

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

**Date** : 2024-08-31  
**Student** : Bram Groenhof | S1103145 | [bram.groenhof@ru.nl](mailto:bram.groenhof@ru.nl)  
**Primary supervisor** : Helmer Strik  
**Secondary supervisor** : Wieke Harmsen  
**Course** : Master's Thesis | LET-TWM400  
**Word count (Abstract to references)** : 18,110

---

## Contents

<b>1. Introduction</b>	<b>1</b>
1.1. Background	1
1.1.1. Reading proficiency of Dutch children	1
1.1.2. Oral Reading Fluency and Models for Reading	1
1.1.3. The Potential of Automatic Speech Recognition (ASR)	4
1.2. Current Study	4
1.2.1. Research Questions and Relevancy	4
1.2.2. Previous Findings	5
1.2.3. Selection of ASR-models for this Study	6
1.3. Hypotheses	7
<b>2. Methodology</b>	<b>9</b>
2.1. Data	9
2.1.1. Participants	9
2.1.2. Pre-Processing	9
2.1.2.1. Reading Materials	9
2.1.2.2. Recording Quality	10
2.1.2.3. Train and Test datasets	10
2.2. Measurements for the Performance of ASR-models	11
2.3. Baseline Results	15
2.4. Experiments to Improve Results	15
2.4.1. Experiment 1: Rule-Based Improvements	15
2.4.2. Experiment 2: Similarity-Based Improvements	16
<b>3. Results</b>	<b>17</b>
3.1. Excluded Participants	17
3.2. Assessor Judgements	17
3.2.1. Error Analysis	18
3.3. ASR-models' Baseline Results	18
3.3.1. Agreement Metrics	19
3.3.2. Error Analysis	20
3.4. Experiment 1: Rule-Based Improvements	24
3.4.1. Agreement Metrics	24
3.4.2. Error Analysis	25
3.5. Experiment 2: Similarity-Based Improvements	26
3.6. Overall Results	28
<b>4. Discussion</b>	<b>31</b>
4.1. Hypothesis 1	31
4.2. Hypothesis 2	32
4.3. Hypothesis 3	32
4.4. Hypothesis 4	33
4.5. General Discussion	34
<b>5: Conclusion</b>	<b>35</b>
5.1. Suggestions for Future Studies	35
<b>6: References</b>	<b>37</b>
<b>Appendices</b>	<b>46</b>
A: Full Similarity-level Based Improvements Tables	46

## Abstract

Dutch children's reading skills have been declining consistently for many years. One of the ways 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 to carry out as teachers have to administer the tests in a one-on-one setting, in which the teacher has 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. However, since the DMT only requires a binary judgement of correct or incorrect, a perfect transcription is not needed.

We explored the performance of two state-of-the-art (SOTA) pre-trained ASR-models: wav2vec2.0-CGN and faster-whisper-v2. We had them carry out assessments on isolated word tasks, similar to the DMT, 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 accuracy, F1-score, and MCC as agreement metrics. We then used two different methods to improve the baseline results, one using manually defined rules and one using a standardized Levenshtein distances.

We found that rule-based improvements obtained the best results for the overall metrics. Faster-whisper-v2 (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 expected the accuracy levels for both models to be lower than the lowest assessor inter-rater accuracy level (.86). However, faster-whisper-v2 performed better than expected on accuracy (.89). We found this result in particular to be misleading as the high accuracy scores could be a result of the imbalanced dataset.

We conclude that the performance of pre-trained ASR-models is promising, but probably not yet good enough for use in primary schools. Future research could aim to improve the performance of these ASR-models through methods like fine-tuning and training, and through collaborative research with teachers. Furthermore, diagnostic information from the DMT are currently underused. ASR-based assessment could allow teachers to make more use of diagnostic information to help familiarize children with the types of words they struggle with most. Prominent models for reading such as the dual route cascaded model (DRC), the triangle model, and the connectionist dual process model of reading aloud (CDP++) all predict that this would improve the oral reading proficiency of the children.

Keywords: ASR, wav2vec2.0, Whisper, children's speech, oral reading, drie-minuten-toets, drie-minuten-toets (DMT)

# 1. Introduction

## 1.1. Background

### 1.1.1. Reading proficiency of Dutch children

In the Netherlands, the reading skills of children have been declining for many years (OECD, 2023). This negative trend has been reported on by Dutch media (De Vries, 2023; NOS Nieuws, 2023) as well as government bodies that underline the importance of this issue (Inspectorate of Education, 2024). This points to a problem that the entire Netherlands faces, with no single clear way to improve reading proficiency at once.

One of the most common ways in which oral reading skills of Dutch primary school children are tested is through the three-minute-exam (DMT). Dutch children aged 6-12 take this test at least once every academic year until the end of primary school. They are provided with different word lists that vary in difficulty and must read all of the words out loud as quickly and as correctly as possible (Van Til et al., 2018). The test itself is administered and marked by teachers manually. 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. Thus, it would be fruitful to look into possible ways to improve the assessment process of the DMT. This could free up much time for teachers to focus on other aspects of reading, or even for them to look at the DMT-results in more detail so that they can provide more feedback to students beyond the single score.

### 1.1.2. Oral Reading Fluency and Models for Reading

To ground oral reading fluency in theory, we must look at models of reading. Van Til et al. (2018) describe that the DMT aims to measure the decoding skills of primary school pupils. According to them, decoding skills refer to 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).

Based on these two phases, it can be said that there are two essential processes for technical reading: decoding and word recognition. Decoding refers to the conversion of graphemes into sounds. Word recognition refers 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.

Since we try to define oral reading fluency, which we will use to refer to the more specific skill of technical reading that the DMT aims to measure from now on, it makes sense to try and find a suitable model to represent it. However, there is no agreement 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 model, as human behavior is always different from a model representation. They mention that the focus should be on what the models have in common. We will follow their approach, using three models to exemplify this for this thesis: the dual-route cascaded model (DRC), the triangle model, and the connectionist dual process model of reading aloud (CDP++). All three of these models are mentioned in van Til et al. (2018) and they are the main computational models for reading (Castles et al., 2018). We will introduce these models very briefly and explain how they help us define oral reading fluency.

**Figure 1**

*The DRC-model (Coltheart et al., 2001)*

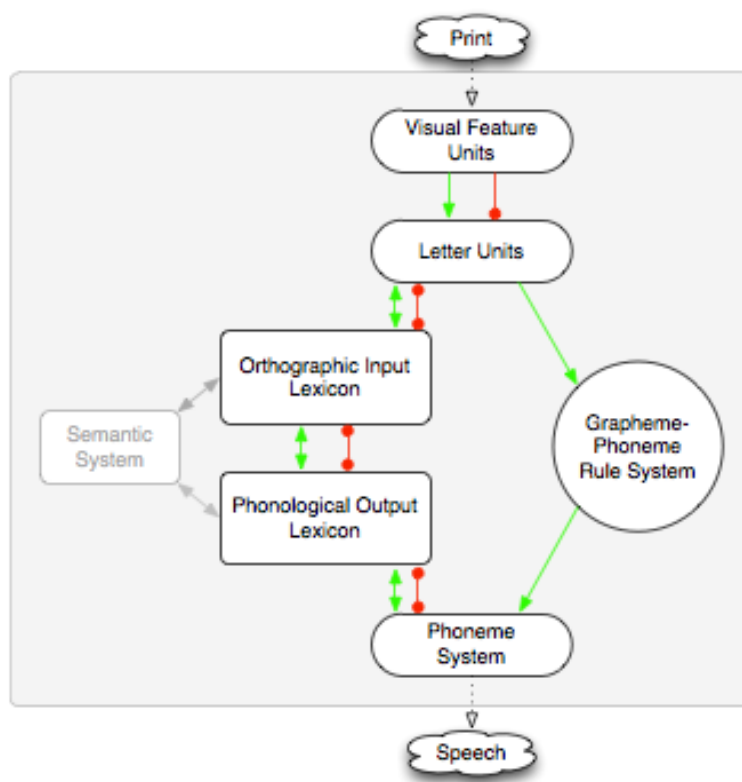
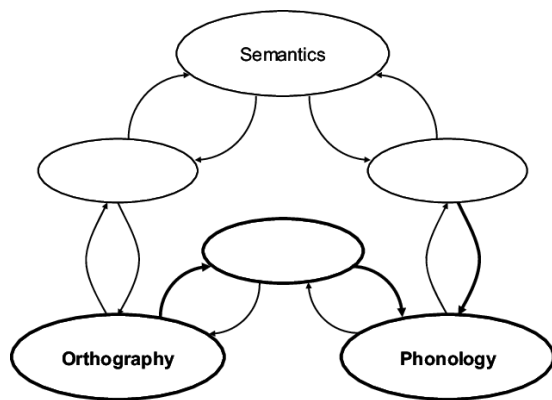


Figure 1 shows the DRC-model. This model states that the process of reading a word happens through one of two routes: phonological or lexical. The mental lexicon plays a big role in this model as well. The lexicon is an internal system where important information about words is stored; including orthographic, 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).

**Figure 2**

The triangle model (Harm & Seidenberg, 2004)



The triangle model is shown in figure 2. Each oval represents a processing layer that is active when a word is read. The large, named ovals represent a certain type of information that is being processed. Each of these layers has an in- and output layer. The smaller, unnamed ovals represent hidden units. These facilitate more complex connections between the larger, named ovals. According to this model, part of learning how to read a certain word is to know how much each processing layer should weight in for specific situations (Chang et al., 2020; Harm & Seidenberg, 2004).

**Figure 3**

The CDP++ model (Perry et al., 2013)

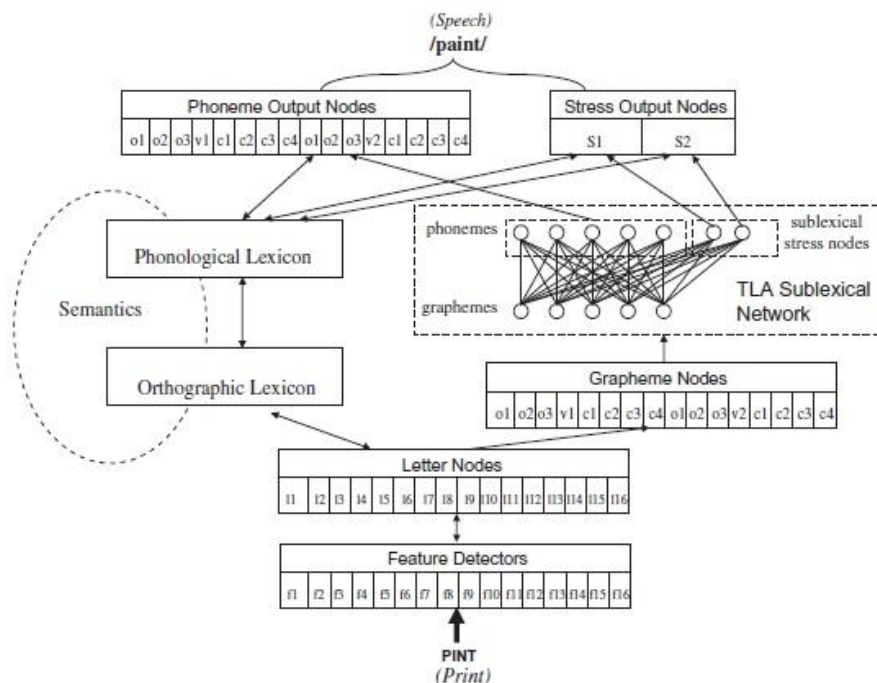


Figure 3 shows the CDP++ model. Here a division of labor exists between the lexical and non-lexical processes within a neural network. The model works by using two different routes: a direct route and a route using a hidden layer (Perry et al., 2013).

It is impossible to argue for the support of one of these models over the other, as numerous studies have shown advantages and disadvantages for each (Perry et al., 2013, 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 two important processes that also apply to the DMT. Firstly, the process of decoding a printed letter string (input) into a pronunciation. Secondly, the process of word recognition, which refers to the activation of a word as a whole. Word recognition is only possible if the word is stored in the mental lexicon (Castles et al., 2018). All 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).

Thus, for the DMT the familiarity with the words and its parts and how strongly the words are part of the child's lexicon are crucial. 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.

### 1.1.3. The Potential of Automatic Speech Recognition (ASR)

One potential way of helping teachers carry out the DMT is to use automatic speech recognition (ASR). ASR-models have been improving for many years, but progress for atypical speech has been lacking (Ngueajio & 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, 2024; Yeung & 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 thesis will take the findings of the previous literature and add to this body of knowledge, by focusing on the DMT and the possible use of different pre-trained ASR-models in the assessment process.

An important aspect to this is that ASR-models only provide a transcription of the audio file. The DMT requires judgements to be made about the words that children read. Accordingly, it is important to utilize the transcriptions ASR-models provide for this specific purpose. The most important step in this process is the alignment of the ASR-transcription and the prompts of the DMT. 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). We will return to this in the methods section.

