHOREC contains data from two types of oral word reading tasks: real word reading tasks (RWRT) and pseudoword reading tasks (PWRT). If we were to include the PWRT, the data from this task is only relevant for decoding skills, but not word recognition skills. Pseudowords do not actually exist, and are not represented in the lexicon as a result (Chuang et al. 2021). Therefore, only the data in CHOREC for which children performed the real RWRT is relevant.

Oral word reading tasks are often assessed by assessors who make correctness judgements for each word a child reads; a word is either read correctly or incorrectly. This holds true for the DMT as well (Cito B.V. 2017). This means that the task for the ASR models is binary classification: either a word was read correctly (0) or it was read incorrectly (1). CHOREC does not contain orthographic transcriptions. However, there is a reading error layer. If the teacher judged the word as read correctly, they did not annotate anything. If they judged the word as read incorrectly, they used codes representing reading errors provided in the annotation protocol (Cleuren et al. 2008). According to the teachers, the word was read correctly when no annotation was made (0) and read incorrectly when one or more error codes were made (1). All annotations were made manually by the teachers. Not all files were annotated with a reading error layer. For this reason, the audio recordings of 15 children had to be excluded from our research.

CHOREC contains three types of word lists, 1LG, 2LG, and 3+4LG, consisting of 40 words each. The 1LG, 2LG, and 3+4LG lists each contained only 1-syllable, 2-syllable, and 3- or 4-syllable words respectively (Cleuren et al. 2008). Not all children read all word lists, 1LG (N = 377) was read most, followed by 2LG (N = 359), and 3+4LG (N = 320). Each word list is more difficult than the other, which explains why the fewest number of children read the 3+4LG list. This means that the words in the easier words are overrepresented.

3.2 Defining Validation and Test Datasets

We assessed the quality of all recordings by calculating the signal-to-noise ratio (SNR) of all recordings using a Python script with Librosa (McFee et al. 2015). We did this to ensure that no poor-quality audio recordings would be present in the dataset. While there is no consensus on what is considered to be a high value for SNR, 20dB is often used as a reference for high SNR values (Hu et al. 2020, Sadeghi et al. 2024). We used this as an initial threshold. Audio files that had an SNR below 20dB were listened to manually to check if there was a lot of background noise. If this was the case, the recording was excluded from the research. If there was not a lot of noise, the recording remained as part of the data.

The justification for the manual check is that the creators of CHOREC mention that all of their data was recorded in a controlled environment with good equipment (Cleuren et al. 2008), which should prevent any recordings from being of poor quality.

Once the data had been gathered, we separated the full dataset into a validation and test set. This is a well-known practice within machine learning. Usually, a model is trained and/or fine-tuned using a training set. The training set is used to explore how modifications to the model affect the results. No modifications to the model are allowed to be made once it is applied to the test set (Galarnyk 2022). While we did not train or fine-tune the ASR models, we did intend to improve upon the ASR models' baseline results. For this reason, we defined a validation set instead of a training set. This allowed us to do error analysis on part of the data, while still leaving data to generate final results on.

Table 1 shows an overview of speaker characteristics in the validation and test sets. For some speakers, the school year annotations were missing. For our research, this is not problematic because we did not intend to look at the results split by school year. We ensured that the validation set was balanced, hence the students for whom the school year annotations were missing were all part of the test set. There was an exact even split of gender for each school year. The years 2, 3, and 4 account for most data in CHOREC, which is why more participants from these years were selected. Overall, the validation set contained 27.18% of all relevant data in CHOREC, leaving 82.82% for the test set. While it is uncommon to have this large of a validation set, we chose to do this because we did manual error analysis instead of machine learning. We justified this choice in two ways. First, manual error analysis is time-consuming. Had we defined a larger validation set, we would have had to spend far more time on this process for diminishing returns. Second, a larger test set allowed us to draw conclusions from the results that were more generalizable than a smaller test set would be.

Table 1: Description of participants in validation and test sets

School year	**Gender**	**Number of participants in validation set (N)**	**Number of participants in test set (N)**
1	Female	2	1
1	Male	2	1
2	Female	15	19
2	Male	15	27
3	Female	15	18
3	Male	15	21
4	Female	15	19
4	Male	15	23
5/6	Female	1	2
5/6	Male	1	3
Unknown	Female	0	23
Unknown	Male	0	67
Total		96	224

3.3 Wav2Vec2.0-CGN and Faster-Whisper-v2

For our research, two pre-trained ASR models will be used. The first model will be referred to as Wav2Vec2.0-CGN (GroNLP 2023). This is a pre-trained on the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) (Taalunie 2014). The second model we will use will be referred to as Faster-Whisper-v2. This is a modified version of Whisper (Klein 2023, Radford et al. 2023).

We mentioned earlier that the performance of end-to-end models are generally better than hybrid models (Parikh et al. 2023, Shraddha et al. 2022). Thus, it is no surprise that many papers of recent years have been using either Wav2Vec2.0, (Ahn et al. 2024, Baevski et al. 2020), Whisper (Jain et al. 2024, Van der Klis et al. 2023), or both (Fan et al. 2024) as representatives of current SOTA ASR models. However, we will only consider Whisper a SOTA ASR model for this paper, referring to Wav2Vec2.0 as commonly used. Its more recent release and better performance on

ASR tasks relating to children's speech compared to Wav2Vec2.0 justifies this distinction (Fan et al. 2024, Van Gompel 2023).

For Wav2Vec2.0, we opted to use the model pre-trained on CGN (GroNLP 2023). This choice was made because, to our knowledge, it is the largest Wav2Vec2.0 model pre-trained on Dutch speech. CGN has seen widespread use when dealing with Dutch speech for ASR purposes (Dyck et al. 2021, Poncelet and Van Hamme 2023).