## 1.2. Current Study

### 1.2.1. Research Questions and Relevancy

The question this thesis aims to answer is: *How well can current state-of-the-art (SOTA) pre-trained ASR-models perform judgements on word list oral reading tasks by children akin to the three-minute exam?*

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). By writing a program that can utilize different pre-trained ASR-models and make them take on the role of the teacher in the context of the DMT, the

possibilities of using these models can be assessed quantitatively. Note that we do not intend to test whether ASR-models could take on the role of the teacher wholly by themselves, but that we explore how they can be used as tools. We do not advocate for the replacement of teachers by ASR-models in the judgement of the DMT, but for their use as tools for teachers. Since the DMT only requires teachers to state whether each word was read correctly or incorrectly, the task that the ASR-models have to perform is quite clear. 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. This would save teachers a lot of time as parts of the DMT can be automated.

Beyond this, the focus in this thesis will not be on simply quantitative metrics that showcase the ASR-models' performance. The quantitative results will be looked at in a more detailed way to spot what types of errors the ASR-models make through error analysis. This does not only provide insight into the performance of the ASR-models used in this thesis, but can also help future research in two ways. First, it can aid the development of fine-tuned ASR-models for the DMT specifically. Second, the types of errors can help find common errors the children make in their oral reading. Applying this to the DMT could help teachers use diagnostic information the DMT provides to help pupils struggling with specific types of words or sounds. An additional question is thus *What are the most prominent types of errors current SOTA ASR-models make when judging isolated word lists read aloud by children?*

### 1.2.2. Previous Findings

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.

Cleuren et al. (2008) describe the Children's Oral Reading Corpus (CHOREC), a corpus that contains data from Flemish primary school children, aged 6-12 years old, reading aloud word lists similar to the DMT. We stated that ASR-models tend to struggle with judging children's read speech for isolated word lists, but this does not insinuate that different teachers are always in agreement about judgements. In their publication on 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 are developed to be like those in the DMT. Teachers were instructed to assess the children's oral reading 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, p. 14). A main advantage of using ASR-models for judgements is that it will make the same judgements objectively and consistently. For this to be successful however, the ASR-model must be making these judgements correctly and reliably.

The aforementioned key issue of ASR-models struggling with children's speech is prevalent, but 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, p. 4576) 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 training 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. For this to be possible, the ASR-transcriptions must be aligned and utilized for this purpose. For the DMT, this means that the ASR-transcriptions must be compared to the prompt and judged as correctly or incorrectly read. While the fine-tuning of models is not within the scope of this thesis, it is known to improve the ASR-transcriptions and in turn the performance of the ASR-model for the task it was fine-tuned for. This will be kept in mind and touched on when discussing results. Especially because for the DMT the words that children need to read are always known, making fine-tuning a realistic possibility in practice.

### 1.2.3. Selection of ASR-models for this Study

In this thesis, two SOTA ASR-models will be used. The first model we will use will be referred to as wav2vec.20-CGN (GroNLP, 2023). This is a pre-trained on the 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 called faster-whisper-v2 (Klein, 2023; Radford et al., 2022).

In order to justify the choice for these models, it is first important to note that Whisper and wav2vec2.0 are SOTA ASR-models. 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.

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 & 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.

### 1.3. Hypotheses

The previous literature, as discussed in previous sections, allows us to formulate the following four hypotheses. The first three aim to answer the main research question: *How well can current state-of-the-art (SOTA) pre-trained ASR-models perform judgements on word list oral reading tasks by children akin to the three-minute exam?*

1: We expect the ASR-models’ judgements to show moderate agreement with the judgements of assessors in CHOREC. This prediction is based on the fact that our ASR-models and approach will share commonalities with the best-performing ASR-model in Harmsen et al. (2023), but they will also lack aspects that made the best-performing ASR-model perform the way it did. Again, we cannot expect to obtain the same results in our study as in the previous studies, because of the inherent differences between the study’s data, participants, and tasks. However, we do use them as rough estimations of what we can expect, which is why we expect moderate correlations. Harmsen et al. (2023, pp. 13-14) found that ASR-models performed better when they could predict non-existing words. The models we will use can do this. Furthermore, we will use the results of the ASR-models to manipulate the decision-making of the models. This is comparable to what teachers do, as they tend to be lenient when certain specific reading errors are made e.g., teachers tend to allow schwa-insertions in consonant clusters). This should help the ASR-models perform better than the lesser-performing models in this study. However, we will not be fine-tuning ASR-models or providing them with the expected lists of prompts. Fine-tuning or training of models has been proven to benefit ASR-models’ performance in general (Jain et al., 2023; Klebanov et al., 2020; Mich et al., 2020; Molenaar et al., 2023).

2: We expect the ASR-models’ inter-rater agreement to assessors to be lower than the lowest inter-rater agreement value of assessors in CHOREC. Numerous studies have pointed out that ASR-models struggle with children’s speech (Feng et al., 2024; Jain et al., 2023, 2024; Yeung & Alwan, 2018). Since we will use the ASR-transcriptions to base the judgements off, we expect that their judgements will be less similar to the assessors in CHOREC than the assessors in CHOREC are similar to each other.

3: We expect that faster-whisper-v2 will outperform wav2vec2.0-CGN. Previous research has emphasized the potential of Whisper-based models for children’s speech specifically (Jain et al.,

2024). Furthermore, Van Gompel (2023) showed that faster-whisper-v2 performs much better on children’s speech than many of its contemporaries, including other Whisper-based models. This is important, because even though other research has shown that wav2vec2.0 can outperform Whisper (Barcowski et al., 2023; Jain et al., 2023), there is no direct comparison between faster-whisper compared to wav2vec2.0.

The final hypothesis aims to answer the sub question: *What are the most prominent types of errors current SOTA ASR-models make when judging isolated word lists read aloud by children?*

4: We expect that the most common type of errors made by the ASR-models can be grouped similar to the error group types outlined by the assessors in CHOREC. We expect that when we categorize the errors made by the ASR-models to be similar to the reading error categories defined in CHOREC (Cleuren, Duchateau, & Sips, 2008). We expect this because the authors of CHOREC were able to define categories that could encompass all types of reading errors they found. When the ASR-models makes a mistake, we should be able to classify it into a general category. Due to the comprehensive list of error categories in CHOREC, we expect that when we categorize the errors of the ASR-models, they will be in categories comparable to those in CHOREC. The results section will show these error categories and the comparisons to the error categories we defined ourselves.

## 2. Methodology

### 2.1. Data

The data used for this thesis came from the CHOREC corpus (Cleuren et al., 2008). We developed two pipelines for this thesis in Python 3.11 (Python Software 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, 2024b).

#### 2.1.1. Participants

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) and elementary schools for children with specific learning disabilities (N = 126). All children were between 6 and 12 years old.

#### 2.1.2. Pre-Processing

##### 2.1.2.1. Reading Materials

For our research, not all the participant data is relevant. The DMT only contains existing words (Cito B.V., 2017). This means that only the data in CHOREC for which children performed the real word reading task (RWRT) is relevant. As mentioned in the introduction, this data is fitting for our purposes because the structure of the reading material for the RWRT in CHOREC is very similar to that of the DMT.

The DMT asks teachers to mark whether each word was read correctly or not. While the teacher marks words that were read incorrectly or skipped, no feedback is given to the pupils beyond a numerical score (Cito B.V., 2017). This means that the task at hand is binary: either a word was read correctly (0) or it was read incorrectly (1). A problem with CHOREC is that there is not a full orthographic transcription available for recordings of children's oral speech. Luckily, the inclusion of a reading error layer in the annotations will suit our purpose well. If the teacher judged the word as read correctly, they did not annotate anything. If they judged the word as read incorrectly, they annotated the words using codes provided in the annotation protocol (Cleuren et al., 2008). The reading error layer can therefore be referenced to see how the assessors classified each word: read correctly when no annotation was made (0) and read incorrectly when one or more error codes were made (1). 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. Thus, the total number of children whose recordings we used in this thesis is 385.

Table 1 shows the number of RWRT recordings available, separated by word list. Each of the word lists in CHOREC consists of 40 words that were presented to the child in isolation from one another. The 1LG, 2LG, and 3+4LG each contained only 1-syllable, 2-syllable, and 3- or 4-syllable words respectively (Cleuren et al., 2008). In short, they increase in difficulty. While this is not the exactly the same as the DMT, it does align with the types of words children are expected to read for the DMT. We can see that we have a different number of available recordings per word list, the most complex word list (3+4LG) has the fewest recordings available while the simplest word list (1LG) has the most. All word lists were equal in length, containing 40 words each.

**Table 1***Available recordings by word list*

<b>Word list</b>	<b>Recordings (N)</b>
1LG	377
2LG	359
3+4LG	320
<b>Total</b>	<b>1056</b>

### *2.1.2.2. Recording Quality*

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 because 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.

### *2.1.2.3. Train and Test datasets*

Once the data had been gathered, we separated the full dataset into a training 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 test how modifications to the model affect the results. When the best possible outcomes are achieved in the training set, the model is applied to the test set to see if the training and/or fine-tuning has improved the results (Galarnyk, 2022). While we did not train or fine-tune the ASR-models, we did intend to improve upon the baseline results. For this reason, we defined a training to test changes on to find the best possible results before applying them to the test set.

The most important rule we followed when we defined the training and test sets was to ensure mutual exclusivity on speaker level: no child’s speech appeared in both sets. Had we not done this, it could have led to over-fitting, which hurts the generalization of the results (Moon et al., 2015; Simonnet et al., 2018). Imagine that the recordings of a particular child are split among the sets. If we had improved the results based on the recordings in the training set, it may have led to an inflated amount of improvement in the test set. The improvements in the test set could be a result of a bias for this child’s way of oral reading. In turn, the results could be much worse when used on a different dataset in which that particular child is not present. In this case, the results cannot be generalized. For this reason, we made sure that no same child’s recordings were present in both training and test sets.

In the same vein, we wanted to make sure that the training set was balanced. Table 2 shows an overview of speaker characteristics in the training 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 training set contained 27.18% of all relevant data in CHOREC, leaving 82.82% for the test set.

**Table 2***Description of participants in training set*

School year	Gender	Number of participants in training set (N)
1	Female	2
1	Male	2
2	Female	15
2	Male	15
3	Female	15
3	Male	15
4	Female	15
4	Male	15
5/6	Female	1
5/6	Male	1
<b>Total</b>		96

## 2.2. Measurements for the Performance of ASR-models

If we want to obtain results that can show how well ASR-models can take on the task of assessing a DMT, we require measurements of how human-like their assessments are. Because teachers are the norm for the DMT, we consider the assessors in CHOREC to be the ground truth. To answer our research questions, we must have metrics that can quantify the level of agreement between the ASR-models and the assessors. Because we take the assessors' judgements as the ground truth due to the fact that this is how the DMT is assessed (Cito B.V., 2017), an ASR-model performs the best when its judgements deviate the least from the assessors.

$$\text{WCPM} = \frac{\text{Total number of words read correctly}}{\left( \frac{\text{Time taken in seconds}}{60} \right)} \quad (1)$$

A possible metric for this is WCPM, which we mentioned before as it was used in the study done by Bernstein et al. (2017). Equation 1 shows how this metric is calculated (Kim et al., 2021). All equations in this thesis were coded into pipelines using the NumPy library (Harris et al., 2020). This metric seems well-fitting for the DMT especially, as the duration for each word list would be exactly 60 seconds long and the child's score is equivalent to a WCPM score (Cito B.V., 2017). The WCPM of assessors and ASR-models could be calculated and compared to see if there are significant differences. Bolaños et al. (2011) used this method to show that their ASR-based system, FLORA, performed similarly to teachers when asked to assess passages of text read out by children.

We do not think this measure is in any way poor or unsuited for our research. However, previous work that we want to compare our results to did not use WCPM (Cleuren et al., 2008; Harmsen et al., 2023). Since we want to be able to compare it directly to previous work and we do not want to bloat our results with too many metrics, we opt for other overall measures instead.

In this thesis, all metrics will be calculated using a confusion matrix. 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 3 explains the meaning of these four metrics applied to the data of this thesis. For our data, a ‘negative’ (0) corresponds to a word that was read correctly and a ‘positive’ (1) corresponds to a word that was read incorrectly.

**Table 3**

*Overview of possible outcomes in a confusion matrix*

<b>Outcome</b>	<b>ASR-model judged word as...</b>	<b>Assessors judged word as...</b>
True negative (TN)	Correctly read	Correctly read
True positive (TP)	Incorrectly read	Incorrectly read
False negative (FN)	Correctly read	Incorrectly read
False positive (FP)	Incorrectly read	Correctly read

Table 4 shows how we obtained the judgements from assessors in CHOREC. For the assessors, a word was noted as correctly read (marked as a “0”) when the reading error layer was empty for a word (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 transcriptions, the transcription for each 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”).

**Table 4**

*Determination of judgements from assessors*

<b>Prompt</b>	<b>Reading Error</b>	<b>Judgement</b>
groen		0
groen	13	1

In order to use the ASR-transcriptions to make the same type of judgements as the assessors in CHOREC, we needed to be able to compare the ASR-transcription to the prompts. This required alignment of the ASR-transcriptions to the prompts to ensure that the prompt was compared to the child’s attempt at reading that prompt. To do this, we aligned the ASR-transcriptions to the prompts using ADAGT (Harmsen et al., 2024). We opted for this alignment algorithm as it provides two-way alignment: forwards and backwards. Children often stutter and restart words. However, only the last attempt counts for the DMT (Cito B.V., 2017). We noticed that the additional backwards alignment that ADAGT offers aligned the ASR-model transcriptions better consistently than other commonly used alignment methods such as SCLITE (National Institute of Standards and Technology, 2021). The alignment allowed us to compare the ASR-transcriptions to the prompts. Table 5 shows examples of how we determined the judgements of the ASR-models. If the ASR-transcription for a word differed from its prompt, the word was judged as incorrectly read (marked as a “1”). If the ASR-transcription and its prompt matched, the word was marked as correctly read (marked as a “0”).