For Whisper, we opted for a modified version called Faster-Whisper-v2 (Klein 2023). Earlier research, which compared the word error rate (WER) of different ASR models on Dutch children's read speech from the JASMIN corpus (Taalunie 2008), showed that the best performing models were as follows: Faster-Whisper v2 w/VAD, Whisper v2 w/VAD, and Faster-Whisper-v2. Their respective WER values were: 19.1%, 20.1%, and 20.3% (Van Gompel 2023). VAD stands for voice activity detector and it is used to filter out parts of audio files with no speech. However, when we tried to use Faster-Whisper v2 w/VAD and Whisper v2 w/VAD, the transcriptions were often incomplete. This was too problematic to use, because up to half of the recording could be missing. For this reason, we opted for Faster-Whisper-v2.

3.4 Alignment

ASR models generate a transcription of what is said from an audio file. For our purposes, the ASR transcriptions must be compared to the prompt. Based on this comparison, we can use the ASR transcription to judge if each word was read correctly or not. For this to be possible, the ASR transcription and the prompts must be aligned. A child's attempt at reading a word can only be judged if the correct part of the ASR transcription is looked at for the corresponding prompt. A common way of doing this is through the use of forced aligners such as SCLITE or ADAGT (Harmsen et al. 2024, National Institute of Standards and Technology 2021). In this research, we use ADAGT for alignment, since it provides two-way alignment: forwards and backwards. Children often stutter and restart words. The backwards alignment that ADAGT offers aligns the final reading attempt more consistently with the prompt than regular forwards alignment alone.

3.5 Measurements for the Performance of ASR models

In order to assess the performance of ASR models, we needed to select metrics to represent their performance. We considered the assessors in CHOREC to be the ground truth in this paper, because assessors are often used in real life for oral word reading tasks including the DMT (Cito B.V. 2017). The less the correctness judgements based on the ASR model's transcription deviated from the assessors in CHOREC, the better we considered their performance. However, previous work that did not use WCPM (Cleuren et al. 2008, Harmsen et al. 2023). Since we intended to compare our results directly to previous work, we opted for measures based on a confusion matrix instead.

A confusion matrix can be defined as a contingency table which summarizes the performance of a binary classifier. It does this by comparing the predictive labels, the ASR model's judgements, to the actual labels of the data, the assessors' judgements. In doing so, the data is categorized into four key metrics based on whether the predictive and actual labels align (Stehman 1997). Table 2 explains the meaning of these four metrics applied to the data of this paper. A 'negative' (0) corresponds to a word that was read correctly and a 'positive' (1) corresponds to a word that was read incorrectly.

Table 2: Overview of possible outcomes in a confusion matrix

Outcome	**Assessors judged word as**	**ASR model judged word as**
True negative (TN)	Correctly read	Correctly read
True positive (TP)	Incorrectly read	Incorrectly read
False negative (FN)	Incorrectly read	Correctly read
False positive (FP)	Correctly read	Incorrectly read

Table 3 shows how we obtained the judgements from assessors in CHOREC and the ASR models. For the assessors, a word was judged as correctly read (marked as a "0") when the reading error layer in the annotations was empty (i.e., there was no error). All remaining words were marked as an incorrectly read word (marked as a "1"), as any note in the reading error layer indicates a reading error according to the assessors. For the ASR model, the transcription was aligned to the prompt using ADAGT first (Harmsen et al. 2024). Once the transcription was aligned, every transcribed word was compared to its prompt. If they were identical, it was marked as a correctly read word (marked as a "0"). In all other cases it was marked as an incorrectly read word (marked as a "1"). Following this, we compared the correctness judgements of the assessors and ASR model gives us one of the four possible outcomes described in Table 3.

Table 3: Examples of outcomes based on assessor correctness judgements

Prompt	**Reading error**	**Assessor judgement**	**ASR model transcription**	**ASR model judgement**	**Outcome**
groen		0	groen	0	TN
groen	13	1	groo	1	TP
groen		0	krom	1	FP
groen	13	1	groen	0	FN

After we obtained the outcome of all read words, we calculated the most relevant agreement metrics from (Chicco and Jurman 2023). An overview of these metrics, how they are calculated, and an explanation of what they represent is provided in Table 4 . All of these metrics were important for our research, as they enable us to interpret the results by looking at different aspects of agreeability between assessors and ASR models. Accuracy and F1-score provided us overall metrics that show the overall performance of the selected ASR model. Specificity and recall were chosen because they represent the performance of the ASR model for cases where the word was judged as being read correctly and incorrectly by assessors respectively. They allowed us to assess the performance at a more detailed level. The higher the value for either, the better the ASR model performs. However, an increase in specificity should not lead to a decrease in recall or vice versa.

A special note must be made about the agreement metric precision. Its inclusion in Table 4 is only because it is used to calculate the F1-score. We consciously chose to not use it as a metric by itself. As discussed above, specificity and recall are more suited for our purposes than precision. For this reason, we did not report on precision in the results section.

Table 4: Explanation of confusion matrix metrics used in this paper (Chicco & Jurman, 2020, 2023).

Metric	Formula	Explanation
Accuracy	(TN+TP)/(TN+ TP+FN+FP)	Proportion of words that were judged in the same way by both the ASR model and the assessors.
Precision, also referred to as positive predictive value (PPV)	TP/(TP+FP)	Proportion of words that were judged as incorrectly read by the ASR model that were also judged as incorrectly read by the assessors.
Specificity, also referred to as true negative rate (TNR)	TN/(TN+FP)	Proportion of words that were judged as incorrectly read by the assessors that were also judged as incorrectly read by the ASR model.
Recall, also referred to as true positive rate (TPR) or sensitivity	TP/(TP+FN)	Proportion of words that were judged as correctly read by the assessors that were also judged as correctly read by the ASR model.
F1-score	2*(Precision*Recall)/ (Precision+Recall)	A measure of predictive performance, representing both precision and recall in a single metric.
Matthew's Correlation Coefficient (MCC)	((TP*TN)-(FP*FN))/ sqrt((TP+FP)*(TP+FN)* (TN+FP)*(TN+FN))	A measure of predictive performance. For imbalanced datasets, MCC is more appropriate than accuracy and F1-score.