**Table 5***Determination of judgements from ASR-transcriptions*

Prompt	Aligned ASR-transcription	Judgement
groen	groen	0
groen	groon	1

After the confusion matrix has been generated, we calculate several metrics following Chicco & Jurman (2023). An overview of these metrics, how they are calculated, and an explanation of what they represent is provided in Table 6. These metrics provide quantitative values that aid us in explaining the performance of the ASR-model with regards to specific cases. For example, a low precision value would indicate that the ASR-model judges a lot of words as incorrectly read when the assessors judge them as correctly read.

**Table 6***Explanation of confusion matrix metrics used in this thesis (Chicco & Jurman, 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.
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.

The metrics in Table 6 provide different insights into the performance of the ASR-models, since they all provide quantitative measure of agreeability between the ASR-model and the assessors. However, while these metrics are helpful for insights, we argue that Matthew’s Correlation Coefficient (MCC) is a better representation of overall agreement.

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (2)$$

Other metrics of overall agreement, such as Cohen’s Kappa (Cleuren et al., 2008) and F1 score (Gao et al., 2024) are widely used instead of MCC (Equation 2). However, a recent body of work has argued strongly for the use of MCC when dealing with binary classification instead. Other metrics, including accuracy and F1-score often have inflated results. This is because, unlike MCC, they can

give high scores even if the prediction does not perform well in all confusion matrix categories. In other words, MCC considers all four elements of the confusion matrix (TN, TP, FN, and FP) and considers them proportionally to the size of positive and negative elements in the data set. MCC is thus more robust than both accuracy and F1-score, as they do not account for this proportionality (Chicco & Jurman, 2020, 2023).

Consequently, MCC is more suitable for our research than accuracy and F1-score as they do not work well with imbalanced datasets; they are often too optimistic about the predictive performance (Chicco et al., 2021; Chicco & Jurman, 2023; Lakshmi & Prasad, 2014; Mudadla, 2023). This point is crucial, because the oral reading data in CHOREC is imbalanced due to there being far more correctly read words than incorrectly read words (Cleuren et al., 2008).

Thus, MCC will be used as the overall measure of agreement between ASR-models and assessors. This value will indicate the performance of a given ASR-model: the closer to 1 this correlation gets, the better the ASR-model performs.

A question that may arise at this point is why we decide to include accuracy and F1-score regardless. The main reason for this is interpretation and comparability with previous studies. While MCC is preferable for inter-rater agreement, the initial CHOREC paper does not use it (Cleuren et al., 2008). Accuracy allows for a direct comparison of performance between inter-rater agreement of the different assessors in CHOREC and the ASR-model compared to the assessors. Similarly, at the moment of writing, F1-score score sees much more widespread use compared to MCC as stated by Chicco and Jurman (2021; 2020, 2023). Thus, while accuracy and F1-score will not be as important for our results compared to MCC, they will allow us to compare results to previous studies and improve the interpretability of the results.

While it is hard to compare our results to earlier studies due to differing datasets, speaker characteristics, and tasks, this provides us with a rough reference point of what could be expected. Since we intend to only use pre-trained ASR-models and to not do any fine-tuning, we expect that the results will not be as good as the best performing models in the mentioned studies that were fine-tuned (Harmsen et al., 2023; Molenaar et al., 2023). We also do not expect the ASR-models to perform as well as assessors. Consequently, we expect the inter-rater agreement between assessors in CHOREC to be higher than the inter-rater agreement between the ASR-model and the assessors (Cleuren et al., 2008). As discussed in earlier sections, previous studies have shown that ASR-models tend to struggle with children's speech. Furthermore, the best results have been acquired through the use of fine-tuning and training of models, which is not something we will do for this thesis. Therefore, the hypotheses we will make will be quite conservative about how well the ASR-models we will use will fare.

The interpretation of MCC we use here is based on Powers (2008). According to them, MCC can be interpreted in the same way as Pearson Product-moment Correlation Coefficient (PCC). We will interpret our results using the following way to describe the strengths of the correlations, in absolute values, is as follows: 0 is no correlation, .01-.39 is a weak correlation, 0.4-.79 is a moderate correlation, .8-.99 is a strong correlation, and 1 is a perfect correlation (Loughbrough University, n.d.). Obtaining higher MCC values is much more difficult than it is for other well-known correlation metrics, such as Pearson's  $r$ , as it takes more variables into account; such as proportionality of positive and negative values in the total dataset.

### 2.3. Baseline Results

The baseline results provided us with all agreement metrics previously described: accuracy, precision, recall, F1-score, and MCC. Accuracy, F1-score, and MCC showed overall performance of the ASR-models. Precision and recall allowed us to zoom in more: if one of them was much lower than the other, this enabled us to say more about the types of errors the ASR-model made.

### 2.4. Experiments to Improve Results

After obtaining the baseline results, we wanted to explore in what way and how much the performance of ASR-models could be improved. For this, we would ideally fine-tune the ASR-models for Dutch oral read children’s speech. However, this was out of the scope for this thesis. Thus, we chose explicit data post-processing instead. 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. Based on this, we postulated that if we allowed the ASR-models to be more lenient, the results would improve. 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). We wanted to allow the ASR-models to let children deviate from the prompt in certain cases as well. We conducted two experiments to do this, using different methods: a rule-based method using confusion pairs (section 2.4.1.) and a method based on Levenshtein distances (section 2.4.2.).

#### 2.4.1. Experiment 1: Rule-Based Improvements

For this experiment, we tried to minimize the most common disagreements between the ASR-model and assessors. First, we categorized the most frequent disagreements and then evaluated the changes in MCC when we allowed the ASR-transcription to deviate from the prompt in the ways described by these categories.

We first 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. Table 7 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 7**

Example of a confusion pair with error category

<b>Error category name</b>	<b>Prompt</b>	<b>ASR-model transcription</b>	<b>Assessor judgement</b>	<b>ASR-model judgement</b>	<b>Occurrences (N)</b>
Substitution k/c	cola	kola	Correct	Incorrect	85

By looking at the most frequently occurring confusion pairs, we could categorize the types of errors made by the ASR-model. This use of confusion pairs is a common error analysis approach for obtaining a better understanding of the errors (Hussein et al., 2021; Prasad & Jyothi, 2020; Tejedor-García et al., 2022). Once the errors had been categorized, we used these categories to change ASR-model’s judgements in post-processing. If the difference in prompt and hypothesis was only due to one of these categories, we changed the ASR-model’s judgement from incorrectly read to correctly read. This allowed us to experiment with the categories to see the effects on the agreement metrics. This introduction of lenience is further justified by the fact that assessors allow readings to deviate from the prompts in certain ways (Harmsen et al., 2023). By looking at the most frequent confusion

pairs, we could find deviations which the assessors allow. In turn, this allowed us to post-process the data so that the ASR-model’s transcription could take certain deviations into account; allowing them to be judged as correct readings of the prompt.

Taking the example shown in Table 7, we could take all confusion pairs of type “substitution k/c” and change the ASR-model’s judgement from read incorrectly to read correctly, changing the overall agreement metrics. By allowing different categories of confusion pairs to be judged as correctly read, the ASR-model’s judgements became more lenient. We computed and evaluated the agreement metrics using this more lenient approach. We tried to find and categorize as many of the most common confusion pairs to maximize improvements in accuracy, F1-score, and MCC.

### 2.4.2. Experiment 2: Similarity-Based Improvements

The second approach used a standardized form of Levenshtein distances. Table 8 shows different deviations between prompts and ASR-model transcriptions. The Levenshtein distance was calculated by counting the total number of insertions, deletions, and substitutions in the ASR-model transcription compared to the prompt. A higher Levenshtein distance indicates a larger deviation from the prompt (Nam, 2019). In the baseline results, only a Levenshtein distance of 0 would lead to the ASR-model judging a word as read correctly.

**Table 8**

*Examples of possible Levenshtein distances and similarity values*

Prompt	ASR-model transcriptions	Insertions (N)	Deletions (N)	Substitutions (N)	Levenshtein distance	Similarity (standardized)
cola	kola	0	0	1	1	.75
moeder	moder	0	1	0	1	.83
moeder	moedder	1	0	0	1	.83
sneeuw witje	sneeuw witje	0	1	0	1	.91

$$\text{Similarity} = 1 - \left( \frac{\text{Levenshtein distance}}{\text{Length of prompt}} \right) \quad (3)$$

An issue with the Levenshtein distance is that it is not standardized. This problem is demonstrated in Table 8. All examples have a Levenshtein distance of 1, but depending on the length of the prompt this can represent a small or large deviation. Because of this, we used a standardized version of the Levenshtein distance described by Higuera & Mico (2008). In their paper, they describe a version of the Levenshtein distance which considers the length of a string, standardizing its value as a proportion. Equation 3 shows how we calculated the standardized Levenshtein distances for each prompt, as shown in Table 8. From this point onward, we will refer to this concept as ‘similarity’. The ASR-models’ judgements become more lenient the lower the similarity value is.

Once the similarity values were calculated, we could introduce leniency by letting the ASR-model judge a word as read correctly if it met a similarity level rather than only allowing transcriptions which were identical to the prompt. Changing this similarity threshold affected the agreement metrics. The ideal similarity level was the one for which the MCC value is highest. By calculating and interpreting all agreement metrics for all similarity levels using steps of .05 on the training sets, we could find the similarity level at which the highest MCC could be obtained. This similarity level was then applied to the test sets.

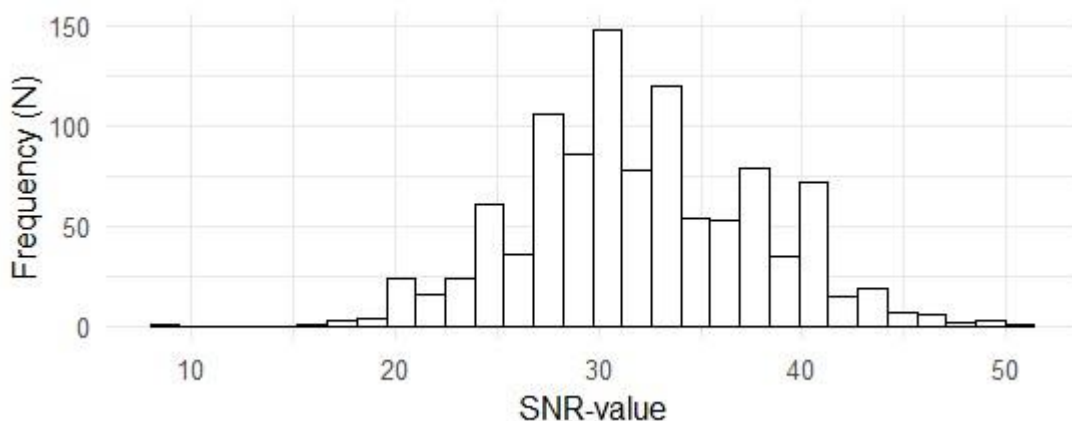
### 3. Results

In this section, we will present all of the relevant results. Note that in this results section, we analyzed and interpreted the data at face value. The discussion section will delve deeper into the interpretation and possible explanations of the results, which will in turn affect our conclusion on whether or not the results support the hypotheses or not.

#### 3.1. Excluded Participants

**Figure 4**

*Distribution of SNR-values*



The results of the SNR-analysis, as shown in figure 4, show that almost all recordings have an SNR-value of at least 20dB ( $N = 926$ ,  $M = 32.07$ ,  $SD = 5.80$ ). After the recordings with SNR-values under 20dB ( $N = 11$ ) were manually checked to assess the audio quality, none were judged as poor quality. For this reason, we decided to exclude no recordings because of poor audio quality.

#### 3.2. Assessor Judgements

Since we are trying to assess ASR-models' performance on this task by comparing the ASR-model judgements to assessor judgements, it should first be made clear how the assessors in the datasets judged the children's oral reading.

Table 9 shows descriptive statistics for the assessor judgements. We calculated the error rate by counting the number of times words were judged as read incorrectly and dividing that by the total number of words in the dataset. We can see that the test dataset consisted of 30,760 read words, of which 11.59% were judged as being read incorrectly. The training dataset, consisted of 11,480 read words, of which 5.35% were judged as being read incorrectly. The difference in error rate between training and test datasets is quite large, which we will touch on in the discussion. In addition, the overall error rate (8.97%) shows how imbalanced the dataset is; 91.03% of the words were read correctly and 8.97% were read incorrectly.

**Table 9***Descriptive statistics for assessor judgements.*

<b>Dataset</b>	<b>Words judged (N)</b>	<b>Correctly read (N)</b>	<b>Error rate (%)</b>
Training	11,480	10,897	5.35
Test	30,760	27,556	11.59
Total	42,240	38,453	8.97

### 3.2.1: Error Analysis

The assessor judgements were based off the information in the reading error labels. We used these to define the most common error categories children made when reading. These errors are made in 95.26% and 96.14% of all erroneously read words in the training and test datasets respectively. While these were the most common, they do not represent this percentage of all errors. The reading error layer often consisted of multiple error codes, meaning that a child could make multiple errors at once.