MCC will be used as the overall measure of agreement between ASR models and assessors in this paper, because the ASR models had to perform a binary classification task and data in CHOREC is imbalanced. This value will indicate the performance of a given ASR model: the closer to 1 this correlation gets, the better the ASR model performs. The inclusion of F1-score and accuracy was a deliberate choice. It allowed for interpretation and comparability with previous studies. Most importantly, Cleuren et al. (2008) used accuracy to represent the inter-rater agreement between assessors. Together, specificity, recall, accuracy, F1-score, and MCC represent the agreement between ASR models' and assessor correctness judgements in this paper.

3.6 Improvements Using Error Categories

After obtaining the baseline results, we explored in what way and how much the performance of ASR models could be improved. For the baseline results, a word was only judged as read correctly by the ASR model if its hypothesis was identical to the prompt. Any deviation led to it being judged as read incorrectly. It has been shown that assessors can still assess a word as being read correctly when the reading deviates from the prompt in certain cases (Harmsen et al. 2023). Because of this, we postulated that if we allowed the ASR models to be more lenient, the results would improve.

First, we performed error analysis to gain an understanding of what types of errors were commonly made in the baseline results. We compared the ASR transcriptions to the prompts in cases where the judgement differed between ASR model and assessors and sorted them by frequency. Table 5 shows an example of such a case. This is what we will refer to as a confusion pair from this point onwards (Tillemans 2007). In this example, it would mean that there were 85 instances where the word "cola" was transcribed as "kola" by the ASR model, causing it to be judged as read incorrectly when assessors judged it as read correctly.

Table 5: Example of a confusion pairs based on which we defined error categories

Prompt	**ASR model transcription**	**Assessor judgement judgement**	**ASR model judgement**	**Occurrences (N)**
cola	kola	Correct	Incorrect	85
cola	Koola	Correct	Incorrect	3

By looking at the most frequently occurring confusion pairs, we categorized the types of errors made by the ASR model in the validation set. The example in the first row of Table 5 was defined as the error category "k/c confusion". This use of confusion pairs is a common error analysis approach for obtaining a better understanding of the errors (Hussein et al. 2021, Prasad and Jyothi 2020, Tejedor-García et al. 2022). We stopped trying to define new rules when the frequency of uncategorized confusion pairs got lower and we had to define extremely specific rules instead of general ones.

After we defined the error categories through these rules, we used them to change ASR model's judgements in post-processing. If the difference in prompt and hypothesis was only due to one of these error categories, we changed the ASR model's correctness judgement from incorrectly read to correctly read. This allowed us to experiment with the error categories to see the effects on the agreement metrics in the validation set.

However, we noticed that this approach was imperfect. Only one error category could be checked for at a time. Looking at Table 5, the first row would fall under the error category "substitution k/c". The second row also contains this error, but was not found using this method. This is because it occurs together with another error category we would define later: "substitution long/short vowels". To account for cases in which the ASR model's transcription deviated from the prompt across multiple error categories, further analysis was conducted to determine whether this would lead to substantial improvements.

Because of this, we also generated the results for when we allowed words to be judged as read correctly where this occurs. We therefore have two sets of results on top of the baseline results: one applying error categories in isolation and one applying error categories simultaneously. While this latter method did not provide improved results, it does provide insight into how changing the ASR model's correctness judgements affect agreement metrics specifically for imbalanced datasets in which the vast majority of words are read correctly.

4. Results

4.1 Audio Quality

First, the results of the SNR-analysis showed that almost all recordings have an SNR-value of at least 20dB (N = 926, M = 32.07, SD = 5.80). We checked recordings with SNR-values under 20dB (N = 11) manually to assess the audio quality, none were judged as having poor audio quality. No participants were excluded from the results.

4.2 Baseline Results

In order to see how well the ASR models performed on making correctness judgements, we assessed them by calculating agreement metrics between them and assessor judgements. The more alike their judgements were, the higher the values for the agreement metrics. Table 6 shows how often each confusion matrix metric was found in the results separated by ASR model. Table 7 shows the agreement metrics that were calculated from this.

Table 6: Confusion matrix metrics for baseline results

ASR model	**Dataset**	**TN (%)**	**TP (%)**	**FN (%)**	**FP (%)**
Wav2Vec2.0-CGN	Validation	55.28	4.28	0.73	39.71
Wav2Vec2.0-CGN	Test	**47.39**	**9.74**	**0.64**	**42.22**
Average		51.34	7.01	0.69	40.97
Faster-Whisper-v2	Validation	83.27	3.07	2.01	11.66
Faster-Whisper-v2	Test	**76.5**	**7.62**	**2.76**	**13.12**
Average		79.89	5.36	2.39	12.39

Table 7: Agreement metrics for baseline results

ASR model	**Dataset**	**Specificity**	**Recall**	**Accuracy**	**F1-score**	**MCC**
Wav2Vec2.0-CGN	Validation	.58	.85	.60	.18	.19
Wav2Vec2.0-CGN	Test	**.53**	**.94**	**.57**	**.31**	**.29**
Faster-Whisper-v2	Validation	.88	.60	.86	.31	.30
Faster-Whisper-v2	Test	**.85**	**.73**	**.84**	**.49**	**.44**

This data shows that Wav2Vec2.0 performs worse than Faster-Whisper-v2 overall. While we consider MCC the most important overall agreement metric, Faster-Whisper-v2 outperformed Wav2Vec2.0-CGN for all three overall agreement metrics: accuracy (.84 vs. .57), F1-score (.49 vs. .31), and MCC (.44 vs. .29). Even though the test dataset results are what we will consider final, the same is true for the validation dataset: accuracy (.86 vs. .60), F1-score (.31 vs. .18), and MCC (.30 vs. .19). This means that when we use these ASR models out-of-the-box, without any post-processing, Faster-Whisper-v2 is better at making correctness judgements for this oral word reading task than Wav2Vec2.0-CGN.