Table 10 shows that the most common type of errors children made in their oral reading, according to assessors, was the replacement of existing words into pseudowords. Both ASR-models used in this thesis can generate pseudowords, which should allow them to catch these reading errors. Furthermore, most other categories are a result of either substitutions, insertions, or deletions. Once again, the ASR-models' capability of generating pseudowords is helpful to catch these reading errors, as these types of errors often lead to either different words in the read language or pseudowords.

**Table 10***Most common errors according to assessor judgements. Error category descriptions taken from Cleuren, Duchateau, & Sips (2008)*

<b>Error category</b>	<b>Part of total errors in training set (%)</b>	<b>Part of total errors in test set (%)</b>
An existent word is replaced by a pseudoword	31.22	41.70
An existent word is replaced by another existent word that is orthographically similar	24.32	26.86
Decoding that is not followed by an attempt at synthesized reading	8.92	16.12
Insertion of an arbitrary consonant	8.78	-
Deletion of an arbitrary consonant	7.63	-
Substitution of vowels	7.48	11.46
Deletion of the suffix “-en” or “-er”	6.91	-
Total	95.26	96.14

*Note. We only found the most common errors in each set, which is why we did not enter all values for the test set. When the errors covered 95%, we considered them to be comprehensive enough.*

### 3.3. ASR-models' Baseline Results

In this section, we will first look at the baseline results. These results were generated from the raw ASR output and ADAGT-alignment. A word was only judged as correctly read if the ASR-transcription matched the prompt exactly.

Table 11 shows the descriptive statistics of both ASR-models’ judgements. The error rates were calculated in the same way as for the assessor judgements: we counted the number of times words were judged as read incorrectly and divided that by the total number of words in the dataset. Only this time, we did not look at the assessor judgements, but the ASR-models’ judgements. The error rates are higher for all models and datasets than the assessor judgements. In other words, the error rates indicate that the ASR-models judged the children’s oral reading of words as incorrect more often than the assessors did. Additionally, wav2vec2.0-CGN model has a much higher overall error rate (52.17%) than the faster-whisper-v2 model (19.11%). This means that the latter is less likely to produce FPs than the former, which is extremely beneficial for an imbalanced dataset such as this due to there being relatively few incorrectly read words to begin with (8.97%, see Table 9). Regardless, even though faster-whisper-v2 outperforms wav2vec2.0-CGN here, its judgements lead to a much higher error rate compared to the assessor judgements.

**Table 11**

*Descriptive statistics for baseline results*

ASR-model	Dataset	Words judged (N)	Correctly read (N)	Error rate (%)
Wav2Vec2.0-CGN	Training	11,480	6,430	56.01
Wav2Vec2.0-CGN	Test	30,760	14,775	48.03
Total/Overall		42,240	20,205	52.17
Faster-Whisper-v2	Training	11,480	9,790	14.72
Faster-Whisper-v2	Test	30,760	24,379	20.74
Total/Overall		42,240	34,169	19.11

### 3.3.1. Agreement Metrics

In order to see how the ASR-models perform, 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. In Table 12 we can see an overview of all agreement metrics. From here we can see that the results for faster-whisper-v2 are generally better than those for wav2vec2.0-CGN. Accuracy, precision, F1-score, and MCC are all higher for faster-whisper-v2 than wav2vec2.0-CGN. The only exception is recall, in which wav2vec2.0-CGN outperforms faster-whisper-v2.

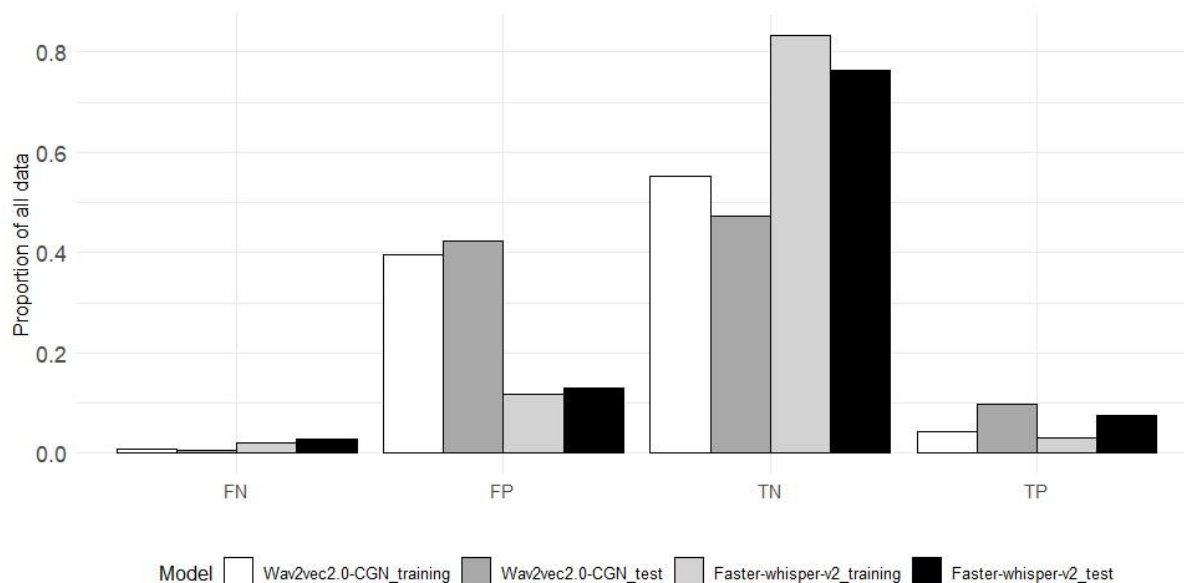
**Table 12**

*Agreement metrics for baseline results*

ASR-model	Dataset	Accuracy	Precision	Recall	F1-score	MCC
Wav2vec2.0-CGN	Training	.60	.10	.85	.17	.19
Wav2vec2.0-CGN	Test	.57	.19	.94	.31	.29
Faster-whisper-v2	Training	.86	.21	.60	.31	.30
Faster-whisper-v2	Test	.84	.37	.73	.49	.44

**Figure 5**

*Proportion of FNs, FPs, TNs, and TPs for each ASR-model and dataset*



*Note. For FNs and FPs, the ASR-model performs better when this number is low. For TNs and TPs, the ASR-model performs better when this number is high.*

Figure 5 shows the proportionate number of FNs, TPs, TNs, and TPs for each ASR-model and dataset. Faster-whisper-v2 performed better than wav2vec2.0-CGN, as the former produced more TNs/TPs and fewer FNs/FPs than the latter. This helps explain why faster-whisper-v2 generated better precision values but lower recall values than wav2vec2.0-CGN. Precision only looks at TPs and FP: more FPs leads to a lower precision value. Wav2vec2.0-CGN generated many more FPs than faster-whisper-v2, explaining the difference in precision values. Similarly, recall only takes into account TPs and FNs. Since wav2vec2.0-CGN generated fewer FNs than faster-whisper-v2, it generated higher values for recall. Wav2vec2.0-CGN generated more FPs, while faster-whisper-v2 generated more FNs.

Subsequently, we can see how the accuracy values represent this. Accuracy is the proportion of TNs and TPs relative to all FPs, FNs, TNs, and TPs. Figure 5 shows a larger impact of FPs than of FNs due to how many were generated. This explains why faster-whisper-v2 has higher accuracy ratings than wav2vec2.0-CGN.

Regarding the MCC values, we can also see that faster-whisper-v2 outperforms wav2vec2.0-CGN in all cases. For the training set, wav2vec2.0-CGN (.19) and faster-whisper-v2 (.30) both obtained weak correlations. For the test the faster-whisper-v2 obtained a moderate correlation (.44), while wav2vec2.0-CGN obtained a weak correlation (.29).

### 3.3.2. Error Analysis

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 general categories, similar to what was done to the reading errors found in CHOREC (Cleuren et al., 2008; see Table 10). This was done by looking at the confusion pairs sorted by frequency and finding patterns therein. Of course, we only did this for the training set as the test set only functioned as a way to test our final improvements. Our focus was on reducing the

number of FPs as much as possible, as these accounted for the vast majority of errors made by both models.

For example, both ASR-models often generated FPs because they inserted spaces in compound words. A word like *ruziemaken* (‘to argue’) would be transcribed as *ruzie maken*. Therefore, the ASR-model’s transcription did not match the prompt exactly and it was judged as read incorrectly by the ASR-model.

Table 13 shows an overview of the error categories that we defined by looking at the most frequently occurring confusion pairs. When defining these, we did not base them on CHOREC but on patterns we saw in the data. These will be used in the remainder of the results section for error analyses. Since we based these off of the training set, this section will only focus on the error categories made in them.

**Table 13**

*Error category categories*

<b>Error category name</b>	<b>Example prompt</b>	<b>Example ASR-output</b>	<b>Explanation</b>
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 14 shows the total number of errors we caught by defining different error categories. These categories caught the greatest number of errors in the training set we could find. We did this using a well-known technique called all-pairs testing (Berger, 2003). For example, we allowed the ASR-models to judge words as read correctly if they fell into the “Ch confusion” and the “Insertion spaces” error categories. We did this for all possible combinations of error categories. We found that enabling all of the error categories shown in Table 14 yielded the best results. In total, these error categories accounted for 34.13% of the FPs in wav2vec2.0-CGN’s and 42.98% of FPs in faster-whisper-v2’s training sets.

**Table 14**

*Error category distribution for the training 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	.30
Substitution double/single consonants	64	1.40	17	1.27
Substitution k/c	55	1.21	0	0
Insertion	48	1.05	95	7.10
Substitution nasals	45	1.00	3	.22
Substitution au/ou	41	1.00	2	.15
Substitution fricative voice	36	.80	1	.08
Deletion liquids	33	.79	12	.90
Substitution i/y	11	.72	0	0
Total	1526	34.13	575	42.98

We tried to compare our error categories (Table 14) to the most common error categories defined in CHOREC (Table 10). However, a major issue with the categorization of errors in CHOREC is that many error categories overlap. For example, the most common type of error in CHOREC according to the assessors was that an existent word was replaced by a pseudoword. However, many of the other categories would lead to pseudowords as well; such as the insertion or deletion of an arbitrary consonant. Furthermore, the definition of the second most common type of error in CHOREC, replacement of an existent word by another existent word is just as unclear. This made it hard to analyze these results using these categories, so any conclusions we draw from them are with due caution. We attempted to group our error categories in line with those from CHOREC as much as possible, but this process was definitely not perfect and will be touched upon in the discussion. Here, we only looked at the training set. However, later we will also analyze the test set in the same way.

In Table 15 we can see that there are numerous mismatches between the error categories we created and those in CHOREC. Many of our categories did not fit in those defined in CHOREC and vice versa. The best-fitting category seems to be substitution of vowels, where many of our defined categories fit in. For this category, wav2vec2.0-CGN judged more words as read incorrectly due to this error (9.70%) than the assessors in CHOREC (7.48%), and faster-whisper-v2 (4.34%).

**Table 15**

*Defined error categories in training dataset grouped by most common error category categories according to CHOREC (Cleuren et al., 2008)*

<b>CHOREC error category</b>	<b>Defined error categories</b>	<b>Part of total errors in wav2vec2.0-CGN (%)</b>	<b>Part of total errors in faster-whisper-v2 (%)</b>	<b>Part of total errors in CHOREC (%)</b>
Insertion of an arbitrary consonant	Insertion	7.68	7.10	8.78
Deletion of an arbitrary consonant	Deletion liquids	.79	.90	7.63
Substitution of vowels	Substitution long/short vowels	9.70	4.34	7.48
	Substitution oe/oo			
	Substitution au/ou			
	Substitution i/y			
Deletion of the suffix “-en” or “-er”	None	None	None	6.91
Not a common error in CHOREC	Ch confusions	15.96	30.65	-
	Insertion spaces			
	Substitution plosive voice			
	Deletion final			
	Substitution double/single consonants			
	Substitution k/c			
	Substitution nasals			
	Substitution fricative voice			

Additionally, it is important to note on the defined error categories in Table 15 is that these were defined by looking at confusion pairs in both ASR-models’ training set results. It is therefore possible that an error category may have been common for one ASR-model, but not the other. For example, the error category “Deletion final” was the fourth most common type of error for wav2vec2-CGN, while faster-whisper-v2 did not make this error at all (see Table 14).

### 3.4. Experiment 1: Rule-Based Improvements

In this section, we will look at the performance of the ASR-models with rule-based manipulations. These manipulations allowed ASR-transcriptions that differed from the prompt in a way described by the earlier defined error categories (see Table 15) to be judged as correctly read. We set out to do this to reduce the number of FPs, in turn increasing the TNs. Since our changes only affect FPs and TNs, we will only be looking at the changes therein for this section.