A large contributing factor to Wav2Vec2.0-CGN's worse performance is its tendency to generate fewer TNs (47.39% vs. 76,5%) and FNs (0.64% vs. 2.74%), while generating more TPs (9.74% vs. 7.62%) and FPs (42.22% vs. 13.12%) than Faster-Whisper-v2. Once again, this is true for the validation set as well for TNs (55.28% vs. 83.27%), FNs (0.73% vs. 2.01%), TPs (4.28% vs 3.07%) and FPs (39.71% vs. 11.66%). Because of these tendencies, it is no surprise that Wav2Vec2.0-CGN had higher recall values (.94 vs. .73 in the test dataset and .85 vs. .60 in the validation dataset) since obtaining a high value requires a high number of TPs and a low number of FNs. Similarly, Faster-Whisper-v2 had higher specificity values (.85 vs. .53 for the test dataset and .88 vs. .58 for the validation dataset) as obtaining a high value requires a high number of TNs and a lower number of FPs. Seeing as the overall metrics are much better for Faster-Whisper-v2 than Wav2Vec2.0-CGN, having a higher specificity value is more important than having a higher recall value for the CHOREC dataset. In other words, there are more words that were read correctly than read incorrectly according to assessors. We will return to this point in the overall results.

4.3 Error Categories

To get a better understanding of the types of errors made by the ASR models, we analyzed the errors and attempted to group them into error categories. We did this by looking at confusion pairs sorted by frequency and attempting to find patterns therein. Of course, we only did this for the validation set as the test set only functioned as a way to test our final improvements. Our goal was to introduce leniency in the ASR models' correctness judgements to help reduce the number of FPs, as these were far more prominent in the baseline results than FNs for both models. Table 8 shows an overview of the error categories that we defined. These will be used in the remainder of the results section for error analyses.

Table 8: Overview of error categories

Error category name	**Example prompt**	**Example ASR output**	**Explanation**
Insertion spaces	Ruziemaken (to argue)	Ruzie maken	Addition of one or more spaces into the prompt
Insertion	Dichtbij (close)	Dichtsbij	Addition of a letter that is not part of the prompt.
Deletion final	Huis (house)	Hui	Removal of final letter
Deletion liquids	Groei (growth)	Goei	Removal of a liquid inside a consonant cluster
Substitution oe/oo	Groen (green)	Groon	Replacement of 'oe' by 'oo' or vice versa
Substitution k/c	Kleuren (colors)	Cleuren	Replacement of 'k' by 'c' or vice versa
Substitution au/ou	Auto (car)	Outo	Replacement of 'au' by 'ou' or vice versa
Substitution i/y	Reis (travel)	Reys	Replacement of 'i' by 'y' or vice versa
Substitution nasals	Groen (green)	Groem	Replacement of a nasal by a different nasal
Substitution double/single consonants	Stoppen (to stop)	Stopen	Replacement of a double consonant by a single one or vice versa
Substitution long/short vowels	Feest (party)	Fest	Replacement of long vowels by a short one or vice versa
Substitution fricative voice	Zacht (soft)	Sacht	Replacement of a voiced fricative by a voiceless one or vice versa.
Substitution plosive voice	Duur (duration)	Tuur	Replacement of a voiced plosive by a voiceless one or vice versa
Ch confusions	Chocolade (chocolate)	Shocolade	Replacement or deletion of "ch".

Table 9 shows the total number of errors that the error categories represent in the validation set. These categories caught the greatest number of errors in the validation set we could find. We found

that the error categories shown in Table 9 accounted for 34.13% of the FPs in wav2vec2.0-CGN's and 42.98% of FPs in faster-whisper-v2's validation sets.

The most frequently found error categories were the same for both Wav2Vec2.0-CGN and Faster-Whisper-v2: Ch confusions (7.68% and 15.25%), substitution long/short vowels (6.47% and 3.89%), insertion spaces (5.22% and 9.12%), and substitution plosive voice (3.66% and 4.71%) were found to represent the largest part of the total errors. We can also that some error categories were only found for Wav2Vec2.0-CGN: Deletion final (1.62%), substitution k/c (1.21%), and substitution i/y 0.72%), which could have contributed to the worse overall performance of Wav2Vec2.0-CGN in the baseline results.

Table 9: Error category distribution for the validation datasets

Error category	Wav2vec2.0-CGN		Faster-Whisper-v2	
	Frequency (N)	Part of total errors (%)	Frequency (N)	Part of total errors (%)
Ch confusions	350	7.68	204	15.25
Substitution long/short vowels	295	6.47	52	3.89
Insertion spaces	238	5.22	122	9.12
Substitution plosive voice	167	3.66	63	4.71
Deletion final	74	1.62	0	0
Substitution oe/oo	69	1.51	4	0.30
Substitution double/single consonants	64	1.40	17	1.27
Substitution k/c	55	1.21	0	0
Insertion [n]	48	1.05	95	7.10
Substitution nasals	45	1.00	3	0.22
Substitution au/ou	41	1.00	2	0.15
Substitution fricative voice	36	0.80	1	0.08
Deletion liquids	33	0.79	12	0.90
Substitution i/y	11	0.72	0	0
Total	1526	34.13	575	42.98

4.4 Error Category Application Results

Figures 1 and 2 show how the post-correction of ASR model correctness judgements based on identified error categories affected the metrics compared to the baseline results. Figure 1 shows the change in confusion matrix metrics, and Figure 2 the changes in agreement metrics. In these figures, we can observed that the baseline results are mostly improved by application of error categories in isolation. In addition, the performance of Wav2Vec2.0-CGN increased more than that of Faster-Whisper-v2.