Table 16 shows the descriptive statistics of both ASR-models’ judgements after we applied the manually defined rules. We can immediately see that the error rates are lower in all cases compared to the baseline results (see Table 11). However, the error rates still indicate that the ASR-models judged the children’s oral reading of words as incorrect more commonly than assessors did. The only exception to this is faster-whisper-v2’s performance on the training set (8.66%), which has a lower error rate than assessors judgements (8.97%, see Table 9). Regardless, these errors indicate, much like the baseline results, that faster-whisper-v2 outperforms wav2vec2.0-CGN in this regard.

**Table 16**

*Descriptive statistics for results with rule-based improvements*

ASR-model	Dataset	Words judged (N)	Correctly read (N)	Error rate (%)
Wav2Vec2.0-CGN	Training	11,480	8,363	27.15
Wav2Vec2.0-CGN	Test	30,760	21,230	30.98
Total/Overall		42,240	29,593	29.94
Faster-Whisper-v2	Training	11,480	10,486	8.66
Faster-Whisper-v2	Test	30,760	27,388	10.96
Total/Overall		42,240	37,874	10.34

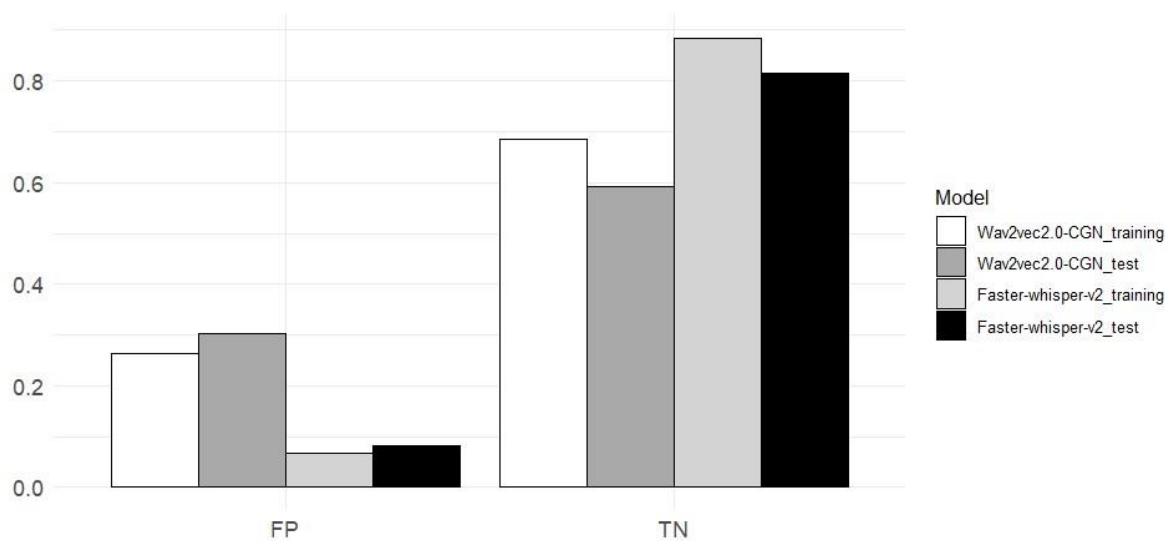
#### 3.4.1. Agreement Metrics

We once again assessed the performance of the ASR-models by calculating their agreement metrics with the assessor judgements. In this section, all results represent the values obtained after improving the results with rule-based improvements.

Table 17 shows an overview of all obtained agreement metrics. Here we can see the exact same trends as for the baseline results (see Table 12). Faster-whisper-v2 outperforms wav2vec2.0-CGN in all metrics but recall. The improvements show in all metrics for both ASR-models except for recall. Since the improvements only focused on reducing the number of FPs this is not a surprise, as these improvements do not affect recall. Turning to MCC, Wav2vec2.0-CGN still showed weak values in both training (.27) and test (.37) sets despite the improvements. Faster-whisper-v2 showed moderate MCC values in both training (.40) and test (.54) sets after the improvements.

**Table 17***Agreement metrics for results with rule-based improvements*

ASR-model	Dataset	Accuracy	Precision	Recall	F1-score	MCC
Wav2Vec2.0-CGN	Training	.73	.14	.85	.24	.27
Wav2Vec2.0-CGN	Test	.69	.24	.94	.39	.37
Faster-Whisper-v2	Training	.91	.32	.60	.41	.40
Faster-Whisper-v2	Test	.89	.48	.73	.58	.54

**Figure 6***Proportion of FPs and TNs each ASR-model and dataset after applying rule-based improvements*

*Note. This figure only shows FPs and TNs because the number of FNs and TPs were unaffected.*

As mentioned before, this method of improvement only reduced the number of FPs and increased the number of TNs. Therefore, it is only relevant to look at the differences here. In Figure 6, we can see how the proportion of FPs and TNs changed between the initial and improved results. Note that if the proportion of FPs lowers by an amount, the number of TNs increases by the same amount. There was more improvement for wav2vec2.0-CGN (.14 and .12 more TNs and fewer FPs in training and test datasets respectively) compared to faster-whisper-v2 (.05 more TNs and fewer FPs in both datasets). Despite this, faster-whisper-v2 still outperformed wav2vec2.0-CGN. As was seen in the baseline results (see Table 12 & Figure 5), this may be largely due to faster-whisper-v2 generating fewer FPs in the first place.

### 3.4.2. Error Analysis

We did not carry out in-depth error analysis for the final results. This would have required us to try and categorize the remaining errors into new, previously unidentified ones. However, we noticed two trends when looking at the results which could constitute to a large number of the errors we failed to catch.

First, we saw numerous examples of errors that were made by the ASR-models that did not fit into a single defined error category, but into multiple error categories at once. Table 18 shows two examples

of this. These prompts were erroneously transcribed in the same way. They both contain a substitution long/short vowels error as well as a substitution plosive voice error. Our pipeline was not able to find these types of errors, leading to many errors not being caught.

**Table 18**

*Example of multiple errors in a single ASR-model transcription*

<b>Prompt</b>	<b>ASR-model transcriptions</b>
Potlood (pencil)	Botlod
Deur (door)	Ter

Second, the alignment of ASR-transcriptions to the prompts may have led to issues. We justified the use of ADAGT for alignment, and in many cases the availability of both forward and backward alignment was beneficial as either method would sometimes be more successful. Despite this, as Table 19 shows, we did spot occasions 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. This could hurt further analysis, because the reading mistake may be different from the one indicated by the alignment.

**Table 19**

*Example of ADAGT-alignment going wrong*

<b>Prompt</b>	<b>ADAGT forward alignment</b>	<b>ADAGT backward alignment</b>
Appel (apple)	Spelen (to play)	Spelen (to play)
Auto (car)	Ouders (parents)	Ouders schilder (parents painter)

### 3.5. Experiment 2: Similarity-Based Improvements

An overview of all metrics at all similarity level intervals for both ASR-models in the training dataset can be found in Appendix A. Here, we use a figure to portray the most important results in a concise manner.

**Figure 7**

*MCC values by similarity levels*

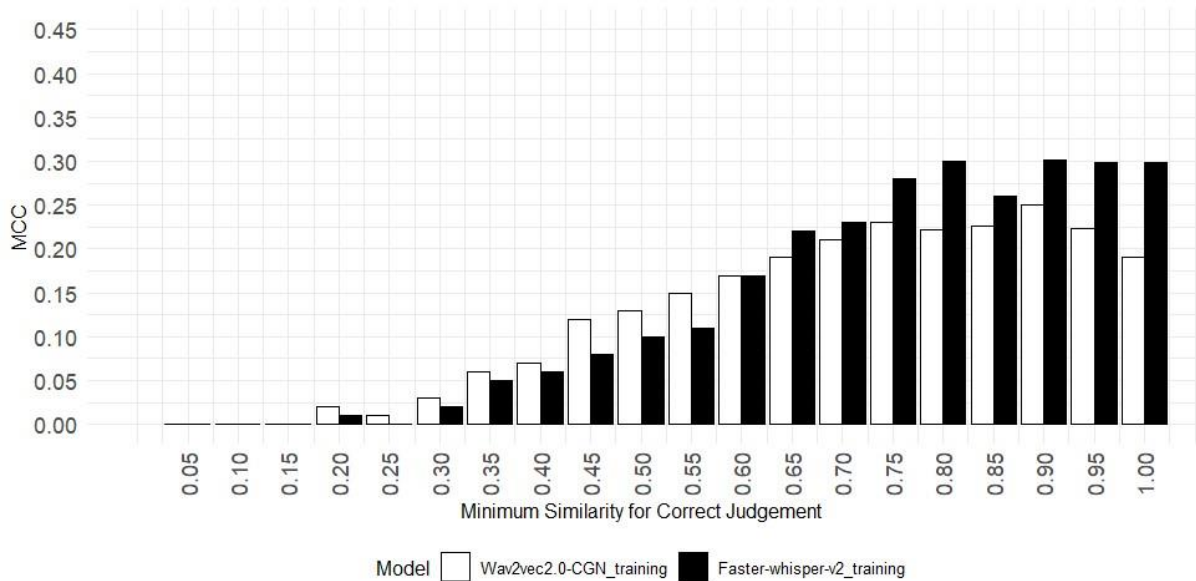


Figure 7 shows all the obtained MCC values per similarity level separated by model. For both models and, we can see that the MCC values steadily increases as the similarity level increases as well. There are however slight differences in results depending on the model used.

For the wav2vec2-CGN, the results show higher MCC values at similarity levels lower than 1. The highest MCC value was obtained at similarity level .9 (.25). This is a noticeable improvement over the baseline results which allowed only for a similarity level of 1, as this similarity obtains an MCC value of .19 (see Table 12).

For faster-whisper-v2, we can see that the highest MCC values were obtained by the highest similarity scores. The highest MCC value was obtained at similarity levels 1, .95, .90, and .80 (.30). This essentially means that this method of improving the results was ineffective for faster-whisper-v2. Similarity level 1 is identical to the baseline results.

Using these results, we applied the similarity-based improvements to the test sets by using the similarity levels that obtained the highest MCC values. Table 20 shows these results. We compared these results to the baseline results (see Table 12). Wav2vec2.0-CGN showed improvements in all metrics but recall. However, the MCC values remained weak. For faster-whisper-v2, the best results were obtained at similarity levels that had no effect on the agreement metrics. The highest MCC after similarity-based improvements (.44) was obtained at similarity levels .95 and 1. The results of these similarity levels are identical to those of the baseline results. While lower similarity levels .9 and .8 also showed the highest MCC on the training set (see Figure 7), they obtained lower MCC values (.42 & .4) here. This experiment was more successful for wav2vec2.0-CGN than faster-whisper-v2.

**Table 20**

Similarity-based improvements results (test set)

ASR-model	Similarity	Accuracy	Precision	Recall	F1-score	MCC
Wav2vec2.0-CGN	.9	.75	.22	.89	.35	.36
Faster-whisper-v2	1	.84	.37	.73	.49	.44
Faster-whisper-v2	.95	.84	.37	.73	.49	.44
Faster-whisper-v2	.9	.85	.37	.66	.47	.42
Fater-whisper-v2	.8	.89	.47	.45	.46	.4

In these results, we noticed a trend in the accuracy values that has implications for one of our hypotheses. Looking at the accuracy values in Table 20, we can see that as the similarity decreases the accuracy increases. We decided to look at whether this trend also showed in the training set.

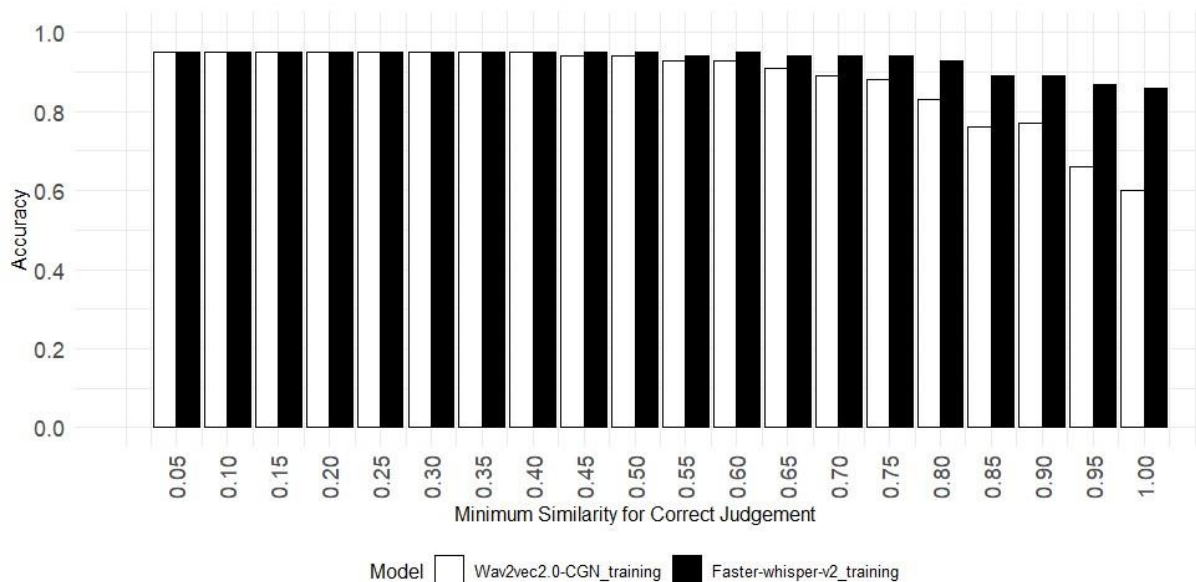
**Figure 8***Accuracy values by similarity levels*

Figure 8 shows that the accuracy values are higher at lower similarity levels. This is a result of the imbalanced dataset. At some point, the similarity value is so low that it will judge all words as being read correctly. In Figure 8, this is represented by an accuracy value of .95, which was obtained by both models at numerous low similarity levels. In context, this value makes sense. When we looked at assessor judgements, we saw that the error rate in the training set was 5.35% (see Table 9). These high accuracy values at lower similarity levels are a result of this. By judging all words as read correctly, even if they differ from the prompt greatly, the ASR-models' accuracy values will always be .95. This is a key issue of the metric which we will return to in the discussion.