Applying the rules in isolation increased the percentage of TNs and reduced the percentage of FPs. Table 10 shows the best obtained results for each ASR model. From this table, we can see that this resulted in higher values for all agreement metrics but recall for all models. Most importantly, the overall agreement metrics for the test datasets show improvements over the baseline results. Faster-Whisper-v2's accuracy, F1-score, and MCC increased by .05, .09, and .10, while they increased by .12, .08, and .09 for Wav2Vec2.0-CGN respectively.

Despite the larger improvements, Wav2Vec2.0-CGN did not perform as well as Faster-Whisper-v2. It only has a better recall value than Faster-Whisper-v2 (.94 vs. .72). Specificity (.67 vs. .91), accuracy (.69 vs. .89), F1-score (.39 vs. 58), and MCC (.38 vs. .54) were all lower for Wav2Vec2.0-CGN than Faster-Whisper-v2.

Allowing simultaneous application of error categories did not improve the results. For Wav2Vec2.0-CNG's test set it reduced performance for all metrics to a point below the baseline results. For Faster-Whisper-v2's test set, the same can only be said for recall, all other metrics show improvements. Compared to the isolated application of error categories, accuracy (.88 vs. .89) was lower. Specificity (.93 vs. .85), F1-score (.62 vs. .49), and MCC (.55 vs. .44) all reached higher values. However, recall (.60 vs. .73) was lower than the baseline value.

Despite these improvements for Faster-Whisper-v2, this is only true for the test dataset. The validation dataset shows a performance worse than the baseline for all metrics for Faster-Whisper-v2 as well. The reason seems to be the drop in TNs and subsequent increase in FNs. Faster-Whisper-v2's test dataset is the sole exception where this does not happen.

Figure 1: Changes in confusion matrix metrics after applying error categories

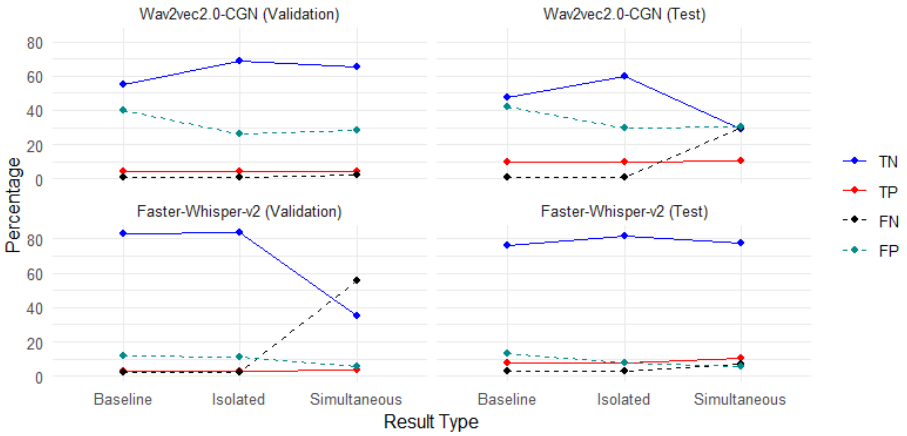

Figure 2: Changes in agreement metrics after applying error categories

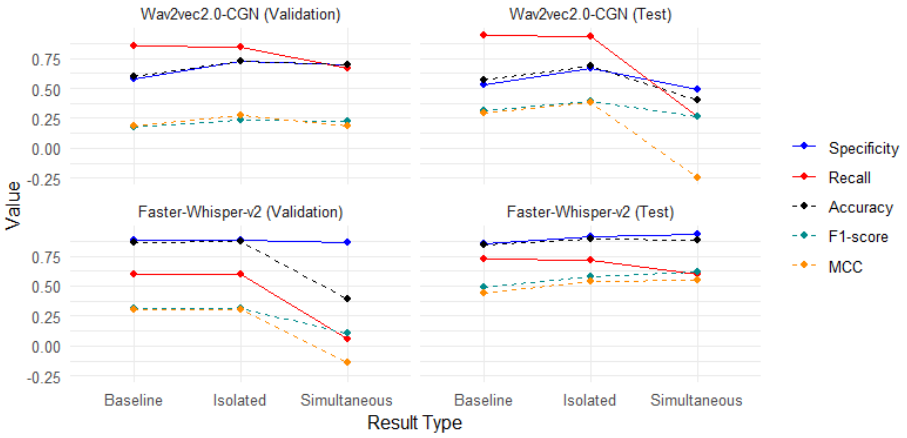

Table 10: Best obtained results for test datasets, change compared to baseline results given in parentheses

ASR Model	Dataset	Specifity (change)	Recall (change)	Accuracy (change)	F1-score (change)	MCC (change)
Wav2Vec2.0-CGN	Test	.67 (.14)	.94 (0)	.69 (.12)	.39 (.08)	.38 (.09)
Faster-Whisper-v2	Test	.91 (.06)	.72 (-.01)	.89 (.05)	.58 (.09)	.50 (.10)

5. Discussion

In the current study, we investigated to what extent commonly used pre-trained ASR models can be employed to automate the assessment of oral word reading tests made by children. To this end, we developed an ASR-based procedure that can automatically assess whether a word is read correctly. We investigated the performance of two different ASR models for this task, Wav2Vec2.0-CGN and Faster-Whisper-v2. In addition, we analyzed the recognition errors of the ASR models and investigated whether rule-based correction of these errors in a post-processing procedure improved the performance.

We found that applying error categories in isolation resulted in higher agreement between the automatic assessments and the human assessments. On the test set, the agreement was higher for Faster-Whisper-v2 (MCC = .54) than for Wav2Vec2.0-CGN (MCC = .38).In addition, we found that the simultaneous application of error categories led to slightly stronger agreement for Faster-Whisper-v2 (MCC = .55) on the test set. However, simultaneous application of error categories did not result in robust increases of agreement with human assessors, since these results were not found for the validation set. It is unclear why this is the case. It might have been because of dividing the data in validation and test sets. In future research, it might be good to pay more attention to intermediate results when applying this method., since in the current pipeline we are only able to inspect the results when all error categories are applied simultaneously. Perhaps one error category, or a specific combination of error categories, led to these noticeably different results. Future research could aim to improve this method by trying to find out if there is a 'sweet spot' at which the ideal number and types of error categories are enabled.

Faster-Whisper-v2 always outperformed Wav2Vec2.0-CGN. It seems that Faster-Whisper-v2 is more suited for children's speech, especially for word lists. This might be because the data in CHOREC is imbalanced, only 8.97% of the words were judged as being read incorrectly by assessors (Cleuren et al. 2008). The tendency of Faster-Whisper-v2 to produce more TNs and FNs, while producing fewer TPs and FPs than Wav2Vec2.0-CGN works in its favor in this context. As touched upon in the results section, these tendencies lead to higher specificity values for Faster-Whisper-v2 and higher recall values for Wav2Vec2.0-CGN. The vast majority of words are judged as read correctly by the assessors. This results in specificity being more important for the CHOREC dataset than recall, since it is much harder to generate FNs when there are so few words read incorrectly in the first place.

Furthermore, the imbalance of the data in CHOREC proves that having a more robust agreement metric, such as F1-score or preferably MCC, is crucial. If we had looked at just accuracy, we could have based conclusions on the fact that Faster-Whisper-v2's agreement (.89) with the assessors is higher than the inter-rater agreement in CHOREC (.86) when applying error categories isolation alone. The high accuracy scores are partially due to the imbalanced dataset. For instance, an even higher level of accuracy (.90) can be reached by simply judging everything as correctly read. F1-score and MCC are affected far less by this imbalance in the data, which is why we strongly suggest that future research always makes use of these more robust measures.

While the results that we obtained with commonly used pre-trained ASR models are in some ways better than expected, caution should be exercised in the use of the procedures. Still, these

procedures can be valuable tools for teachers. Three of the most prominent theoretical models for oral reading proficiency, the DRC-model, the triangle model, and the CDP++ model, all predict that children's oral reading skills will increase as they become more familiar with the letters, clusters of letters, or full words that they are asked to read (Castles et al. 2018, Coltheart et al. 2001, Harm and Seidenberg 2004, Perry et al. 2013). While the commonly used pre-trained ASR models can not yet detect completely correct which types of sounds or words the pupils struggle with, the results can still provide the teachers with useful information on frequently made errors, which would allow them to focus on these sounds and words specifically. In this way, pupils become more familiar with the letters, clusters of letters, or full words that they struggle with reading leading to improved reading skills.

From a pedagogical viewpoint, it is far more important to avoid FPs than FNs, as FPs can lead to unneeded frustration and stress for learners (Cucchiarini et al. 2009). This is what we observe for Faster-Whisper-v2, it has a tendency to generate fewer FPs than FNs. Put otherwise, a high specificity is more important than a high recall. The fact that we find a specificity of .91 (see 10) is thus an interesting result, indicating Faster-Whisper-v2 shows promise for oral word reading tasks such as the RWRT in CHOREC or DMT in general.

5.1 Future Research

Bernstein et al. (2017) showed that self-administered oral reading assessment is feasible for children as young as five. In future studies, researchers could collaborate with teachers and assessors to see if the results from the ASR models are usable to them, so that these accuracy judgements can partly be automated. One possible way in which this could be explored is through a large-scale experimental study in which one group of teachers use the results of ASR models to help them assess oral word reading texts for children. These findings could then be compared to the results of a group of teachers using a procedure without ASR. The results could be compared to see if the use of the results of ASR models leads to valid judgements. Furthermore, the teachers who used the results of ASR models could share their experiences. This could then help researchers find the most suitable way of implementing ASR for teachers.

Finally, we underline an issue which remains a challenge for the application of ASR models is alignment. The use of ADAGT for alignment in this paper was due to the availability of both forward and backward alignment. This proved beneficial, as either method would sometimes be more successful. Despite this, as Table 11 shows, there were many instances where neither of these alignment directions could correctly align the ASR output with the prompt. The reason this could be problematic is that the alignment may ascribe an attempt at reading a specific word to the wrong prompt. In the recordings of CHOREC, children read the words unnaturally; they often tried to read as fast as possible. This led to stuttering, mumbling, and restarts. While ADAGT's backwards alignment helped, this still made it difficult to align the words to the prompts correctly. In future, researchers should be open to testing new or improved existing alignment algorithms, as these could lead to more valid and reliable alignment.

Table 11: Example of ADAGT-alignment going wrong

Prompt	**ADAGT forward alignment**	**ADAGT backward alignment**
Appel (apple)	Spelen (to play)	Spelen (to play)
Auto (car)	Ouders (parents)	Ouders schilder (parents painter)

6. Conclusion

The aim of this study was to investigate how well commonly used pre-trained ASR models performed at making correctness judgements for oral reading tests made by children. If it is possible to use these

ASR models for this purpose, they could potentially be incorporated in practice and aid teachers in the assessment process of these types of exams. Furthermore, we aimed to explore the use of commonly used pre-trained ASR models in the context of oral word reading tasks because this is an understudied context, as most studies applying ASR models on children's speech focus on sentences and/or stories.

For the commonly used, 'out-of-the-box' pre-trained ASR models that we used in the current study, the results were sometimes already better than we expected beforehand. For instance, we found an accuracy of .89, but noted that this high accuracy might be partially due to the imbalanced data. The metrics F1 and MCC are more robust to imbalanced data, and they provide a more realistic picture: the ASR models show mild agreement with the assessors. One thus has to be cautious to apply these procedures. Faster-Whisper-v2 shows potential, and esp. a specificity of .91 is an interesting result. If one takes into account the limitations and possibilities of these procedures, and if it is possible to improve the performance of such procedures, they can be useful to assist teachers, e.g. to provide them with additional information. It then becomes interesting to study how these procedures can be used to assist teachers. Obviously, collaboration with teachers is crucial for this.