### 3.6. Overall Results

In this final section of the results, we present only the most crucial results from the attempted improvements. We will discuss their relevance to the hypotheses for our research questions in the discussion section. Therefore, this is not done here. The information presented here has already been provided, but here it is shown with less context to summarize it more concisely. The tables will

mention metrics and in how far they improved after applying each method. Improvement refers to the increase in value of that metric in the test set before and after improvements were applied.

Table 21 shows the overall results from the rule-based improvements. Here we can see that the accuracy, F1, and MCC scores improved for both ASR-models. Most notably, the improvements seemed to have a larger effect on MCC for faster-whisper-v2 than wav2vec2.0. Both showed similar improvements in terms of F1-score. In terms of accuracy, there was a larger improvement for wav2vec2.0-CGN compared to faster-whisper-v2. Despite this, faster-whisper-v2 outperformed wav2vec2.0-CGN for all three metrics in all cases.

**Table 21**

*Overview of most important rule-based improvements results compared to base results, test set only for final results*

<b>ASR-model</b>	<b>Accuracy (improvement)</b>	<b>F1-score (improvement)</b>	<b>MCC (improvement)</b>
Wav2vec2.0-CGN	.69 (.12)	.39 (.08)	.37 (.08)
Faster-whisper-v2	.89 (.05)	.58 (.09)	.54 (.1)

For the similarity-based improvements, results are slightly different. This method did not affect the baseline results of faster-whisper-v2, as can be seen in Table 22. For faster-whisper-v2, the best MCC was obtained at similarity score 1. This means that these results are identical to the baseline results. For wav2vec2.0-CGN, we can see that there was a slight improvement in F1-score and MCC at the best-performing similarity level of .90. A much larger improvement can be seen in accuracy. Interestingly, in wav2vec2.0-CGN's case this method improved the results more than the rule-based improvements method for all metrics except for F1-score.

**Table 22**

*Overview of most important similarity-based improvements results, test set only for final results*

<b>ASR-model</b>	<b>Similarity level</b>	<b>Accuracy (improvement)</b>	<b>F1-score (improvement)</b>	<b>MCC (improvement)</b>
Wav2vec2.0-CGN	.9	.75 (.18)	.35 (.04)	.36 (.07)
Faster-whisper-v2	1	.84 (0)	.49 (0)	.44 (0)

*Note. values taken from the similarity value with the highest MCC.*

Finally, we looked at the precision values for each improvement method compared to the baseline results. The results are shown in Table 23. We found that precision was the most important metric underlying accuracy, F1-score, and MCC. For both ASR-models, this was by far the lowest scoring metric. This should come as no surprise since both ASR-models generated many FPs. Despite the improvements, there is much to gain by finding ways to improve it further. The more FPs can be accounted for, the better the accuracy, F1-score, and MCC would be. This would of course only be so if the metrics are not negatively impacted when trying to get rid of more FPs, as too much leniency would introduce more FNs instead.

**Table 23***Overview of results for precision*

<b>ASR-model</b>	<b>Improvement method</b>	<b>Precision (improvement)</b>
Wav2vec2.0-CGN_test	Rule-based	.24 (.05)
Wav2vec2.0-CGN_test	Similarity-based	.22 (.03)
Faster-Whisper-V2_test	Rule-based	.48 (.11)
Faster-Whisper-V2_test	Similarity-based	.37 (0)

*Note: values for similarity-based improvement models taken from the similarity value with the highest MCC.*

## 4. Discussion

To reiterate, our research question is as follows: *How well can current state-of-the-art (SOTA) pre-trained ASR-models perform judgements on word list oral reading tasks by children akin to the three-minute exam?* In order to answer this question, set out three hypotheses. We also defined a sub-question, *What are the most prominent types of errors current SOTA ASR-models make when judging isolated word lists read aloud by children?* We defined a separate hypothesis to answer this sub-question.

### 4.1. Hypothesis 1

Using MCC scores to represent the correlations between ASR-model and assessor judgements, we found that faster-whisper-v2's results show moderate correlations, supporting the first hypothesis. However, wav2vec2.0's results only showed weak correlations even after the experiments, which does not support the first hypothesis.

In the baseline results, the MCC was weak for wav2vec2.0-CGN (.29), but moderate for faster-whisper-v2 (.44). Faster-whisper-v2 performed as well as we had hypothesized in the baseline results. While both experiments improved the metrics, they did not lead to moderate MCCs for wav2vec2.0 (.36 for rule-based improvements, .37 for similarity-based improvements). Faster-whisper-v2's results only improved using rule-based improvements, leading to an MCC of .54. While this is an improvement compared to the baseline results (.44), it is still a moderate MCC. These findings do not support the hypothesis, due to wav2vec2.0's results. However, as a stand-alone, faster-whisper-v2's results do support the hypothesis.

A natural question that arises from this is what could help explain this difference in performance. Section 4.3 will provide more insight as to why we think faster-whisper-v2 outperformed wav2vec2.0-CGN. Here, we want to focus on why we think the inclusion of wav2vec2.0-CGN over other ASR-models may have been the core issue.

One aspect that could play a major role is the trust we put in wav2vec2.0-CGN due to it being an end-to-end model. Recent literature has spoken highly of end-to-end models, with evidence showing that they outperform hybrid-based models (Parikh et al., 2023). This leads to an impactful limitation in this study. We have not included any hybrid-based models, such as Kaldi, in this thesis because it is not considered a SOTA ASR-model as end-to-end models are currently representatives of this title. As a result, we may have overlooked a potential model that could have done extremely well for the task at hand.

Two recent studies that help clarify how problematic this exclusion is are those done by Mich et al. (2020) and Molenaar (2023). Both of these studies show that Kaldi-based models can perform well on oral children's speech. The latter study is especially relevant, as it focuses on Dutch children's oral reading. The findings show that a Kaldi-based model with prompts and orthographic transcriptions in its language model outperformed even two Whisper-based models. What's more, the Kaldi-based model that only included CGN in its model shows a weak correlation as it has an MCC of .24. This is close to the baseline result for wav2vec2.0 in our study. While we cannot compare our MCC values to this study directly, it shows that a Kaldi-based model has obtained results in another study similar to the results we obtained for this study. In our study, we should have given more thought to hybrid-based models like Kaldi and included them, so that we could directly compare the results.

Future studies should bear this in mind when selecting which ASR-models to use and include in their research. Excluding Kaldi, or any other ASR-model, from a study because it is not considered SOTA anymore is not enough reason to do so.

Another insight that is applicable to both wav2vec2.0-CGN and faster-whisper-v2 comes from the same study (Molenaar et al., 2023). The study includes two different Whisper models: Whisper-Lv2 and Whisper-PR. The former's language model is Whisper large-v2, the same pre-trained ASR-model that underlies faster-whisper-v2. The latter uses the same language model but also includes the prompts. The results show that the MCC for Whisper-Lv2 (.01) was much lower than for Whisper-PR (.28), showing that the inclusion of prompts leads to a large improvement of performance. It would have been interesting to see if the effects on the results of this study had we provided prompts to the ASR-models' language models.

Furthermore, providing prompts to the ASR-model is feasible for tasks that include children reading out predetermined lists of words. To illustrate, the DMT has two versions of three word lists, each containing 150 prompts. This means that the DMT has a total of 900 prompts (Cito B.V., 2017). Since all of these are known, they can be added to the language model of ASR-models when using them on audio recordings of the DMT.

In terms of future implications, there are two important aspects with regards to the inclusion of prompts in the language model of ASR-models. First, they can lead to improvements in the performance of the ASR-models. Second, since we are dealing with predetermined word lists, providing prompts is a realistic method of further improving the results. Therefore, future studies on the use of ASR-models to help assess oral reading exams could provide the language models of the ASR-models they include with the prompts.

## 4.2. Hypothesis 2

We found that the obtained accuracy values showed that the second hypothesis is supported by the results for wav2vec2.0-CGN, but not by faster-whisper-v2. Both ASR-models' accuracy values are lower than the lowest inter-rater agreement value of assessors in CHOREC in the baseline and similarity-based improvements results ( $<.86$ ). For the rule-based improvements, wav2vec2.0-CGN's accuracy value also supports this hypothesis (.69), but faster-whisper-v2's does not (.89). Faster-whisper-v2 performed better than we hypothesized when the rule-based improvements were applied.

## 4.3. Hypothesis 3

We found that faster-whisper-v2 outperformed wav2vec2.0-CGN in all overall metrics (accuracy, F1-score, and MCC) in the baseline, rule-based improvements, and similarity-based improvements results. Therefore, the results support this hypothesis. There are multiple factors that can help explain why faster-whisper-v2 outperformed wav2vec2.0-CGN.

When we look at the results underlying the overall metrics, precision and recall, we can see that wav2vec2.0-CGN outperforms faster-whisper-v2 in recall. This is a consequence of the type of errors that each model was more likely to make. Wav2vec2.0-CGN made more FP errors, while faster-whisper-v2 made more FN errors. Having more FPs hurts precision, while having more FNs hurts recall. This explains the difference in the baseline results for precision and recall between the ASR-models. Since the values for precision were noticeably lower than recall for both ASR-models, we focused our improvement methods on reducing the number of FPs as much as possible. We expected these improvements to affect wav2vec2.0 more than faster-whisper-v2, because there was much more to improve. This was confirmed in the final results, as both improvement methods reduced the proportional number of FPs more for wav2vec2.0-CGN than faster-whisper-v2. Despite this, the precision value for wav2vec2.0 did not reach the value faster-whisper-v2 obtained in the baseline results, let alone after the post-improvements value.

This proneness of making FPs may be the single most important reason as to why wav2vec2.0-CGN obtained lesser results than faster-whisper-v2. For the task at hand, assessment of word correctness of children's oral reading using word lists, faster-whisper-v2 may be inherently better because of this. We know that faster-whisper-v2 is well-suited for children's oral read speech (Van Gompel, 2023).

However, the imbalanced nature of word lists may have played a much larger role in why we obtained the results in this thesis. When children are asked to read word lists aloud, it seems the vast majority of words are read correctly. This is true for both CHOREC and DART (Cleuren et al., 2008; Strik et al., 2022). If faster-whisper-v2 is less prone to making FP errors, it is inherently better at judging children's oral reading for word lists for this reason. However, In section 4.1. we mentioned that the inclusion of prompts in the language model of ASR-models could potentially lead to large improvements in their performances based on Molenaar (2023). Since we did not do this, we do not know what the results would be for either ASR-model in this study if this was applied. It could be that wav2vec2.0-CGN would generate far fewer FPs if it is provided with the prompts, which would improve its performance drastically. We already made a call for future research to provide the ASR-models with prompts in section 4.1., which is why we refrain from doing so again. However, it is just as relevant for this hypothesis as it is for hypothesis 1.

Future studies that aim to assess ASR-models' performances on oral word reading tasks should bear in mind that the data will most likely be imbalanced. It is likely that more words will be read correctly than incorrectly. If certain ASR-models are known to generate few FPs or deal well with imbalanced datasets specifically, they should be considered for inclusion in these studies.

#### 4.4. Hypothesis 4

Upon analyzing and interpreting the results, we found them and the way in which we obtained them problematic. For this reason, we find the results to be ambiguous, although comparison of the error categories we created and the most common types of errors according to the assessors in CHOREC seem to not support the hypothesis.

A flaw in our approach is that our defined error categories only found errors if there was one type of error, the ASR-transcription could only differ from the prompt in the way described by the error category. This is not how the error categories in CHOREC were created, as they allow for different types of errors to be present in the same word. An incorrectly read word can, for example, have a vowel insertion error, a consonant deletion error, or both (Cleuren, Duchateau, & Sips, 2008). We were not able to adjust our pipeline to be able to do this for this thesis. Had we been able to do this, it would have led to less ambiguous results as the comparison of our error categories to those in CHOREC would be more straightforward. An additional benefit of changing the pipeline so that it can find multiple error categories in the same word is that more confusion pairs would be found. This would make the rule-based improvements more impactful than they are in this thesis.

Although we consider the results to be ambiguous, they do not seem to support the hypothesis. For example, substitution errors of consonants accounted for 16.54% of all FPs in wav2vec2.0-CGN's and 22.35% of all FPs in faster-whisper-v2's training sets respectively. The equivalent error category in CHOREC, which is made up of six categories that have some overlap, was 14.82% (Cleuren, Duchateau, & Sips, 2008). While this is close to the value in wav2vec2.0-CGN, the fact that there is overlap may have inflated this number. This could be an indication that this type of error, which was common according to both wav2vec2.0-CGN and faster-whisper-v2, was not as common in CHOREC according to the assessors. This is not the only case in which the commonality of an error category differed in the ASR-models compared to in CHOREC. A large part of errors we found in wav2vec2.0-CGN (15.96%) and faster-whisper-v2 (30.65%) could not be categorized into an error category common in CHOREC.

In future studies using CHOREC, we advise researchers to orientate to see whether it is possible for them to design a pipeline that can take multiple error categories into account for single words. Without this, the results are too ambiguous to strongly advocate for or against the hypothesis.

## 4.5. General Discussion

Beyond specific hypotheses, there are some points of discussion regarding the approach and results of this thesis. Some of these raise issues that are a threat to the validity and generalisability of the results.

In the introduction, we mention that CHOREC was chosen as the dataset since its RWRT is structured in a similar way as the DMT (Cleuren et al., 2008). While this is a major upside, there are also drawbacks to using CHOREC. First, the children, whose recordings make up the data, are Flemish. It would be preferable to use a dataset that contains speech of native Dutch children such as DART (Strik et al., 2022). The main reason for this is that while both Dutch and Flemish children speak Dutch natively, they speak different kinds of Dutch which can be distinguished from each other (van Halteren & Oostdijk, 2018). It is not within the scope of this thesis to delve into specific differences, but ASR-models' performance would most likely be different if applied to Dutch children. Ideally, recordings of DMTs should be used for this type of research. Second, the metadata of CHOREC turned out to be incomplete. For example, the school grade the child was in at the moment of recording was missing for some of the data. This made it more difficult to define the training set, as we wanted to have as even of a split as possible. While we could have inferred the school grade by calculating the children's age, this would be imperfect as age can vary slightly in every school grade. These drawbacks of CHOREC hurt the generalisability of our results.

For future researchers doing similar research on the DMT, we recommend the use of DART instead of CHOREC when assessing Dutch children's oral reading skills of word lists. This is especially important if the school grade of the child is an important variable, since this is not present for all children in the metadata. However, it would be ideal to record DMTs and use this data, though this would require a large amount of work to obtain the data.

A problematic aspect of our approach is that we did not separate data by type of school the children were attending. At the time recordings were made for CHOREC, children were attending a regular primary school or a primary school for children with learning disabilities (Cleuren et al., 2008). In our research, we only generated results for all children within the training and test datasets. It is unlikely that one dataset contains many more children attending primary schools for children with learning disabilities than the other dataset. However, it still could happen, which would lead to an overrepresentation of these children in one dataset and an underrepresentation in the other. We could have separated all children who were attending regular primary schools and children who were attending primary schools for children with learning disabilities. Alternatively, we could also have made sure that these children were represented in both datasets in a balanced way.

Another potentially problematic aspect of how we defined the training and test sets is the percentage of errors they contained according to the assessors. The training set contains considerably fewer errors (5.35%) than test set (11.59%). Because of this, neither of them is representative of the percentage of errors in all data (8.97%). This hurts the generalisability of the results. For future research, the aim should be to keep the error rates of the training and test sets as close to the error rate of all data. This should be possible, since the error rates can be calculated based on the data available in CHOREC alone (Cleuren, Duchateau, & Sips, 2008).

## 5: Conclusion

The goal of this thesis was find out how well SOTA pre-trained ASR-models could perform judgements on word list oral reading tasks by children. Wav2vec2.0-CGN and faster-whisper-v2 performed as expected by the hypotheses. both showed moderate agreement with the assessors using MCC, indicating decent overall performance. Furthermore, faster-whisper-v2 outperformed wav2vec2.0-CGN in both MCC as well as accuracy. While wav2vec2.0-CGN's accuracy supports the hypothesis, faster-whisper-v2 performed better than expected.

We cannot provide a clear overview of the most prominent types of errors the ASR-models made, as our approach to error categorisation led to ambiguous results. Error categories common according to ASR-models were not common according to assessors in CHOREC.

While the results are promising, especially those for faster-whisper-v2, the use of pre-trained ASR-models is not advisable based on these results. While the accuracy of faster-whisper-v2 may look like "human-like" performance at face value, the accuracy value can be misleading. In terms of MCC only moderate correlations were found, which leave much room for improvement. In order to comfortably use ASR-models in practice, the resulting metrics need to be improved.

While the use of these pre-trained ASR-models may not be advisable yet, better results can be achieved by methods provided in this thesis. The error categories could be defined more in-line with how they were in CHOREC, which would allow for multiple error categories to be spotted in a single word. The list of prompts could be provided to the ASR-models' language models to help them judge words better. Furthermore, hybrid-models should not be ruled out simply because they are no longer considered SOTA.

### 5.1. Suggestions for Future Studies

In future, it could be beneficial for researchers to work together with teachers to see if the results from ASR-models are usable for them in practice. Bernstein (2017) showed that self-administered oral reading assessment is feasible for children as young as 5. Perhaps the DMT could be self-administered by using ASR-based assessments. This would alleviate a lot of time and energy primary school teachers are currently putting into the DMT. One possible way in which this could be tested is through a large-scale experimental study in which one group of teachers use the results of ASR-models to help them assess oral children's word reading and one group of teachers uses the standard DMT-procedure. 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.

If this were to be possible, it could lead to improvements in the DMT and even reading education in general. Teachers would have more time to either teach more reading, or to obtain more diagnostically relevant information from the DMT results. We know that the three most prominent 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 & Seidenberg, 2004; Perry et al., 2013).

If this is successful, teacher could use this extra time to help pupils become more familiar with the types of letters, clusters of letters, and/or full words they struggle with most. This would require the ASR-based assessment to be able to categorise the errors. For example, if these errors show that a child only makes errors in words with consonant clusters, the teacher could use this information to

help familiarise this child with consonant clusters more through the reading materials used in class. Van Til et al. (2018) even states that teachers could take note of common errors in the current DMT to help find patterns in the errors being made, though it is not considered necessary.

The potential of ASR-based assessment to aid teachers is there. It would increase the amount of time teachers would have for other parts of the teaching process instead of conducting the DMT. Furthermore, there is much diagnostic information hidden in the results of the DMT that are not currently used. This information could help teachers accurately tackle the parts of reading that children specifically struggle with. Future research could explore these possibilities.

## 6: References

- Ahn, T., Hong, Y., Im, Y., Kim, D. H., Kang, D., Jeong, J. W., Kim, J. W., Kim, M. J., Cho, A., Jang, D.-H., & Nam, H. (2024). *Automatic Speech Recognition (ASR) for the Diagnosis of pronunciation of Speech Sound Disorders in Korean children* (arXiv:2403.08187). arXiv. <http://arxiv.org/abs/2403.08187>
- Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* (arXiv:2006.11477). arXiv. <http://arxiv.org/abs/2006.11477>
- Barcovschi, A., Jain, R., & Corcoran, P. (2023). A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition. *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 42–47. <https://doi.org/10.1109/SpeD59241.2023.10314867>
- Bell, N. (2023). *A deep dive into phonemic proficiency* [Research Report]. Macquarie University.
- Berger, B. (2003). *Efficient Testing with All-Pairs*. STAREast.
- Bernstein, J., Cheng, J., Balogh, J., & Rosenfeld, E. (2017). Studies of a Self-Administered Oral Reading Assessment. *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2017)*, 172–176. <https://doi.org/10.21437/SLaTE.2017-30>
- Bolaños, D., Cole, R. A., Ward, W., Borts, E., & Svirsky, E. (2011). FLORA: Fluent oral reading assessment of children’s speech. *ACM Transactions on Speech and Language Processing*, 7(4), 1–19. <https://doi.org/10.1145/1998384.1998390>
- Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest*, 19(1), 5–51. <https://doi.org/10.1177/1529100618772271>
- Chang, Y.-N., Taylor, J. S. H., Rastle, K., & Monaghan, P. (2020). The relationships between oral language and reading instruction: Evidence from a computational model of reading. *Cognitive Psychology*, 123, 101336. <https://doi.org/10.1016/j.cogpsych.2020.101336>

- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6.  
<https://doi.org/10.1186/s12864-019-6413-7>
- Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 4. <https://doi.org/10.1186/s13040-023-00322-4>
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE Access*, *9*, 78368–78381. IEEE Access. <https://doi.org/10.1109/ACCESS.2021.3084050>
- Cito B.V. (2017). *Handleiding DMT* (Cito volgsysteem). Cito B.V.  
[http://www.goloca.org/nt2/dmt/cito\\_dmt\\_handleiding\\_groep\\_3-8.pdf](http://www.goloca.org/nt2/dmt/cito_dmt_handleiding_groep_3-8.pdf)
- Cleuren, L., Duchateau, J., Ghesquière, P., & Van hamme, H. (2008). Children's Oral Reading Corpus (CHOREC): Description and Assessment of Annotator Agreement. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Eds.), *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](http://www.lrec-conf.org/proceedings/lrec2008/pdf/254_paper.pdf)
- Cleuren, L., Duchateau, J., & Sips, A. (2008). *Annotation Protocol for CHOREC* (p. 41). K.U. Leuven.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256.  
<https://doi.org/10.1037/0033-295x.108.1.204>
- De Vries, D. (2023, December 5). 'Onverwacht snelle daling' leesvaardigheid Nederlandse tieners; één op drie 15-jarigen nu 'onvoldoende geletterd.' *de Volkskrant*.  
<https://www.volkskrant.nl/nieuws-achtergrond/onverwacht-snelle-daling-leesvaardigheid-nederlandse-tieners-een-op-drie-15-jarigen-nu-onvoldoende-geletterd~b0f38d8f/>
- Dyck, B. V., BabaAli, B., & Compernelle, D. V. (2021). A Hybrid ASR System for Southern Dutch. *Computational Linguistics in the Netherlands Journal*, *11*, 27–34.

- Fan, R., Shankar, N. B., & Alwan, A. (2024). *Benchmarking Children's ASR with Supervised and Self-supervised Speech Foundation Models* (arXiv:2406.10507). arXiv.  
<http://arxiv.org/abs/2406.10507>
- Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84, 101567.  
<https://doi.org/10.1016/j.csl.2023.101567>
- Galarnyk, M. (2022). *Train Test Split: What it Means and How to Use It*. Built In.  
<https://builtin.com/data-science/train-test-split>
- Gao, L., Tejedor-Garcia, C., Strik, H., & Cucchiari, C. (2024). *Reading Miscue Detection in Primary School through Automatic Speech Recognition* (arXiv:2406.07060). arXiv.  
<http://arxiv.org/abs/2406.07060>
- Groenhof, B. (2024a). *GroenhofBram/wav2vec-CHOREC* (Version 1.0) [Python].  
<https://github.com/GroenhofBram/wav2vec-CHOREC>
- Groenhof, B. (2024b). *GroenhofBram/whisper-CHOREC* (Version 1.0) [Python].  
<https://github.com/GroenhofBram/whisper-CHOREC>
- GroNLP. (2023) *Wav2vec2-dutch-large-ft-cgn · Hugging Face* [Computer software].  
<https://huggingface.co/GroNLP/wav2vec2-dutch-large-ft-cgn>
- Harm, M. W., & Seidenberg, M. S. (2004). Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. *Psychological Review*, 111(3), 662–720. <https://doi.org/10.1037/0033-295X.111.3.662>
- Harmsen, W., Cucchiari, C., van Hout, R., & Strik, H. (2024). A Joint Approach for Automatic Analysis of Reading and Writing Errors. In K. Gorman, E. Prud'hommeaux, B. Roark, & R. Sproat (Eds.), *Proceedings of the Second Workshop on Computation and Written Language (CAWL) @ LREC-COLING 2024* (pp. 8–17). ELRA and ICCL.  
<https://aclanthology.org/2024.cawl-1.2>
- Harmsen, W., Hubers, F., Van Hout, R., Cucchiari, C., & Strik, H. (2023). Measuring Word Correctness in Young Initial Readers: Comparing Assessments from Teachers, Phoneticians,

- and ASR Models. *9th Workshop on Speech and Language Technology in Education (SLaTE)*, 11–15. <https://doi.org/10.21437/SLaTE.2023-3>
- Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- Higuera, C. de la, & Mico, L. (2008). *A Contextual Normalised Edit Distance* (p. 361). <https://doi.org/10.1109/ICDEW.2008.4498345>
- Hu, G., Determan, S. C., Dong, Y., Beeve, A. T., Collins, J. E., & Gai, Y. (2020). Spectral and Temporal Envelope Cues for Human and Automatic Speech Recognition in Noise. *Journal of the Association for Research in Otolaryngology*, *21*(1), 73–87. <https://doi.org/10.1007/s10162-019-00737-z>
- Hussein, A., Watanabe, S., & Ali, A. (2021). Arabic speech recognition by end-to-end, modular systems and human. *Computer Speech & Language*, *71*, 101272. <https://doi.org/10.1016/j.csl.2021.101272>
- Inspectorate of Education (2024). *De Staat van het Onderwijs* (p. 140). Inspectorate of Education.
- Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., & Cucu, H. (2023). *Adaptation of Whisper models to child speech recognition* (arXiv:2307.13008). arXiv. <https://doi.org/10.48550/arXiv.2307.13008>
- Jain, R., Barcovschi, A., Yiwere, M. Y., Corcoran, P., & Cucu, H. (2024). Exploring Native and Non-Native English Child Speech Recognition With Whisper. *IEEE Access*, *12*, 41601–41610. <https://doi.org/10.1109/ACCESS.2024.3378738>
- Kim, H., Hannah, L., & Jang, E. (2021). *Using acoustic features to predict oral reading fluency of students with diverse language backgrounds*.
- Klebanov, B. B., Loukina, A., Lockwood, J., Licalalde, V. R. T., Sabatini, J., Madnani, N., Gyawali, B., Wang, Z., & Lentini, J. (2020). Detecting learning in noisy data: The case of oral reading