References

Ahn, T., Y. Hong, Y. Im, D. H. Kim, D. Kang, J. W. Jeong, J. W. Kim, M. J. Kim, A. Cho, D.-H. Jang, and H. Nam (2024), Automatic speech recognition (ASR) for the diagnosis of pronunciation of speech sound disorders in korean children, *Clinical linguistics & phonetics* p. 1–14.

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed, and Michael Auli (2020), wav2vec 2.0: a framework for self-supervised learning of speech representations, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Curran Associates Inc., Red Hook, NY, USA.

Bernstein, J., J. Cheng, J. Balogh, and E. Rosenfeld (2017), Studies of a self-administered oral reading assessment, *Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2017)*, pp. 172–176.

Castles, A., K. Rastle, and K. Nation (2018), Ending the reading wars: Reading acquisition from novice to expert, *Psychological Science in the Public Interest* **19** (1), pp. 5–51.

Chang, Y.-N., J. S. H. Taylor, K. Rastle, and P. Monaghan (2020), The relationships between oral language and reading instruction: Evidence from a computational model of reading, *Cognitive Psychology* **123**, pp. 101336.

Chicco, D. and G. Jurman (2023), The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification, *BioData Mining* **16** (1), pp. 4.

Chuang, Y.-Y., M. L. Vollmer, E. Shafaei-Bajestan, S. Gahl, P. Hendrix, and R. H. Baayen (2021), The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning, *Behavior Research Methods* **53** (3), pp. 945–976.

Cito B.V. (2017), Handleiding dmt (cito volgsysteem), *Technical report*, Cito B.V. `http://www.goloca.org/nt2/dmt/cito_dmt_handleiding_groep_3-8.pdf`.

Cleuren, L., J. Duchateau, P. Ghesquière, and H. Van Hamme (2008), Children's oral reading corpus (CHOREC): Description and assessment of annotator agreement, *in* Calzolari, N., K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias, editors, *Proceedings of the Sixth*

International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/254_paper.pdf.

Coltheart, M., K. Rastle, C. Perry, R. Langdon, and J. Ziegler (2001), Drc: A dual route cascaded model of visual word recognition and reading aloud, *Psychological Review* **108** (1), pp. 204–256.

Cucchiarini, C., A. Neri, and H. Strik (2009), Oral proficiency training in dutch l2: The contribution of asr-based corrective feedback, *Speech Communication* **51** (10), pp. 853–863.

Dyck, B. V., B. BabaAli, and D. V. Compernolle (2021), A hybrid asr system for southern dutch, *Computational Linguistics in the Netherlands Journal* **11**, pp. 27–34.

Fan, Ruchao, Natarajan Balaji Shankar, and Abeer Alwan (2024), Benchmarking children's asr with supervised and self-supervised speech foundation models, *Interspeech 2024*, pp. 5173–5177.

Feng, S., B. M. Halpern, O. Kudina, and O. Scharenborg (2024), Towards inclusive automatic speech recognition, *Computer Speech & Language* **84**, pp. 101567.

Foundation, Python Software (2022), Python release python 3.11. Accessed: 2024-12-18.

Galarnyk, M. (2022), Train test split: What it means and how to use it. Accessed: 2024-12-18.

Groenhof, B. (2024a), Groenhofbram/wav2vec-chorec (version 1.0). Accessed: 2024-12-18.

Groenhof, B. (2024b), Groenhofbram/whisper-chorec (version 1.0). Accessed: 2024-12-18.

GroNLP (2023), Wav2vec2-dutch-large-ft-cgn · hugging face [computer software]. Accessed: 2024-12-18.

Harm, M. W. and M. S. Seidenberg (2004), Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes, *Psychological Review* **111** (3), pp. 662–720.

Harmsen, W., C. Cucchiarini, R. van Hout, and H. Strik (2024), A joint approach for automatic analysis of reading and writing errors, *in* Gorman, K., E. Prud'hommeaux, B. Roark, and R. Sproat, editors, *Proceedings of the Second Workshop on Computation and Written Language (CAWL) @ LREC-COLING 2024*, ELRA and ICCL, pp. 8–17. https://aclanthology.org/2024.cawl-1.2.

Harmsen, W., F. Hubers, R. Van Hout, C. Cucchiarini, and H. Strik (2023), Measuring word correctness in young initial readers: Comparing assessments from teachers, phoneticians, and asr models, *Proceedings of the 9th Workshop on Speech and Language Technology in Education (SLaTE)*, pp. 11–15.

Hu, G., S. C. Determan, Y. Dong, A. T. Beeve, J. E. Collins, and Y. Gai (2020), Spectral and temporal envelope cues for human and automatic speech recognition in noise, *Journal of the Association for Research in Otolaryngology* **21** (1), pp. 73–87.

Hussein, A., S. Watanabe, and A. Ali (2021), Arabic speech recognition by end-to-end, modular systems and human, *Computer Speech & Language* **71**, pp. 101272.

Jain, Rishabh, Andrei Barcovschi, Mariam Yahayah Yiwere, Peter Corcoran, and Horia Cucu (2024), Exploring native and non-native english child speech recognition with whisper, *IEEE Access* **12**, pp. 41601–41610.

Jain, Rishabh, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu (2023), Adaptation of whisper models to child speech recognition, *INTERSPEECH 2023*, pp. 5242–5246.

Kendeou, P., P. van den Broek, M. J. White, and J. S. Lynch (2009), Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills, *Journal of Educational Psychology* **101** (4), pp. 765–778.

Klebanov, B. B., A. Loukina, J. Lockwood, V. R. T. Liceralde, J. Sabatini, N. Madnani, B. Gyawali, Z. Wang, and J. Lentini (2020), Detecting learning in noisy data: The case of oral reading fluency, *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pp. 490–495.

Klein, G. (2023), Whisper-large-v2 · hugging face [computer software]. Accessed: 2024-12-18.

Loukina, A., B. B. Klebanov, P. Lange, B. Gyawali, and Y. Qian (2017), Developing speech processing technologies for shared book reading with a computer, *Proceedings of the 6th Workshop on Child Computer Interaction (WOCCI 2017)*, pp. 46–51.

McFee, B., C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto (2015), Librosa: Audio and music signal analysis in python, *Majora* p. 24.

Mich, O., N. Mana, R. Gretter, M. Matassoni, and D. Falavigna (2020), Automatically assess children's reading skills, *in* Gala, N. and R. Wilkens, editors, *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, European Language Resources Association, pp. 20–26. https://aclanthology.org/2020.readi-1.4.

Molenaar, B., C. Tejedor-Garcia, C. Cucchiarini, and H. Strik (2023), Automatic assessment of oral reading accuracy for reading diagnostics, *INTERSPEECH 2023*, pp. 5232–5236.

National Institute of Standards and Technology (2021), Sctk [python]. Accessed: 2024-12-18.

Ngueajio, M. K. and G. Washington (2022), Hey asr system! why aren't you more inclusive?, *in* Chen, J. Y. C., G. Fragomeni, H. Degen, and S. Ntoa, editors, *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, Springer Nature Switzerland, pp. 421–440.

OECD (2023), *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, OECD.

Parikh, A., L. ten Bosch, H. van den Heuvel, and C. Tejedor-Garcia (2023), Comparing modular and end-to-end approaches in asr for well-resourced and low-resourced languages, *in* Abbas, M. and A. A. Freihat, editors, *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, Association for Computational Linguistics, pp. 266–273. https://aclanthology.org/2023.icnlsp-1.28.

Perfetti, C. A. and T. Hogaboam (1975), Relationship between single word decoding and reading comprehension skill, *Journal of Educational Psychology* **67** (4), pp. 461–469.

Perry, C., J. C. Ziegler, and M. Zorzi (2013), A computational and empirical investigation of graphemes in reading, *Cognitive Science* **37** (5), pp. 800–828.

Piton, T., E. Hermann, A. Pasqualotto, M. Cohen, M. Magimai-Doss, and D. Bavelier (2023), Using commercial ASR solutions to assess reading skills in children: A case report, *INTERSPEECH 2023*, pp. 4573–4577.

Poncelet, J. and H. Van Hamme (2023), Learning to jointly transcribe and subtitle for end-to-end spontaneous speech recognition, *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 182–189.

Prasad, A. and P. Jyothi (2020), How accents confound: Probing for accent information in end-to-end speech recognition systems, *in* Jurafsky, D., J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 3739–3753.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever (2023), Robust speech recognition via large-scale weak supervision, *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, JMLR.org.

Rapcsak, S. Z., M. L. Henry, S. L. Teague, S. D. Carnahan, and P. M. Beeson (2007), Do dual-route models accurately predict reading and spelling performance in individuals with acquired alexia and agraphia?, *Neuropsychologia* **45** (11), pp. 2519–2524.

Sadeghi, M. E., H. Sheikhzadeh, and M. J. Emadi (2024), A proposed method to improve the wer of an asr system in the noisy reverberant room, *Journal of the Franklin Institute* **361** (1), pp. 99–109.

Seidenberg, M. (2005), Connectionist models of word reading, *Current Directions in Psychological Science - CURR DIRECTIONS PSYCHOL SCI* **14**, pp. 238–242.

Shraddha, S., J. L. G, and S. K. S (2022), Child speech recognition on end-to-end neural asr models, *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pp. 1–6.

Stehman, S. V. (1997), Selecting and interpreting measures of thematic classification accuracy, *Remote Sensing of Environment* **62** (1), pp. 77–89.

Taalunie (2008), Jasmin-spraakcorpus (version 1.0) [dataset]. `https://taalmaterialen.ivdnt.org/download/tstc-jasmin-spraakcorpus/`.

Taalunie (2014), Corpus Gesproken Nederlands—CGN (Version 2.0.3) [Dataset]. `https://taalmaterialen.ivdnt.org/wp-content/uploads/documentatie/cgn_website/doc_Dutch/topics/index.htm`.

Tejedor-García, C., B. van der Molen, H. van den Heuvel, A. van Hessen, and T. Pieters (2022), Towards an open-source Dutch speech recognition system for the healthcare domain, *in* Calzolari, N., F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, pp. 1032–1039. https://aclanthology.org/2022.lrec-1.110.

Tillemans, M. (2007), *Dissolving the 'd/dt' disambiguation problem*, number 07-01.

van den Bosch, K. P., B. J. A. de Groot, and J. R. de Vries (2019), Klepel-R — revised. `https://www.pearsonclinical.nl/klepel-r-revised`.

Van der Klis, A., F. Adriaans, M. Han, and R. Kager (2023), Using open-source automatic speech recognition tools for the annotation of Dutch infant-directed speech, *Multimodal Technologies and Interaction*.

Van Gompel, M. (2023), Dutch open speech recognition benchmark. `https://opensource-spraakherkenning-nl.github.io/ASR_NL_results/UT/Jasmin/jasmin_res.html`.

Van Til, A., F. Kamphuis, J. Keuning, M. Gijsel, J. Vloedgraven, and A. de Wijs (2018), Wetenschappelijke verantwoording LVS-toetsen DMT, *Technical report*, Cito.

Weijnman, H. C. (2013), *The role of word decoding and language comprehension in reading comprehension*, Master's thesis, Utrecht University. https://studenttheses.uu.nl/handle/20.500.12932/15317.

Wentink, W. M. J. (1997), *From graphemes to syllables: The development of phonological decoding skills in poor and normal readers*, PhD thesis, Radboud University. https://repository.ubn.ru.nl/handle/2066/265053.

Woollams, A., M. Ralph, D. Plaut, and K. Patterson (2007), SD-squared: On the association between semantic dementia and surface dyslexia, *Psychological Review* **114**, pp. 316–339.

Yeung, G. and A. Alwan (2018), On the difficulties of automatic speech recognition for kindergarten-aged children, *Interspeech 2018*, pp. 1661–1665.