- fluency. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 490–495. <https://doi.org/10.1145/3375462.3375490>
- Klein, G. (2023). *Whisper-large-v2 · Hugging Face* [Computer software].  
<https://huggingface.co/guillaumekln/faster-whisper-large-v2>
- Lakshmi, T. J., & Prasad, Ch. S. R. (2014). A study on classifying imbalanced datasets. *2014 First International Conference on Networks & Soft Computing (ICNSC2014)*, 141–145.  
<https://doi.org/10.1109/CNSC.2014.6906652>
- Loughbrough University. (n.d.). *Numeracy, Maths and Statistics—Academic Skills Kit*. Retrieved August 28, 2024, from <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/strength-of-correlation.html>
- Loukina, A., Klebanov, B. B., Lange, P., Gyawali, B., & Qian, Y. (2017). Developing speech processing technologies for shared book reading with a computer. *6th Workshop on Child Computer Interaction (WOCCI 2017)*, 46–51. <https://doi.org/10.21437/WOCCI.2017-8>
- McFee, B., Raffel, C., Liang, D., Ellis, D., Mcvicar, M., Battenberg, E., & Nieto, O. (2015). *librosa: Audio and Music Signal Analysis in Python* (p. 24). <https://doi.org/10.25080/Majora-7b98e3ed-003>
- Mich, O., Mana, N., Gretter, R., Matassoni, M., & Falavigna, D. (2020). Automatically Assess Children’s Reading Skills. In N. Gala & R. Wilkens (Eds.), *Proceedings of the 1st Workshop on Tools and Resources to Empower People with READING Difficulties (READI)* (pp. 20–26). European Language Resources Association. <https://aclanthology.org/2020.readi-1.4>
- Molenaar, B., Tejedor-Garcia, C., Cucchiarini, C., & Strik, H. (2023). Automatic Assessment of Oral Reading Accuracy for Reading Diagnostics. *INTERSPEECH 2023*, 5232–5236.  
<https://doi.org/10.21437/Interspeech.2023-1681>
- Moon, T., Choi, H., Lee, H., & Song, I. (2015). RNNDROP: A novel dropout for RNNS in ASR. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 65–70.  
<https://doi.org/10.1109/ASRU.2015.7404775>
- Mudadla, S. (2023). Is accuracy a good performance metric? When does it fail to capture the performance of an ML model? *Medium*. <https://medium.com/@sujathamudadla1213/is->

accuracy-a-good-performance-metric-when-does-it-fail-to-capture-the-performance-of-an-ml-model-dd04a74c56ed

Nam, E. (2019). Understanding the Levenshtein Distance Equation for Beginners. *Medium*.

<https://medium.com/@ethannam/understanding-the-levenshtein-distance-equation-for-beginners-c4285a5604f0>

National Institute of Standards and Technology (2021). *SCTK* [Python]. National Institute of Standards and Technology. <https://github.com/usnistgov/SCTK>

Ngueajio, M. K., & Washington, G. (2022). Hey ASR System! Why Aren't You More Inclusive? In J.

Y. C. Chen, G. Fragomeni, H. Degen, & S. Ntoa (Eds.), *HCI International 2022 – Late*

*Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence* (pp. 421–440).

Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-21707-4\\_30](https://doi.org/10.1007/978-3-031-21707-4_30)

NOS Nieuws (2023, December 5). Leesvaardigheid Nederlandse 15-jarigen verder achteruitgegaan.

*NOS Nieuws*. <https://nos.nl/artikel/2500415-leesvaardigheid-nederlandse-15-jarigen-verder-achteruitgegaan>

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

<https://doi.org/10.1787/53f23881-en>

Parikh, A., ten Bosch, L., van den Heuvel, H., & Tejedor-Garcia, C. (2023). Comparing Modular and

End-To-End Approaches in ASR for Well-Resourced and Low-Resourced Languages. In M.

Abbas & A. A. Freihat (Eds.), *Proceedings of the 6th International Conference on Natural*

*Language and Speech Processing (ICNLSP 2023)* (pp. 266–273). Association for

Computational Linguistics. <https://aclanthology.org/2023.icnlspp-1.28>

Perry, C., Ziegler, J. C., & Zorzi, M. (2013). A Computational and Empirical Investigation of

Graphemes in Reading. *Cognitive Science*, 37(5), 800–828.

<https://doi.org/10.1111/cogs.12030>

Piton, T., Hermann, E., Pasqualotto, A., Cohen, M., Magimai. -Doss, M., & Bavelier, D. (2023).

*Using Commercial ASR Solutions to Assess Reading Skills in Children: A Case Report*. 4573–

4577. <https://doi.org/10.21437/Interspeech.2023-928>

- Poncelet, J., & Van Hamme, H. (2023). Learning to Jointly Transcribe and Subtitle for End-To-End Spontaneous Speech Recognition. *2022 IEEE Spoken Language Technology Workshop (SLT)*, 182–189. <https://doi.org/10.1109/SLT54892.2023.10022420>
- Prasad, A., & Jyothi, P. (2020). How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3739–3753). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.345>
- Python Software Foundation (2022). *Python Release Python 3.11*. Python.Org. <https://www.python.org/downloads/release/python-3110/>
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*.
- Rapcsak, S. Z., Henry, M. L., Teague, S. L., Carnahan, S. D., & Beeson, P. M. (2007). Do Dual-Route Models Accurately Predict Reading and Spelling Performance in Individuals with Acquired Alexia and Agraphia? *Neuropsychologia*, 45(11), 2519–2524. <https://doi.org/10.1016/j.neuropsychologia.2007.03.019>
- Sadeghi, M. E., Sheikhzadeh, H., & Emadi, M. J. (2024). A proposed method to improve the WER of an ASR system in the noisy reverberant room. *Journal of the Franklin Institute*, 361(1), 99–109. <https://doi.org/10.1016/j.jfranklin.2023.11.039>
- Seidenberg, M. (2005). Connectionist Models of Word Reading. *Current Directions in Psychological Science - CURR DIRECTIONS PSYCHOL SCI*, 14, 238–242. <https://doi.org/10.1111/j.0963-7214.2005.00372.x>
- Shraddha, S., G, J. L., & S, S. K. (2022). Child Speech Recognition on End-to-End Neural ASR Models. *2022 2nd International Conference on Intelligent Technologies (CONIT)*, 1–6. <https://doi.org/10.1109/CONIT55038.2022.9847929>
- Simonnet, E., Ghannay, S., Camelin, N., & Estève, Y. (2018, May 7). *Simulating ASR errors for training SLU systems*. LREC 2018. <https://univ-lemans.hal.science/hal-01715923>

- Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1), 77–89. [https://doi.org/10.1016/S0034-4257\(97\)00083-7](https://doi.org/10.1016/S0034-4257(97)00083-7)
- Strik, H., Cucchiarini, C., Hubers, F., Bakker, M., Irausquin, R., Enter, M., & Van Schooten, E. (2022). *DART | Dutch Automatic Reading Tutor*. <https://www.ru.nl/en/research/research-projects/dart-dutch-automatic-reading-tutor>, <https://www.ru.nl/en/research/research-projects/dart-dutch-automatic-reading-tutor>
- 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](https://taalmaterialen.ivdnt.org/wp-content/uploads/documentatie/cgn_website/doc_Dutch/topics/index.htm)
- Tejedor-García, C., van der Molen, B., van den Heuvel, H., van Hessen, A., & Pieters, T. (2022). Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 1032–1039). European Language Resources Association. <https://aclanthology.org/2022.lrec-1.110>
- Tillemans, M. (2007). *Dissolving the ‘d/dt’ disambiguation problem* (07–01; Technical Report Series). University of Tilburg.
- Van der Klis, A., Adriaans, F., Han, M., & Kager, R. (2023). Using Open-Source Automatic Speech Recognition Tools for the Annotation of Dutch Infant-Directed Speech. *Multimodal Technologies and Interaction*, 7(7), Article 7. <https://doi.org/10.3390/mti7070068>
- Van Gompel, M. (2023). *Dutch Open Speech Recognition Benchmark*. Dutch Open Speech Recognition Benchmark. [https://opensource-spraakherkenning.nl.github.io/ASR\\_NL\\_results/UT/Jasmin/jasmin\\_res.html](https://opensource-spraakherkenning.nl.github.io/ASR_NL_results/UT/Jasmin/jasmin_res.html)
- Van Halteren, H., & Oostdijk, N. (2018). Identification of Differences between Dutch Language Varieties with the VarDial2018 Dutch-Flemish Subtitle Data. In M. Zampieri, P. Nakov, N.

- Ljubešić, J. Tiedemann, S. Malmasi, & A. Ali (Eds.), *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)* (pp. 199–209). Association for Computational Linguistics. <https://aclanthology.org/W18-3923>
- Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., Vloedgraven, J., & de Wijs, A. (2018). *Wetenschappelijke verantwoording LVS-toetsen DMT*. Cito.
- Wentink, W. M. J. (1997). *From graphemes to syllables: The development of phonological decoding skills in poor and normal readers* [Radboud University]. <https://repository.ubn.ru.nl/handle/2066/265053>
- Woolams, A., Ralph, M., Plaut, D., & Patterson, K. (2007). SD-Squared: On the Association Between Semantic Dementia and Surface Dyslexia. *Psychological Review*, *114*, 316–339. <https://doi.org/10.1037/0033-295X.114.2.316>
- Yeung, G., & Alwan, A. (2018). On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children. *Interspeech 2018*, 1661–1665. <https://doi.org/10.21437/Interspeech.2018-2297>

## Appendices

### A: Full Similarity-level Based Improvements Tables

The following tables show all results of the similarity-based improvements.

**Table A.1**

*Similarity-based improvements results for wav2vec2.0-CGN (training set)*

Similarity	TN	TP	FN	FP	Accuracy	Precision	Recall	F1-score	MCC
.05	10904	0	575	1	.95	0	0	0	0
.1	10904	0	575	1	.95	0	0	0	0
.15	10904	0	575	1	.95	0	0	0	0
.2	10903	1	574	2	.95	.33	0	0	.02
.25	10894	2	572	12	.95	.14	0	0	.01
.3	10883	5	570	12	.95	.29	.01	.02	.04
.35	10875	11	564	30	.95	.27	.02	.04	.06
.4	10862	16	559	43	.95	.27	.03	.05	.07
.45	10810	38	537	95	.94	.29	.07	.11	.12
.5	10784	48	527	121	.94	.28	.08	.12	.13
.55	10638	81	494	267	.93	.23	.14	.17	.15
.6	10523	111	464	382	.93	.23	.19	.21	.17
.65	10264	159	416	641	.91	.2	.28	.23	.19
.7	10015	206	369	890	.89	.19	.36	.25	.21
.75	9824	250	325	1081	.88	.19	.43	.26	.23
.8	9265	306	269	1640	.83	.16	.53	.25	.22
.85	8345	395	180	2560	.76	.13	.69	.22	.23
.9	7716	455	120	3189	.77	.12	.79	.21	.25
.95	7128	485	90	3777	.66	.11	.84	.19	.22
1	6346	491	84	4559	.6	.1	.85	.18	0.19

**Table A.2***Similarity-based improvements results for faster-whisper-v2 (training set)*

<b>Similarity</b>	<b>TN</b>	<b>TP</b>	<b>FN</b>	<b>FP</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>MCC</b>
.05	10877	1	582	20	.95	.05	0	0	0
.1	10877	1	582	20	.95	.05	0	0	0
.15	10873	1	582	24	.95	.04	0	0	0
.2	10873	2	581	24	.95	.08	0	0	.01
.25	10871	2	581	26	.95	.07	0	0	0
.3	10868	5	578	29	.95	.15	.01	.02	.02
.35	10866	9	574	31	.95	.23	.02	.04	.05
.4	10866	12	571	31	.95	.28	.02	.04	.06
.45	10854	17	566	43	.95	.28	.03	.05	.08
.5	10851	24	559	46	.95	.34	.04	.07	.1
.55	10802	35	548	95	.94	.27	.06	.1	.11
.6	10791	58	525	106	.95	.35	.1	.16	.17
.65	10757	87	496	140	.94	.38	.15	.22	.22
.7	10664	116	467	233	.94	.33	.2	.25	.23
.75	10626	153	430	271	.94	.36	.26	.3	.28
.8	10537	189	394	360	.93	.34	.32	.33	.3
.85	9982	257	326	915	.89	.22	.44	.29	.26
.9	9776	321	262	1121	.89	.22	.55	.31	.3
.95	9591	349	234	1306	.87	.21	.6	.31	.3
1	9559	352	231	1338	.86	.21	.6	.31	.3